RoleNet: Treat a Movie as a Small Society

Chung-Yi Weng[1], Wei-Ta Chu[1,3] , and Ja-Ling Wu[1,2]

[1]Department of CSIE
[2]Graduate Inst. of Networking and Multimedia
National Taiwan University

[3]Department of CSIE
National Chung Cheng University

{chunye,wtchu,wjl}@cmlab.csie.ntu.edu.tw

ABSTRACT

We present a brave new way to analyze movie content, from the perspectives of the relationships between roles rather than low-level audiovisual features. Interactions between roles in a movie resemble human behaviors in a society. Roles' actions lead the story and make viewers understand what directors want to present. In this paper, we introduce the idea of social network analysis to model the relationships of actors/actresses as a network, called RoleNet. Through analyzing this network, the proposed approach automatically determines the leading roles and the communities embedded in movies. We also describe an implementation framework to realize the proposed model. The experimental results show that the proposed methods can effectively capture social characteristics in movies. It's believed that this idea provides a different way to approach movie understanding.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *indexing methods.* I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding –*Representations, data structures, and transforms; Modeling and recovery of physical attributes.*

General Terms

Algorithms, Management, Human Factors.

Keywords

RoleNet, movie analysis, social network analysis.

1. 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.

Many studies have been proposed to analyze movies based on audiovisual features. They can be roughly categorized into the following fields: genre classification, story segmentation, and video abstraction. Rasheed et al. [1] exploit color, motion, and shot information to classify movies into comedies, action, dramas, or horror films. Adams et al. [2] evaluate 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][5][6] have been proposed to represent movie content in a compact manner, such as automatic summarization for action movies [7][8].

Some studies are conducted from the perspective of 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.

Over the past decade, researches on movie analysis attempt to solve the most notorious problem: bridging the semantic gap. However, it seems that approaches based on audiovisual features face an unbreakable impediment. In this work, we try to analyze movie from a different perspective. 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 for discovering 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 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 content analysis. Figure 1 shows the resemblance in these two research fields. Computers are just able to calculate or extract low-level observations and never understand the hidden nature of digital content. In this work, we try to model relationships between computational observations as a complex network, and introduce SNA techniques to discover hidden semantic information.

Humans understand stories conveyed by a movie because they learn the mutual relations between roles from watching the movie. How the roles interact or conflict leads the story. Therefore, we treat a movie as a small society, which is constructed by the roles and their interactions. We model the relationships between roles as a role's social network, which is named as RoleNet. Based on

MIR'07, September 28-29, 2007, Augsburg, Bavaria, Germany.

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.

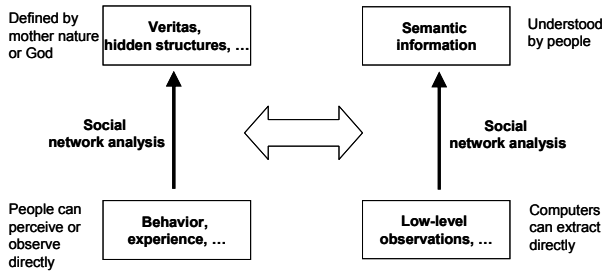

Figure 1. The distinction of social network analyses in different research fields.

The contributions of this work are summarized as follows:

- Introducing the idea of SNA to movie analysis: We elaborately introduce the concept of SNA to achieve semantic movie analysis. An implementation framework is proposed to realize the idea and practically bridge computational observations and semantic information.
- Proposing an approach to model roles' interrelationship as a network: We explicitly address how to evaluate roles' interrelationships and transform relationships into a network (RoleNet). The proposed construction methods are general to various types of movies.
- Proposing algorithms to analyze social relationship in movies: Based on RoleNet, several algorithms are designed to discover hidden semantics. Two important issues are addressed, including leading roles determination and community identification. They are social characteristics especially existing in movie videos, and are good clues to approach movie understanding.

This paper is organized as follows. In Section 2, we describe how to model roles' interrelationship and construct the RoleNet. Based on RoleNet, we start analysis from the most popular type of movies, the so-called bilateral movies, in Section 3. We generalize the proposed model to most types of movies and perform finer analysis in Section 4. To realize the proposed idea, we construct a framework and describe details of implementation in Section 5. Section 6 describes the experimental results. We give some discussions in Section 7, and Section 8 states the concluding remarks.

2. ROLENET

2.1 Definition of RoleNet

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

- Representing relationships effectively: There may be many roles in a movie, and relationships between them are often intricate. In addition, closeness between different pairs of roles varies. How to effectively represent these characteristics is the first key point.
- 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 facilitate systematic analysis.

