韓國컴퓨터情報學會 論文誌
第14卷 第10號, 2009. 10.

2009-14-10-4-9

대화 참여자 결정을 통한 Character-net의 개선

김 원 택*, 박 승 보**, 조 근 식***

Improvement of Character-net via Detection of Conversation Participant

Wontaek Kim*, Seung-Bo Park**, Geun-Sik Jo***

요 약

동영상 검색이나 축약과 같은 동영상 분석을 위해 동영상 어노테이션 기술이나 동영상 정보 표현에 대한 다양한 연구가 있어왔다. 이를 위해 본 논문은 대화 참여자 결정을 위한 영상적 요소와 이러한 요소를 이용하여 Character-net 표현을 개선하는 방법을 제안한다. 기존 Character-net이 자막이 뜨는 시간에 나타나는 등장인물들만을 대화참여자로 고려하므로 일부의 청자를 제외시키는 문제점이 있다. 대화 참여자는 대화상황 파악의 극히 중요한 요소로 동영상 검색 시에 기준이 될 수 있으며 동영상의 이야기 전개를 이끌어 나간다. 대화 참여자를 결정하기 위한 영상적 요소에는 자막의 유무, 장면, 인물 등장순서, 시선방향, 패턴, 입의 움직임 등이 있다. 본 논문에서는 이러한 영상적 요소에 근거하여 대화 참여자를 판단하고 동영상 표현방법인 Character-net을 개선하고자 한다. 제안한 여러 요소들이 결합되고 일정한 조건이 만족되었을 때 대화참여자를 정확히 검출할 수 있다. 따라서 본 논문에서는 대화참여자를 결정하기 위한 영상적 요소들을 제안하고 이를 통해 Character-net의 표현성능을 개선하고 실험을 통하여 제안된 방법론이 대화 참여자 판단의 정확성과 Character-net의 표현성능을 제고함을 증명하였다.

Abstract

Recently, a number of researches related to video annotation and representation have been proposed to analyze video for searching and abstraction. In this paper, we have presented a method to provide the picture elements of conversational participants in video and the enhanced representation of the characters using those elements, collectively called Character-net. Because conversational participants are decided as characters detected in a script holding time, the

• 제1저자 : 김원택
• 투고일 : 2009. 09. 04, 심사일 : 2009. 09. 15, 게재확정일 : 2009. 10. 07.
* 인하대학교 정보공학과 석사과정 ** 인하대학교 정보공학과 박사과정 *** 인하대학교 컴퓨터정보공학부 교수
※ 본 연구는 한국과학재단을 통해 교육과학기술부의 세계수준의 연구중심대학육성사업(WCU)으로부터 지원받아 수행되었습니다 (R33-2008-000-10109-0).
※ 이 논문은 2009년 한국컴퓨터정보학회 제40차 하계학술대회에 발표한 "대화 참여자를 결정하는 영상적 요소"를 확장한 것임[10].

previous Character-net suffers serious limitation that some listeners could not be detected as the participants. The participants who complete the story in video are very important factor to understand the context of the conversation. The picture elements for detecting the conversational participants consist of six elements as follows: subtitle, scene, the order of appearance, characters' eyes, patterns, and lip motion. In this paper, we present how to use those elements for detecting conversational participants and how to improve the representation of the Character-net. We can detect the conversational participants accurately when the proposed elements combine together and satisfy the special conditions. The experimental evaluation shows that the proposed method brings significant advantages in terms of both improving the detection of the conversational participants and enhancing the representation of Character-net.

▸ Keyword : 대화(Conversation), 화자(Speaker), 청자(Listener), 캐릭터-넷(Character-net)

Ⅰ. 서 론

현재 인터넷에는 방대한 양의 동영상들이 존재한다. 이러한 동영상들 중에서 원하는 정보를 찾는 것은 상당히 어려운 문제이다. 그래서 다양한 접근방식을 가지고 원하는 동영상을 검색하는 기술에 대한 많은 연구가 진행되고 있다.

기존의 동영상들은 검색을 위해 제목에 주제어나 출연자의 이름을 넣어주어 검색에 활용한다. 하지만 기하급수적으로 증가하는 동영상들에서 이러한 검색수단을 사용하는 것은 정확한 검색결과를 가져올 수 없다. 찾고자 하는 동영상의 제목이나 출연자의 이름과 같은 특정 정보를 모를 경우 검색할 수 없으며 검색하였다 해도 동영상의 특정장면이 아닌 동영상의 유무를 검색할 따름이다. 이런 문제점을 해결하기 위하여 동영상의 숏(shot)에 어노테이션(annotation) 하거나 동영상의 내용을 기술할 수 있는 방법이 필요하다. 숏(shot)이란 한 번의 연속 촬영으로 찍은 프레임들의 집합을 지칭한다[1].동영상 어노테이션은 동영상의 콘텐츠의 내용을 추출하여 기술함으로서 검색에 활용할 수 있게 한다. 동영상 어노테이션을 거치면 찾고자 하는 동영상 콘텐츠의 한 부분에 대하여 접근할 수 있을 것이다. 또한 동영상의 내용 기술은 특정 사물이나 인물을 인식하여 표시하거나 장면들 간의 관계를 기술하여 검색이나 축약에 활용한다.

