IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 20, 2018, accepted April 25, 2018, date of publication May 1, 2018, date of current version June 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2832087

StoryRoleNet: Social Network Construction of Role Relationship in Video

JINNA LV, BIN WU, LILI ZHOU, AND HAN WANG
Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Jinna Lv (lvjinna@bupt.edu.cn)

This work was supported in part by the National High-Tech Research and Development Program (863 Program) under Grant 2015AA050204, in part by the National Social Science Foundation of China under Grant 16ZDA055, and in part by the Special Fund for the Beijing Common Construction Project.

ABSTRACT The automatic extraction of social relationship among individuals from massive quantities of video data is an important aspect of information extraction. However, most existing studies have focused on rough information, which result in inaccurate social network of role relationship. In this paper, the StoryRoleNet model is proposed for constructing an accurate and integral network representing the relationships among roles. First, to avoid the redundancy calculation of the relationships on the segmentation points of neighboring story units, we measure the weights of relationships by a weighted-Gaussian method in each story unit. More importantly, a new story segmentation method for long video is proposed by analyzing hierarchical features of the video. Then, we combine relationship networks constructed from the video and subtitle text. Some missed relationships can be complemented by this way. At last, the final network is analyzed to discover communities and important roles. Comprehensive evaluations were conducted using three movies and one television drama. The results demonstrate that the proposed method outperforms state-of-the-art methods in terms of the F_1 accuracy measure and the normalized mutual information value.

INDEX TERMS Relationship network construction, story segmentation, social network analysis, community discovery.

I. INTRODUCTION

With the proliferation of social media, an immense quantity of video data is produced. A large proportion of these videos are role-centric. More importantly, the mining and extraction of social relationship among the roles have been increasingly important tasks because of their wide-ranging applications [1], [2]. The relationships are significant clues that can help understand the video story [3]. Furthermore, by gaining a better understanding of the implicit relationships, users can be protected from potential privacy exposure [4]. Conversely, such information can assist in searches for criminals [5].

Most current methods of constructing social networks for roles are built on structured data [6] and text data [7], [8]. Such methods cannot meet the rapidly increasing needs of extracting relationships among roles from the unstructured video. Owing to the complexity of roles' interactions and the variety of scenes or stories in long videos, social network construction has some challenges.

On the one hand, it is difficult to measure the weights for relationships among the roles appearing in a video. The reason is that relationships among roles are often intricate and can vary with time. Many studies have been conducted on methods for determining the weights of relationships, such as co-occurrence-based [3], [6] and the weighted-Gaussian method [9]. However, the methods of co-occurrence only roughly measured the relationships. In addition, the weights of relationships were computed over the whole video without taking the segmentation of the scenes or stories into account. Redundancy calculation may be produced at the segmentation point, because roles are in adjacent story units who may have no relationship. Owing to the complexity of roles' interactions and varies in different scenes of stories, simple methods cannot adequately measure the relationships of roles.

On the other hand, a lot of methods consider only visual features of video, few works take subtitle features into consideration. Roles often talk about other people who may have relationships with them. To improve the integration of the relationship network, method that combined with subtitle text information is advisable. In recently, some methods employ conversations between roles to analyze the interaction of roles [10], [11]. Conversations between people can well reflect the behavior of roles, thus their roles of the group can be properly recognized. In addition, relationship extraction from text has attractively been researched by scholars [12], [13]. However, these methods ignored the temporal feature of the conversation between people. As stories and events develop temporally, interactions and conversations of people change. Therefore, how to effectively take advantage of the subtitle text for extracting relationships remains a challenge.

In this paper, we propose the StoryRoleNet model to address these two challenges of social network construction. In order to accurately measure the weights of relationships between people, not only shot information but also story information is taken into consideration. A new story segmentation method is proposed for long video using the multi-level features. After story segmentation, the weights of relationships are computed by a weighted-Gaussian method in each story unit. In this way, our model can avoid the redundancy calculation of the weights of the story boundaries. More importantly, to achieve a more integrated network, we incorporate the relationship network extracted from the subtitle text into the network extracted from video. We employ the natural language processing method to identify the names of roles. Taking advantage of the temporal features of subtitle, we measure the weights of relationships of roles when they are in the same story unit. At last, social network analysis is conducted on the constructed network. It enables the discovery of hidden structures and properties that cannot be directly perceived or manually measured [14].

The main contributions of this paper are summarized as follows.

- We employ multi-level features of the video to accurately measure the weights of relationships. A new method of story segmentation for long video is proposed using hierarchical features of the video. After stories are segmented, the relationships based on temporal information are obtained by weighted-Gaussian method for each story unit.
- The relationships extracted from subtitle text are integrated into the relationship network constructed from visual features of the video. In this way, some missed relationships that having been ignored by visual features are obtained.
- In order to facilitate deeper investigation, extra evaluations, including community analysis and the identification of important roles, are conducted on the constructed network.

The remainder of this paper is organized as follows. Section II reviews three categories of related work on social network construction from three aspects. In Section III, we present definitions and notations. Our proposed method is described in Section IV, including video preprocessing, the StoryRoleNet model, and the social network analysis. Section V describes the experiments and results, and provides discussions. We conclude our research and discuss future work in Section VI.

II. RELATED WORK

There are many recent studies on the extraction of relationships between people and the construction of the corresponding network. Such studies can be classified into three categories based on the type of data used for the extraction: structured data, text data, and video data.

A. SOCIAL RELATIONSHIP NETWORK OF ROLES EXTRACTION FROM STRUCTURED DATA

Several studies have built real-world networks from massive quantities of social network data. Chen *et al.* [6] extracted social relationship network from transaction logs of a system for managing student cards. Using the co-occurrence between active students as determined by spatiotemporal attributes inferred from these logs, the authors determined a reasonable co-occurrence threshold for constructing the relationship network. He *et al.* [15] built social relationship network based on project cooperation data, addressing the issues of inaccurate entity identification and the inconvenience of data updates. In addition, an academic social relationship network was built by Tang *et al.* [16]. The above studies were based on structured data; however, the rapid growth of unstructured data in recent years has brought new challenges to the construction of relationship networks.

B. SOCIAL RELATIONSHIP NETWORK OF ROLES EXTRACTION FROM TEXT DATA

The extraction of entity relationship is a critical area in the field of text analysis. Peng *et al.* [7] proposed a tree-based method for the extraction of role relationship. Their method can handle numerous roles in a relational corpus. Warren *et al.* [12] presented the Six Degrees of Francis Bacon project and described the natural language processing tools and statistical graph learning techniques used to extract names and infer relationships from the Oxford Dictionary of National Biography. Srivastava *et al.* [13] proposed a general model that combines evidence from linguistic and semantic features of a text as well as features based on the structure of the social community in the text. Despite the above efforts, there remain major challenges in the disambiguation of names of persons in text-based relationship extraction [17], resulting in omissions and phenomena arising from mistakes during network construction.

C. SOCIAL RELATIONSHIP NETWORK OF ROLES EXTRACTED FROM VIDEO DATA

With the advent of video data, some researchers have extracted social relationship between people based on video and image content. Scholars have analyzed the contents of pictures and short videos to determine friendly, hostile, paternal, and other kinds of relations among roles according to gender, age, posture, and facial expression [4], [18]. Ramanathan *et al.* [19] analyzed the features of social roles in video events and proposed a weakly supervised social role recognition method. Nan *et al.* [20] adopted deep concept hierarchies and used a convolutional-recursive neural network to establish the social network between characters in a video.