With these requirements, roles' social network, i.e. RoleNet, is defined as follows:

Definition: A RoleNet is a weighted graph expressed by

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

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

2.2 Construction of RoleNet

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

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

The matrix presentation of Figure 2(a) is shown in Figure 2(c). More specifically, the *i*th row vector, $\boldsymbol{a}_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$, of A denotes the scenes where the *i*th role appeared. Based on this occurrence matrix, we can identify the co-occurrence of the *i*th role and the *j*th role by

$$w_{ij} = \sum_{k=1}^{n} a_{ik} a_{jk} = \boldsymbol{a}_i \boldsymbol{a}_j^T . \quad (2)$$

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 between roles in a movie can be expressed by

$$W_{m \times m} = AA^T. \quad (3)$$

In the example of Figure 2, the co-occurrence status between roles is expressed by Figure 2(d), and corresponding graphical representation, i.e. RoleNet, is shown in Figure 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 two roles. The thicker an edge is (larger weight), 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.

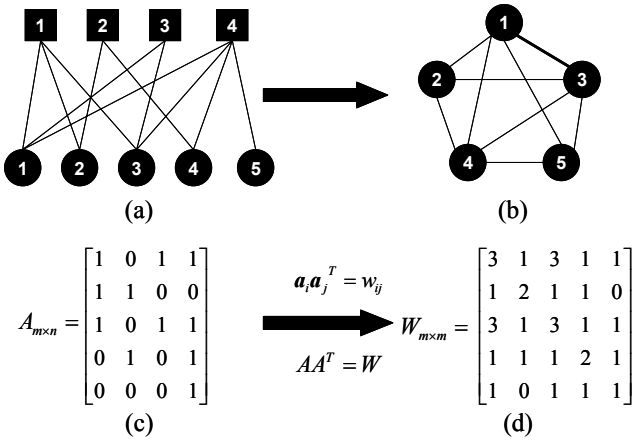

Figure 2. A graphical example to show the relationships between roles.

3. BILATERAL MOVIE ANALYSIS

3.1 Introduction

To demonstrate the effectiveness of the proposed idea, we take the most popular type of movie, named "bilateral movies", as a pilot instance. In a bilateral movie, there are two apparent leading roles (usually the hero and the heroine). Other roles support the progress of story and primarily form two major groups, which are respectively led by two leading roles. For example, there are justice and evil groups in most action movies. As for romance movies, it is very common to find two groups that belong to the hero and the heroine, respectively.

Figure 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 1 shows the true casts and the corresponding positions in this movie. Roles 1 and 2 are the hero and heroine, respectively. Other roles can be clearly separated into two groups, which are respectively led by the hero and the heroine. The RoleNet in Figure 3 consists of intricate edges so that the finer structure cannot be directly observed.

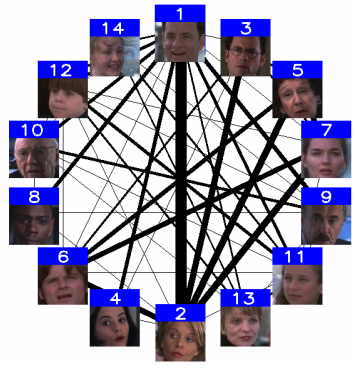

Figure 3. RoleNet of the movie "You've Got Mail".

To facilitate deeper investigation, we propose a process depicted in Figure 4 to perform RoleNet analysis. Given a RoleNet, we first determine the leading roles, and then identify the hidden communities. Leading roles are the persons who have the most significant impacts in movies. They dominate the progress of stories. 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 audiences can clearly perceive that they are "at the same side". Humans understand the conveyed stories through knowing the relationships between roles and their interactions.

Table 1. 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.

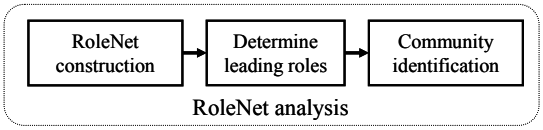

Figure 4. The proposed diagram of RoleNet analysis.

3.2 Determining Leading Roles

In SNA, evaluating the impact of each individual is one of the earliest issues. It is known as the *centrality problem* [11]. Centrality of a node can be evaluated as the number of connected edges. However, this measurement doesn't faithfully reflect the information in RoleNet, which is a weighted graph and convey much information in edge weights. Therefore, based on RoleNet, we evaluate the centrality c_i of a node (role) i as

$$c_i = \sum_{j \neq i} w_{ij}, \tag{4}$$

where the w_{ij} is the edge weight defined in Section 2.2.

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

3.3 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 is expressed as follows.

