Journal of Korea Multimedia Society Vol. 12, No. 9, September 2009(pp. 1252-1261)

화자인식을 이용한 대화 상황정보 어노테이션

박승보[†], 김유원[††], 조근식[†††]

요 약

효율적인 영상의 검색과 동영상의 축약을 위해 선행되어야 하는 것이 동영상 정보에서 의미를 추출하여 영상 정보를 어노테이션 하는 작업이다. 어노테이션을 위한 동영상의 의미 정보는 다양한 방식에 의해 얻어질 수 있다. 동영상의 의미정보는 영상의 개체들의 단순한 정체 정보를 추출하는 방식과 개체들이 만들어 내는 상황정보를 추출하는 방식으로 구분될 수 있다. 하지만 개체들의 단순 정보만으로 어노테이션을 진행하기 보다는 개체들 간의 상호작용이나 관계에 대한 표현을 개체 정보와 함께 고려하여 대화 상황에 대한 온전한 의미를 어노테이션 하는 것이 더욱 좋다. 본 논문은 영상으로부터 화자정보를 추출하고 대화상황을 구성하여 어노테이션 하는 것에 대한 연구이다. 인식된 얼굴 정보로부터 현재 영상에 누가 있는 지 알아낸 후 입의 움직임을 분석하여 화자가 누구인지 알아내고, 화자와 청자 및 자막의 유무를 통해 대화 상황을 추출하여 XML로 변환하는 방법을 본 연구에서 제안한다.

Conversation Context Annotation using Speaker Detection

Seung-Bo Park[†], Yoo-Won Kim[††], Geun-Sik Jo[†††]

ABSTRACT

One notable challenge in video searching and summarizing is extracting semantic from video contents and annotating context for video contents. Video semantic or context could be obtained by two methods to extract objects and contexts between objects from video. However, the method that use just to extracts objects do not express enough semantic for shot or scene as it does not describe relation and interaction between objects. To be more effective, after extracting some objects, context like relation and interaction between objects needs to be extracted from conversation situation. This paper is a study for how to detect speaker and how to compose context for talking to annotate conversation context. For this, based on this study, we proposed the methods that characters are recognized through face recognition technology, speaker is detected through mouth motion, conversation context is extracted using the rule that is composed of speaker existing, the number of characters and subtitles existing and, finally, scene context is changed to xml file and saved.

Key words: Annotation(어노테이션), Speaker Detection(화자인식), Face Recognition(얼굴인식), Video(동영상), Context(상황정보), Multimedia(멀티미디어)

1. 서 론

UCC 사이트와 휴대폰 동호회 사이트와 동영상 포탈 등의 인터넷 사이트에는 방대한 양의 동영상이 존재한다. 이러한 동영상들은 검색을 위해 제목에 주제어나 특징을 표현 할 수 있는 텍스트를 넣어주어

※ 교신저자(Corresponding Author) : 박승보, 주소 : 인천광역시 남구 용현동 253(402-751), 전화 : 032)875-5863, FAX : 032)875-5863, E-mail : molaal@eslab.inha.ac.kr
접수일 : 2009년 3월 20일, 수정일 : 2009년 6월 16일
완료일 : 2009년 7월 27일
[†] 준회원, 인하대학교 정보공학과 박사과정
[††] 준회원, 인하대학교 정보공학과 박사과정
(E-mail : yoowon@eslab.inha.ac.kr)
[†††] 정회원, 인하대학교 컴퓨터정보공학부 교수
(E-mail : gsjo@inha.ac.kr)
※ 본 연구는 한국과학재단을 통해 교육과학기술부의 세계수준의 연구중심대학육성사업(WCU)으로부터 지원받아 수행되었습니다(R33-2008-000-10109-0).

검색 시 활용하게 된다. 최근 들어 ISO/IEC MPEG그룹에서 제안한 멀티미디어 컨텐츠(Contents) 데이터를 서술하는 표준인 MPEG-7과 같은 방식을 이용한 동영상 내의 내용과 관련된 글들을 어노테이션(Annotation) 하는 방법들이 활발히 연구되고 있다. 현재 동영상을 어노테이션 하는데 사용되는 정보는 동영상 주변의 관련 글이나 자막 같은 텍스트 정보와 영상에서 추출되는 시각기술자(Visual Descriptor)를 기반으로 하는 시각정보 등이 있다[1,2].

