Computational Visual Media
DOI 10.1007/s41095-015-0015-3 Vol. 1, No. 2, June 2015, 129–141

Research Article

Semantic movie summarization based on string of IE-RoleNets

Wen Qu[1], Yifei Zhang[1,2], Daling Wang[1,2], Shi Feng[1,2], and Ge Yu[1,2] (✉)

Abstract Roles, their emotion, and interactions between them are three key elements for semantic content understanding of movies. In this paper, we proposed a novel movie summarization method to capture the semantic content in movies based on a string of IE-RoleNets. An IE-RoleNet (interaction and emotion rolenet) models the emotion and interactions of roles in a shot of the movie. The whole movie is represented as a string of IE-RoleNets. Summarization of a movie is transformed into finding an optimal substring with user-specified summarization ratio. Hierarchical substring mining is conducted to find an optimal substring of the whole movie. We have conducted objective and subjective experiments on our method. Experimental results show the ability of our method to capture the semantic content of movies.

Keywords movie summarization; content analysis; movie understanding

1 Introduction

Due to the proliferation of movie videos on websites, it is hard for users to find interesting movies. Most users turn to web applications, such as movie recommendation and movie retrieval, for help. But these applications still provide hundreds and thousands of hours of video content to users. Therefore, there is a strong demand for a mechanism that allows users to gain an overview of a movie without watching the entire video. Movie abstraction techniques facilitate the browsing and navigation of a movie collection for users. Using movie abstraction as a preview, the user can quickly access each movie and evaluate if it is important, interesting, or enjoyable.

Compared to other kind of videos (e.g., news, sports video, etc.), movies have a large variety of scenarios and styles, which makes it challenging to summarize them. Generally, movie abstraction can be grouped into two categories based on movie key frames and movie skims [1]. Key frame abstraction selects several key frames of a movie to represent the content of the whole movie. Movie skims select interesting excerpts of a movie to provide a shortened video to represent the movie. Key frame based abstraction is effective for content indexing and retrieval. But from the viewpoint of user, movie skims are more vivid and informative, keeping audio and video content of movies. In this paper, we focus on the problem of extracting semantic movie skims from movies.

Movie skimming aims to select the important video segments from movies. To identify important content in movies, low-level features, such as color, texture, and motion features are used in existing work. Audio or visual saliency [2, 3] is detected to find interesting information in temporal and visual space. Since low-level feature based movie content analysis fails to capture the semantics of movie content, recent research has explored high-level features, at the role-level [4–6] and affective-level [7], for movie summarization.

In our daily experience, the audience of movies pays attention to how characters relate to one another, the interactions between characters, and the emotional development of characters. Thus, shots that include personal interaction and emotions are

1 School of Information Science and Engineering, Northeastern University, Shenyang 110819, China. E-mail: quwen@research.neu.edu.cn.

2 Key Laboratory of Medical Image Computing (Northeastern University), Ministry of Education, Shenyang 110819, China. E-mail: zhangyifei@mail.neu.edu.cn, wangdaling@mail.neu.edu.cn, fengshi@mail.neu.edu.cn, yuge@mail.neu.edu.cn (✉).

Manuscript received: 2015-02-06; accepted: 2015-05-01

always the highlights of a movie. Such content plays an important role in promoting the development of the storyline of a movie. Previous role-based approaches only model co-occurrence relationships of roles. Interactions between them, such as hugs and handshakes, have the same weight during the modeling process. However, the emotions in roles are not considered in existing role-based methods, but they are a key factor in movie content understanding. In this paper, we extract information of roles, their emotions, and interactions between them as the semantic content of thestoryline. We described roles' relationships in an IE-RoleNet, and modeled dynamic development of the storyline using a string of IE-RoleNets.

Existing methods mostly provide structure-level summarization in which the segments in skim videos have the same order as the original videos. Such summarization fails to satisfy the demands of users who are only interested in some special roles in a movie. Our method provides both structure-level and role-level summarization. We extract and organize the movie skim according to the dramatic structure. The main contributions of our work are: (1) we explore both interaction and emotional cues of movie characters, and propose the IE-RoleNet as a means of modeling the semantic information in movies; (2) we use a string of IE-RoleNets to model the dynamic storyline of a movie, and to extract summarization, by mining optimal substrings from the string of IE-RoleNets; (3) we provide both structure-level and role-level summarization.

The rest of this paper is organized as follows. In Section 2, we review related works on video and movie abstraction. In Section 3, we define IE-RoleNets, strings of IE-RoleNets, and their construction. Section 4 details summarization of movies based on the string of IE-RoleNets. Experiments are illustrated in Section 5 followed by conclusions in Section 6.

2 Related work

First, we review abstraction techniques for generic videos, then review state-of-the-art techniques for film skimming.

Video skims have been produced for the various kinds of video: sports, news, documentaries, movies, music videos [8], and home videos. Sports video summarization focuses on selecting highlights of sports, such as shots at basketball and goals in football. Babaguchi et al. [9] detected significant events by matching text in the image frame with game statistics. Takahashi et al. [10] temporally compressed video data to generate summaries of large sports video archives. Chen et al. [11] proposed a hybrid summarization framework that combines adaptive fast-forwarding and content truncation to generate summaries for soccer videos and surveillance videos. Audio features are often used in sports video analysis to detect applause and cheering. News videos usually consist of an anchor person and story units, giving them a fixed structure. Ide et al. [12] used the character of news videos to compose video stories on the topic thread. User-generated videos have also been widely investigated. MyVideos [13] is a prototype system for managing home videos, providing video segmentation, summarization, and editing. Zhao et al. [14] used both audio and video content for home video abstraction, selecting video segments containing special audio features, mainly based on speech, laughter, song, and applause. Qiu et al. [15] considered spatial–temporal characteristics to propose a new video attention analysis method for home video. Lee et al. [16] presented an egocentric video summarization approach focused on the most important objects and people that interact with the camera operator. All such videos have particular characters with respect to either content or structure. Movies are different from other kinds of video in having a large variety of scenarios and styles, but they also have particular characteristics, focusing on stringing several scenes together to tell a story.