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

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

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

$$\Delta = \{\delta_i,\ i = 1, 2, ..., m\}, \tag{7}$$

$$\begin{cases} \delta_i = 0 & \text{if } v_i \text{ is assigned to the community led by } v_p \\ \delta_i = 1 & \text{if } v_i \text{ is assigned to the community led by } v_q \end{cases}$$

where m is the number of roles, $\mathbf{\Lambda}$ is a set of labels. $C(\mathbf{\Lambda})$ 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 (δ_p=0), and the second leading role v_q is labeled as 1 (δ_q=1). For 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] in this work. Additional discussions about applying SNA to bilateral movies please refer to our previous work [20].

Figure 5 shows the community identification result of the network in Figure 3. Node 1 and node 2 are correctly detected as the leading roles (with the dash-line circle). The roles identified as the members of G_p are marked by 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 1.

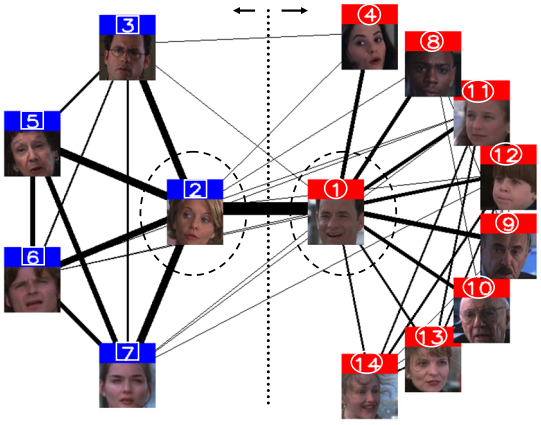

Figure 5. Results of community identification for the movie "You've Got Mail".

4. GENERALIZATION

The bilateral case has shown that RoleNet well describes the relationship between roles and the proposed methods effectively discover the hidden structure. In this section, we would like to generalize the idea to any kind of movies. The generalization process tackles with the following challenges:

- Automatically determining the number of leading roles: The prescribed process is conducted under the condition that we knew there are two leading roles beforehand. The prescribed community identification method can only be applied to binary labeling cases. Therefore, we try to develop a method that automatically determines the number of leading roles, and therefore remove the limitation that only bilateral movies can be analyzed.
- Analyzing finer communities: In the identified communities, there are actually finer structures within them. For example, although the roles no. 4 and no. 8 are both identified in the community led by the hero, as shown in Figure 5, they have totally different positions in the movie. Table 2 shows the true information at finer granularity. In this work, we call the rough communities identified in Section 3 as macro-communities, and call the finer structure in Table 2 as micro-communities.

Table 2. True communities 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 friends.
	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.

4.1 Determining Leading Roles

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

$$\Gamma^* = \arg\max_{\Gamma} \left[\min \Theta_1 - \max \Theta_0\right], \tag{8}$$

where $\Theta_1 = \{c_i \,|\, \ell_i = 1\}$,

$\Theta_0 = \{c_i \,|\, \ell_i = 0\}$,

$\Gamma = \{\ell_i, i = 1, 2, ..., m\}$,

$$\begin{cases} \ell_i = 1, & \text{if the } i\text{th role is assigned as a leading role,} \\ \ell_i = 0, & \text{otherwise,} \end{cases}$$

where m is the number of roles, Γ is a set of binary labels representing which roles are assigned as leading roles. Θ_1 represents the set of centrality values from the roles assigned to leading roles. The physical meaning of [minΘ_1 - maxΘ_0] is the centrality difference between the least important leading role and the most important supporting role.

To solve this problem, we still take "You've Got Mail" as an example, and propose a determination method in the following.

Step 1: Calculate the centrality value of each role, as shown in equation (4). Figure 6(a) shows a real example.

Step 2: Sort the centrality value in descending order, as shown in Figure 6(b).

Step 3: Calculate the centrality difference between two adjacent roles. Figure 6(c) shows the centrality difference distribution, in which each point represents the boundary characteristic between two roles.

Step 4: Find the maximum point in the difference distribution, which represents the largest gap in centrality. Then the selected boundary determines the number of leading roles. In the example of Figure 6, this method automatically determines that the roles no. 2 and no. 1 should be leading roles.