In some studies, social relationship networks were constructed based on video content, and the relationships between roles were then analyzed from a social relationship network perspective. Ding and Yilmaz [21] utilized video and audio features to construct a network of individuals based on video scenes. Tran and Jung [1] considered a relationship network as a weighted graph wherein the weights of the relationships between two roles were determined based on their co-occurrence time: the longer the co-occurrence time of the two roles, the greater the weights of their relationships. Weng *et al.* [3] noted that there exist relationships between roles that appear in the same scene. Accordingly, they constructed a weight matrix based on the number of same-scene co-occurrences. Yuan *et al.* [9] considered that the weights of the relationships between roles aligns with a weighted Gaussian distribution. Thus, they calculated, layer by layer, the weight between roles in accordance with the shots and scenes in which the characters appear.

The above methods quantify the weights of human relationships according to co-occurrence times or the frequency of same-scene co-occurrences while ignoring the interaction of the roles in adjacent scenes. Yuan *et al.* [9] believe that closer frames in a shot indicate closer relationships between the characters appearing in them. However, there are cases in which roles are present in adjacent video frames who have no relationship between them. This is because these particular adjacent frames are ahead of and behind a story segmentation point, respectively. The roles in the two stories are not related to each other or have not yet established a relationship because they are in different relational communities. Therefore, to construct an accurate social relationship network, accurate quantification of the weights of relationships between roles is needed because it determines whether there exists a connected edge between them.

III. PROBLEM DEFINITION

A. RELATED DEFINITIONS

To construct the network of social relationship of roles in a video, that is, to extract the roles as a set of network nodes, the interactions of the roles in the video are analyzed to determine whether there exist edges between the corresponding nodes.

Definition 1. Set C of role nodes. Roles are the nodes in the relationship network. They are identified and marked in a video through face detection and recognition algorithms to form the set of role nodes $C = \{c_1, c_2, \ldots, c_n\}$.

Definition 2. Matrix W of relationship weights. Matrix W is an $n \times n$ matrix $[w_{ij}]_{n \times n}$, where element w_{ij} is the weight of the edge between nodes i and j, $i \in C, j \in C$. A weight is determined by analyzing the interactions of the corresponding roles appearing in the video. It is a measure of the relative strength of the relationships between roles.

Definition 3. Initialization relationship network G'. Network G' is defined as a weighted network and denoted as $G' = \langle C, E, W\rangle$, where C is the set of role nodes, E is the set of edges that connect two roles, and W is the weight matrix.

Definition 4. Social relationship network of roles. The social relationship network of the roles is denoted as $G = \langle C, E\rangle$, where C is the set of role nodes and E is the set of edges connecting the roles. As two roles in the video may interact multiple times in different scenes and stories, noise and redundancy may be introduced in the quantization process. We employ a threshold to filter out noise from the relationships.

B. PROBLEM FORMULATION

The StoryRoleNet model first divides video V into a collection of story units $V = \{s_1, s_2, \ldots, s_m\}$. Then the relationship weights can be statistically obtained in each story unit by employing a weighted-Gaussian-based method. Thus, we gain the initial social networks G'_V and G'_T based on the video and subtitle contents, respectively. The final relationship network $G = G_V \cup G_T$ is generated using weight thresholds. Specifically, the problem is defined as follows:

$$\left.\begin{array}{l} V- > V = \{s_1, s_2, \ldots, s_m\} \\ C = \{c_1, c_2, \ldots, c_n\} \end{array}\right\}$$
$$- > \left\{\begin{array}{l} G'_V =< C, E_V, W_V > \\ G'_T =< C, E_T, W_T > \end{array}\right\}$$
$$- > G = G_V \cup G_T =< C, E >,$$

where E_V and E_T denote the edges of G'_V and G'_T, and W_V and W_T denote the weights of edges in G'_V and G'_T, respectively.

IV. PROPOSED METHOD

To solve the above problems, we propose a construction method for social relationship of roles in videos. As shown in Fig. 1, the architecture for the framework consists of three parts: video preprocessing, StoryRoleNet model including story segmentation and social network construction, and social network analysis. The video preprocessing is used to identify the roles in the video in the first place. Then, the StoryRoleNet model includes the methods of story segmentation and construction of social network. Finally, the social network analysis further investigates communities and mines the constructed network for important roles.

A. VIDEO PREPROCESSING

The main task of the video preprocessing is to extract and identify the set of role nodes C from the video. Role recognition is the foundation of the role relationship analysis, as shown in Fig. 1. Set C is obtained using the following steps. First, using the methods of Douse *et al.* [22] and Ngo *et al.* [23], we extract the middle frame from each shot to serve as the keyframe for that shot. Consequently, each scene can be represented as a sequence of keyframes. Second, the method of Sun *et al.* [24] is employed to detect and recognize the face images in the keyframes. Because a

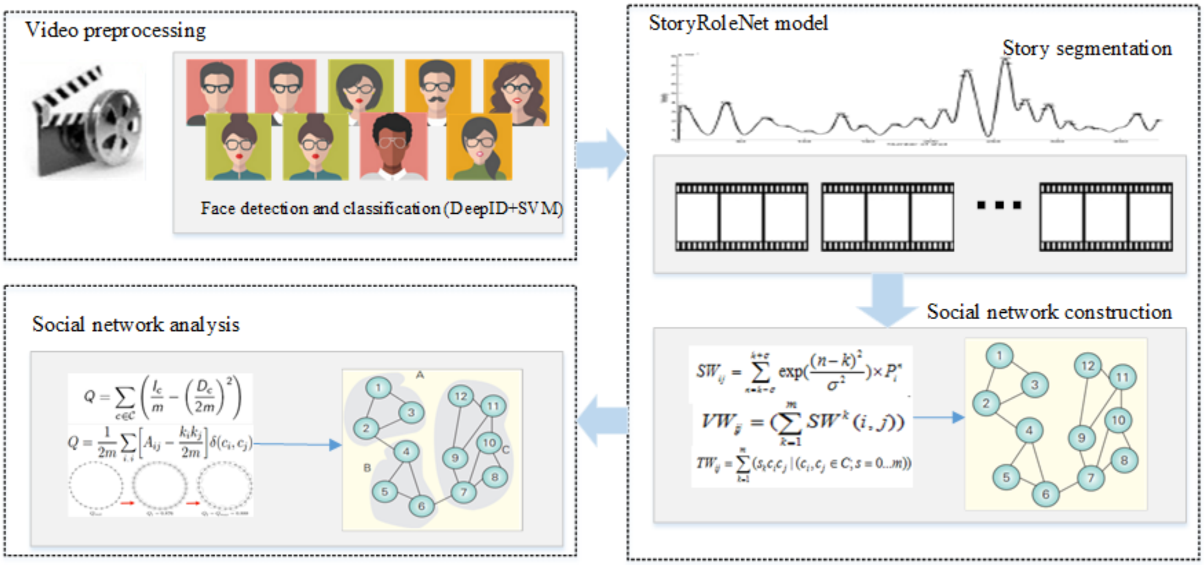

Video preprocessing
Face detection and classification (DeepID+SVM)
StoryRoleNet model
Story segmentation
Social network construction
$SW_{ij} = \sum_{n=k-e}^{k+e} \exp(\frac{(n-k)^2}{\sigma^2}) \times P_i^*$
$VW_{ij} = (\sum_{k=1}^{m} SW^k(i,j))$
$TW_{ij} = \sum_{k=1}^{m} (s_k c_i c_j \mid (c_i, c_j \in C; s = 0...m))$
Social network analysis
$Q = \sum_{c \in C} \left(\frac{l_c}{m} - \left(\frac{D_c}{2m} \right)^2 \right)$
$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$
A
B
C

FIGURE 1. Framework for constructing a social network from a video. The framework is divided into three parts: video preprocessing, the StoryRoleNet model, and social network analysis.

human face occupies only a small part of an image containing it, the detected face image is first cropped to increase the scale of the face. Third, the deep feature representation is extracted using the DeepID algorithm [25]. Last, a support vector machine (SVM) classifier is used to label the detected faces.