Previous works on movie skim summarization can be broadly classified as being low-level feature based, or semantic feature based. Low-level feature based methods use video content such as visual and audio saliency to select extracts. Chen et al. [17] used audio features for story unit extraction, and proposed an action movie segmentation and summarization framework based on movie tempo. Chen et al. [18] utilized four high-level concepts: who, what, where, and when to describe video content, and explored the associations between them to detect meaningful shots and relations. Zhu

et al. [19] incorporated movie scripts into movie content analysis, and conducted storyline analysis to generate movie abstractions. Ren et al. [20] used Wikipedia as a middle layer to bridge the descriptive gap between general user statements and movie subtitles.

Since role-based or character-based methods have an advantage in understanding semantic relationships between roles, many works [4–6, 18] have analyzed movie content based on roles or characters. Scene analysis and character interaction features were used in Ref. [6] to summarize movies. Tsai et al. [5] proposed a two-stage scene-based movie summarization method using a role-community network. Existing role-based methods only model simple relationships between roles. For example, Refs. [4, 5] analysed the co-occurrence of roles while Ref. [6] explored dialog relationships. Furthermore, they focused on modeling roles in movies as characters in a social network, which ignored the dynamic development of roles' relationships over time. A movie is modeled as a graph of unique role communities (urcs) in Ref. [5], where each urc is a set of unique roles. Weng et al. [4] described roles' relationships within a role's social network. These representations of movies are static and global, which is more suitable for leading role determination and community identification.

3 IE-RoleNet based movie content analysis

In this section, the process of movie content analysis is elaborated. Figure 1 illustrates the modeling process of our method. A movie is segmented into shots, and each shot is represented as an IE-RoleNet. Then the movie is represented as a string of IE-RoleNets. Finally, string-based summarization is applied to generate movie skims.

3.1 Definition of IE-RoleNet

RoleNets were proposed in Ref. [4] to describe the relationship between roles in movies. The roles are analogous to users in social network analysis. In a RoleNet, two roles are considered to have a relationship if they appear in the same scene. The strength of the relationship is quantified by the number of co-occurrences between roles.

RoleNet based movie content analysis has two limits: firstly, it fails to reveal and describe fine-grained relationships between roles. Two roles appearing in the same scene have the same relationship no matter whether they are close or not. Secondly, it does not take the dynamics of roles' relationships into account. As the storyline proceeds, the relationship between roles changes. A RoleNet only describes the static and global relationships between roles in a movie.

We propose as an alternative the interaction and emotion RoleNet (IE-RoleNet). Compared to a RoleNet, an IE-RoleNet differs in three ways. It adds emotional information about roles, it models the strength of interactions in a fine-grain way, and it describes the roles' relationships at shot-level rather than at whole movie-level. To incorporate these characters, an IE-RoleNet is defined as follows.

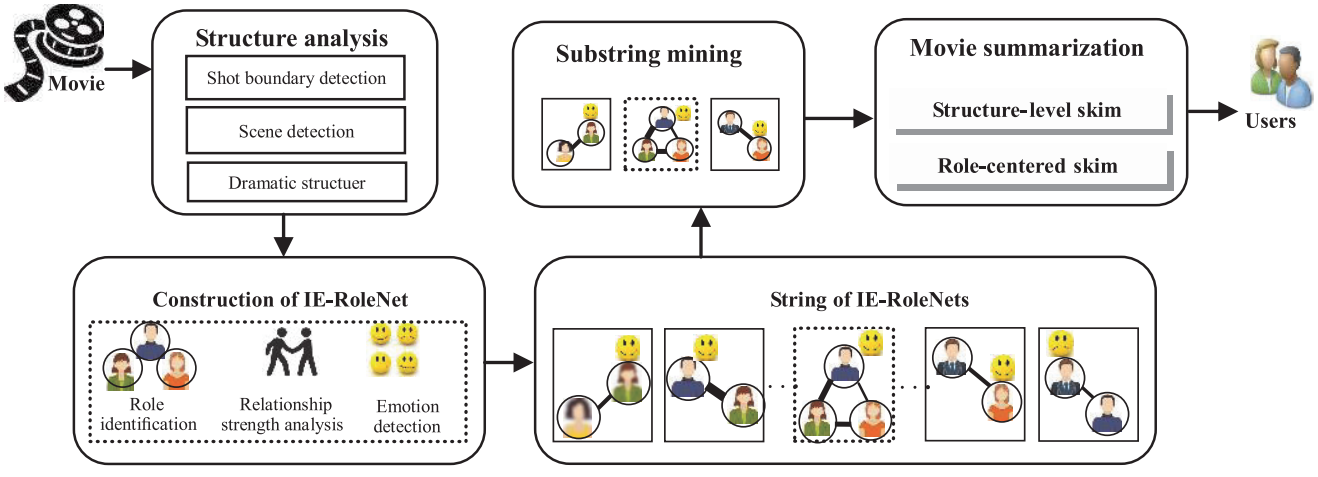


Fig. 1 Movie summarization based on a string of IE-RoleNets.

Definition: an IE-RoleNet is a weighted graph represented by

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

where $V = \{v_1, \cdots, v_m\}$ represents the set of roles in a movie shot, $E = \{e_{ij} | \text{ if } v_i \text{ and } v_j \text{ have interaction}\}$ indicates an interaction between roles, A is the emotion of the speaker role in the shot taken from the set {happiness, sadness, fear, anger, disgust, surprise, natural}, and W represents the weight of interaction between roles, from 1 (co-occurrence) to 3 (close interaction).

3.2 Construction of an IE-RoleNet

3.2.1 Structure analysis

The structure of a typical movie can be expressed as a hierarchy: frame, shot, scene, and movie. A shot is a group of consecutive frames recorded with the same background and a single camera. It is the basic unit of a video. A movie sequence that alternates between views of two people consists of multiple shots. A scene is a series of adjoining shots that are semantically correlated. For example, three camera shots showing three different people talking might be one scene if they are talking together in the same room. Compared with shots and scenes, the storyline of a movie is based on higher-level concepts, characterised by semantic analysis and content understanding. One of the most common and universal approaches to dramatic structure analysis is based on "three act structure" [21]. This divides a movie into three distinct parts: the beginning (exposition), the middle (conflict), and the end (resolution). The exposition establishes the main characters, their relationships, and the world they live in. The story then develops via a series of conflicts between the protagonists, which is known as the first turning point. In the second part, the protagonists attempt to resolve the problem initiated by the first turning point. Finally, the story's climax, in the third part, provides the resolution. Figure 2 shows the hierarchical structure of movies.