동영상 주변에 관련 글이 없는 경우 영상 자체의 시각정보만을 활용하여 어노테이션을 작성하여야 한다. 동영상 어노테이션을 위해 시각정보를 추출할 경우 프레임에 대한 영상 신호적인 특징, 영상에 나타나는 로고나 물체, 영상에 나타나는 의미와 같은 다양한 정보를 추출해야 한다[2-10]. 이에 대한 다양한 연구가 이루어지고 있고 그 연구의 한 분야로 영상에서 얼굴을 인식하여 누구인지를 보여주거나 얼굴들 중에서 화자가 누구인지 판단하는 연구가 있다[8,9]. 이러한 연구는 동영상에 있는 화자가 누구인지 파악하는 내용으로 프레임이나 장면(Scene)의 의미 정보를 추출하는 연구 분야에 속한다. 하지만 이보다 더 나아가 인식된 얼굴들 간의 상황정보를 추출하여 어노테이션 하면 영상의 축약이나 검색 시 양질의 정보를 제공할 수 있다. 얼굴들 간의 상황정보는 화자인식과 등장인물들 간의 대화 형태를 통해 판단할 수 있다. 자막의 시간정보를 이용해 대화의 경계를 정하고 자막을 통해 대화 형태를 결정할 수 있다. 본 논문에서는 얼굴 인식 기술을 통하여 화자를 결정하고 인식된 얼굴과 자막과의 관계를 통해 대화의 상황정보를 추출하여 어노테이션 하는 방법을 제안한다. 또한 어노테이션을 위해 상황정보가 저장될 XML 파일의 형식을 제안하고 이를 구현한 결과를 보여줄 것이다.

본 논문은 제안한 방법을 설명하기 위해 2장에서 관련된 연구들을 설명하고 3장에서 본 논문에서 제안하는 화자인식을 이용한 대화의 상황정보 추출 방법을 설명하고 4장에서 제안한 방법으로 구현한 동영상 어노테이션 과정과 결과를 보여줄 것이다. 그리고 이어서 본 논문의 결론을 서술하겠다.

2. 동영상 어노테이션을 위한 정보 추출과 관련된 연구

동영상 어노테이션 문제는 프레임이나 장면에서 중요한 정보를 추출하여 주석으로 추가하는 문제로 동영상의 정보를 추출하기 위한 다양한 접근 방법들이 있다. 이야기 구조를 가진 동영상의 경우 등장인물에 의해 주로 이야기가 진행된다. 특히나 상업용 영화나 TV 드라마의 경우 인물 위주의 진행이 나타난다. 동영상에서 정보를 알아내기 위한 방법은 색상 정보나 에지(Edge) 또는 히스토그램(Histogram)과 같은 영상의 특징을 이용하는 방법[2-5]과 프레임에서 특정한 이미지나 문자를 추출하는 방법[6-10]과 영상의 장면들이 갖는 상황정보를 추출하는 방법[11-13] 등이 있다.

첫 번째, 영상 자체의 특징을 파악하는 방식은 내용곡선이나 색정보를 이용하여 영상의 변화를 파악해 숏(Shot)을 구별하거나 비슷한 숏을 추출하는 방법으로 동영상 검색이나 동영상 요약에 주로 사용된다[2,3]. 또는 영상 내에 포함된 해나 달, 꽃과 같은 이미지들을 추출하여 인덱싱(Indexing)하는 방식으로 영상의 색 정보, 질감정보, 모양정보 등을 이용하여 영상의 특징을 파악하는데 사용된다[4,5].

두 번째, 프레임에서 특정한 이미지나 문자를 추출하는 방법은 로고 이미지, 운동선수의 운동모습, 사람의 얼굴을 추출하거나, 뉴스의 자막이나 간판 등의 글씨 정보를 오브젝트(Object) 형태로 분리하여 문자의 내용을 파악하는 것을 목적으로 한다[6-10]. 이렇게 추출된 정보는 동영상 분류에 사용할 수 있으며, 여기에 해당하는 다양한 연구들이 있다. Mark 등은 그들의 논문에서 동영상에 등장하는 인물들을 얼굴인식과 화자인식 기술 그리고 자막과 인터넷 상에 존재하는 대본을 활용하여 배역명을 알아내어 표시하는 방법을 제안하였다[10]. 이외에도 여러 연구자들에 의해 영상에 존재하는 로고 이미지를 분리하여 영상의 특징을 파악하거나 화자인식 기술을 이용하여 자막을 화자부근에 표시하여 자막과 영상에 대한 가시성을 높여주는 연구들도 진행되어 왔다[6,9]. 하지만 이 연구들은 영상으로부터 단순한 의미를 추출하여 활용하는 연구들로 개체들 간의 관계와 같은 상황정보를 추출하지는 못한다.

마지막으로, 영상이 갖고 있는 컨텐츠의 의미를 파악하여 각 장면들의 상황을 추출하는 방법이 있는데, 이 방법은 장면의 의미요소들 간의 관계를 컴퓨터가 자동으로 알아내기 힘들기 때문에 아직은 수동적으로 영상의 상황정보나 의미를 사람이 알아내어