B. STORYROLENET MODEL: THE METHOD OF VIDEO STORY SEGMENTATION

The appearance and interrelationship of roles convey the stories in a movie or television series. The development of a storyline often reflects the interactions among the roles. Most existing story segmentation methods have been designed for news videos [26], [27], whose video and audio content differ substantially from clip to clip. However, for entertainment and social videos, story segmentation points are not as clear. Therefore, a story segmentation method is herein proposed that is based on the multi-level features of the video. The method extracts the video content hierarchically and segments the story by using a watershed algorithm.

A video's content can be described using a five-level hierarchy [28], as shown in Fig. 2. A keyframe is a frame that highlights the content of a shot; a shot is a sequence of frames that are continuously obtained by the camera; and a scene is a collection of semantically related shots that represent a meaningful story unit. In addition, a story is a video sequence composed of consecutive scenes, and a series of stories constitutes a long video.

We proposes a video story segmentation method utilizing the multi-level features of the video content [29]. The main steps are as follows: 1) Detect the shots (the bottom layer of video features) and extract the middle frame of each shot as the keyframe [30] for that shot. We extract the local and global visual features of each keyframe image and merge them to generate a visual-feature vector [31]. 2) Apply the method in [32] to detect the scene segmentation points, utilizing time-based features of the scene to detect major changes in the content. 2) Calculate the difference in the visual features between adjacent scenes, and use the watershed algorithm to segment the video into stories [3].

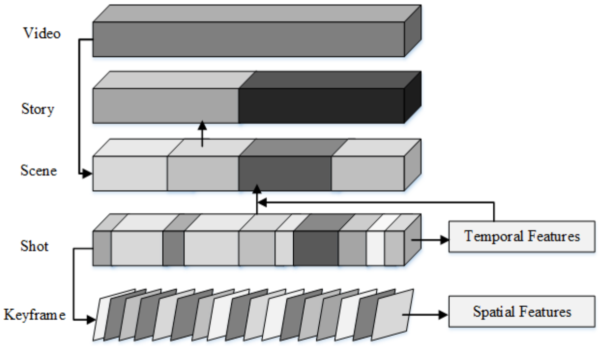


FIGURE 2. Hierarchical representation of a video.

1) EXTRACTION OF KEYFRAME FEATURES

The visual features of the keyframes are represented by the scale-invariant feature transform (SIFT) and global color name (CN). These are combined using the bag-of-visual-words (BoVW) method to generate a visual-feature dictionary for obtaining the feature vector representation of the video keyframes. The SIFT local feature descriptor includes a local feature point detector using the difference-of-Gaussian (DoG) and Hessian-affine detection methods. Each feature point is mapped to a 128-dimensional feature vector. The SIFT feature vector, denoted by BoF_{SIFT}, is obtained using K-means clustering and soft-weighting. The CN global feature descriptor is used to extract the global color feature. The keyframe image is first normalized. Then, the CN feature is extracted by dense sampling over an 8 × 8 pixel block and using a sampling step eight pixels to keep the samples from overlapping. Finally, an average feature is calculated for each sample block. Similarly, the CN feature vector, denoted by BoF_{CN}, is obtained using the K-means clustering method combined with BoVW. The captured visual features FS are obtained by combining BoF_{SIFT} and BoF_{CN} feature vectors. The pseudo-code of the keyframe segmentation visual-feature-extraction algorithm is shown as Algorithm 1.

Algorithm 1 Generations of BoF_{SIFT} and BoF_{CN}

```
Input: keyframes $\{kf_1, kf_2, \ldots, kf_n\}$
Output: features of keyframes $BoF_{SIFT}$ and $BoF_{CN}$
 Initialize each of $u \in U$
 Stage1:
 for $i = kf_1$ to $kf_n$ do
     extract $f_{SIFT}$ and $f_{CN}$ features
 end for
 Stage2:
 ($f_{SIFT}$ and $f_{CN}$ features are clustered into $U$ centers respectively)
 for each $f \in (f_{SIFT}, f_{CN})$ do
     for all $(t, f)$ do
         $bof \leftarrow 0$
         for all $p \in P$ do
             $U' \leftarrow f(U, p, k)$
             for $i = 1$ to $k$ do
                 $bof_i = bof_i + \frac{1}{2^{i-1}} sim(p, U'_i)$
             end for
         end for
     end for
     return $BoF$
 end for
 return $BoF_{SIFT}$ and $BoF_{CN}$
```

2) SCENE SEGMENTATION

A scene consists of multiple consecutive shots that are related in time. Multiple sets of shots in the same scene have similar visual features, and those from different scenes have different features. Therefore, we segment scenes based on the degree of coherence among video frames.

The similarity SS_{ij} of adjacent shots i and j is computed by cosine similarity as

$$SS_{ij} = \frac{\sum_{k=1}^{m} w_k(FS_i) \times w_k(FS_j)}{\sqrt{\sum_{k=1}^{m} w_k(FS_i)^2 \sum_{k=1}^{m} w_k(FS_j)^2}}, \tag{1}$$

where w_k denotes the feature value of the k-th dimension and m denotes the feature dimensionality.

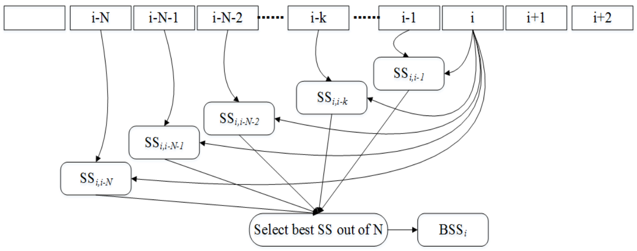


FIGURE 3. **Computation of the shot coherence.**

A scene consists of a series of consecutive shots. We use BSS_i to denote the degree of continuity between shot i and the previous shot. Its value is determined by the maximum similarity of the pairs of shots within a window of length N. We compute BSS_i as

$$BSS_i = \max_{1 \le k \le N} (SS_{i,i-k}). \tag{2}$$

Fig. 3 shows the computation of shot coherence. If the similarity between shots in the same scene is relatively high, the BSS value of the shots is correspondingly high because of the better continuity. When a new scene begins with a major difference in the visual content between the initial shot and the shot from the previous scene, the value of BSS decreases in accordance with the discontinuities between shots. We designate these shots as the scene segmentation points.

3) STORY SEGMENTATION

A story segmentation point is also a scene segmentation point. The feature of a scene is represented by the average of the features of all its keyframes. Therefore, the feature of the k-th scene, FC_k, can be denoted as

$$FC_k = \frac{1}{r} \sum_{s=1}^{r} FS_s, \tag{3}$$

where FS_s denotes the keyframe's frame feature of the s-th shot, and r denotes the number of shots in the k-th scene.

