Character-Net: Character Network Analysis from Video

Seung-Bo Park, Yoo-Won Kim, Mohammed Nazim Uddin, Geun-Sik Jo
Department of Computer & Information Engineering, Inha University, Incheon, Korea
{molaal, yoowon, nazim}@eslab.inha.ac.kr, gsjo@inha.ac.kr

Abstract

Managing the video content for searching and summarizing has become a challenging task. Extracting semantics from video scenes enables information to be presented in a more understandable manner. Finding the semantics between video contexts is a difficult task; much recent research has focused on this issue. Most videos, such as TV serials and commercial movies, are character- centric. Therefore, the context and relationship between characters needs to be organized systematically to analyze the video. So, it is necessary to identify the contextual relationships between characters in the scene and the video. We propose Character-Net, a network structure. It finds characters in a group of shots, extracts the speaker and listeners in the scene, represents it with character-based graphs and draws the relationship between all characters by accumulating the character-based graphs at video. In this paper, we describe how to build Character-Net. Experimental results show Character-Net is an effective methodology to extract the major characters in videos.

1. Introduction

This paper introduces the concept of Character-Net. Character-Net can represent relationships between characters in the video based on the context. We describe how to build the network.

The search system needs to have related information about the objects or characters in the video. Discovering semantic-based representation of the video content is necessary to achieve this goal. Various researchers have experimented on approaches in relation to this issue. Until now, the research in the various streams on semantic-based representation has progressed independently. However, without applying these methods individually we can combine methods to extract semantic content from video that could be more efficient than to extract only the object information or to find only the relationship between objects to search or summarize the video. Thus, we propose a novel method that extracts the character information, determines dialog between characters and represents dialog type between characters as a relationship graph in the video. Accordingly, this paper proposes the Character-Net methodology to recognize characters and determine dialog type through speaker detection. Most of the storyline in the video mainly corresponds to characters and the relationship between them. Character-Net expresses the important semantics of the scene and video. Character-Net can be applied to extract the major characters of the video or important scenes.

The remainder of the paper is organized as follows. The necessity of Character-Net will be explained in section 2 covering related work. Section 3 describes the basic concept of Character-Net and the methodology to build Character-Net and how to extract the major characters. Section 4 outlines the experimental evaluation to build Character-Net and its effectiveness is shown. Section 5 summarizes our approach and discusses future work.

2. Related works

The semantic-based representation of video content has been diversely researched. The objects, such as logos or characters, should be detected or recognized to represent the content. The relationship between objects needs to be extracted to understand the context. The researches may be classified into three different areas: extraction of object information, identification the context between objects, and extraction of information about the objects and the relationship between them together.

First, various approaches detect a variety of objects, such as humans face [6] or logos [1]. The research to extract object information is based on representing the semantics of video content, because it can be used to classify the video types, using elements to represent the video content.

Second, an approach represents the context defined based on the event generated between objects [4]. The event tends to be context triggered by movement of

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.54

objects, such as cars and people. They represented the event as an 'and-or' graph in their study. Events are represented as objects, such as stay, stop, move, death or birth, using a graph.

In the third approach, existing work extracts object information and combines the context with objects and the relationships between them [5]. The semantic network can represent objects, such as man, woman or thing, and the event, such as dance, run, or explode in the video and scenes in their research. This approach also proposed a structure to search the scene. However, there is a weakness in that the user must manually prescribe the objects and events.

3. Representation of Character-Net

3.1 Characters Representation

The characters in the shot are the basic element of story structure basic element. They are represented as:

$$Shot_a = \{sa_1, sa_2, ..., sa_n\} \quad (1)$$

Generally, dialogue cannot be indicated by only one shot in the video. A shot is used to indicate a sub-action. Therefore, 1~3 shots are used to create one speech. The characters in the shot are identified by face recognition; the names of the characters can be detected through the matching technology between subtitles and script [2][3]. When character names are detected, whether each character is a speaker or a listener should be distinguished. The speaker can be detected by mouth movement.