동영상 어노테이션을 위해 동영상 내의 상황정보를 파악하여 개체와 개체사이의 관계를 네트워크로 표현하여 검색에 활용하게 하려는 연구들이 있다. 이러한 연구들은 동영상 내의 장면을 기술하는 방법으로 등장인물이나 기타 사물들을 추출하고 이들 사이의 관계를 기술하여 검색에 활용하고 동영상 정보들을 추출하는 것을 목적으로 한다.

TV드라마나 영화와 같은 동영상들은 등장인물들 사이의 대화를 기본으로 동영상의 내용이 진행되므로 이 대화를 기반으로 동영상을 표현할 수 있다. 이러한 대화 기반의 동영상 표현방식을 Character-net이라 하며 이것을 정확히 표현하기 위해서 가장 필수적인 부분이 바로 대화 참여자 판단이다. *Character-net*은 동영상의 주연이나 조연과 같은 중요한 의미정보(semantic information)를 추출할 수 있고 등장인물 간의 관계를 묘사할 수 있는 유용한 방법론이다[2]. Character-net은 등장인물간의 대화를 기반으로 구성되므로 등장인물간의 대화를 정확하게 찾는 것이 가장 기본적인 요구사항이다. 등장인물들의 대화는 상황정보추출의 중요한 기반이며, 극의 등장인물과 대화 참여자를 판단하는 것은 동영상 내용 표현과 어노테이션에서 중요한 과제라 할 수 있다. 기존의 Character-net 연구에서는 단순히 자막의 유지시간과 인물인식에 의하여 대화상대를 판단하였다. 동영상에서 단순히 자막의 유지시간만을 이용한 대화 참여자의 판단은 오류가 많으므로 본 논문에서는 기타 대화요소들을 추가하여 정확성을 높이려 한다. 여기서 자막의 유지시간이란 화면상에 한 자막이 나타나고 이 자막이 사라질 때까지의 지속시간을 지칭한다. 자막 이외에 동영상에는 시선, 인물위치, 장면(scene), 숏 등 기타 영상적 요소들이 있다. 대화참여자를 정확히 판단하여야 등장인물간의 대화상황정보가 정확히 추출될 수 있고 이를 바탕으로 동영상의 등장인물간의 관계가 명확히 표현될 수 있다. 하지만 대화 참여자를 판단하기 위해 다양한 기준들이 존재하고 이 기준들은 서로 보완적이거나 상호 제약적으로 작용한다. 따라서 대화참여자를 정확히 판단하기 위해 기준들을 명확히 제시할 필요가 있으며 기준들이 갖는 값의 범위가 정확히 설정될 필요가 있다. 이렇게 설정된 기준들은 대화 참여자의 판단에 도움을 줄 것으로 판단되며 대화 참여자의 정확한 결정을 통해 Character-net을

개선하고자 한다. 또한 실험과 기존연구와의 비교를 거쳐 제안하는 방법론의 정확성이 제고되는 것을 확인한다.

본 논문은 제안된 요소들을 설명하기 위하여, 2장에서 관련된 연구들에 대하여 기술하고 다음으로 3장에서 Character-net과 제안한 영상적 요소들에 대하여 자세히 설명하고 마지막으로 실험을 통하여 제안한 요소들을 실제 동영상에 적용시켜서 기존의 연구와 비교하고 정확성의 제고를 실험으로 증명한다.

II. 기존 연구

동영상의 장면에서 상황정보를 추출하려는 다양한 연구들이 진행되어 왔다. 상황정보를 추출하여 어노테이션 하면 원하는 장면을 검색하거나 동영상을 축약하는데 이용될 수 있다[3-5]. 이를 위해 장면의 상황정보를 추출하기 위해 등장인물 간의 대화를 파악하여 대화에 대한 상황정보를 추출하려는 시도가 있다[4]. 이 연구에서는 동영상의 등장인물들을 추출하고 등장인물들 간의 대화 형태를 파악하여 대화 형태를 어노테이션 하였다. 하지만 대화에 대한 상황정보를 추출하려는 연구는 대화 상황에서 발생하는 화자와 청자의 관계를 표현하려는 데에 치중된 연구이다. 이 연구는 대화를 구성하는 참여자를 영상인식 기술을 이용하여 화자와 청자(들)로 분류하고 화자와 청자(들)의 상황 규칙으로 판단하는 것을 제안하였다. 하지만 이 연구에서 대화의 경계를 자막이 유지되는 시간으로 결정하여 대화의 참여자를 정확히 추출하지 못하는 문제점이 존재하였다. 따라서 대화의 경계를 정할 수 있는 영상적 요소를 찾아내어 대화 참여자를 모두 추출하는 것이 가능하도록 하는 연구가 진행될 필요가 있다.

이와 관련된 연구로 영화제작이나 편집에서 사용된 대화 장면의 무대화에 대한 연구가 있다[6]. 동영상 내에서 대화 장면을 무대화하는 것은 영화 편집 분야의 중요한 이슈로서 많은 영화학적인 연구가 진행되어왔다.