Assume there are n scenes in the video. The set of scenes boundaries is denoted by $B = b_1, b_2, \ldots, b_{n-1}$, where b_i denotes the boundary between the i-th and the $i+1$-th scenes, and B denotes the set of potential story segmentation points. The scene distance for boundary b_i is measured by the similarity of the content of the scenes on each side. A higher similarity of scenes results in a smaller distance. The feature distance between scenes is obtained by calculating the reciprocal of Equation (1) as $DC = \{d_1, d_2, \ldots, d_{n-1}\}$, where d_i denotes the distance between scenes on each side of the b_i-th boundary. The minimum and maximum scene boundary points are determined according to the distance between the scenes as

$$\begin{cases} b_i \in V. & if\ d_i < d_{i-\alpha 1} \quad and\ d_i < d_{i+\alpha 2} \\ b_i \in P. & if\ d_i > d_{i-\alpha 1} \quad and\ d_i > d_{i+\alpha 2} \\ b_i \in OT. & otherwise, \end{cases} \tag{4}$$

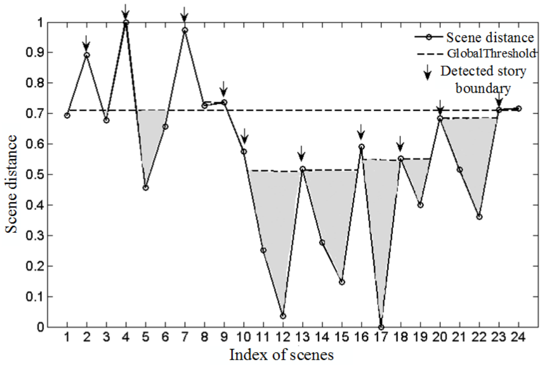


FIGURE 4. Example of a story segmentation result.

where $\alpha 1 = \min\{j|j \in \{k|(d_i - d_{i-k}) \neq 0.1 \leq k \leq i-1\}\}$, $\alpha 2 = \min\{j|j \in \{k|(d_i - d_{i+k}) \neq 0.1 \leq k \leq (n-1-i)\}\}$, and $2 <= i <= n-2$. Furthermore, V denotes the set of minimum scene boundary points, P is the set of maximum scene boundary points, and OT is the set of other scene boundary points. Clearly, $V \cup P \cup OT = B$.

Considering the global features of all scenes in the video, a global threshold GT is assigned as the average of the differences in the content of the peak points. We use SB to denote the set of story segmentation points, $SB = P \cup HF$, where P denotes the maximum point and HF denotes the set of scene boundary points that are greater than the threshold value GT. The watershed concept is analogous to fill a valley with water until the height of the water just floods the closest maximum point or the global threshold. The points midway between the horizon and the maximum points are the story segmentation points.

Fig. 4 shows an example of the story segmentation algorithm. The fifth episode of the television drama *Empresses in the Palace* is used as an example. There are 24 scenes in this video clip. The dotted line in the figure is the global threshold, and the gray part shows the value required for reaching the top of the "valley". The downward arrows in the figure represent the story segmentation points obtained by the story segmentation algorithm. Here, the video clip will be divided into 11 segments.

C. STORYROLENET MODEL: CONSTRUCTION OF THE SOCIAL RELATIONSHIP NETWORK

To build an accurate network representing the social relationship network of roles, a method should explicitly represent the interactions between roles in the video. In any video, there may be complex relationship between multiple roles. In addition, the relationships between roles constantly change as the story develops. Correctly quantifying the weights of the relationships between roles is the key to constructing a network representing the roles' social relationships. The StoryRoleNet model calculates the weights of the relationships between roles and constructs the social relationship network using specified weight threshold.

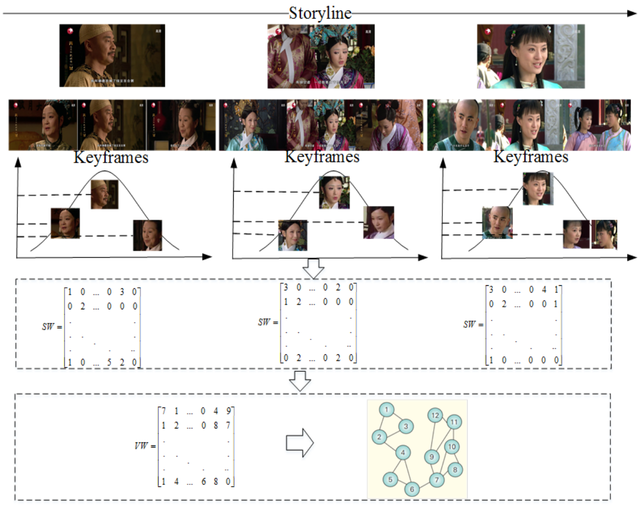


FIGURE 5. Model for extracting the relationship network from a video.

1) EXTRACTION OF ROLE RELATIONSHIP FROM VIDEO

In the method presented in this paper, the weights of relationships between roles are calculated by considering each story as a unit and the keyframe of each shot as the basis, as shown in Fig. 5. The weighted-Gaussian method is applied to determine the weights of the relationships between roles in each story unit. Then, relational weight matrix W is generated to construct the social relationship network.

First, we analyze the weights of relationships between roles in each story unit. The closer the content of the shots, the higher the probability that a relationship exists between roles that appear in them. Therefore, the weight SW_{ik} of the relationship between role i and other roles in the k-th shot are measured by Gaussian distribution $N(k, \sigma)$ as follows:

$$SW_{ik} = \sum_{n=k-\sigma}^{k+\sigma} \exp \frac{(n-k)^2}{\sigma^2} \times P_i^k \times P_j^n, \tag{5}$$

where σ denotes the standard deviation of the Gaussian distribution, that is, how the weight of the relationship between role i in shot k and role j within the shots varies over $(k-\sigma, k+\sigma)$. Here, P_i^k and P_j^n denote whether roles i and j are shown, respectively, in shots k and n. If they are shown, $P_i^k = 1$ and $P_j^n = 1$; otherwise, $P_i^k = 0$ and $P_j^n = 0$.

Therefore, the relationship matrix for roles in the story unit is represented as

$$SW(i,j) = \sum_{k=1}^{s} SW_{ik} \times SW_{kj}, \tag{6}$$

where s denotes the number of shots included in the story unit. Then, according to the weights for the role pairs in the story unit, a relational weight matrix from the video is generated as follows:

$$W_V(i,j) = \sum_{k=1}^{m} SW^k(i,j), \tag{7}$$

where m denotes the number of story segments in the video, and $SW^k(i,j)$ represents the weight of the relationship between roles i and j in the k-th story segment. Therefore, we can obtain the final relationship network extracted from the video, which is denoted as $G'_V = \langle C, E_V, W_V \rangle$.

Finally, the set of relationship between roles is determined according to the threshold as

$$G_V = \{c_i c_j = 1 | W_V(i,j) > T_v; c_i, c_j \in C\}, \tag{8}$$

where $c_i c_j = 1$ represents the existence of a connected edge between roles i and j in the unweighted network G, C denotes the set of all roles (the set of nodes of the relationship network); and T_v denotes the threshold for the relationship weights.

2) CONSTRUCTION OF SOCIAL RELATIONSHIP NETWORK FROM SUBTITLE TEXT

Owing to variations in the face detection and recognition algorithm, the relationship identified between roles may be incomplete if only analyzing the visual content of the video. The roles in a video often describe other people or the events related to them. Hence, valuable information about the relationships among roles may be contained in the subtitle text. In addition to extracting social relationship information of roles from the video content, the StoryRoleNet model extracts social relationship network from the subtitle text.

The subtitle text is first segmented according to the output from the story segmentation algorithm given in Section IV. The subtitle text is accordingly divided into multiple sets. Second, we extract the entities of each role's name using the HanLp[1] or the Stanford Named Entity Recognizer (NER) [33] tool. In this way, a dictionary of roles to be analyzed is obtained.