지정하는 식으로 연구들이 이루어지고 있다[11-13]. Ling 등은 그들의 연구에서 장면의 감정정보를 어노테이션하는 방법을 제안하였는데 감정정보를 5가지의 감정 상태(Angry, Fear, Sad, Happy, Neutral)와 1개의 감정 없는 상태(Absent)로 구분하여 표시하였다[11]. 등장인물들의 대화 장면에서 전체적인 화면의 감정상태를 표시하게 하였고, 이렇게 어노테이션된 결과를 가지고 효율적인 검색이 가능하다는 것을 실험을 통해 증명 하였다. Roth 역시 장면의 정보를 수동으로 사용자가 지정하여 시맨틱 네트워크로 표현하도록 하여 장면 검색에 활용하는 연구를 진행하였다[12]. 하지만 이 연구들의 경우 화면의 상황정보를 자동화된 방식이 아닌 수작업으로 사람이 지정하는 방식이다. 따라서 구현 가능한 어노테이션 방법론에 대한 연구이기 보다는 검색의 효율성에 초점이 맞춰진 방식이다. 반면에 Liang 등은 그들의 연구에서 장면에 나타나는 개체들이 발생시키는 이벤트(Event)들을 일정 규칙에 의해 정의하고 자동으로 인식하도록 하였다[13]. 현실적으로 인터넷에 존재하는 방대한 양의 동영상을 수작업으로 어노테이션하는 것은 불가능하며, 동영상의 모든 장면에 대하여 상황 정보를 추출하는 것은 상당히 난해한 작업이다. 하지만 Liang 등의 연구에서와 같이 일정조건으로 제한된 장면에서 상황을 파악하는 연구가 이루어지고 있으며 장면에서 인식되는 대상물들과 특정 이벤트를 연계하여 장면의 상황정보를 추출할 수 있다.

3. 화자인식을 이용한 대화 상황정보 어노테이션

본 논문은 동영상의 대화 상황정보를 자동으로 추출하는 것에 대한 연구로서 상황정보를 추출하기 위해 자막정보와 화자인식 기술을 활용하였다. 동영상에서 벌어지는 대화 상황에 해당하는 범위를 경계짓기 위해 자막정보의 시간을 사용하였으며 얼굴인식기술과 화자인식 기술을 이용하여 자동으로 대화에 대한 상황정보를 추출하여 동영상을 어노테이션하도록 하였다. 본 논문에서 제안하는 시스템은 크게 3부분으로 구성되는데, 화자를 인식하는 영상 처리부와 자막 정보로부터 시간정보와 자막내용을 활용하는 Timed Text 처리부 및 대화의 상황정보를 생성하여 저장하는 대화상황 생성부로 이루어져 있다.

3.1 영상 처리부

영상 처리부는 영상에서 얼굴들을 검출하고, 검출된 얼굴이 누구인지 인식하는 얼굴 인식과 인식된 얼굴들 중에서 화자가 누구인지 판단하는 화자인식 부분으로 구성된다. 이를 위해 먼저 Timed Text 처리부로부터 자막이 표시되는 시간정보를 가져와 시간에 해당하는 프레임들을 추출하여 대화의 경계로 정하고 이 프레임들에서 얼굴을 검출하고 사전에 얼굴 정보가 저장된 DB(Data Base)에서 검출된 얼굴과 유사한 얼굴을 찾은 후 DB에 등록된 관련 정보를 가져오도록 하였다. 이 과정을 거쳐 대화중인 검출된 얼굴을 인식하고 입의 움직임을 알아내어 검출된 얼굴들 중에서 화자가 누구인지 알아내게 된다.

그림 1의 화자인식 부분에서는 인식된 얼굴이 화자인지 아닌지를 알아내기 위해 입의 움직임을 판단한다. 인식된 얼굴이 화자가 아니라면 입과 입 주변의 움직임이 거의 없어 시간에 따른 프레임 간 입 주변 영역의 히스토그램의 변화가 거의 없을 것이다. 화자일 경우 입이 움직이게 되므로 입과 입 주변 영역의 히스토그램이, 프레임 간에 경계값(Threshold)을 넘는 차이를 나타내게 된다. 이런 방식을 이용하면 인식된 얼굴들 중에서 누가 화자인지를 쉽게 알아낼 수 있다. 입 주변의 영역을 설정하기 위해 그림 2와 같이 검출된 얼굴에서 좌안의 위치 (x_1, y_1)와 우안의 위치 (x_2, y_2)를 알아낸 후 양쪽 눈 사이의 거리 d를 식 1과 같이 계산한다.

$$d = x_2 - x_1 \tag{1}$$

입의 위치는 좌측 눈의 좌표값에서 y값에 d를 더하여 입 주변영역의 좌측 상단 위치로 정하며 우측 하단의 길이는 우측 눈의 좌표값에서 y값에 d를 더하고, 보정을 위한 일정한 상수값을 추가로 더하여 식