In our method, we construct an IE-RoleNet at shot-level. The work in Refs. [2, 4] constructed RoleNets at the scene-level. They focused on analyzing the social network between roles, ignoring the temporal dynamics of movies, while we aim to model the dynamic development of the movie. We detect shot boundaries using a motion estimation

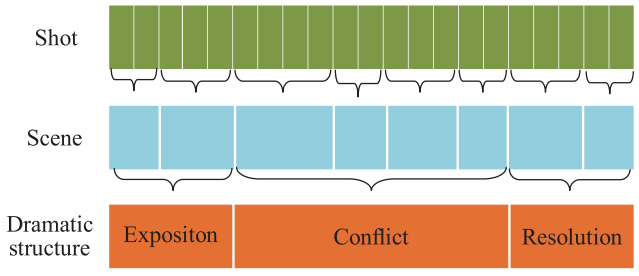


Fig. 2 Structural movie analysis.

based method [22]. The difference between the reconstructed frame and the original frame is computed as the shot boundary detection metric. After the shots have been detected, they are clustered into scenes as in Ref. [23]. To avoid segmentation during speech, we detect pauses in speech and select shot boundaries in the speech pauses.

Based on shots and scenes, we analyse the dramatic structure of the movie and use them to explore summarization. Since the roles and their relationships are built up during the exposition, we extracted the initial part of the movie that includes all the roles within the minimal number of scenes as the exposition part. The last scene with emotions to the end of the movie is viewed as the resolution. The middle scenes between these two parts form the conflict part.

3.2.2 Role identification and tracking

Existing works which identify roles in movies can use supervised or unsupervised methods. We use an unsupervised approach because the supervised method assumes that all roles appear in the preceding part of the movie. Since unsupervised methods using only visual information are sensitive to the variation of characters in scale and pose, we combine visual and audio information together to identify roles. First, we employ the facial detector proposed in Ref. [24]; it can detect both frontal faces and profile views. We approach the problem using tracking-by-detection. Specifically, we initialize tracks by scanning the whole image every fifth frame. The Fisherface [25] of each track is extracted as facial features for face clustering. Following the method in Ref. [26], we extract and track the mouth region of the roles to identify the speaker in the shot. The audio is first down-sampled and split into overlapping frames of 20 ms. For each frame, Mel frequency cepstral coefficients [27] are

computed. The facial features and speech features are used together to cluster roles, where each cluster corresponds to a role in the movie. Faces without speech features are assigned to the nearest roles in facial feature space.

3.2.3 Interaction atom based relationship analysis

Many interactions reveal the relationship between roles. For example, a kiss or hug indicates familiarity, love, or affection. There is much research on human action detection and recognition in Hollywood movies and TV shows [28, 29]. These methods aim to recognize the categories of actions. In our work, we are interested in whether there are interactions between roles and the degree of interactions. We adopt a coarse to fine strategy to localize the interactions in movies. Our action detection approach consists of the following three steps.

Given an input shot, the person area is first detected and tracked, using a state-of-the-art person detector [30] which provides person bounding boxes in static video frames. Detections in different frames are connected using a KLT point tracker. Any isolated detections within a temporal window (empirically set to 15 frames) are discarded.

Secondly, the spatial relations of roles are modeled using structure learning. The spatial relation between two persons is divided into discrete sets following [28]: overlap, adjacent left, adjacent right, near left, near right, and far. The relation between each pair of persons is computed and tracked during each shot. Shots with zero or one person are filtered out as interactions need at least two persons.

Finally, pre-defined interaction atoms are detected from shots. Inspired by the action atoms in Ref. [31], we model the interactions between roles in terms of three atoms: person–person together, person–person interacting, and person–person overlaping. The weights given to them respectively are $\{1, 2, 3\}$. The atom person–person together means co-occurrence of roles. Person–person interacting means two roles have interacting bodies but the centers of the bodies do not overlap. "Person–person overlaping" is the interaction with the largest weight, which indicates the centers of roles' bodies overlap. The presences of atoms are evaluated by tracking the spatial relations between persons through the shots. Figure 3 shows instances of these interaction atoms.

3.2.4 Emotion detection for speakers

Tracking emotions of major roles allows capture of the development of the storyline in a movie. Main cues used for emotion recognition include facial expressions (FER), body gestures, speech acoustics [32], and lexical cues. We combine facial features and audio features to improve the emotion detection, as do Refs. [7, 33]. The effect is described in terms of discrete categories, as conceptualized in psychological research [34]. We use the most popular basic emotion categories as the description: happiness, sadness, fear, anger, disgust, surprise, and natural. The emotion of speakers in each shot is detected, and we set the value of A to one of the above emotion categories.

Emotion detection based on visual cues. Given a shot, facial feature points of speakers are extracted following Ref. [24]. To avoid errors due to variation in scale of the face, features are normalized as in Ref. [35]. We use facial animation parameters (FAPs) [36] as the features to train. To normalize between different people, FAPs are expressed in terms of FAPUs. Details of how to compute FAPs are given in Ref. [36]. Because frontal faces and profile faces have different feature points, we trained frontal faces and profile faces independently. Given a role face in the shot, we first determine whether it is frontal or in profile. We trained seven one-to-rest SVM classifiers for frontal faces and profile faces separately.