The relationships between roles are extracted by determining whether their names appear in the same story unit. If the names of two roles appear together in a story segment, a relationship between them is established. The number of co-occurrences defines the relational weight as follows:

$$W_T = \sum_{k=1}^{m} (s_k c_i c_j = 1 | (c_i, c_j \in C; k = 0, 1, \ldots, m)), \tag{9}$$

where $s_k c_i c_j = 1$ indicates that the names of roles i and j appear together in story segment s_k. Consequently, the initialization network is generated and is denoted as $G'_T = \langle C, E_T, W_T \rangle$. The final relationship network for the subtitle text is determined according to the threshold as follows:

$$G_T = \{c_i c_j = 1 | W_T(i,j) > T_t; c_i, c_j \in C\}. \tag{10}$$

The final social relationship network is the union of the social relationship network extracted from the video content and that extracted from the subtitle text. It is represented as follows:

$$G = G_V \cup G_T. \tag{11}$$

D. ANALYSIS USING THE SOCIAL RELATIONSHIP NETWORK

The discovery of important roles in the video and the detection of communities are important aspects of data mining using massive quantities video data and social networks [34]. To demonstrate the effectiveness of the proposed method, we also provide processes to perform community analysis and to identify important roles. When viewers watch a video, they tend to focus their attention on the relationships between the protagonists and the villains, the intimacy between different roles, and similar story elements. The roles in the video often form small communities or small gangs. As the story develops, the relationships between roles change, and the communities evolve.

In this method, an analysis for discovering communities of roles can be deployed on the constructed social relationship network. The Lovain method for community discovery [35] is applied to discover the relational communities of roles in the constructed relationship network. Modularity, which reflects the closeness of relationship between roles, is an efficient means of measuring the strength of a community structure. Modularity is defined as

$$Q = \frac{1}{2m} \sum_{i,j} [w_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j), \tag{12}$$

where w_{ij} represents the weight of the edge between i and j, $k_i = \sum_j w_{ij}$ is the sum of the weights of the edges attached to node i and c_i is the community to which node i is assigned. Furthermore, $\delta(c_i, c_j)$ is 1 if $c_i = c_j$ and 0 otherwise and $m = \frac{1}{2} \sum_{ij} w_{ij}$.

The modularity increment is defined as

$$\Delta Q = [\frac{\sum_{in} + k_{i,in}}{2m} - (\frac{\sum_{tot} + k_i}{2m})^2] - [\frac{\sum_{in}}{2m} - (\frac{\sum_{tot}}{2m})^2 - (\frac{k_i}{2m})^2], \tag{13}$$

where $\sum_{in}$ is the sum of the weights of the links inside C, $\sum_{tot}$ is the sum of the weights of the links that are incident to nodes in C and $k_{i,in}$ is the sum of the weights of the links from i to nodes in C.

The algorithm is divided into two phases. In the first phase, each of the role nodes is independent of one class. The algorithm strives to add traversed node i to the community to which a neighboring node belongs and that will result in the largest modularity increment ΔQ. This process is repeated until the communities of all nodes no longer change, which indicates that the roles have a closer relationship with the communities they have joined. Therefore, the modularity of the overall social relationship network division may have increased making community discovery more stable.

[1] http://hanlp.linrunsoft.com/

In the second phase, the communities obtained from the first phase are treated as a node, and the first phase is repeated.

The importance of a node in a social relationship network can be measured by its degree in the network. The greater its degree, the more important the node. In this method, we calculate the degree of each node in the network to discover the important people in the social relationship network. In an undirected network G, the degree of a node is the number of nodes that are directly connected to it. The degree of node i is denoted as

$$k_i = \sum_{j}^{N} a_{i,j}, \tag{14}$$

where $a_{i,j}$ denotes whether there exists a connected edge between nodes i and j and N denotes the number of role nodes in the social relationship network.

E. COMPLEXITY ANALYSIS

In this subsection, we discuss the time complexity of the core algorithms of the StoryRoleNet model. In the story segment part, the upper bound of time complexity is $O(k * n^2)$, where n and k are the number of keyframes and the length of visual words, respectively. In which, the time complexity values of scene and story segmentation are both $O(n^2)$. Especially, the most time-consuming stage is the extraction of keyframe features, which is $O(k * n^2)$. In the network construction part, the worst-case time complexity is $O((s_1 + s_2 + \ldots + s_m) * {C_n}^2)$, where s_m represents the s-th story unit, and C_n is the number of roles appearing the n-th keyframe which value usually is 1, 2, or 3. Our method can support the analysis of massive video, more importantly, the model can be parallel implemented based on the cloud computing platform, such as Hadoop and Spark [30].

V. EXPERIMENTS AND RESULTS

This section presents the experimental setup, which including the dataset description, parameter selection, methods compared, and evaluation metrics. It additionally includes analysis of the experimental results, specifically an evaluation of the work from a network construction perspective and a relationship-network-analysis perspective. Our experiment environment was comprised of a CPU of 32 Intel(R) Xeon(R) E5-2620 v4 processors running at 2.10GHz and the system was Ubuntu 16.04.

A. EXPERIMENT SETUP

1) DATASETS

The experiment datasets are list in Table 1. They include a television drama and three movies. There are many roles with complex relationships in the television series and movies. New roles and new relationship emerge as the story plots develop. Moreover, existing roles disappear and existing relationships are destroyed. The experiment analyzed the relationships among the total of 149 main roles.

TABLE 1. Datasets description.

Dataset	Name	Length of video(min)	Number of roles
TV drama	Empresses in the Palace	3040	100
Movie	Book of Love	142	17
	Forrest Gump	129	14
	Sissi	105	18

2) PARAMETER SELECTION

The parameters for the network construction are the following: (i) σ, the standard deviation of the Gaussian distribution; (ii) T_v, the weight threshold for the relationship matrix extracted from the video content; and (iii) T_t, the weight threshold for the relationship matrix extracted from the subtitle text. To determine the values of the threshold parameters, 75% of the dataset for each episode was selected and analyzed. In the result, the highest F_1 value (described below) was obtained when the parameters were set to $\sigma = 4$, $T_v = 0.7$, and $T_t = 0.4$. Therefore, we used these values in the following experiment.

3) COMPARISON METHODS

In our experiment, we compare seven methods that construct social relationship networks of roles in a video:

- **PlotNet**: This method, which is our baseline method, extracts the relationship network from the plot summary text and connect roles with a relational edge when they appear together in the same sentence.
- **RoleNet [3]**: This method utilizes the relationships between roles and shots to generate the role relational weights and thus builds a social relationship network.
- **CoCharNet [1]**: Tran *et al.* contend that roles that appear simultaneously have a closer relationship, and thus a higher relational weight and build a network based on that premise.
- **ICASSP10 [9]**: Yuan *et al.* hierarchically analyze the relationships between roles using a weighted Gaussian method.
- **StoryRoleNet(+SS [27])**: We employ a story segmentation method developed for news videos and then construct a relationship network using our StroyRoleNet model.
- **StoryRoleNet(+SS_Ours)**: This is our proposed method, excluding the use of subtitle information. By including this method in the comparison, we can better evaluate the influence of the story segmentation algorithm and the integration of subtitle text analysis on the network construction.
- **StoryRoleNet(+SS_Ours+Subtitle)**: This is the complete version of our proposed method.

4) EVALUATION INDICATORS

A standard relationship network G^* for each dataset was manually generated. For each video, we tasked three persons

TABLE 2. Comparison of F_1 values for different methods.

Methods	Empresses in the Palace			Book of Love			Forrest Gump			Sissi			Average F_1
	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	
PlotNet	0.1018	0.5111	0.0576	0.3529	0.8571	0.222	0.1905	0.3333	0.1333	0.1670	0.4213	0.1042	0.2031
RoleNet [3]	0.2166	0.1393	0.2990	0.2795	0.2581	0.2795	0.0517	0.05	0.0667	0.1481	0.1818	0.125	0.1740
CoCharNet [1]	0.3149	0.4730	0.3115	0.3868	0.7612	0.2593	0.1053	0.25	0.0667	0.5	0.4808	0.5208	0.3268
ICASSP10 [9]	0.6878	0.5891	**0.8364**	0.7451	0.7917	0.7037	0.3997	0.3333	0.5	0.2105	0.6667	0.125	0.5108
StoryRoleNe (+SS [27])	0.71	0.6621	0.7654	0.6636	0.6128	0.7235	0.5071	0.4675	0.5542	0.6248	0.665	0.5891	0.6264
StoryRoleNe (+SS_Ours)	0.7384	0.6835	0.8031	0.8549	0.8347	0.8762	0.6757	0.6356	0.7211	0.6714	0.6852	0.6583	0.7351
StoryRoleNe (+SS_Ours+Subtitle)	**0.7678**	**0.7300**	0.8307	**0.8727**	**0.8571**	**0.8889**	**0.6875**	**0.6471**	**0.7333**	**0.6947**	**0.7021**	**0.6875**	**0.7557**

with labeling the relationships between roles. Then, the video annotations for each dataset were summarized to obtain the final G^*.