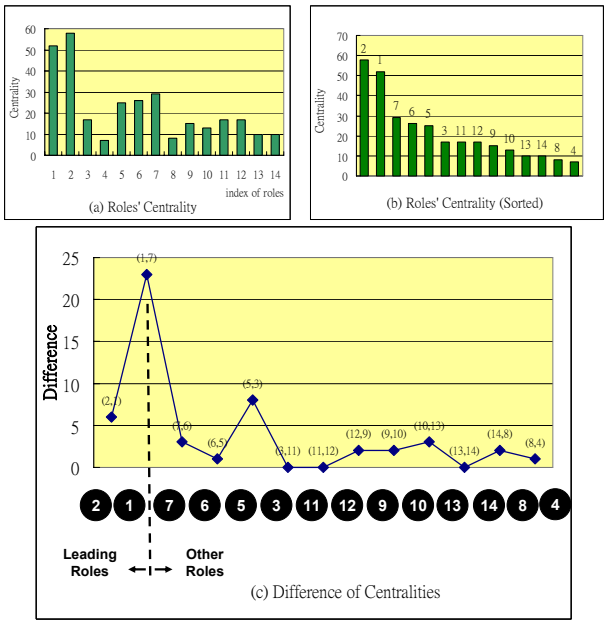

Figure 6. Centrality distribution of roles in the movie "You've Got Mail".

4.2 Community Identification

4.2.1 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 all edges linked to them from the RoleNet. Then, the following iterative algorithm is applied to the modified RoleNet. We use the value t to index the community situation in the progress of the algorithm. The value t is initialized as 0 in the beginning, and increases by one when the community situation changes.

Step 1: Initialize every individual node as a community. The set of community is denoted as $\Pi_t=\{T_1^t, T_2^t, \ldots, T_k^t\}$, $t = 0$, if there are initially k individual nodes. The size of the pth community in Π_t is denotes as $|T_p^t|$, which is the number of nodes included in this community.

Step 2: 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

- If $|T_p^t| \geq 1$ and $|T_q^t| = 1$, $T_p^{t+1} = T_p^t \cup T_q^t$, $\Pi_{t+1} = \Pi_t - \{T_q^t\}$, and $t = t+1$.
- If $|T_p^t| > 1$ and $|T_q^t| > 1$, keep current community situation.

Step 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 in each step. For example, as shown in Figure 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). The same process can be iteratively applied until all nodes are determined.

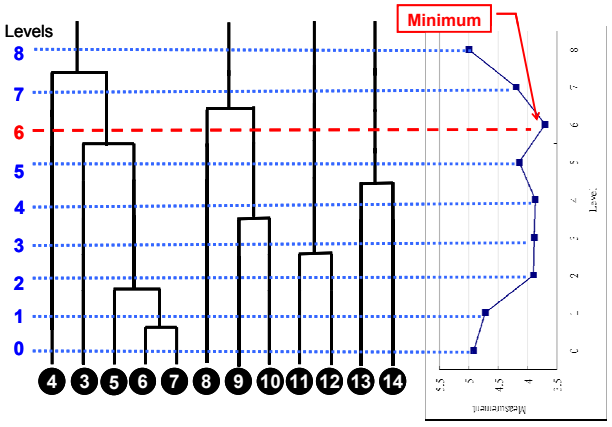

Figure 7. Dendrogram of the movie "You've Got Mail".

Each level in the dendrogram represents a case of community identification. Now the problem is how 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 calculated by

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

where Π_t denotes the community case 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 Figure 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 Figure 7, the minimal $AvgW$ value occurs at the level six, in which the communities are {4}, {3, 5, 6, 7}, {8}, {9, 10}, {11, 12}, and {13, 14}. The roles in the same brace are classified into the same communities. This result is very close to the true state listed in Table 2.

4.2.2 Macro-community identification

The communities described above show finer structures of a movie and are called micro-communities. 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.

Assume that L represents the set of leading roles. For the micro-community T_p, the assigning process is as follows.

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

Where the value w_{ij} denote 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.

Figure 8 simultaneously shows the results of micro-community and macro-community identification. 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 algorithms. Note that the result of macro-community identification is the same as that in Figure 5.

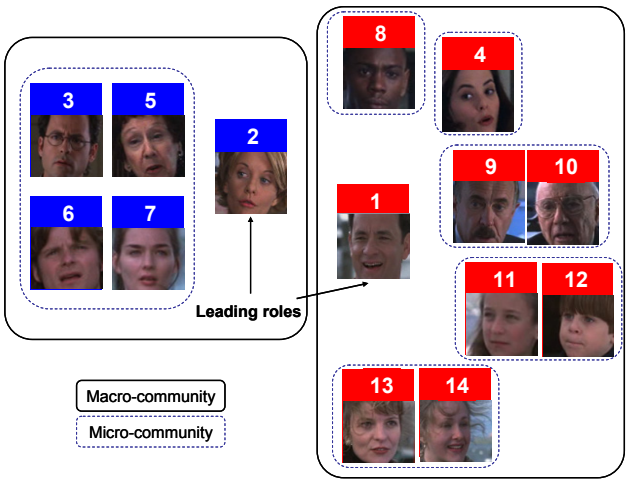

Figure 8. Results of micro-community and macro-community identification.