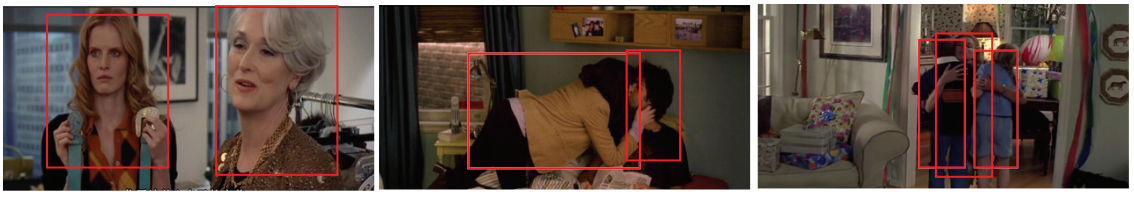


Fig. 3 Instances of interaction atoms person–person together, person–person interact, and person–person overlap in the movies The Devil Wears Prada and The Butterfly Effect.

Emotion detection based on audio cues. Emotion detection based on audio cues considers both speech and background music. First, audio of a shot clip is classified into music and speech [37]. Then affective related features are extracted from music and speech respectively. Most existing approaches to vocal affect recognition use multiple acoustic features. The popular features are prosodic features and spectral features. For each speech audio segment, 4 features and their respective statistics are extracted. In particular, pitch, loudness of speech (energy), Mel frequency cepstral coefficients (MFCC), and zero crossing rate are extracted. We used the Belfast emotion dataset collected by Cowie et al. [38, 39] as the training dataset. For music audio, we extract pitch, Mel frequency cepstral coefficients (MFCC), short-term energy, timbre, rhythm, and zero crossing frequency.

4 Movie summarization method

After a movie has been represented as a string of IE-RoleNets, the problem of extracting summarization from a movie can be represented as selecting a substring which maximally preserves the content of the movie given a user-specified summarization ratio.

4.1 Problem formulation

Give a movie M contains t shots, represented by a string of IE-RoleNets $M = \{G_1, \cdots, G_t\}$, the summarization problem is to select a set of IE-RoleNets to cover the information in the string of IE-RoleNet given the length constraint. Our method selects the substring having the minimum distance from the IE-RoleNet string. This is formulated as the following optimization problem:

$$ms = \arg\min \{D, (M, R)\}, \text{ s.t. } |ms| = \sigma|M| \quad (1)$$

where R denotes a substring of $\{G_1, \cdots, G_t\}$ and ms is the movie skim generated based on R. D is the distance between the strings of two IE-RoleNets. $|ms|$ and $|M|$ are the lengths of the movie skim and the original movie in terms of time. σ is the summarization ratio specified by the user, which controls the length of the generated movie skim.

4.2 Substring mining

To find an optimal substring from M, we need to define the distance between substring R and M. We first define matching and distances between two IE-RoleNets. We find the optimal substring using dynamic time warping distance.

4.2.1 Matching between two IE-RoleNets

Given two IE-RoleNets G and G', the similarity measure includes the distance between nodes, edges, and emotion. Let a movie have p roles. The role sets V and V' can be represented as vectors of length p, in which the i-th element is 1 if the corresponding role appears in the shot and 0 otherwise. The matrices W and W' of size $p \times p$ correspond to the interaction weight between each pair of roles in the shots. The distance of two IE-RoleNets is measured by

$$D(G, G') = D_n(V, V') + D_w(W, W') + D_\varepsilon(A, A')\delta(G, G') \quad (2)$$

where D_n is the Manhattan distance between two role sets, and D_w is the similarity of two interaction weight matrices. D_ε is the distance between the emotion vectors of the two IE-RoleNets, where $\delta(G, G')$ is 1 if the speakers of two IE-RoleNets are the same.

4.2.2 Mining the optimal substring

Suppose we have a string $S = g_1, \cdots, g_n$. We attempt to find a substring $S' \subset S$, which satisfies Formula (1). We use dynamic time warping (DTW) [40] as the distance measure D between strings.

First, we initialize S' to be the same as S. A matrix of n rows and n columns gives the distance between the elements of two strings, and the optimal warping path between them is the diagonal path as shown in Fig. 4(a). Each time we select an element g_i in S' for deletion, using the criterion:

$$g_i^* = \arg\min_{g_i}(D(S, S'/g_i)) \quad (3)$$

where S'/g_i is the string after deleting g_i from S'. There are two situations: either the neighbors of g_i are not deleted before this step, as illustrated in Fig. 4(b), or the elements around g_i are deleted before this step, as shown in Fig. 4(c). The warping path changes between substring $\{g_{i-j}, g_{i+k}\}$ and $\{g_{i-j}, g_{i-j+1}, \cdots, g_i, \cdots, g_{i+k-1}, g_{i+k}\}$, where j and k are the numbers of elements deleted before and after g_i, respectively. The problem of selecting g_i is the same as finding the element in S' which satisfies:

$$g_i^* = \arg\min_{g_i}(\text{DTW}(\{g_{i-j}, g_{j+k}\}, \{g_{i-j}, \cdots, g_i, \cdots, g_{i+k}\})) \quad (4)$$

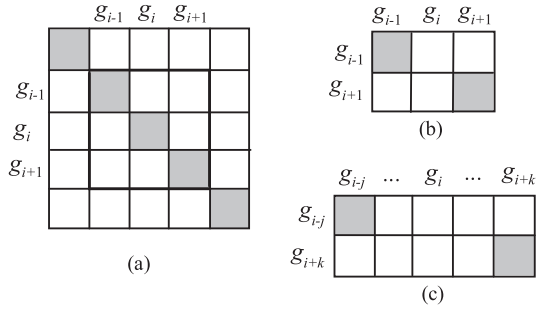


Fig. 4 Deleting an element from the string of an IE-RoleNet.

where DTW is the dynamic time warping distance between two strings. The process of deleting element is repeated until the length of S' satisfies the user-specified summarization ratio. The process is shown in Algorithm 1.

4.3 String of IE-RoleNets

A movie video M consisting of t shots is represented as $M = \{S_1, \cdots, S_t\}$ where the indices denote successive times. For each shot S_i, a graphical description IE-RoleNet is used to describe the collection of roles, their interactions and emotions, denoted as G_i. This representation of a video naturally leads to a string representation $M = \{G_1, \cdots, G_t\}$, where the IE-RoleNets follow the temporal order of the shots. The process is illustrated in Fig. 1.