A group is the combination of sequential shots corresponding to the action. The dialogue sequence is separated into 1~3 shots. A speaker and the listeners may appear simultaneously in a shot or in each shot separately. A given speaker or listener may appear in the each shot repeatedly. Therefore, when shots are merged to form a group, the same character appearing in several shots is reduced to a single character instance. Therefore, we can define a group as follows.

$$Group_i = Shot_a \oplus Shot_b = \{a_1, a_2, \cdots, a_n\} \quad (2)$$

$\oplus$: denotes reducing the character appearing more than once to a single instance.

3.2 Character-based Graph

Once the speakers and listeners of a group have been determined, the conversational relationships should be defined. The conversational relationships can be classed in five categories, based on the number of speakers and listeners, as in table 1.

Table 1. Categorization of Conversation Relationships

The number of		Conversational relationship
Speakers	Listeners	
1	1	(speaker) talking to (listener)
0	1	(listener) listen
1	0	(speaker) speaking alone
1	≥ 2	(speaker) talks to (listeners)
0	≥ 2	(listeners) listen

The conversational relationships in table 1 may be represented as a graph in figure 1.

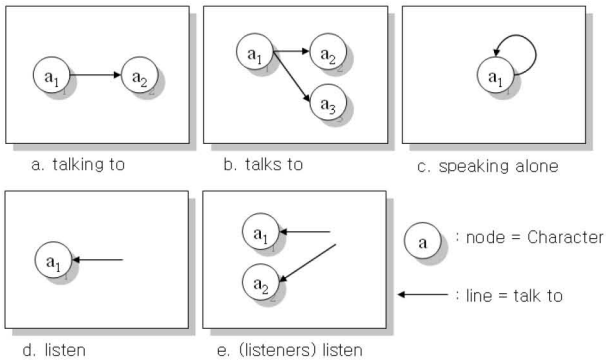

Figure 1. Character-based Graphs

Figure 1 is the character-based graph that represents relationships between characters based on conversation. It is the basic Character-Net graph. Circles in the graph are nodes and represent the characters extracted from a group. The line between nodes denotes the action of 'talk to' and is stands for the directed action. Since the conversation relationship has the direction and talk time from speaker to listener, the valued direction is selected. Therefore, the starting node of the line denotes the speaker and the target node denotes the listener.

3.3 Character-Net Construction

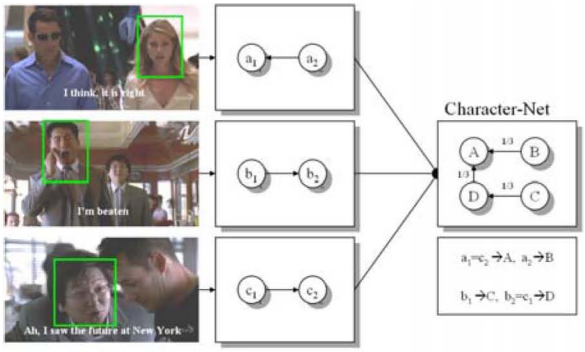

Figure 2. Character-Net Construction

A graph for every scene or video is first drawn to depict the character based graph for a group. In a

group, repeated characters are treated as a single character. An example Character-Net graph is shown in figure 2. The images in the boxes in the left side of figure 2 indicate the speaker. The middle section of the figure represents character-based graphs for each group with speaker and listeners detected. Finally, Character-Net is drawn, combining the character-based graphs of each group and suppressing repeated characters, as in the right side of figure 2. When graphs are converted to Character-Net, weights are assigned based on the direction of conversational frequency between two nodes. The valued graph is used in Character-Net.