5. EXPERIMMENTAL SETUP AND IMPLEMENTATION FRAMEWORK

5.1 Evaluation Data

We use three Hollywood movies to evaluate the proposed model. As shown in Table 3, these movies belong to different genres and have different numbers of leading roles.

Table 3. Information of the evaluation data.

ID	Movie Title	Genres	# of leading roles
M1	The Devil Wears Prada	Comedy / Drama	1
M2	You've Got Mail	Comedy / Romance	2
M3	21 grams	Crime / Thriller	3

We evaluate RoleNet analysis algorithm based on two kinds of input data: "Clean Data" and "Realistic Data". In Clean Data, roles present in a specific scene are manually labeled. In Realistic Data, the data of role's occurrence are obtained through applying a real face recognition module, and the detail collection process will be explained in Section 5.2.

In short, the analysis results based on Clean Data are used to faithfully show the effectiveness of the proposed RoleNet analysis method, and the results based on Realistic Data would give us a picture about how well the model works in realistic situation.

5.2 Implementation Framework

5.2.1 Overview

In order to collect Realistic Data successfully, we propose a framework to integrate existing face recognition module, as shown in Figure 9.

In the beginning, we apply the method proposed in [15] to perform scene detection, and manually correct errors to get precise scene boundaries. Next, a face processing module is built to perform face detection, data processing, and face recognition. Then, Realistic Data are collected and can be taken as input data for the following RoleNet analysis.

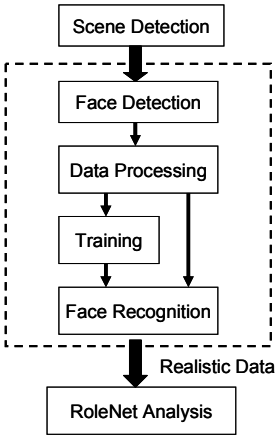

Figure 9. The proposed implementation framework.

5.2.2 Face Detection

We exploit OpenCV face detector [16] to detect the locations of faces in each scene. The information obtained from the face detection module includes the locations and regions of faces.

5.2.3 Data Processing

Obtaining satisfactory face detection/recognition results from movies is a very challenging problem. There may be many errors in face detection. Therefore, we have to process face location data further to obtain more reliable bases. Two processes are developed for this task.

- Step 1: Noise filtering

Many non-face objects are mis-detected as faces. Because non-face objects are usually just like a face in certain view angle, we can remove these noises based on the information of adjacent frames. Figure 10 shows an illustrative example about face data processing. The horizontal-axis denotes the frame number, while the vertical-axis denotes the x coordinates of detected faces. We connect faces that are in adjacent frames and are similar in location and region size into a list. Generally, non-face objects are intermittently mis-detected as faces. Therefore, we remove the list of detected faces that doesn't last longer than F_1 successive frames. In this work, the threshold F_1 is set as 15.

- Step 2: Grouping

The noise filtering process resolves the false alarm problem in face detection, while the grouping process resolves the miss problem. Two lists of faces, which both last longer than F_1 frames, will be connected if the faces are spatially near to each other and

the time distance between two lists is smaller than F_2. In this work, F_2 is set as 15. This operation is illustrated in Figure 10.

Because the final result we want to know is which roles ever appeared in which scene, we don't need to recognize every detected face. As shown in Figure 10, we can see that the faces in the same list belong to one role. Therefore, we just randomly choose one face from a list to perform face recognition. Grouping of adjacent lists of faces can effectively reduce the number of faces to be recognized.

Note that we can also equally sample several faces from a face list to perform face recognition, and vote to determine which face is presented in this list. In current work, we simply recognize one face from a list for computation efficiency.

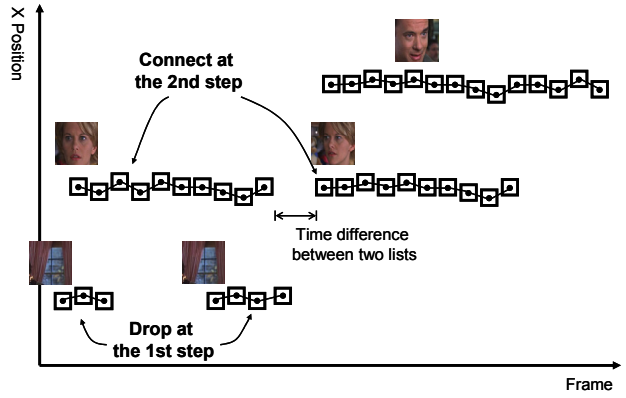

Figure 10. An illustrative example of face data processing.