5 Movie skim construction

The mined substring gives a compact representation of the movie, but it is still not a skim of the movie. Since each element in the substring corresponds to one shot in the video, we producde the video skim using the shots in the substring. The audio and video information of the shots are assembled in synchrony according to their time order.

Our method provides skim visualization in two ways: structure-level skim and role-centered skim. For the first kind of skim, video excerpts added to the skim are concatenated following the temporal order of the original movie. Role centered skims allow browsing of the movie from the perspective of roles. The algorithm of movie summarization is given in Algorithm 2.

Algorithm 1: IE-RoleNet substring mining
Input: String of IE-RoleNets $S=\{g_1,\cdots, g_n\}$.
Output: SubString S'.
1) Initialization $S'=S$, $j=1$, $k=1$
2) Each time we select an element g_i in the S' to delete using the criterion
3) **while** $videoLength(S')>\sigma \cdot videoLength(S)$ **do**
4) **for** $i=1$ to n **do**
5) $D_i=$ DTW$(\{g_{i-j}, g_{i-k}\}, \{g_{i-j}, g_{i-j+1},\cdots, g_i,\cdots, g_{i+k-1}, g_{i+k}\})$
6) $g^*=\arg\min(D_i)$
7) $S'=S/g^*$

5.1 Structure-level movie summarization

A structure-based skim organizes the selected video segments in their original temporal order. Most existing summarization systems provide movie skims of this kind.

Since a movie usually has a large number of shots, there is correspondingly a long string of IE-RoleNets. It is time consuming to find the optimal substring from the string directly. We employ a hierarchical movie summarization method. First, substring mining is conducted on each scene. Then the strings of scenes are concentrated according to the three act structures. Substring mining is conducted on each act. Finally, we mine substrings on the string of the whole movie formed from the substrings of the three acts.

5.1.1 Scene-level summarization

Many shots include only one person, particularly dialog shots. Correpsonding elements in the string of IE-RoleNets have only one node. Therefore, we first conduct substring mining on scene level to reduce the redundancy.

Algorithm 2: IE-RoleNet substring mining
Input: Movie video M.
Output: Structure-level skim $ms_{\text{structure}}$, role centered skim ms_{role}.
1) **IE-RoleNet construction from movie**
2) $G_1,\cdots,G_t=IERoleNetConstruction(M)$
3) **Structure-level skim construction**
4) **for** each scene i of M **do**
5) $S_i=sceneSubstring(G_1,\cdots,G_t)$
6) $R_i=substringMining(S_i)$
7) $M^{\text{scene}}=M^{\text{scene}}\cup R_i$
8) **for** each storyline part of M **do**
9) $S_i=storylineSubstring(M^{\text{scene}})$
10) $R_i=substringMining(S_i)$
11) $M^{\text{storyline}}=M^{\text{storyline}}\cup R_i$
12) $R=substringMining(M^{\text{storyline}})$
13) $ms_{\text{structure}}=SkimConstruction(R)$
14) **Role centered skim construction**
15) IE-RoleNetsGroup $(\{G_1,\cdots,G_t\}, \text{group}_{\text{type}})$
16) **for** each role **do**
17) $S_i=roleSubstring(G_1,\cdots,G_t)$
18) $R_i=substringMining(S_i)$
19) $M^{\text{role}}=M^{\text{role}}\cup R_i$
20) $ms_{\text{role}\,i}=StructureLevelSkimConstruction(M^{\text{role}})$

Given an IE-RoleNet string for movie M, it is cut into multiple substrings according to scene boundaries. The movie is represented as $M^{\text{scene}} = \{S_1^L, \cdots, S_n^L\}$, where $L =$ scene and n is the number of scenes. S_i^L is the string of IE-RoleNets for the i-th scene. We find the optimal substring R_i^L for each scene using the process in Section 4.2.2. The summarization ratio for each scene S_i^L is $\alpha_i k_{\text{scene}} \sigma$, where k_{scene} is the summarization weight for the scene and α_i is the summarization weight for the i-th scene. The higher the weight, the more compact the skim is.

5.1.2 Dramatic structure-level summarization

After get substrings at scene level, the movie is represented as $M^{\text{scene}} = \{R_1^L, \cdots, R_n^L\}$. The movie is represented as $M^{\text{storyline}} = \{S_1^L, S_2^L, S_3^L\}$, where $L =$ storyline; the movie is divided into three acts or parts. Then, for the string of IE-RoleNets for each part S_i^L, we find the optimal substring using the process in Section 4.2.2. The summarization ratio for each part S_i^L can be set according to user preference. For example, if a user wants to pay more attention to the outcome and is not interested in the character setting, the weight can be set higher for the third act and lower for the first act.

5.1.3 Movie-level summarization

Finally, the substring at the movie level can be mined from the strings at the dramatic structure level, denoted by $M = \{R_1^L, \cdots, R_n^L\}$, where $L =$ storyline. We find the optimal substring for the whole string again following the process in Section 4.2.2. Video clips of each shot are concatenated according to their temporal sequence in the original movie.

5.2 Role centered movie summarization

The role centered skim organizes the video segments accroding to roles. The string of movie M is split into multiple substrings corresponding to roles in the movie. Then, for each substring, movie summarization is conducted following structure-level movie summarization.

5.2.1 Role centered IE-RoleNet grouping

To construct role centered movie summarization, IE-RoleNets are first grouped according to the roles. Three kinds of grouping strategy are used in our approach.

The first strategy uses the occurrence of roles as the metric. An IE-RoleNet is assigned to the roles that have appeared in the shot. When a shot has multiple roles, it is assigned to multiple role categories. The second strategy uses interaction as the metric. An IE-RoleNet is assigned into role categories that meet two conditions: the role has appeared in the shot and the role has interactions with other roles. The third strategy assigns an IE-RoleNet into role categories according to the speaker in the shot.

5.2.2 Role centered substring mining

Given the substring of each role S_i^{role}, the movie is represented as $M^{\text{role}} = \{S_1^L, \cdots, S_p^L\}$, where $L =$ role and p is the number of roles in the movie. For each S_i^{role}, we conduct substring mining and obtain the corresponding substring R_i^{role}. The skim ratio σ for each role is set by the user. Using each R_i^{role}, the movie video can be summarized into p video segments.