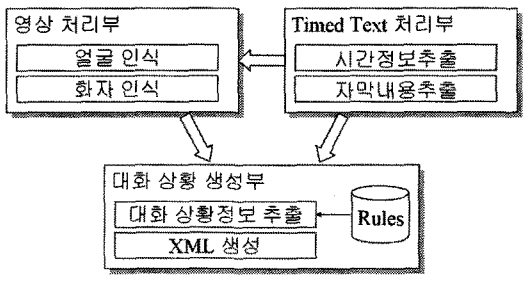


그림 1. 동영상 상황정보 어노테이션 구조도

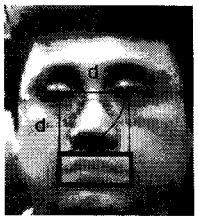


그림 2. 얼굴 영역 중 입 주변영역

2와 3과 같이 결정한다. 여기서 $Mouth_l$은 입 영역의 좌측 상단의 좌표고, $Mouth_r$은 입영역의 우측 하단의 좌표이다. C값은 상수 값으로 d 값보다 작은 값 중에서 결정하여 사용한다.

$$Mouth_l = (x_1, y_1 + d) \tag{2}$$

$$Mouth_r = (x_2, y_2 + d + C) \tag{3}$$

이렇게 설정한 결과 영역이 그림 2와 같은 얼굴 하단의 4각형 모양의 영역이 된다. 그리고 이 영역에 대한 프레임 간의 히스토그램을 비교하여 값의 차이가 기준값을 넘는 얼굴을 화자로 인식하여 입이 움직이는 것으로 판정된 얼굴은 화자, 그렇지 않은 얼굴은 청자로 구분하여 인식한다.

픽셀에 대한 히스토그램은 식 4에 의하여 계산되어 진다. 식 4에서 R은 적색(Red), G는 녹색(Green), B는 파랑색(Blue)의 값을 의미한다.

$$Y = (0.299 \times R) + (0.587 \times G) + (0.114 \times B) \tag{4}$$

입술 영역에 대한 히스토그램 그래프는 256 Gray-Level 영상으로 표현하였으며 각 명암 값의 빈도수를 조사하여 그래프의 높이로 나타내었다. 입이 열릴 때 입 안쪽이 어둡게 보여 입술이 움직일 경우 명암값 차이가 프레임 간에 차이를 나타내게 되는 된다. 반면에 인접한 프레임의 경우 조명으로 인한 명암 차이가 무시할 수 있을 정도이므로 입술의 움직임에 의한 명암 차이만이 프레임 간 히스토그램 차이에 영향을 미치는 요소가 된다. 입술영역의 픽셀에 대해 식 4에 의해 Y값을 계산하여 256 Gray-Level 중에 해당하는 Y값을 1만큼 증가 시켜준다.

식 5와 같이 입술영역에 대해 계산된 현재 n 프레임의 히스토그램 값($HistoArr_n$)과 바로 이전 $n-1$ 프레임의 입술영역에 대한 히스토그램 값($HistoArr_{n-1}$) 간의 차이에 의해 경계값($Value$)을 계산하여 경계 기준치(α)를 넘는 지의 여부에 따라 입술이 움직이는

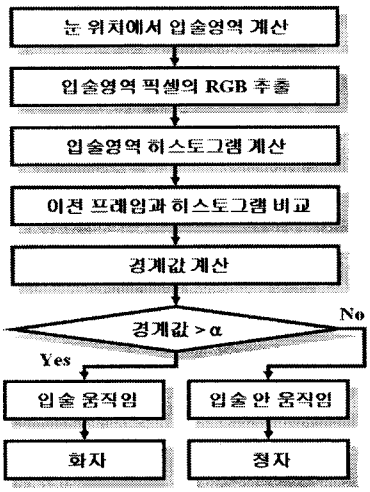


그림 3. 입술 움직임 판단 흐름도

지 또는 안 움직이는지를 판단한다. 히스토그램 값 간의 차이는 히스토그램의 동일 Level에 대해 차이를 계산하여 특정 값 이상의 차이를 보이는 경우 경계값($Value$)을 1 증가 하는 것을 256 Level별로 진행하여 경계값을 누적 계산 한다. 이 경계값이 일정한 값의 경계 기준치보다 높은 경우 입이 움직이는 것으로 판단하게 된다.

$$Value_n = HistoArr_n - HistoArr_{n-1} \tag{5}$$

검출된 얼굴의 입술이 움직이는 것으로 판단되는 경우 화자로 인식하며, 입술이 움직이지 않는 경우 청자로 판단한다.

이 과정을 정리하면 그림 3과 같이 표현된다.

3.2 Timed Text 처리부

Timed Text는 동영상과 함께 저장된 자막 정보로 DVD의 자막정보나 컴퓨터의 동영상 재생기인 윈도우 미디어 플레이어(Windows Media Player)의 SMI 형식이나 리얼 플레이어(RealPlayer)의 RealText가 여기에 해당된다. 또한 월드와이드웹(WWW) 국제 표준화 기관인 W3C(World Wide Web Consortium)에서도 DFXP라는 Timed Text 표준을 제안하고 있다. Timed Text는 그림 4에서 보여 지는 SMI형식의 내용처럼 기본적으로 시간 정보와 자막 내용으로 구성된다. 그림 4의 경우 <SYNC> 태그의 Start 값이 자막이 표현되는 시작시간을 나타내며, 표현된 자막은 다음 자막의 <SYNC> 태그의 Start 값의 시간이

```
<SAMI>
<BODY>
<SYNC Start=30862>
<P Class=EGCC><i>옛날 옛적</i>
<SYNC Start=31500>
<P Class=EGCC><i>고대 중국에 선과악 사이에
</i>
 ...
```

그림 4. SMI 형식의 Timed Text

될 때까지 화면에 유지되게 된다.