대화 장면들을 무대화할 때 감독은 인간관계의 성실한 표현과 이런 관계를 관객에게 제시하여야 한다. 이러한 관계를 관객에게 제시하는 일은 무대화, 촬영, 그리고 편집에 의해 결정된다. 따라서 감독의 무대화, 촬영, 편집에 대한 목표가 본연구의 관점과 관련된 목표이다. 이 중에서 무대화는 장면을 시각화하는 방법으로 관객들에게 대화를 이해하고 줄거리를 연결하는 기본요소로서 작용한다. 무대화를 위한 기초적 요소는 대화 참여자들에 대한 무대화 패턴들이다. 패턴의 의미는 한 프레임 내에서 인물을 배치하는 방법으로서 이러한 패턴의 목적은 동작선에 따라서 배우들을 가장 간단히 배열시켜 대화 장면을 구성하기 위한 것이다. 패턴은 관객들이 대화 상황을 이해할 수 있게 잘 표현 할 수 있으므로 대부분의 동영상 내 대화 장면에서 사용된다.

본 연구에서는 대화 참여자를 정할 수 있는 영상적 요소와 대화 장면의 무대화에 사용되는 기본적인 방법을 사용하여 대화 참여자를 판단하고 이를 이용하여 Character-net 표현을 개선하고자 한다.

III. 대화 참여자 결정을 통한 Character-net의 개선

3.1 용어정의

본 논문에서 사용한 동영상 내에서의 대화와 관련된 용어를 정의하면 다음과 같다.

- 화자: 대화를 시작하는 발신자를 화자라고 한다.
- 청자: 발신된 대화를 받아들이는 수신자를 청자라 한다.
- 1회 대화: 한 문장으로 구성되어 있고, 화자와 청자가 있다. 화자는 1인, 청자는 0명에서 다수가 있다. 화자 또는 청자는 존재하지만 화면에 나타나지 않을 수 있다.
- 3자: 3자는 화자도 청자도 아닌 대화 화면에 나타나지만 대화에는 참여하지 않는 대화 비참여자를 지칭한다.

본 논문에서는 1회 대화중에서 대화 참여자로써 화자와 청자를 판단하는 것을 목적으로 한다.

3.2. 대화 참여자를 판단하는 영상적 요소

본 논문에서는 동영상 내에서 대화 참여자를 판단하기 위하여 아래와 같은 요소들을 제안한다.

A. 자막의 유무
B. 장면
C. 인물 등장순서
D. 시선방향
E. 패턴
F. 입의 움직임

위의 요소들이 동영상 내의 대화 참여자를 어떻게 결정하며 다음으로 어떻게 서로 결합되어 보다 정확한 판단을 할 것인가에 관한 설명은 아래와 같다.

A. 자막의 유무

- 정의 : 자막은 동영상 내의 대화를 글로 표현한 것
- 값 : On / Off
- 설명 : 자막이 나오면 장면 속에 대화가 이루어지고 있음을 알 수 있다. 단순한 자막에만 의거하여 대화의 경계를

설정하면 표현방식이 다양한 동영상에서 정확한 대화상황을 추출할 수 없다[4].

예를 들어 그림 1의 첫 숏에서 화자가 말을 하였을 때 자막은 다음 숏까지 이어지지 않았다. 인접되어있는 청자를 나타내는 숏에서 다른 인물의 말을 표현한 자막이 나타나거나 혹은 자막이 없을 경우, 기존의 연구에서는 화자가 혼잣말을 하고 있다고 판단하게 된다[4]. 자막만으로 결정하는 대화의 경계는 자막의 유지시간으로 그 경계가 어디까지인지 정해지게 된다. 이 경우에 자막이 지속되는 시간 내에 화면상에 나타나는 인물들은 대화 참여자이고, 지속시간 이외의 화면상에 나타나는 인물들은 대화의 참여자가 아니라고 판단한다. 이러한 판단은 한 화면상에 대화 참여자들이 전부 등장하였을 때는 정확하게 판단할 수 있지만, 그렇지 않은 경우에는 대화 참여자를 놓칠 수 없다. 따라서 본 논문에서는 1회 대화라는 개념으로 대화경계를 명백히 설정하였고 장면을 바탕으로 여러 요소들을 결합하여 대화 참여자를 판단한다.

그림 1. 단순히 자막에 근거하는 경우
Fig 1. Conversation Range by a Script

B. 장면(scene)

- 정의 : 동영상의 이야기 의미 단위이며 시간적으로 이웃된 숏들의 모임이다.
- 값 : 0, 1, 2, ... , N (N : 마지막 장면 번호)
- 설명 : 한 장면 안의 대화는 한 의미를 둘러싸고 발생한 대화라고 볼 수 있다. 따라서 본 논문에서는 한 장면 속에 등장하는 인물들이 전부 대화 참여자라고 전제하고 다른 요소들과 제약조건을 적용하여 이들 중에서 대화 참여자를 추출한다. 동일 장면을 대화 참여자 판단을 위한 기본적인 경계로 하면 화자가 판단되었을 때 동일 장면에 등장하는 다른 등장인물을 청자로 전제할 수 있다. 이렇게 판단된 인물들은 본 장면 내의 모든 1회대화의 참여자의 바탕이 될 수 있다.