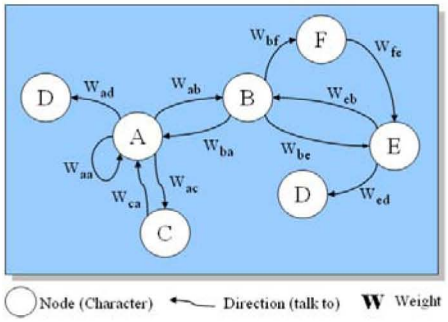

Figure 3. Character-Net

The Character-Net is completed after constructing Character-Net from groups, as shown in figure 3. At here, the major character is recognized as connected to many other nodes and has a high weight; the centrality may be calculated as a high value. Various forms of Character-Net can depict the idea of kind and genre of the video. This weight can be assigned using two approaches. First, count the number of times two characters make conversation, using equation (3). Second, the elapsed talking time between two characters, using equation (4).

$$w_{ab}=NumTalk_{ab} \div TotalNumTalk \quad (3)$$

$$w_{ab}=TalkTime_{ab} \div TotalTalkTime \quad (4)$$

In equations (3) and (4), w_{ab} denotes the weight from node *a* to node *b*. In equation (3), $NumTalk_{ab}$ denotes how many times node *a* talked to node *b* and *TotalNumTalk* denotes the total amount of talk between all characters in the video. In equation (4), $TalkTime_{ab}$ denotes the accumulated talk time from node *a* to node *b*, and *TotalTalkTime* denotes the total talk time between all characters in the video.

3.4 Major Character Extraction

The major character in a video has a central role amongst other characters. In our approach, we recognize the major character if she/he has greater connectivity to others and has a high conversational frequency. This person is represented as the central node in the Character-Net. The major character can be extracted by centrality analysis. The degree centrality (DC) for each node is represented in equation (5).

$$DC_a=\sum_{i=1}^{n} w_{ia}+\sum_{j=1}^{m} w_{aj}-w_{aa} \quad (5)$$

Here, *n* is the number of in-coming nodes to node *a* and *m* denotes the number of out-going nodes from node *a*. The value of *DC* is between 0 and 1, because it is less than the summation of all weights. In equation (5), the first factor on the right side denotes the in-degree, the second factor denotes the out-degree; the last factor denotes the self loop. w_{aa} is subtracted in DC, since w_{aa} is duplicated in the in-degree and the out-degree components of the equation. And the major character nodes can be extracted comparing the *DC* value to a specified threshold.

4. Experiment and discussion

Character-Net was implemented using Microsoft's Windows XP operating system on the PC. Face recognition was applied using Nurotechnology's Verilook Face Identification SDK API. Speaker detection and conversational relationship detection were tested using Visual Basic 6. The construction of Character-Net of the video was implemented using Borland's Delphi 7.0. For experiment, videos selected were *Ace Ventura, Heroes episode 3 from season 1, Friends episode 1 in season 3*. The major characters for these three videos were extracted. Weights are assigned for these Character-Nets using equation (4).

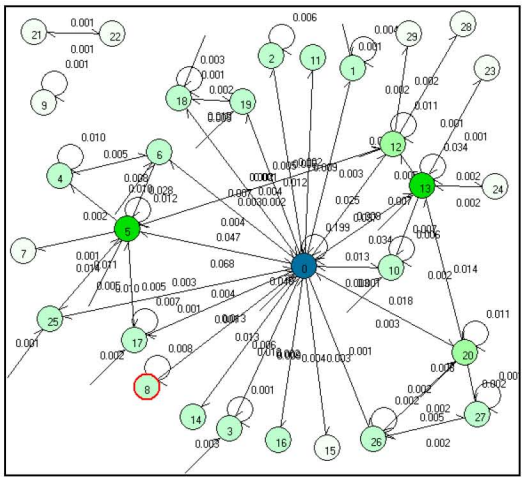

Figure 4. Ace Ventura Character-Net

In *Ace Ventura*, node 0 plays an important role at this video; it was connected to almost all nodes and had the very high DC, 0.687. Nodes 5 and 13 had DCs of 0.216 and 0.204 respectively; they were extracted as the major characters. Nodes 20 and 12 are the side characters that help major characters. In this analysis, we set the threshold that determine major characters at 0.1, that is, the character has more than 10% of the total conversation. Nodes whose degree centrality range from 0.05 to 0.10 are extracted as side characters.