시간정보는 동영상이 시작되는 처음을 기준으로 하여 경과시간 또는 프레임 수를 표기한다. 일반적으로 경과시간을 표기 하는데 시간단위는 밀리세컨드(Millisecond) 단위를 사용한다. 자막내용은 글씨체 형식이나 언어 종류별로 해당 자막을 적어 넣는다. 자막 형식에 따라 자막의 위치를 추가적으로 넣어주는 형식도 있다. W3C에서 제안한 DFXP 자막형식이나 리얼 플레이어(RealPlayer)의 RealText 자막형식이 여기에 해당한다.

Timed Text 처리부는 Timed Text로부터 시간정보를 추출하는 부분과 자막을 가져오는 부분으로 나누어진다. 시간정보추출 부분은 Timed Text에서 자막이 나타나는 시작 시간과 자막이 사라지는 시간을 추출하여 영상 처리부에 전달하여 대화의 프레임들을 추출하는 대화 범위 시간으로 활용한다. SMI 형식의 경우 자막이 시작되는 시간은 그림 4의 <SYNC> 태그의 Start 값에서 알아내며, 자막이 사라지는 시간은 다음 자막이 나타나는 <SYNC> 태그의 Start 값을 사용할 수 있다. 뒤의 본 논문의 실험에서는 현재 국내에서 주로 사용되고 있는 SMI 형식의 Timed Text를 활용하여 진행하였으나, 다양한 형식의 Timed Text로부터 시간정보를 추출하는 추가 연구를 통해 호환성을 높일 필요가 있다. 자막내용은 <P> 태그 이후에 오는 내용으로 화면에 표현되는 대사정보가 된다. 이렇게 추출된 자막시작시간과 자막종료시간 및 자막 내용은 대화 상황 생성정보부에 전달하여 상황정보를 생성할 때 각각 자막시작시간(start_time)과 자막종료시간(end_time) 및 자막내용(subtitle)로 활용된다.

3.3 대화 상황 생성부

대화 상황 생성부는 대화상태에 대한 상황정보를 생성하는 대화 상황정보 추출 부분과 생성된 상황정보를 XML로 변환하여 저장하는 XML 생성 부분으로 구성된다. 대화 상황정보 추출부분에서는 영상 처리부에서 인식된 화자와 얼굴 정보를 수신하여 Timed Text 처리부에서 받은 자막과 연계하여 대화에 대한 상황정보를 생성한다. 이때 상황정보의 중요한 요소인 동작형태의 생성은 표 1에서 보는 바와 같이 등장인물수, 화자수, 청자수, 그리고 자막유무에 의해 이루어지는 여러 가지 규칙들(Rules)에 의해 결정된다.

대화 상황 생성부는 대화상태에 대한 상황정보를 생성하는 대화상황정보추출부분과 생성된 상황정보를 XML로 변환하여 저장하는 XML 생성부분으로 구성된다. 대화 상황정보 추출부분에서는 영상처리부에서 인식된 화자와 얼굴 정보를 수신하여 Timed Text 처리부에서 받은 자막과 연계하여 대화에 대한 상황정보를 생성한다. 이때 상황정보의 중요한 요소인 동작형태의 생성은 표 1에서 보는 바와 같이 등장인물수, 화자수, 청자수, 그리고 자막유무에 의해 이루어지는 여러 가지 규칙들에 의해 결정된다.

표 1의 규칙들을 살펴보면 다음과 같다. 화자가 인식되고 다른 얼굴도 인식된 상태에서 자막내용이 있는 경우 화자가 청자에게 말하고 있는 상태로 판단된다. 이 경우 등장인물이 2명이면 번호 1의 'say to' 상태가 되며, 청자가 2인 이상인 경우 화자 1명에 청자가 다수이므로 연설 또는 토론과 같은 상황으로 번호 4의 서로 대화하는 상태인 'talk with' 상태라고 판단한다. 그리고 번호 2와 같이 화자가 인식되지 않고 얼굴만 인식된 경우 자막이 있으므로 영상에 나타나지 않는 화자가 존재하는 것으로 화면에 보이는 얼굴이 청자가 되며 청자가 화자의 말을 듣고 있는

표 1. 등장인물 간 동작형태 판단 규칙

번호	등장 인물(명)	화자수 (명)	청자수 (명)	자막 유무	동작형태
1	2	1	1	O	(화자) say to (청자)
2	1	0	1	O	(청자) listen
3	1	1	0	O	(화자) speak alone
4	3이상	1	2이상	O	(화자) talk with (청자들)
5	2이상	0	2이상	O	(청자들) listen
6	0	0	0	O	nothing
7	-	-	-	X	nothing