C. 인물 등장순서

- 정의 : 한 장면에서 인물이 나타나는 숏 번호
- 값 : 0, 1, 2, ... ,m (m : 한 장면에서 마지막 숏 번호)
- 설명 : 대화상대는 동영상 촬영기법상 연속적으로 나오거나 한 화면에 함께 나온다. 이러한 규칙에 의해 서로 다른 화면이거나 하나의 롱컷(long cut)안에 인물들이 가까운 순서에 등장하면, 대화 참여자인 청자라고 판단할 수 있다. 위에서 장면을 바탕으로 대화 참여자를 선정한 기초 상에서 인물 등장순서 요소는 특정된 한 1회 대화에 대하여 효율적으로 대화 참여자의 범위를 줄일 수 있다. 따라서 실험을 거쳐 정확성이 가장 높은 인물등장 순서범위를 선정한다. 본 논문에서는 화자가 인식된 후 인접하여 등장하는 앞뒤로 2번째 인물까지만 대화 참여자라고 판단한다.

D. 시선방향

- 정의 : 등장인물들이 바라보는 방향
- 값 : 좌(←), 정(⊙), 우(→)
- 설명 : 대화 중 화자와 청자는 서로 바라보게 된다. 따라서 시선방향은 참여자를 판단하는 요소로써 위에 기술한 인물 등장순서 요소와 결합되어 화자와 청자를 더욱 정확하게 판단하는데 사용될 수 있다.

그림 2. 인접된 숏에서 대화 참여자들의 시선방향
Fig 2. Sight Line in Neighbor Shot

일반적이고 정확한 촬영기법에서는 관객들이 동영상 내의 대화에 대한 이해를 돕기 위하여 한 화면에 나타난 대화 참여자는 다른 대화 참여자의 얼굴방향과 마주보게 된다. 실험에서 사용한 영화의 대부분의 대화에서 화자와 청자 즉 대화 참여자들은 그림 2와같이 인접한 화면에서 서로 마주보고 있음을 파악할 수 있다. 시선방향은 눈동자 인식이 가능한 경우에는 눈동자의 방향으로 정하고, 눈동자 인식이 가능하지 않는 경우에는 얼굴방향으로 대화상대를 결정한다.

E. 패턴

- 정의 : 화면상에서 인물들의 상대적인 위치
- 값 : 추출(on) / 미추출(off)
- 설명 : 영화감독들이 영화나 기타 동영상을 촬영함에 있어서 대화 참여자를 배치하는 기법이다. 동영상 내에서 사람들의 위치를 판단할 수 있는 화면이 나오면 패턴을 이용하여

그 위치로 대화 참여자들을 판단할 수 있다. 추출된 기타 인물들이 화자와 같은 패턴에 속하여 있으면 기타 요소의 참여 없이도 대화 참여자임을 판단할 수 있다[6]. 영화에서 대화를 표현하기 위해 사용되는 패턴은 일반적으로 A패턴, L패턴, I 패턴이 있다. A패턴과 L패턴은 3인 이상대화에서 사용되며 I 패턴은 2인 대화에서 주로 사용된다. 이러한 패턴으로 대화 참여자를 위치시켜 화자와 청자가 현실감 있게 화면에 나타나도록 구성한다.

F. 입의 움직임

- 정의 : 인식된 얼굴에서 입 부근의 영상의 변화
- 값 : On / Off
- 설명 : 화자와 청자를 나누는 기준으로 입의 움직임이 사용된다. 자막의 유무와 화면에서 추출된 인물의 입모양을 가지고 화면상의 인물이 말을 하고 있는지 아닌지를 판단한다. 이러한 방식은 화면에 나타난 인물이 말을 하는 것처럼 입을 움직이더라도 자막의 유무에 근거하여 판단하기 때문에 대화를 하고 있는지 아닌지를 정확히 판단할 수 있다[4,7,8]. 입의 움직임은 얼굴에서 입의 영역을 판단한 후 해당하는 영역의 히스토그램 변화를 파악하거나 입의 특징점을 찾아 특징점간의 거리 변화를 통해 판단한다. 따라서 본 연구의 요소인 입의 움직임 역시 위의 방법을 통해 파악할 수 있다.

3.3. 대화 참여자 판단 기준

대화 참여자를 판단하기 위하여 아래의 기준에 따라서 위의 요소들을 사용한다.

표 1. 화자를 판단하는 규칙
Table 1. The Speaker Detection Rule

A	B	C	D	E	F
on	-	-	-	-	on

- : 값에 상관없음

표 1의 화자를 판단하는 규칙은 자막의 유무와 입의 움직임만으로 판단된다. 즉 다음의 2가지 규칙을 한꺼번에 만족시키는 경우이다.

1. 화면상에 자막이 나타난다.
2. 화면상의 인물의 입의 움직임이 추출되었다.

이렇게 화자로 판단되는 인물이 영상에 나타날 수 있는 경우를 분류하면 표 2처럼 6가지 경우가 된다. 이 분류는 시선방향과 패턴의 유무에 의해 나누어진다.