5.2.4 Face Recognition

The face recognition method proposed in [17] is adopted at this stage. For recognizing roles in a movie, we choose and label the detected faces in the earliest *S* scenes for training data. From the beginning of a movie, the value *S* is determined by the number of scenes in which all roles have appeared at least one time. For example, all fourteen roles appear at least one time before the eleventh scene in the movie "You've Got Mail". The value of *S* is 11, therefore. The length of the earliest 11 scenes is about 20 minutes. Faces in these scenes are used to train the face recognition model.

For the remaining scenes, the faces selected by the processes described in Section 5.2.3 are recognized based on the trained model. Finally, we can know which roles presenting in which scenes. The data that are practically obtained from the described face detection, processing, and recognition modules are the Realistic Data.

5.2.5 Accuracy of Face Recognition

Accuracy of the face recognition model represents the characteristics of Realistic Data. More accurate the recognition module is, more reliable the Realistic Data are. Table 4 shows the recognition accuracy of the adopted module in these three movies. Due to large variations in luminance and pose, the recognition accuracy is not very good. With this kind of recognition performance, the Realistic Data are interfered by many noises. In the evaluation, we want to understand how many the hidden properties of role's relationships can be maintained, and whether the proposed model is robust to successfully extract them.

Table 4. Recognition accuracy of the adopted module.

Movie ID	# of faces recognized correctly / # of faces recognized	Recognition accuracy
M1	165 / 306	53.9%
M2	200 / 427	46.8%
M3	102 / 255	40.0%

6. RESULTS

6.1 Ground Truth

Based on the three movies described in Table 3, ground truths of leading roles and communities should be defined. For each movie, we ask three persons to manually label which ones are the leading roles after they watch this movie. The roles that are labeled as the leading ones by all the three persons are treated as the ground truth.

Ground truths of community identification are in two parts. For macro-community, we ask three persons to label which roles belong to which macro-communities (led by each leading role, respectively). If the position of a certain role is controversial, we assign him/her according to voting.

These three persons are also asked to label which roles belong to the same micro-community. Only the roles labeled to the same micro-community by all three persons are viewed as in the same micro-community.

Tables 5 and 6 show the ground truths of the macro- and micro-communities in the movies "The Devil Wears Prada" and "21 grams". The community truth in the movie "You've Got Mail" has been shown in Table 2.

Table 5. Ground truth of macro- and micro-communities in the movie "The Devil Wears Prada".

Macro	**Micro**	**Meaning of roles**
1	The 1st leading role	
2,3,4,5,6, 7,8,9, 10,11,12	2,3,4,9,11	The 1st leading role's boss and colleagues
	5,6,7	The 1st leading role's friends
	10	The 1st leading role's father
	11	The man who has affair with the 1st leading role.

Table 6. Ground truth of macro- and micro-communities in the movie "21 grams".

Macro	**Micro**	**Meaning of roles**
1	The 1st leading role	
2	The 2nd leading role	
6	The 3rd leading role	
8,9,18, 20	8,9,18	The 1st leading role's girlfriend and her doctors
	20	A detective employed by the 1st leading role
3,4,5,14, 17,19	3,4,5	The 2nd leading role's husband and daughters
	14,17	The 2nd leading role's father and sister
	19	The 2nd leading role's friend
7,10,11, 12,13,16	7	A boy taught by the 3rd leading role
	10,11,12, 13,16	The 3rd leading role's family and their friends
	15	The 3rd leading role's boss

Table 7. Performance of leading role determination.

Movie ID	Set of leading roles - Ground truth	Determined leading roles (Clean Data)	Recall (Clean Data)	Precision (Clean Data)	Determined leading roles (Realistic Data)	Recall (Realistic Data)	Precision (Realistic Data)
M1	{1}	{1}	100%	100%	{1}	100%	100%
M2	{1,2}	{1,2}	100%	100%	{1,2}	100%	100%
M3	{1,2,6}	{1,2,6}	100%	100%	{1,2,6}	100%	100%

Table 8. Performance of macro-community identification.

Movie ID	# of roles labeled correctly / # of roles (Clean Data)	Precision (Clean Data)	# of roles labeled correctly / # of roles (Realistic Data)	Precision (Realistic Data)
M1	12 / 12	100 %	12 / 12	100 %
M2	14 / 14	100%	14 / 14	100 %
M3	20 / 20	100%	18 / 20	90%

Table 9. Performance of micro-community identification.