We use the F_1 value to measure the similarity of the constructed relationship network G to the standard relationship network G^*. The F_1 value reflects the accuracy of the constructed network. F_1 is calculated as follows:

$$F_1(G, G^*) = 2\frac{P(G, G^*) \bullet R(G, G^*)}{P(G, G^*) + R(G, G^*)}, \tag{15}$$

where P (precision) and R (recall) are defined as

$$P(G, G^*) = \frac{|G \cap G^*|}{|G|}, \quad R(G, G^*) = \frac{|G \cap G^*|}{|G^*|}. \tag{16}$$

The community discovery method is evaluated using the NMI, which are between 0 and 1. The higher the value is, the more effectively divided the communities are. We denote the set of real communities as T^* and the set of communities divided according to the relationship network division algorithm as T. Additionally, r and k denote the number of communities in T^* and T, respectively. N is a mixed matrix, rows representing the real communities and the columns representing the found communities. N_{ij} denotes the numbers of nodes both in real community i and found community j. $N_{i\cdot}$ and $N_{\cdot j}$ is the sum of the i row and j column in N, respectively. NMI is calculated as follows:

$$NMI = \frac{-2\sum_{i=1}^{r}\sum_{j=1}^{k} N_{ij}\log\frac{N_{ij}N}{N_{i\cdot}N_{\cdot j}}}{H(T^*) + H(T)}, \tag{17}$$

where $H(T^*)$ and $H(T)$ are

$$H(T^*) = \sum\nolimits_{i=1}^{r} N_{i\cdot}\log(\frac{N_{i\cdot}}{N}), \tag{18}$$

$$H(T) = \sum\nolimits_{j=1}^{k} N_{\cdot j}\log(\frac{N_{\cdot j}}{N}). \tag{19}$$

B. EXPERIMENT RESULTS

1) ANALYSIS OF THE NETWORK CONSTRUCTION RESULTS

Table 2 compares the F_1 values, precision (P), and recall (R) of the different approaches to construct the relationship network. The results show that our method can substantially improve the results. The average F_1 value of the StoryRoleNet(+SS_Ours+Subtitle) model in the *Empresses in the Palace* dataset, a total of 76 episodes, reached 0.7678, which is higher than those of the other methods. Its performance in terms of recall is lower than that of the ICASSP10 method by 0.0057; its precision, however, is increased by 0.1409. Its F_1 value is 0.5512 higher than that of RoleNet and 0.6660 higher than that of the PlotNet model. On the other three datasets, the F_1 value of the proposed method is higher than that of the best performing among the other methods by 0.1276, 0.2878 and 0.1947, respectively.

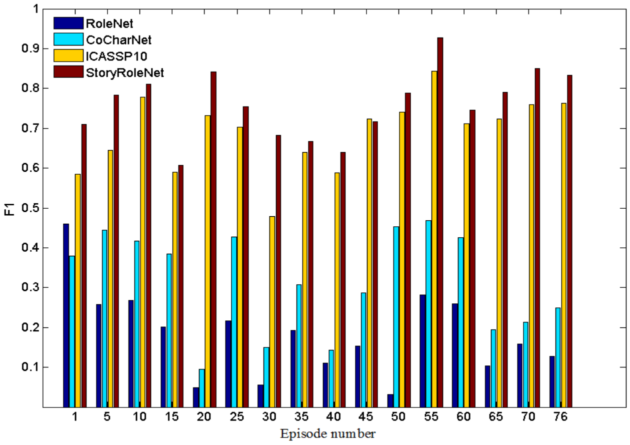


FIGURE 6. F_1 values for social networks on different episodes with different methods.

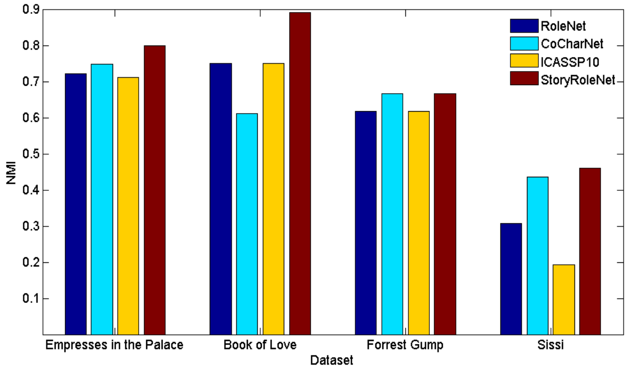


FIGURE 7. Comparison of the performance of different methods.

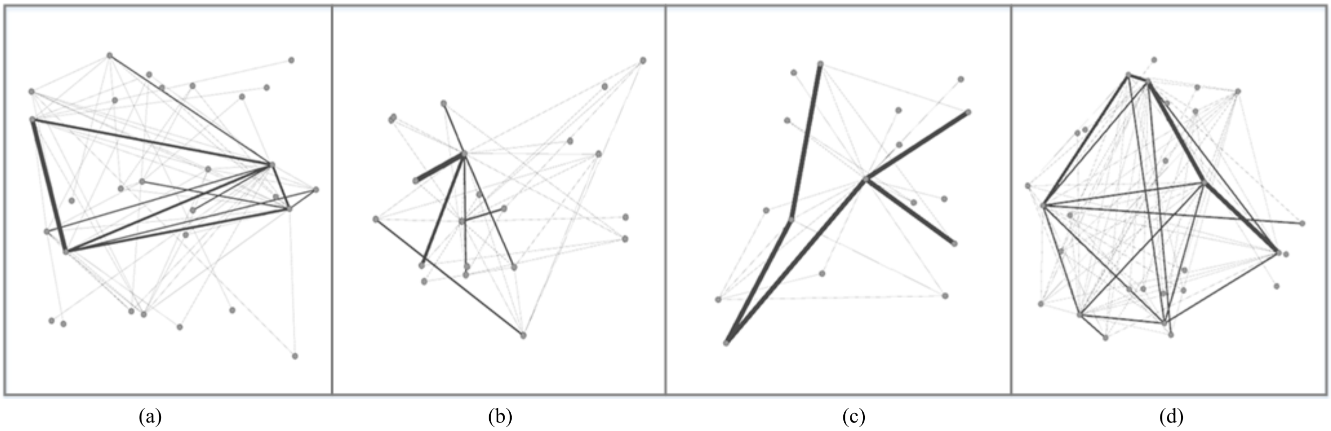


FIGURE 8. Visualization of weighted networks. (a) Empresses in the Place (Episode 1). (b) Book of Love. (c) Forrest Gump. (d) Sissi.