표 2. 화자가 나타날 수 있는 경우 6가지
Table 2. The six occasions of Speaker detection

화자경우	A	B	C	D	E	F
1	on	SCi	SHj	→	on	on
2	on	SCi	SHj	←	on	on
3	on	SCi	SHj	⊙	on	on
4	on	SCi	SHj	→	off	on
5	on	SCi	SHj	←	off	on
6	on	SCi	SHj	⊙	off	on

SCi : 장면 번호, SHj : 숏 번호

화자의 시선방향은 3가지 방향 좌(←), 정(⊙), 우(→) 중에 1가지 값을 가질 수 있으며, 패턴은 추출(on)되거나 미추출(off)되는 2가지 값을 가질 수 있다.

화자의 말을 듣고 있는 청자는 표 3의 화자 경우에 따른 청자의 판단규칙에 의해 결정된다. 여기서 표 2의 화자경우와 표 3의 화자 경우는 동일하다.

표 3. 화자의 경우에 따른 청자의 판단 규칙
Table 3. Listener detection rules

화자경우	A	B	C	D	E	F
1	-	x	i±2 이내	← or ⊙	On	-
2	-	x	i±2 이내	→ or ⊙	On	-
3	-	x	i±2 이내	→ or ← or ⊙	On	-
4	-	x	i±2 이내	← or ⊙	-	-
5	-	x	i±2 이내	→ or ⊙	-	-
6	-	x	i±2 이내	→ or ← or ⊙	-	-

표 3의 청자로 선택되는 판단 규칙은 화자의 판단보다 복잡하고 다양한 경우가 존재한다.

1. 화자와 한 장면에 속하는 등장인물이어야 한다.
2. 청자는 화자와 연속적으로 등장한 인물이여야 하며 이때 화자를 중심으로 앞, 뒤로 2번째 숏의 인물까지만 선택한다.
3. 화자가 판단되었으면 청자의 시선방향은 화자와 반대되거나 정면을 바라보고 있어야 한다. 화자의 시선방향이 정면을 바라보고 있으면 청자의 시선방향은 좌, 우, 정면을 바라 볼 수 있다.
4. 대화 중 패턴이 추출되었을 때 패턴에 속하는 인물이 아니면 위의 조건을 만족하여도 청자가 아니다.

3.4. Character-net

Character-net은 등장인물 간의 대화를 누적하여 사회연결망 형태로 표현한 것이다. Character-net은 정점(vertex)

이 등장인물을 의미하고 등장인물 간의 대화 정도가 간선(edge)으로 표현되는 그래프 $G(V,E)$이다. 등장인물은 얼굴인식을 통해 식별되며[7] 간선은 1회 대화를 누적하여 계산한다. 1회 대화는 3가지 요소로 구성되는 데 화자와 청자와 대화량으로 구성된다. 이를 그래프로 표현하기 위해 그림 3과 같이 2개의 정점과 1개의 간선이 필요하다. 간선은 방향성을 가진 계량화된 화살표를 사용한다. 그림 3의 화살표가 나가는 C_1은 화자이며 그 화살표가 도달하는 C_2는 청자이다. 간선이 갖는 가중치 w_{12}는 C_1과 C_2사이의 대화량이 된다.

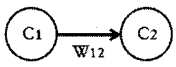

그림 3. 1회 대화의 그래프 표현
Fig 3. Graph of a Conversation

동영상에서 1회 대화는 그림 4처럼 여러 개의 숏들로 표현되는데 이 숏들을 묶어서 그룹이라고 정의한다. 화자와 청자를 판단하기 위해 얼굴인식 기술과 화자 인식 기술을 이용하였다. 얼굴인식 기술을 이용하여 현재 검출된 얼굴이 누구인지를 알아내고 입의 움직임이 있는 경우 3.3절에서 정의한 규칙에 의해 화자인지 아닌지를 판단한다. 그리고 화자가 아닌 다른 얼굴은 3.1절부터 3.3절까지에서 정의한 규칙들에 의해서 청자나 대화 비참여자로 판정한다[4,7,8].

그림 4. 숏들의 그룹
Fig 4. A Group of Shots

1회 대화에 해당하는 숏들의 얼굴들(Cg_h)을 연합할 때는 식 1에 의해 숏에 나타나는 얼굴들(Cs_k)을 합하게 된다.

$$Cg_h = Cs_k \oplus Cs_{k+1} \oplus \cdots \oplus Cs_{k+l} = \sum_{i=1}^{l} cg_{hi} \quad \cdots\cdots (1)$$

⊕: 숏들에서 반복되는 동일얼굴을 1명으로 취급하여 합함

단, 그룹의 숏들에서 화자나 청자가 반복적으로 나오는 경우 1명으로 인식하여 등장인물을 결정하게 된다. 인식된 얼굴들은 입의 움직임에 의해 화자와 청자로 분리가 되어 그림 2처럼 정점과 화살표로 표현된다. 여기서 화살표 상에 대화량이 설정되는데 이는 대화 회수와 대화 시간이라는 2가지 종류가 있다. 대화 회수는 1회 대화마다 가중치 값을 1의 가중치로 설정하는 방식이고, 대화 시간은 1회 대화의 시간을 가중치 값으로 할당하는 방식으로 동영상에 자막이 나타난 시간을 의미한다.