Movie ID	Ground Truth	Result of Micro-Community Identification	
M1	Leading roles: {1} Other Roles: {2,3,4,9,11},{5,6,7}, {8},{10},{12}	(Clean Data)	{2,3,4,9},{5,6,7},{8}, ,{10},{11},{12}
		(Realistic Data)	{2,3,4,12},{5,6,7},{8}, {9},{10},{11}
M2	Leading roles: {1,2} Other Roles: {3},{4},{5,6,7},{8}, {9,10},{11,12},{13,14}	(Clean Data)	{3,5,6,7},{4},{8}, {9,10},{11,12},{13,14}
		(Realistic Data)	{3,6,7}{4},{5},{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}	(Clean Data)	{3,4,5},{7,10,11,12,13,16}, {8,9,18},{14,17},{15},{19},{20}
		(Realistic Data)	{3,4,5},{7,14,17,19},{8,9},{18}, {10,11,12,13},{15},{16} ,{20}

6.2 Performance of Leading Role Determination

The performance of leading role determination is shown in Table 7. The numbers in braces denote the indices of roles. For example, {1,2,6} in the "determined leading roles (Clean Data)" represent that the roles no. 1, 2, and 6 are automatically determined as the leading roles. We can see that perfect performance can be achieved no matter in Clean Data or Realistic Data.

The promising performance comes from two reasons:

- Leading roles pass through most scenes in a movie and have close relationship with other ones. Although the Realistic Data are interfered by many recognition noises, the major trends of relationships are largely retained.
- The proposed algorithm effectively captures the characteristics of leading roles. Based on the representation of RoleNet, leading roles can be clearly identified by measuring the impacts of roles.

6.3 Performance of Community Identification

6.3.1 Macro-Community Identification

Table 8 represents the performance of macro-community identification. It shows that the proposed method works perfectly for "Clean Data". Besides, the proposed method still provides good results for "Realistic Data" despite the performance of face recognition is not so satisfactory. This not only explains that the hidden macro-community structures are still maintained under poor face recognition condition, but also demonstrates the robustness of our algorithms.

6.3.2 Micro-Community Identification

The experimental results of micro-community identification are listed in Table 9. The numbers in each brace denote the indices of roles, which are categorized into the same micro-community. To briefly show the performance of this task, we design a method to quantify the massive experimental results.

We transform the community structures of the ground truth and the identified results into the relationship 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 C_2^k possible pairs if there are k roles to be identified. Among the C_2^k 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 by

$$R = \frac{\sum_{i=1}^{k}\sum_{j\neq i}\delta_{ij}}{2\times C_2^k}, \tag{11}$$

$$\begin{cases} \delta_{ij}=1, & \text{if } \zeta_{ij}^{g}=1 \text{ and } \zeta_{ij}^{v}=1, \\ \delta_{ij}=1, & \text{if } \zeta_{ij}^{g}=0 \text{ and } \zeta_{ij}^{v}=0, \\ \delta_{ij}=0, & \text{otherwise,} \end{cases} \tag{12}$$

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

Based on this measurement, we compare identification results derived from Clean Data and Realistic Data, respectively. 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 community alone. Figure 11 shows the performance comparison based on the proposed quantification method. Performance derived from Clean Data is reasonably superior to others. Due to the interference of face recognition errors, the performance of identification results from Realistic Data degrades. The degree of degradation is larger than that in leading role determination and macro-community identification. The reason is that role's relationship to maintain micro-communities are relatively weaker so that they are more sensitive to recognition errors. As the accuracy of face recognition increases, the performance of identification results is supposed to increase.

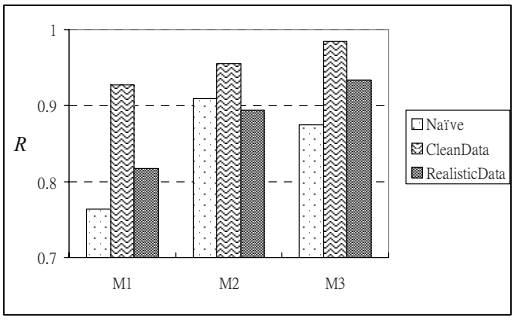

Figure 11. Performance comparison based on the proposed quantification method.

7. DISCUSSION

We have realized the proposed idea and shown the effectiveness of the proposed models/algorithms. Based on the results of RoleNet analysis, we address several potentials of RoleNet.

- Community-based hierarchical browsing system

As we know the community structure and leading roles, a community-based browsing system can be built. This kind of browsing scheme is totally different from hierarchical shot-based browsing, like the ones in [18][19]. Users can browse the story made by the hero and his relatives, for example, by exploring the corresponding macro-community. More specifically, users can explore deeper to see the stories related to the hero's family or the story specifically related to an individual. Figure 12 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, 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". A new browsing scenario can be elaborately developed on the basis of the proposed RoleNet analysis.