상태인 'listen' 으로 판단된다. 이때 청자가 1인이면 청자는 'listen' 상태에 있는 것이며 전화를 받거나 범인이 심문을 당하는 상황 등에 해당하며, 청자가 2인 이상이면 번호 5에 해당하는 상태로 청자들 모두 'listen' 상태가 된다. 그리고 번호 3과 같이 화자가 인식되었지만 다른 얼굴은 검출되지 않았고 자막내용이 있는 경우 화자가 혼자 말하는 상태인 'speak alone'으로 추론할 수 있다. 전화 통화중인 송신자나 연설하는 사람이 이 경우에 해당한다. 번호 6과 같이 화자나 여타의 얼굴이 인식되지 않은 채로 자막만 있는 경우 얼굴과 관계된 어떠한 상황도 존재 하지 않는 'nothing' 상태인 것을 의미한다. 이 경우 자막은 오픈 캡션(Open Caption)과 같이 화면에 대한 단순한 설명이거나 화면에 나오는 문자들에 대한 번역 정보일 수도 있다. 물론 음성에 대한 분석을 실행한다면 좀 더 정확한 추론이 가능하겠으나 본 연구에선 동영상의 화상정보만을 그 분석 대상으로 하기 때문에 음성 인식에 대한 경우는 고려하지 않는다. 그리고 마지막의 번호 7의 규칙은 화자나 얼굴의 인식과 관계없이 자막이 없는 경우로 둘 간의 대화가 이루어지고 있지 않다는 'nothing' 상태를 의미한다. 서로 마주보고 있거나 여러 사람이 모여서 대기하고 있는 상태가 여기에 해당될 수 있다. 등장인물의 출현만으로 중요한 의미를 전달할 수도 있지만 본 논문에선 대화 상황만을 어노테이션 하기로 하였으므로 등장인물들의 대화가 없는 것은 배제 시켰다.

대화 상황정보 생성부분은 표 1의 규칙에 의해 판단된 대화상황의 정보를 누적하여 대화에 대한 상황정보를 생성한다. 대화에 대해 생성된 상황정보(context)는 어노테이션을 위해 XML 생성부분에서 그림 5와 같이 XML로 변환되어 저장된다. XML은 먼저 Timed Text의 시간정보를 대화상황정보(context)의 id로 지정하여 생성하게 된다. 대화상황

```
<context id="1000">
  <start_time>1000</start_time>
  <end_time>2100</end_time>
  <speaker id="A001">SeungBo Park</speaker>
  <face id="A002">John Kim</face>
  <action_type>say to</action_type>
  <subtitles>Where are you going?</subtitles>
</context>
```

그림 5. XML 파일형식

정보는 해당하는 자막으로부터 넘겨받은 자막시작시간(start_time), 자막종료시간(end_time), 자막내용(subtitles) 정보들과 인식된 화자(speaker)의 정보(id, name), 청자들(face)에 대한 정보(id, name), 규칙에 의하여 판단된 동작형태 정보(action_type)와 같이 5개의 요소로 이루어져 있다.

4. 실험결과 및 토의

본 논문에서 제안된 화자인식을 통한 어노테이션 방법에 따라 동영상의 대화상황을 어노테이션하는 소프트웨어를 프로토타입으로 구현하였다. 그리고 동영상에서 나타나는 대화 상황을 인식하여 'say to'와 'speak alone'에 대해 동작형태를 추출하여 XML로 어노테이션 하는 것을 실험하였다. 동영상은 총 5개의 대화상황을 사용하였으며 PC 카메라로 촬영한 2개와 영화 "해리가 샐리를 만났을 때..."에서 추출한 3개의 대화상황을 실험에 사용하였다.

동영상의 상황정보는 3절에서 설명한 과정을 거쳐 생성된 후 XML 형식으로 동영상을 어노테이션하게 되는데, 본 논문에서 제안한 동영상 어노테이션 방법론의 구현을 위해 윈도우 XP 운영체제에서 Visual Basic 6.0을 이용 하였다. 또한 실시간 다중 얼굴 인식 및 검색이 가능한 VeriLook Face Identification SDK에서 제공하는 API를 사용 하였으며, 입술 영역을 설정하고 히스토그램을 분석하여 화자인식을 처리하는 부분과 동작형태를 추론하는 부분 그리고 XML 생성 부분 등은 직접 구현하였다. 상기 SDK는 필드에서 검증된 상용툴로 실시간으로 전면의 얼굴을 다중으로 추출 및 인식 할 수 있다. 대화 상황에는 다수의 얼굴이 동시에 나오기 때문에 빠르게 얼굴을 인식하여 동작형태를 파악하여야 한다. 이를 위해 본 실험에선 산업계에서 얼굴인식용 툴로 활용되어 검증된 상기 SDK를 사용하였다. 상기한 SDK는 동영상에서 얼굴부분을 자동으로 검출한 후 Enrollment 기능을 이용하여 얼굴 이미지와 이름을 함께 저장할 수 있다. 한 사람에 대해 다수의 얼굴 이미지를 저장할 수 있어서 다양한 포즈나 얼굴 상태를 미리 저장하여 얼굴 인식 성능을 좀 더 높일 수 있는 특징이 있다.