The second video of our experiment, *Heroes* is a TV series. Figure 5 depicts the Character-Net for one of the beginning episodes of Heroes season 1. Therefore, this has subgroups neighboring the major characters. In addition, a sub-group surrounding node 2 is isolated from other sub-networks, because this episode is in at beginning of the series and the pattern emerging in this group is not affected by other characters, or not all characters or the major character appear in beginning episodes.

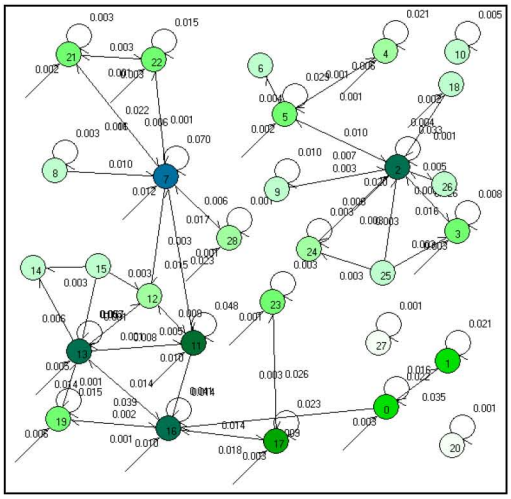

Figure 5. Heroes Character-Net

In the Character-Net depicting figure 5, the degree centralities for the nodes are listed in table 3. Nodes 7, 13, 16, 2 and 11 are classified as major characters and others are classified as side characters.

At the Character-Net of the TV series *Friend,* there are 7 major characters (nodes 4, 0, 1, 2, 5, 3, 6) and several other characters (nodes 8, 9, 7).

5. Conclusion

In this paper, we proposed and implemented a structure to build the Character-Net of a video based on the conversational relationship among characters. We experimentally demonstrated the Character-Net can be an efficient method to detect important characters, such as major and side characters, in any video. In Character-Net, the concept of social network is applied to design the structure and to detect major and side characters. We proposed character-based graphs, a calculated weight and set for the conversational direction, and the calculation of the degree centrality to extract major characters defining the conversational relationship. This Character-Net is a method to express the semantics of the video content based on the context of conversational relationships among characters. We calculated the degree centrality for each node that can extract major characters and side characters.

For the next step of this research, we plan to improve the method by updating the relationship representation among nodes. The Character-Net analysis will extract more semantics from the video.

Acknowledgment

This research was supported by WCU(World Class University) program through the Korea Science and Engineering Foundation funded by the Ministry of Education, Science and Technology (R33-2008-000-10109-0).

6. References

[1] J.R. Cózar, N. Guil, J.M. González-Linares, E.L. Zapata, E. Izquierdo, "Logotype detection to support semantic-based video annotation," Signal Processing: Image Communication, Vol. 22, Issues 7-8, Aug.-Sep. 2007, pp. 669-679.

[2] J. Yang, R. Yan, A.G. Hauptmann, "Multiple instance learning for labeling faces in broadcasting news video," in: Proceedings of the ACM International Conference on Multimedia, 2005, pp. 31-40.

[3] M. Everingham, J. Sivic, A. Zisserman, "Taking the bite out of automated naming of characters in TV video," Image and Vision Computing, In Press, Corrected Proof, Available online, 4 May 2008.

[4] L. Liang, G. Haifeng, L. Li, and W, Liang. "Semantic event representation and recognition using syntactic attribute graph grammar," Pattern Recognition Letters, Vol. 30, Issue 2, 15 Jan. 2009, pp. 180-186.

[5] V. Roth, "Content-based retrieval from digital video," Image and Vision Computing, Vol. 17, No. 7, 1999, pp. 531-540.

[6] D. Cristinacce and T.F. Cootes, "Feature Detection and Tracking with Constrained Local Models," Proc. 17th British Machine Vision Conf., 2006, pp. 929-938.