6 Experiments

To evaluate the performance of the proposed movie summarization method, we conducted objective and subjective tests on Hollywood movies. We selected 11 popular movies as the dataset, as given in Table 1. They include multiple movie genres, such as action, comedy, and so on.

We compared our approach with two other summarization methods based on role analysis: scene-level summarization [5] and the RoleNet based method [6]. During the experiments, all methods used the same skim ratio σ, with values of $\sigma = 10\%, 20\%,$ and 30%. Since there is no standard evaluation framework in video summarization research, we evaluated the performance of the proposed movie summarization method following existing methods. Differing experimental methods exist [41], based on result descriptions, objective metrics, and user studies. Because result descriptions are subjective and hard to justify, we used objective metrics and user studies in our paper.

6.1 Information retrieval metrics

Since objective ground-truth for movie skim is lacking, it is difficult to evaluate the correctness of a video abstract. But performance evaluation

Table 1 Test movies

ID	Movie title	Genre	Length
1	The Lord of the Rings: The Fellowship of the Ring	Action/Adventure/Fantasy	178 min
2	Harry Potter and the Philosopher's Stone	Advanture/Family/Fantasy	152 min
3	The Devil Wears Prada	Comedy/Drama/Romance	109 min
4	Evan Almighty	Comedy/Family/Fantasy	96 min
5	Ratatouille	Animation/Comedy/Family	111 min
6	Up	Animation/Adventure/Drama	96 min
7	The Pursuit of Happiness	Biography/Drama	117 min
8	The Hurt Locker	Drama/Thriller/War	131 min
9	The Shawshank Redemption	Crime/Drama	142 min
10	The Da Vinci Code	Mystery/Thriller	149 min
11	The Butterfly Effect	Sci-fi/Thriller	113 min

for skimming based on the detection of interesting/important segments can be carried out via common information retrieval metrics. In the proposed method, we detect interactions and emotional roles, which can be viewed as important segments. To evaluate the interaction and emotion coverage of the produced summarization, we use information retrieval metrics of precision and recall as objective metrics. Interactions and emotional roles were annotated with five subjects. The subjects were given the same movie videos and asked to annotate ground-truth interaction and emotions. If more than three subjects' annotations overlaped, the average length and center of the points were considered to provide ground-truth. A detected interaction and emotion is claimed as correct if its temporal extension overlaps with any ground-truth annotation. We generated summarization for the movies in the dataset with $\sigma = 10\%, 20\%, 30\%$, and computed recall of the significant segment detections in summarizations. Recall is the proportion of interaction segments included in the summarization skimming. Table 2 gives the interaction coverage and emotion coverage of the results with summary lengths from 30% to 10%. The coverage rate deceases with summary length. This observation shows that our method is effective in keeping interaction and emotion contents of roles.

6.2 User studies

We invited twenty subjects to participate in the subjective tests in our experiments. They were divided into two groups, and two kinds of user studies were conducted, a survey, and one on user-specific interest shot selection.

6.3 Questionnaire

To measure the user satisfaction with summarization, we showed the original movie and the movie skims generated using user-specified summarization ratios. After watching the original movie and the movie skim, the subjects are asked to answer the questions below:

Table 2 Coverage of interaction and emotion segments for different summarization ratios

ID	Interaction coverage			Emotion coverage		
	$\sigma = 10\%$	$\sigma = 20\%$	$\sigma = 30\%$	$\sigma = 10\%$	$\sigma = 20\%$	$\sigma = 30\%$
1	0.420	0.482	0.512	0.509	0.602	0.652
2	0.162	0.204	0.372	0.692	0.692	0.723
3	0.308	0.409	0.470	0.398	0.420	0.543
4	0.439	0.535	0.632	0.352	0.395	0.431
5	0.258	0.301	0.350	0.499	0.567	0.623
6	0.249	0.319	0.409	0.458	0.485	0.509
7	0.153	0.198	0.267	0.500	0.520	0.578
8	0.249	0.352	0.348	0.390	0.409	0.482
9	0.360	0.403	0.464	0.751	0.751	0.842
10	0.194	0.295	0.322	0.360	0.383	0.351
11	0.428	0.487	0.589	0.497	0.497	0.549

(1) Do you think the summarization includes the most important parts of the plot of the movie?
(2) Does the summarization effectively include the interactions between roles?
(3) Does the summarization capture the development of the emotions of the roles?

Since the opinions of each subject may scale differently, we use the rank of each method as the metric, as in Ref. [2]. For each question, subjects were asked to answer with a score 1 (the worst) to 3 (the best) to indicate the ranking of three methods. Figures 5(a)–5(c) depict the evaluation results of the questionnaire survey for the 11 test movies.

6.4 User-specific interest shot selection

The subjects were asked to select shots from the movies which they thought were unnecessary for generating a summarization of the movie.

Figure 6 gives the percentages of interesting shots that were included in our movie skims. The results of this subjective evaluation shows that users retain interactions and shots with emotional content, which validates the effectiveness of our proposed method.

The average percentage of included user-annotated shots illustrates that movie skims generated by our method correspond to human opinion as to which shots are of interest. The results for the second, seventh and tenth movies have poorer performance than other movies due to the low degree of interaction and emotion in these movies.