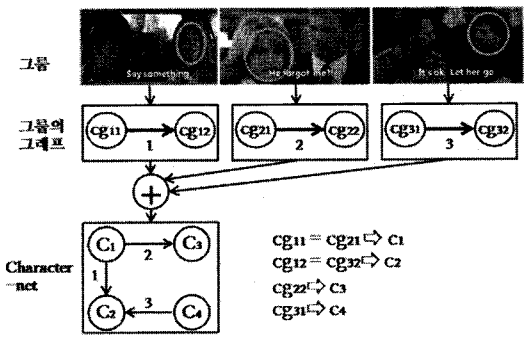

그림 5. Character-net의 구축
Fig 5. The Construction of the Character-net

그룹의 대화 그래프가 그려지면 그림 5와 같이 그래프들을 누적하여 동영상 전체의 등장인물들 간의 대화 그래프를 구축하게 되는데 이렇게 누적된 그래프를 Character-net이라 한다. 그룹에서 화자와 청자를 판단하여 그림 5의 가운데 그룹의 그래프를 그리고 다음 그룹과 연합하면서 그래프를 누적 확장하여 하단에 보이는 Character-net을 그려 나가게 된다. 이때 얼굴인식 기술을 사용하여 동일인 인지를 판단하여 그래프의 정점들이 연계되도록 하고 중복되는 등장인물 간의 간선의 가중치는 값을 누적시킨다.

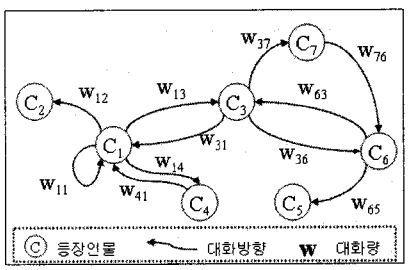

그림 6. Character-net
Fig 6. The Character-net

동영상에 대해 이러한 방식으로 그래프를 누적시키면 그림 6과 같은 결과가 나타날 것이다. 그림 6에서 정점과 간선과 가중치는 각각 등장인물과 대화방향과 등장인물 간의 누적 대화량을 의미한다.

IV. 실험 및 분석

제안한 요소들을 이용하여 대화 참여자를 결정할 수 있음을 실험을 통해 확인하고, 동영상에 대한 Character-net 표현이 개선되는 것을 확인하였다. 기존연구와 비교를 위해 4개의 영화를 선택하여 실험하였으며 영화 목록은 표 4와 같다.

표 4. 실험 영화 목록
Table 4. Movie List for testing

번호	동영상 제목
M1	해리가 샐리를 만났을 때 (1989)
M2	에이스 벤츄라 (1994)
M3	귀여운 여인 (1990)
M4	스타더스트 (2007)

4.1 대화 참여자의 결정

4개의 영화에 대해 각각 20개의 대화를 추출하여 대화 참여자를 정확히 결정하는 지를 실험하였다.

인물 A	
A	on
B	SC3
C	SH2
D	→
E	off
F	on

인물 B	
A	on
B	SC3
C	SH3
D	←
E	off
F	off

그림 7. 인물의 정보 표현
Fig 7. The representation of character information

장면에 나타나는 인물들에 대하여 숏별로 6개의 요소를 사용해서 그림 7과 같이 표시하였다.

자막 내용	인물	A	B	C	D	E	F
1600달러짜리나…	A	on	SC7	SH1	→	on	on
	B	on	SC7	SH1	←	on	off
	C	on	SC7	SH1	⊙	on	off

그림 8. 대화의 영상적 요소의 내용
Fig 8. The contents of image elements in conversation

이렇게 표시된 정보는 각 요소의 값을 비교함으로서 대화 참여자를 판단하게 된다. 이렇게 대화 별로 추출된 표현 정보를 이용하여 대화의 영상적 요소를 표현하면 그림 8과 같이 나타난다. 그림 8의 대화에 대해 추출된 등자인물은 3명(A, B, C)가 되며 각 인물에 대한 영상적 요소의 내용은 그림 8 하단의 표와 같이 표시된다. 여기서 A가 표 1의 화자 판단규칙에 해당하므로 화자로 분류되며 표 2에서 1번 화자 경우에 해당하고 B와 C는 표 3에서 1번에 해당하는 청자로 분류된다.

이러한 과정을 거쳐 4개의 동영상에 대해 대화참여자 결정을 위한 실험을 진행하였으며 그 실험 결과는 그림 9와 같이 나타났다. 그림 9는 제안한 방법론에 의한 결과를 기존의 자막에만 근거하는 결과와 사람이 직접 보고 대화 참여자를 판단하는 경우와 비교한 결과이다.