To demonstrate the effectiveness of utilizing the story segmentation, we compare the results of network construction with and without it. As shown by the results in Table 2, the StoryRoleNet(+SS_Ours) model obtains an average increase of about 4.4% over the ICASSP10 method on all datasets in terms of F_1 when story segmentation is used. This is because the participation of two roles in the same story unit indicate a closer social relationship between them. More importantly, the result of StoryRoleNet (+SS [27]) compared with StoryRoleNet(+SS_Ours) shows that the story segmentation method used in new videos cannot effectively process long movies. We also evaluated the performance of combining the results with the social network extracted from the subtitle text. The F_1 value of the StoryRoleNet(+SS_Ours+Subtitle) method increased by average 2.8% on all datasets over the StoryRoleNet(+SS_Ours) method. This is because the relationships from the subtitle text are further complement the relationship network. However, the average F_1 value increased by 4.0% on the *Empresses in the Palace* dataset, which is greater than that for the other datasets. This result indicates that there is more language-based communication between two roles in television dramas than in movies, therefore more detailed relationship can be mined.

Fig. 6 compares the F_1 values of the different models constructed for each episode in a collection of 16 representative episodes of *Empresses in the Palace*. The results show that the F_1 value of the StoryRoleNet model is higher than those of other methods. However, the F_1 values of our proposed model for different episodes vary. For example, the F_1 value for episode 15 is relatively low, whereas the F_1 value for episode 55 is relatively high. The probable cause is that episode 15 conveys the story of the fake pregnancy of a noble lady, Hui. The emperor, many royal consorts, and doctors are involved in each story segment. Hence, the story segmentation becomes indistinct, with complex relationships between multiple roles. In contrast, the story of episode 55 is clearly segmented, and the relationships between roles are relatively independent in each story segment.

In summary, the experimental results show that the StoryRoleNet model outperformed the existing models. The model analyzed the relationships between roles on the basis of story units to eliminate redundant calculations of the relational weights between roles, thereby achieved higher accuracy. In addition, subtitle text was combined to supplement the incomplete analysis and to address errors in the video-based detection of roles, improving the recall rate. Thus, a more accurate social relationship network was obtained, and a higher F_1 value was achieved.

2) SOCIAL RELATIONSHIP NETWORK ANALYSIS

The classic community discovery algorithm, the Louvain method, was used to determine the communities within the social relationship network of roles obtained by our proposed StoryRoleNet model and by the comparison methods. The results are shown in Fig. 7, where it can be seen that the NMI of our proposed model is higher than those of the other methods, indicating that it better characterizes the relationships between roles in the video. However, the results vary across the different datasets. On the *Book of love* dataset, for example, the NMI value is 18.9% higher than that produced by the other methods, an improvement greater than that found with the other datasets. This is because there is a major difference between roles in different story units in this dataset and therefore the plot can be explicitly segmented. For the *Forrest Gump* dataset, by contrast, the NMI value is similar to those for the other methods because the main roles in this movie appear together in most of the storylines. The NMI values for the *Sissi* are lower than for the other datasets because interactions occur among a majority of the roles in the movie, which causes community clustering to perform poorly.

Figs. 8 and 9 visualize the weighted networks constructed and the results of community and important role discoveries, respectively. In Fig. 8, the dots represent roles, and the connected edges represent the social relationship between the roles. Thicker edges indicate a higher number of interactions. In Fig. 9, the dashed lines encircle the communities of roles.

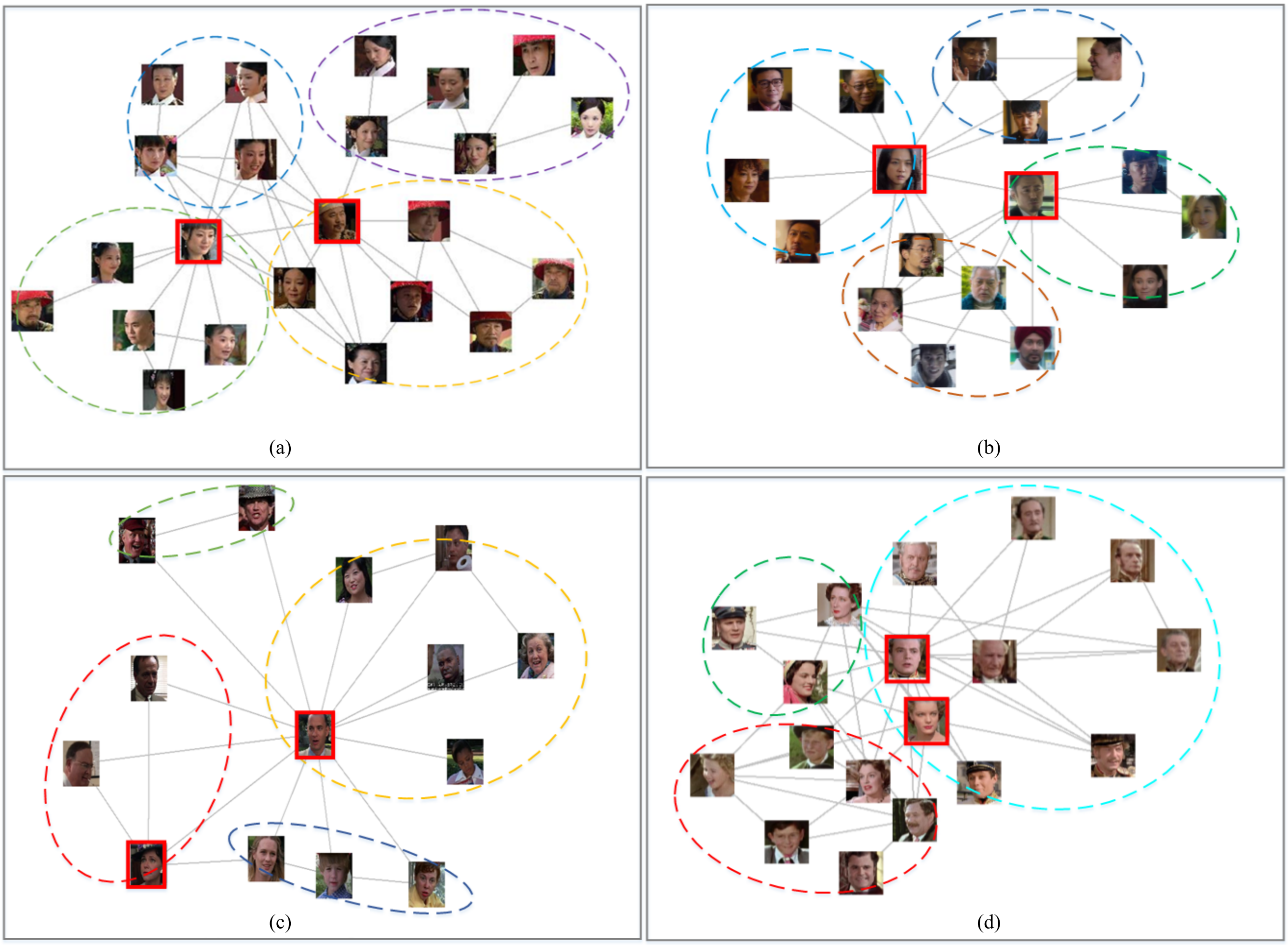


FIGURE 9. Visualization of discovery of communities and important roles. (a) Empresses in the Place (Episode 1). (b) Book of Love. (c) Forrest Gump. (d) Sissi.

In this experiment, by analyzing the degree of the network, we searched for nodes having the highest degree to determine the important roles in the given episode; the red boxes denote the two most important roles in the network.

VI. CONCLUSION

In this paper, the StoryRoleNet model was proposed as a method to construct the social relationship network of roles in a video. First, to obtain accurate weights for the relationships, we analyze the video content hierarchically and apply the watershed algorithm concept to segment the story units. Second, we calculate the weight matrix for each story unit, thereby determining a weight threshold for generating the relationship network. More importantly, the model extracts the network of roles not only from the video, but also from the subtitle text. Finally, a classic community discovery algorithm is used to divide the network into communities. In the experiments, the relationships among 149 main roles in four video datasets were analyzed. The results show that the StoryRoleNet model constructed a more accurate relationship network than similar existing methods in terms of F_1 and NMI values.