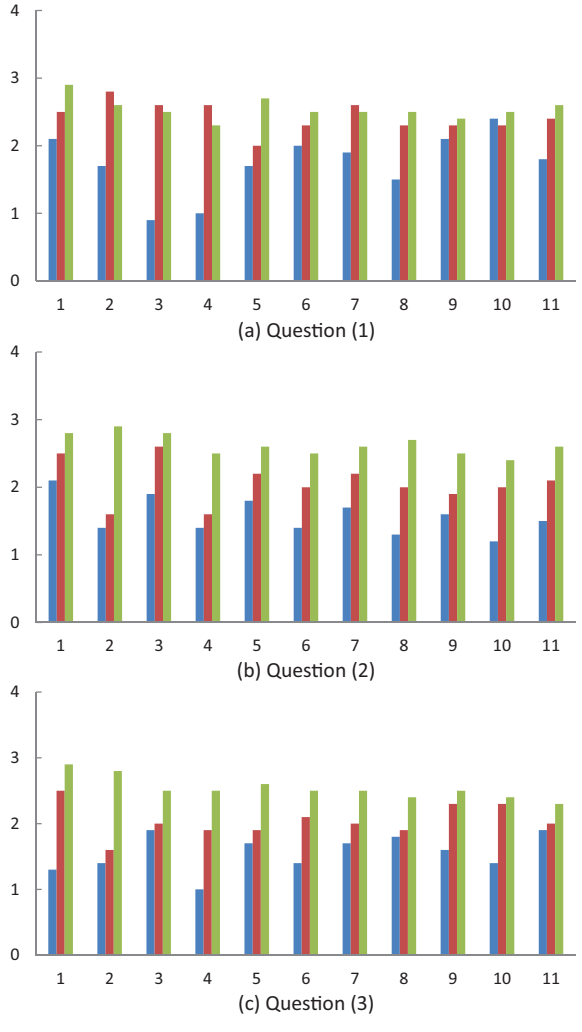


Fig. 5 Questionnaire results for the 11 test movies. The blue, red, and green lines indicated the RoleNet-based method, RCN-based method, and our proposed method, respectively.

7 Conclusions

In this paper, we proposed a novel semantic movie summarization framework based on roles, their interaction and emotion in a movie. Using IE-RoleNets to describe shots, we represent a movie as a string of IE-RoleNets. Summarization of a movie becomes a substring mining problem. Substrings are mined hierarchically from scene level to storyline level, which keeps continuity and completeness of the skims. Furthermore, our method provides both structure-level and role-centered movie skims that meet the preferences of multiple users. Objective and subjective evaluation results demonstrate that the proposed method can capture movie content following the understanding of human subjects. The proposed method preserves the segments of interest effectively.

Acknowledgements

This project was supported by the National Basic Research Program of China under Grant No.

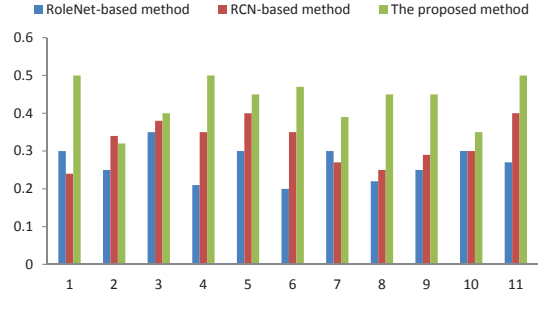


Fig. 6 Correspondence of three summarization methods with user determined interesting shots.

2011CB302200-G, the National Natural Science Foundation of China under Grant Nos. 61370074 and 61402091, and the Fundamental Research Funds for the Central Universities of China under Grant Nos. N120404007 and N140404012.

References

[1] Li, Y.; Lee, S.-H.; Yeh, C.-H.; Kuo, C.-C. J. Techniques for movie content analysis and skimming: Tutorial and overview on video abstraction techniques. *IEEE Singnal Processing Magazine* Vol. 23, No. 2, 79–89, 2006.

[2] Evangelopoulos, G.; Zlatintsi, A.; Potamianos, A.; Maragos, P.; Rapantzikos, K.; Skoumas, G.; Avrithis, Y. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia* Vol. 15, No. 7, 1553–1568, 2013.

[3] Evangelopulos, G.; Rapantzikos, K.; Potamianos, A.; Maragos, P. Zlatintsi, A.; Avrithis, Y. Movie summarization based on audiovisual saliency detection. In: 15th IEEE International Conference on Image Processing, 2528–2531, 2008.

[4] Weng, C.-Y.; Chu, W.-T.; Wu, J.-L. RoleNet: Movie analysis from the perspective of social networks. *IEEE Transactionns on Multimedia* Vol. 11, No. 2, 256–271, 2009.

[5] Tsai, C.-M.; Kang L.-W.; Lin, C.-W.; Lin, W. Scene-based movie summarization via role-community networks. *IEEE Transactions on Circuits and System for Video Technology* Vol. 23, No. 11, 1927–1940, 2013.

[6] Sang, J.; Xu, C. Character-based movie summarization. In: Proceedings of the international conference on Multimedia, 855–858, 2010.

[7] Hanjalic, A.; Xu, L.-Q. Affective video content representation and modeling. *IEEE Transactions on Multimedia* Vol. 7, No. 1, 143–154, 2005.

[8] Shao, X.; Xu, C.; Maddage, N. C.; Tian, Q.; Kankanhalli, M. S.; Jin, J. S. Automatic summarization of music videos. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 2, No. 2, 127–148, 2006.

[9] Babaguchi, N.; Kawai, Y.; Ogura, T.; Kitahashi, T. Personalized abstraction of broadcasted American football video by highlight selection. *IEEE Transactions on Multimedia* Vol. 6, No. 4, 575–586, 2004.

[10] Takahashi, Y.; Nitta, N.; Babaguchi, N. Video summarization for large sports video archives. In: IEEE International Conference on Multimedia and Expo, 1170–1173, 2005.

[11] Chen, F.; De Vleeschouwer, C.; Cavallaro, A. Resource allocation for personalized video summarization. *IEEE Transactions on Multimedia* Vol. 16, No. 2, 455–469, 2014.

[12] Ide, I.; Mo, H.; Katayama, N.; Satoh, S. Exploiting topic thread structures in a news video archive for the semi-automatic generation of video summaries. In: IEEE International Conference on Multimedia and Expo, 1473–1476, 2006.

[13] Wang, Y; Zhao, P.; Zhang, D.; Li, M.; Zhang, H. MyVideos: A system for home video management. In: Proceedings of the tenth ACM international conference on Multimedia, 412–413, 2002.

[14] Zhao, M.; Bu, J.; Chen, C. Audio and video combined for home video abstraction. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, V-620-3, 2003.

[15] Qiu, X.; Jiang, S. Liu, H.; Huang, Q.; Cao, L. Spatialtemporal attention analysis for home video. In: IEEE International Conference on Multimedia and Expo, 1517–1520, 2008.