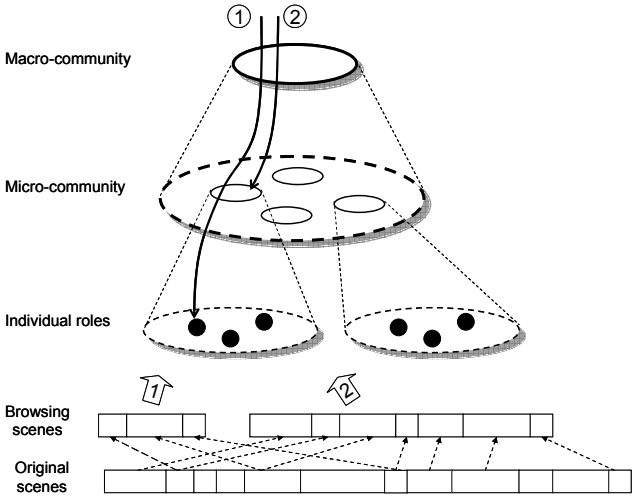

Figure 12. Community-based hierarchical browsing.

- Scene description

We can discover hidden semantics of each scene by observing the community structures. For example, some scenes represent the interaction between two communities, and some scenes enhance the relationships between roles in the same community. Through the characteristics of community, scenes can be described in terms of social relationships.

- Compatibility

It's important to point out that the proposed method and existing content-based ones are not irreconcilable. They can be combined to achieve finer movie understanding. For example, existing scene importance measures can be integrated with the RoleNet-based measures to facilitate advanced highlight extraction.

- Extensibility

The proposed model can not only be applied to movie videos. Other kinds of story-oriented videos that consist of roles' interaction, such as TV drama, can be effectively modeled and analyzed.

8. CONCLUSION

We have introduced the idea of social network analysis to movie content. 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, by the proposed construction method. Based on RoleNet, we use bilateral movies as the pilot instance to show the insights of the proposed idea. The analysis methods are further generalized to automatically determine the number of leading roles and identification of macro- and micro-communities. We also describe an implementation framework to make the idea practice. In experiments, we carefully evaluate the performance of the proposed methods based on three different movies. The promising results demonstrate the effectiveness and feasibility of the idea. With the aid of RoleNet, we approach movie understanding from a perspective different from audiovisual features.

In the future, we will step further to investigate and realize the potentials described in the discussion section. For example, by combining feature-based methods, more interesting and intelligent applications will be built. Furthermore, speaker identification techniques may be incorporated into the process of RoleNet construction.

9. ACKNOWLEDGEMENTS

This work was partially supported by the National Science Council of the Republic of China under grants NSC 95-2622-E-002-018, NSC 95-2752-E-002-006-PAE, and NSC 95-2221-E-002-332. It was also supported by National Taiwan University under grant 95R0062-AE00-02.

10. REFERENCES

[1] Rasheed, Z., Sheikh, Y., and Shah, M. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, 2005, pp. 52-64.

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

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

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

[5] Truong, B.T. and Venkatesh, S. Video abstraction: a systematic review and classification. *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, no. 1, 2007.

[6] Jung, B., Kwak, T., Song, J., and Lee, Y. Narrative abstraction model for story-oriented video. In *Proceedings of ACM Multimedia Conference*, 2004, pp. 828-835.

[7] Smeaton, A.F., Lehane, B., O'Connor, N.E., Brady, C., and Craig, G. Automatically selecting shots for action movie trailers. In *Proceedings of ACM International Workshop on Multimedia Information Retrieval*, 2006, pp. 231-238.

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

[9] Wang, H.L. and Cheong, L.-F. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, 2006, pp. 689-704.

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

[11] Scott, J. *Social network analysis: a handbook*. Newbury Park, 1991.

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

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

[14] Boykov, Y. and Kolmogorov, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, pp. 1124~1137.

[15] Rasheed, Z., and Shah, M. Scene detection in Hollywood movies and TV shows. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. 343-348.

[16] Open Source Computer Vision Library, http://www.intel.com/technology/computing/opencv/

[17] Nefian, A.V., and Hayes, M.H., III. An embedded HMM-based approach for face detection and recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, 1999, pp. 3553-3556.

[18] Zhu, X., Elmagarmid, A.K., Xue, X., Wu, L., and Catlin, A.C. InsightVideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, vol. 7, no. 4, 2005, pp. 648-666.

[19] Zhang, H.J., Low, C.Y., Smoliar, S.W., and Wu, J.H. Video parsing, retrieval and browsing: an integrated and content-based solution. In *Proceedings of ACM Multimedia Conference*, 1995, pp. 15-24.

[20] Weng, C.-Y., Chu, W.-T., and Wu, J.-L. Movie analysis based on roles' social network. In *Proceedings of IEEE International Conference on Multimedia & Expo*, pp. 1403-1406, 2007.