To build on the research reported in this paper, future work may involve extending the method with respect to the following perspectives. First, high-level features of video, text, audio, and other multi-source heterogeneous data can be integrated to further improve the accuracy of the relationship network construction. Second, we can analyze and mine the implicit information and knowledge in the relationship network using deep learning. Third, a parallel algorithm can be implemented to improve the speed of analyzing massive quantities of video data.

REFERENCES

[1] Q. D. Tran and J. E. Jung, "CoCharNet: Extracting social networks using character co-occurrence in movies," *J. Universal Comput. Sci.*, vol. 21, no. 6, pp. 796–815, 2015.

[2] G. Tanisik, C. Zalluhoglu, and N. Ikizler-Cinbis, "Facial descriptors for human interaction recognition in still images," *Pattern Recogn. Lett.*, vol. 73, pp. 44–51, Apr. 2016.

[3] C. Y. Weng, W. T. Chu, and J. L. Wu, "RoleNet: Movie analysis from the perspective of social networks," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 256–271, Feb. 2009.

[4] Q. Sun, B. Schiele, and M. Fritz, "A domain based approach to social relation recognition," in *Proc. IEEE CVPR*, Apr. 2017, pp. 435–444.

[5] L. Feng and B. Bhanu, "Understanding dynamic social grouping behaviors of pedestrians," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 2, pp. 317–329, Mar. 2015.

[6] C. Chen, B. Xu, Y. Xiao, Q. Shi, and W. Wang, "Extracting social network from transaction logs," *J. Comput. Res. Dev.*, vol. 52, no. 11, pp. 2508–2516, 2015.

[7] C. Peng, J. Gu, L. Qian, "Research on tree kernel-based personal relation extraction," in *Natural Language Processing and Chinese Computing* (Communications in Computer and Information Science), vol. 333, M. Zhou, G. Zhou, D. Zhao, Q. Liu, and L. Zou, Eds. Berlin, Germany: Springer, 2012, pp. 225–236.

[8] F. Li, M. Zhang, G. Fu, and D. Ji, "A neural joint model for entity and relation extraction from biomedical text," *BMC Bioinf.*, vol. 18, no. 1, Mar. 2017, Art. no. 198.

[9] K. Yuan, H. Yao, R. Ji, and X. Sun, "Mining actor correlations with hierarchical concurrence parsing," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 798–801.

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

[11] A. Sapru and H. Bourlard, "Automatic recognition of emergent social roles in small group interactions," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 746–760, May 2015.

[12] C. N. Warren, D. Shore, J. Otis, L. Wang, M. Finegold, and C. Shalizi, "Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks," *Digit. Humanities Quart.*, vol. 10, no. 3, 2016.

[13] S. Srivastava, S. Chaturvedi, and T. M. Mitchell, "Inferring interpersonal relations in narrative summaries," in *Proc. AAAI*, 2016, pp. 2807–2813.

[14] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, "Social set analysis: A set theoretical approach to big data analytics," *IEEE Access*, vol. 4, pp. 2542–2571, 2016.

[15] X. He, Y. Chen, D. Li, and Y. Hao, "A construction for social network on the basis of project cooperation," *J. Comput. Res. Dev.*, vol. 53, no. 4, pp. 776–784, 2016.

[16] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. ACM SIGKDD*, Aug. 2008, pp. 990–998.

[17] H. Han, C. Yao, Y. Fu, Y. Yu, Y. Zhang, and S. Xu, "Semantic fingerprints-based author name disambiguation in Chinese documents," *Scientometrics*, vol. 111, no. 3, pp. 1879–1896, 2017.

[18] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning social relation traits from face images," in *Proc. IEEE ICCV*, Sep. 2015, pp. 3631–3639.

[19] V. Ramanathan, B. Yao, and L. Fei-Fei, "Social role discovery in human events," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2475–2482.

[20] C. J. Nan, K. M. Kim, and B.-T. Zhang, "Social network analysis of TV drama characters via deep concept hierarchies," in *Proc. IEEE/ACM ASONAM*, Aug. 2016, pp. 831–836.

[21] L. Ding and A. Yilmaz, "Learning relations among movie characters: A social network perspective," in *Proc. IEEE ECCV*, Sep. 2010, pp. 410–423.

[22] M. Douze, H. Jégou, and C. Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 257–266, Jun. 2010.

[23] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 296–305, Feb. 2005.

[24] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3476–3483.

[25] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE CVPR*, Jun. 2014, pp. 1891–1898.

[26] X. Lu, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Broadcast news story segmentation using latent topics on data manifold," in *Proc. IEEE ICASSP*, May 2013, pp. 8465–8469.

[27] R. Tapu, B. Mocanu, and T. Zaharia, "TV news retrieval based on story segmentation and concept association," in *Proc. SITIS*, Dec. 2016, pp. 327–334.

[28] S. B. Park, K. J. Oh, and G. S. Jo, "Social network analysis in a movie using character-net," *Multimedia Tools Appl.*, vol. 59, no. 2, pp. 601–627, 2012.

[29] R. Harakawa, T. Ogawa, and M. Haseyama, "Extracting hierarchical structure of Web video groups based on sentiment-aware signed network analysis," *IEEE Access*, vol. 5, pp. 16963–16973, 2017.

[30] J. Lv, B. Wu, S. Yang, B. Jia, and P. Qiu, "Efficient large scale near-duplicate video detection base on spark," in *Proc. IEEE Big Data*, Dec. 2016, pp. 957–962.

[31] S. Tippaya, S. Sitjongsataporn, T. Tan, M. M. Khan, and K. Chamnongthai, "Multi-modal visual features-based video shot boundary detection," *IEEE Access*, vol. 5, pp. 12563–12575, 2017.

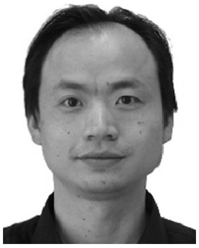
[32] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," in *Proc. IEEE CVPR*, vol. 2. Jun. 2003, pp. II-343–II-348.

[33] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proc. Int. Conf. ACL*, Jun. 2005, pp. 363–370.

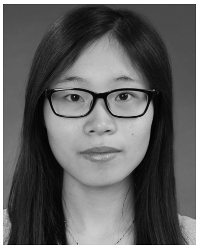
[34] J. Shen, T. Zhou, C. F. Lai, J. Li, and X. Li, "Hierarchical trust level evaluation for pervasive social networking," *IEEE Access*, vol. 5, pp. 1178–1187, 2017.

[35] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory E*, vol. 10, p. P10008, 2008.

JINNA LV received the B.E. and M.E. degrees from Zhengzhou University, Zhengzhou, China, in 2006 and 2009, respectively. She is currently pursuing the Ph.D. degree in computer science and technology with the Beijing University of Posts and Telecommunications. Her research interests include multimedia content analysis, social relation extraction, and social network analysis.



BIN WU received B.S. degree from the Beijing University of Posts and Telecommunications in 1991 and the M.S. and Ph.D. degrees from the ICT of Chinese Academic of Sciences in 1998 and 2002, respectively. He joined the Beijing University of Posts and Telecommunications as a Lecturer in 2002, where he is currently a Professor. He has published over 100 papers in refereed journals and conferences. His research interests include data mining, complex network, and cloud computing.



LILI ZHOU received the B.S. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016, where she is currently pursuing the master's degree. Her research interests include computer vision and machine learning with a focus on relation extraction from multimedia.

HAN WANG is currently pursuing the bachelor's degree in computer science with the Beijing University of Posts and Telecommunications. His research interests include deep learning and object detection.