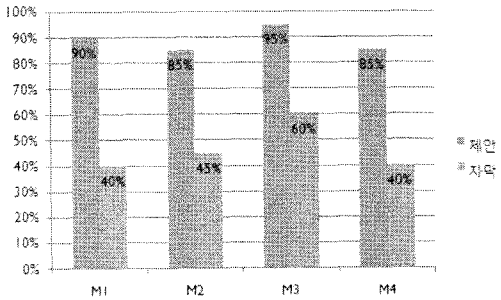

그림 9. 대화 참여자 결정에 대한 실험결과
Fig 9. The result of experiment

실험 결과 기존의 자막에 의한 대화참여자 분류 방법보다 제안한 영상적 요소에 의해 대화 참여자를 분류한 것이 더욱 정확한 결과를 나타내는 것을 확인하였다. 특히 자막에 의한 대화 참여자 분류의 경우 대화에서 화자만 인식되는 경우가 많아 혼잣말의 비중이 높게 나타나는 것으로 판단되었으며 이 부분이 오차의 대부분을 차지하였다. 즉 본 논문에서 제안하는 영상적요소에 의한 분류를 통해 이러한 오류를 대부분 개선할 수 있었다. 하지만 본 논문에서 제안하는 방법론 또한 사람이 인식하는 것보다 약간의 오류를 나타내었는데 그 이유는 얼굴인식을 못하였거나 각 요소별로 약간의 부정확한 분류 기준이 있기 때문이다. 시선방향은 1개의 대화 상황에서 지속적으로 변화해서 대푯값을 설정하기 어려운 경우가 있으며, 패턴은 정확히 추출되지 못하는 상황이 있기 때문이다. 따라서 본 논문에서 제안한 영상적 요소 이외에 추가적인 정보인 음성에 대한 분석이 필요하다. 음성 분석은 얼굴인식이 안 되는 경우에도 화자를 정확히 인식할 수 있을 것으로 추측된다.

4.2 Character-net의 개선

제안된 영상적 요소에 의해 대화 참여자를 판단하여 4개의 영화에 대해 정확한 대화 그래프를 구성한 후 누적하여 3.4절의 Character-net 구축 과정을 거쳐 그래프를 완성하였다. 이 그래프들 중 영화 '해리가 샐리를 만났을 때'에 대해 그려진 그래프는 그림 10과 11과 같다. 나머지 3개의 그래프 또한 유사한 형태로 그려졌다. 그림 10은 본 논문에서 제안한 6가지 영상적 요소를 사용해서 대화 참여자를 판단하여 구축한 그래프이고 그림 11은 기존의 대화 경계를 결정하는 방법인 자막의 유무와 지속시간으로 대화 참여자를 판단하는 기법으로 구축되어진 그래프이다.

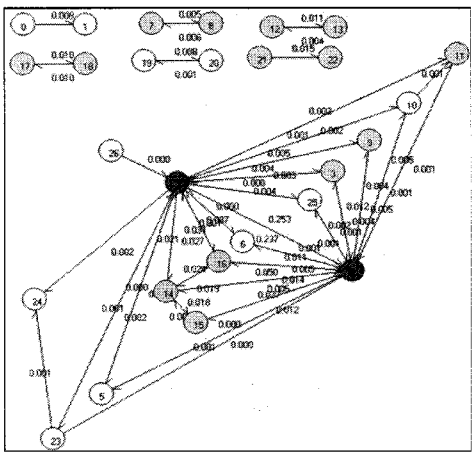

그림 10. 제안된 방법론에 의한 Character-net
Fig 10. The Character-net of proposed methodology

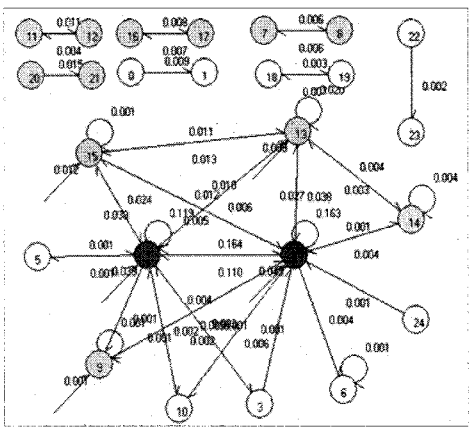

그림 11. 자막에 근거한 대화분류를 통한 Character-net
Fig 11. The Character-net of previous methodology

그림 11은 기존의 방법론인 자막에 의한 대화 참여자 분류를 통해 그려진 Character-net으로 2번과 4번이 주요 등장인물로 표시된다는 면에서는 유사하나 그림 10에 비해 혼잣말이 유난히 높게 나오는 것을 알 수 있다. 즉 4.1절에서 실험한 자막에 의한 대화 참여자 결정 방법론에서 나타난 결과가 Character-net에서도 그대로 나타나는 것을 알 수 있다. 따라서 4개의 동영상에 대해 그려진 Character-net에 대해 성능 평가를 위해 Character-net에서 차지하는 혼잣말 시간의 비율을 비교 분석하였다.

표 5. Character-net에 나타나는 혼잣말 시간 비교
Table 5. Comparison of monologue on the Character-net

비교 방법	영화	제안된 방법론	기존 방법론
$\left\|\frac{T_{tc}-T_{ac}}{T_{tc}}\right\|$	M1	100%	71.98%
	M2	96.65%	86.83%
	M3	99.66%	75.46%
	M4	100%	83.55%

표 5의 비교 방법은 전체 대화시간에서 혼잣말이 적게 포함된 비율을 의미하며 100%에 접근할수록 혼잣말이 적게 추출된 것을 의미한다. 이 기준에서 T_{tc}는 동영상내의 전체 대화시간이고 T_{ac}는 혼잣말이라고 판단된 대화의 시간이다. 대부분의 동영상 내의 대화에는 혼잣말이 전체 대화에서 아주 작은 부분을 차지하며 또한 많은 동영상에서는 혼잣말이 존재하지 않는다. 따라서 표 5에서처럼 제안된 영상적 요소에 의한 방법론이 기존 방법론에 비해 혼잣말 시간이 월등히 줄어든 것을 파악 할 수 있다. 하지만 본 논문에서 제안하는 방법으로 대화 참여자를 판단하였을 때에도 혼잣말이 아니지만 혼잣말이라고 판단하는 경우나 실제적으로 혼잣말인데 대화 참여자인 청자를 추출해내는 오류가 일부분 발생하였다. 이를 개선하기 위한 동영상의 대화 상황정보를 정확히 파악하는 추가적인 연구가 필요하다.