구현을 위해 먼저 SMI 형식으로 저장된 Timed Text의 시간정보를 가져와서 자막에 해당하는 시간

의 대화 영상을 동영상으로부터 추출하였다. 특정 자막이 나타나는 시간 대역을 대화상황의 한 단위로 정하고 초기 프레임에서 얼굴을 검출하였다. 검출된 얼굴들은 대화에 해당하는 영상의 전체 프레임에서 검출하며 이중 대화상황의 전체 프레임들 중 1초 이상의 프레임에서 연속적으로 나타나는 얼굴들만을 검출하도록 하였다. 그림 6에서 보는 바와 같이 검출된 얼굴들은 누구인지 판단하는 과정을 거쳐 얼굴을 인식하게 된다. 얼굴 DB에 미리 등장인물의 얼굴과 이름을 저장하였는데, 한 명의 이름에 대해 다양한 얼굴 조건(포즈, 감정, 조명)에 따라 10개의 얼굴을 저장하였다. 얼굴 인식은 검출된 얼굴영역을 추출하여 저장되어 있는 얼굴 DB와 비교하여 동일한 얼굴이 있는지를 판단하여 얼굴에 대한 이름을 가져온다.

얼굴 인식 후 입이 움직이는 지를 판단하는 화자 인식 과정을 거치게 된다. 입의 움직임을 알아내기 위해 인식된 얼굴 영역 중에서 하단 부분을 분리하여 Grey Y에 대한 히스토그램 그래프를 식 4를 이용하여 구한다. 식 4에 의해 추출된 입 주변 영역에 대한 히스토그램을 표현하면 그림 7의 좌측 그래프와 같이 나타난다.

이렇게 그려진 히스토그램 그래프는 입의 움직임에 따라 그림 8과 같이 여러 가지 형태로 나타난다. 이와 같은 특징을 이용하여 식 5와 같이 이전 프레임의 입 주변 영역 히스토그램과 비교하여 경계값이 일정 수준의 경계 기준치를 넘으면 입이 움직이는

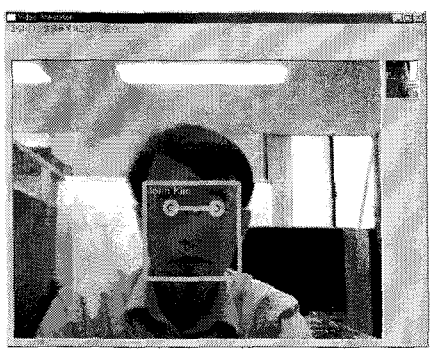


그림 6. 얼굴인식 결과

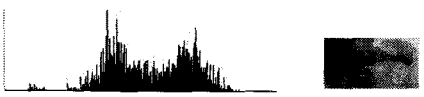


그림 7. 입 주변 영역에 대한 히스토그램

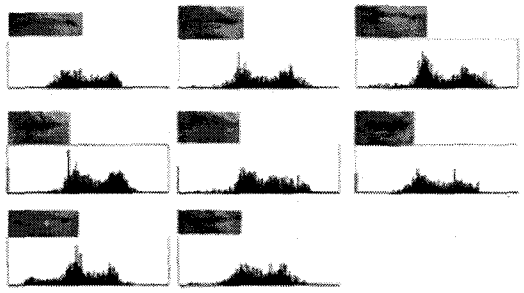


그림 8. 입의 움직임에 따른 히스토그램 변화

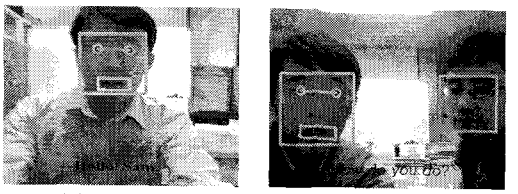
것으로 판단하여 화자로 인식한다. 그림 8은 입의 움직임에 따른 히스토그램 그래프를 시각적으로 그려준 것이며, 그림 8과 같은 히스토그램의 변화를 보이는 경우 입이 움직인 것으로 판단하게 된다. 실험에선 입의 움직임을 판단하는 경계 기준치는 3으로 설정하였다.