[16] Lee, Y. J.; Ghosh, J.; Grauman, K. Discovering important people and objects for egocentric video summarization. In: IEEE Conference on Computer Vision and Pattern Recognition, 1346–1353, 2012.

[17] Chen, H.-W.; Kuo, J.-H.; Chu, W.-T.; Wu, J.-T. Action movies segmentation and summarization based on tempo analysis. In: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, 251–258, 2004.

[18] Chen, B.-W.; Wang, J.-C.; Wang, J.-F. A novel video summarization based on mining the story structure and semantic relations among concept entities. *IEEE Transactions on Multimedia* Vol. 11, No. 2, 295–312, 2009.

[19] Zhu, S.; Zhao, Y.; Liang, Z.; Jing, X. Movie abstraction via the progress of the storyline. *IET Signal Processing* Vol. 6, No. 8, 751–762, 2012.

[20] Ren, R.; Misra, H.; Jose, J. M. Semantic based adaptive movie summarization. *Lecture Notes in Computer Science* Vol. 5916, 389–399, 2010.

[21] Trottier, D. *The Screenwriter's Bible: A Complete Guide to Writing, Formatting, and Selling Your Script*. Silman-James Press, 1998.

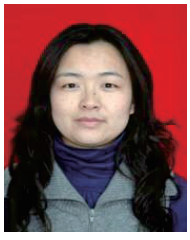
[22] Yusoff, Y.; Christmas, W.; Kittler, J. A study on automatic shot change detection. *Lecture Notes in Computer Science* Vol. 1425, 177–189, 1998.

[23] Zhu, S.; Liu, Y. Automatic scene detection for advanced story retrieval. *Expert Systems with Applications* Vol. 36, No. 3, 5976–5986, 2009.

[24] Zhu, X; Ramanan, D. Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, 2879–2886, 2012.

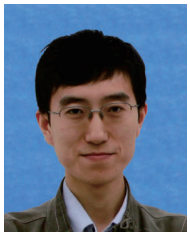
[25] Belhumeur, P. N.; Hespanha, J. P.; Kriegman, D. J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 19, No. 7, 711–720, 1997.
[26] Everingham, M.; Sivic, J.; Zisserman, A. "Hello! My name is . . . Buffy"—automatic naming of characters in TV video. In: Proceedings of the British Machine Vision Conference, 92.1–92.10, 2006.
[27] Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing* Vol. 28, No. 4, 357–366, 1980.
[28] Patron-Perez, A.; Marszalek, M.; Reid, I.; Zisserman, A. Structured learning of human interactions in TV shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 34, No. 12, 2441–2453, 2012.
[29] Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2008.
[30] Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; Ramanan D. Object detection with discriminatively trained partbased models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 32, No. 9, 1627–1645, 2009.
[31] Jiang Y.-G.; Li, Z.; Chang, S.-F. Modeling scene and object contexts for human action retrieval with few examples. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 21, No. 5, 674–681, 2011.
[32] Giannakopoulos, T.; Pikrakis, A.; Theodoridis, S. A dimensional approach to emotion recognition of speech from movies. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 65–68, 2009.
[33] Xu, M.; Xu, C.; He, X.; Jin, J. S.; Luo, S.; Rui, Y. Hierarchical affective content analysis in arousal and valence dimensions. *Signal Processing* Vol. 93, No. 8, 2140–2150, 2013.
[34] Ekman, P. *Emotion in the Human Face*. Cambridge University Press, 1982.
[35] Srivastava, R.; Yan, S.; Sim, T; Roy, S. Recognizing emotions of characters in movies. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 993–996, 2012.
[36] Lavagetto, F. The facial animation engine: Toward a high-level interface for the design of MPEG-4 compliant animated faces. *IEEE Transactions on Circuits and System for Video Technology* Vol. 9, No. 2, 277–289, 1999.
[37] Pikrakis, A.; Giannakopoulos, T.; Theodoridis, S. A speech/music discriminator of radio recordings based on dynamic programming and Bayesian networks. *IEEE Transactions on Multimedia* Vol. 10, No. 5, 846–857, 2008.
[38] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G. Emotion recognition in human–computer interaction. *IEEE Signal Processing Magazine* Vol. 18, No. 1, 32–80, 2001
[39] McGilloway, S.; Cowie, R.; Douglas-Cowie, E.; Gielen, S.; Westerdijk, M.; Stroeve, S. Approaching automatic recognition of emotion from voice: A rough benchmark. In: ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, 2000. Available at http://www.isca-speech.org/archive_open/archive_papers/speech_emotion/spem_207.pdf.
[40] Sakoe, H.; Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* Vol. 26, No. 1, 43–49, 1978.
[41] Truong, B. T.; Venkatesh, S. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 3, No. 1, Article No. 3, 2007.

Wen Qu is a Ph.D. candidate in the Department of Computer Science at Northeastern University, Shenyang, China. She received her B.S. and M.E. degrees from the same institution. Her research interests include video content analysis and multimedia data mining.

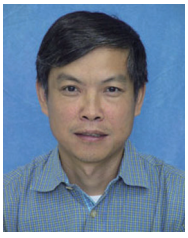


Yifei Zhang is an assistant professor in the Department of Computer Science at Northeastern University, Shenyang, China. She received her B.S. and M.E. degrees from the same institution. Her research interests include image processing and machine learning.

Daling Wang is a professor at School of Information Science and Engineering, Northeastern University, Shenyang, China. She received her Ph.D. degree in computer software and theory from Northeastern University, Shenyang, China, in 2003. Her main research interests include data mining, machine learning, and information retrieval.



Shi Feng is an assistant professor in the Department of Computer Science at Northeastern University, Shenyang, China. He received his B.S. and M.E. degrees from the same institution. His research interests include opinion mining, sentiment analysis, and emotion detection.



Ge Yu is a professor at School of Information Science and Engineering, Northeastern University, Shenyang, China. He received his Ph.D. degree in computer science from Kyushu University, Japan, in 1996. He is a member of both IEEE and ACM and a senior member of China Computer Federation. His research interests include database theory and technology, distributed and parallel systems, embedded software, and network information security.