V. 결론

본 논문에서는 인터넷상에 있는 방대한 양의 동영상에 대한 검색에 사용되는 동영상 어노테이션 기술을 지원하는 대화 참여자 판단을 위한 영상적 요소들을 제안하였다. 각 요소들이 동영상 내의 대화 장면에서의 작용을 설명하였고 이들을 결합한 기준을 통하여 대화 참여자를 결정할 수 있음을 기술하였다. 그리고 제안된 영상적 요소를 사용하여 동영상의 내용을 표현할 수 있는 Character-net의 표현을 개선하였다. 위의 기준은 장르별로 동영상을 분석하면서 찾아내었고 또한 동영상을 통하여 기존의 연구에서의 판단과 사람에 의한 판단

과 비교하여 성능이 뛰어남을 실험으로 증명하였다.

본 논문의 연구는 영화나 TV드라마와 같은 인물 중심의 동영상을 분석하기에 적합하다. 하지만 다큐멘터리와 같이 대화 중심이 아닌 해설 위주의 동영상은 등장인물간의 대화가 적거나 존재하지 않으므로 객체들 간의 관계를 추출할 수 없다. 따라서 다큐멘터리 등과 같은 사건이나 배경 중심의 동영상을 분석하기 위한 적절한 방법에 대한 연구가 필요하다. 또한, 실험에서 언급한 것처럼 정확한 화자를 추출하기 위해 동영상 내 영상적인 요소들뿐만 아니라 음성정보도 추가하여 보다 정확한 대화 참여자를 판단하는 연구가 필요하다[9]. 그리고 향후에 효율적인 검색을 위하여 동영상에서 추출된 정보를 메타데이터 형태로 어토테이션하는 표준화 방법인 MPEG-7의 표현방식과 연계하는 연구를 진행할 예정이다.

참고문헌

[1] Y. Rui, T.S. Huang, S. Mehrotra, "Constructing Table-of-Content for Videos," to appear in ACM Multimedia Systems Journal, Special Issue Multimedia Systems on Video Libraries, Sep. 1999.

[2] S. Park, Y. Kim, M. N. Uddin, G. Jo, "Character-Net: Character Network Analysis from Video," 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence, pp. 305-308, Sep. 2009.

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

[4] 박승보, 김유원, 조근식, "얼굴인식을 이용한 동영상 상황정보 어노테이션," 한국 지능정보시스템학회, 2008 한국지능정보시스템학회 추계학술대회 논문집, 319-324쪽, 2008년 11월.

[5] 이진활, 박승보, 김유원, 조근식, "비디오 배역명 추출을 이용한 자동 어노테이션," 한국정보과학회, 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집, 제36권, 제1호(C),525-530쪽, 2009년 6월.

[6] 스티븐 D. 캐츠, *"영화연출론,"* 시공사, 185-240쪽, 1998년.

[7] 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.

[8] 이경호, 양룡, 이상범, "색상 정보를 이용한 자동 독화 특징 추출," 한국컴퓨터정보학회, 한국컴퓨터정보학회 논문지, 제13권, 제6호, 107-115쪽, 2008년 11월

[9] 김명훈, 이지근, 소인미, 정성태, "얼굴과 음성 정보를 이용한 바이모달 사용자 인식 시스템 설계 및 구현," 한국컴퓨터정보학회, 한국컴퓨터정보학회 논문지, 제10권, 제5호, 353-362쪽, 2005년 11월

[10] 김원택, 박승보, 조근식, "대화 참여자를 결정하는 영상적 요소," 한국컴퓨터정보학회, 2009년도 한국컴퓨터정보학회 하계학술대회 논문집, 제17권, 1호, 81-84쪽, 2009년 6월.

저 자 소 개

김 원 택

2007 : 길림 건축공정학원 공학사
2008 - 현재 : 인하대학교 석사과정.
관심분야 : 시맨틱 웹, 비디오 어노테이션

박 승 보

1995 : 인하대학교 공학사
1997 : 인하대학교 공학석사
1996.12 - 2002.5 : 대우전자 연구소 주임연구원
2003.9 - 현재 : 인하대학교 박사과정
관심분야 : 멀티미디어 정보검색

조 근 식

1982 : 인하대학교 공학사
1991 : City University of New York Computer Science 공학박사
1992.3 - 현재 : 인하대학교 컴퓨터정보공학과 교수
관심분야 : 인공지능, 시맨틱 웹, 지능형 에이전트 시스템