화자를 인식한 후, 화자와 얼굴과 자막 정보를 가지고 표 1의 규칙에 의해 각 대화의 상황을 여러 개의 동영상 클립에 적용하였다. 각 동영상에 대해 판단된 결과를 보면 표 2와 같다

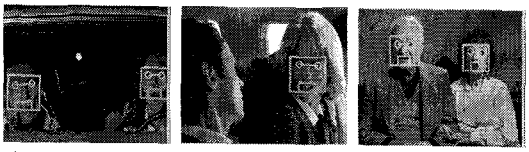
각 동영상에 대하여 그림 9과 같이 판단되었다. (a)의 경우 화자 1인만이 검출되었고 동작형태는 표 1의 규칙에 의해 번호 3의 'speak alone' 상태로 파악되었다. (b)의 경우 얼굴이 2명 인식되었으며 입의 움직임을 통해 화자가 1인, 청자가 1인 있는 것으로 파악되어 표 1에 의해 화자가 청자에게 이야기 하는 'say to'로 판정 되었다. (c)는 2인 검출과 화자인식에 의해 'say to' 라는 것은 알 수 있었으나 얼굴인식에 실패하여 화자의 정보는 알아낼 수 없었다. (d)는 1인 얼굴검출과 화자인식에는 성공하여 'speak alone'인

표 2. 동영상들의 대화 상황

화일명	영상 내용	대화 상황	얼굴 검출	얼굴 인식	화자 인식
PC_1.avi	PC 카메라 활용 #1	speak alone	성공	성공	성공
PC_2.avi	PC 카메라 활용 #2	say to	성공	성공	성공
Harry_1.avi	해리가 샐리를 만났을 때 #1	say to	성공	실패	성공
Harry_2.avi	해리가 샐리를 만났을 때 #2	speak alone	성공	성공	성공
Harry_3.avi	해리가 샐리를 만났을 때 #3	say to	성공	실패	성공



(a) PC_1.avi (b) PC_2.avi



(c) Harry_1.avi (d) Harry_2.avi (e) Harry_3.avi

그림 9. 상황정보가 추출된 동영상의 대화 영상들

것을 찾아내었다. (e)는 2인 얼굴검출과 화자인식에는 성공하였지만 얼굴인식에 실패하여 'say to'만 추출하고 대화참여자가 누구인지는 알아내지 못했다.

인식된 대화상황(context)들은 시간(start_time, end_time), 화자(speaker), 청자(face), 동작형태(action_type), 자막내용(subtitles)으로 구성 하였다. 이렇게 생성된 상황정보는 그림 10에서와 같이 XML 파일로 변환되어 저장되었다.

저장된 XML 파일은 5개의 대화에 대한 상황정보들로 PC카메라 촬영 동영상은 id가 3200과 41013인 상황정보(context)로 표 2와 같이 추출된 정보가 각각 지정되었다. 영화 동영상에 대한 상황정보는 id가 2201과 2000과 3500인 것들이다. 이 중 첫 번째 상황정보인 id가 3200만 설명하면 청자 없이 화자(John Kim)만 검출된 대화에 대한 상황정보로 화자 혼자 이야기하는 'speak alone' 상태로 지정 되었다. 각 상황정보들의 시간 정보는 자막이 나타났다(start_time) 사라지기(end_time)까지의 시간을 표시하였으며, 자막내용(subtitles)은 Timed Text의 자막내용을 그대로 지정하였고 XML 형식으로 별도의 파일로 작성되었다.

하지만 화자 인식이 정확히 되지 않아 대화 상황을 정확히 판단하지 못하는 경우가 발생할 수 있다. 이는 화자인식을 위해 입의 위치를 파악할 때 그림 11과 같이 스크린에 나타나는 얼굴의 각도나 경사에 의해 눈동자와 입술의 거리가 바뀌게 되므로 입술의 위치가 잘못 표시되는 경우이다. 이의 해결을 위해 얼굴에서 입술영역을 추출하여 위치와 모양을 알아내는 추가적인 연구가 필요하다.

```
<?xml version="1.0" encoding="UTF-8"?>
<context id="32000">
  <start_time>32000</start_time>
  <end_time>33500</end_time>
  <speaker id="A001">John Kim</speaker>
  <action_type>speak alone</action_type>
  <subtitles>Hello! Sam</subtitles>
</context>
<context id="41013">
  <start_time>41013</start_time>
  <end_time>42300</end_time>
  <speaker id="A001">John Kim</speaker>
  <face id="A002">SB Park</face>
  <action_type>say to</action_type>
  <subtitles>How do you do?</subtitles>
</context>
<context id="2201">
  <start_time>2201</start_time>
  <end_time>5648</end_time>
  <speaker id="A001">Undefined</speaker>
  <face id="A002">Undefined</face>
  <action_type>say to</action_type>
  <subtitles>Really?</subtitles>
</context>
<context id="2000">
  <start_time>2000</start_time>
  <end_time>3100</end_time>
  <speaker id="A001">Sally</speaker>
  <action_type>speak alone</action_type>
  <subtitles>I think, there is no friends between
    man and woman</subtitles>
</context>
<context id="3500">
  <start_time>3500</start_time>
  <end_time>5231</end_time>
  <speaker id="A001">Undefined</speaker>
  <face id="A002">Undefined</face>
  <action_type>say to</action_type>
  <subtitles>I met her at there</subtitles>
</context>
```

그림 10. 생성된 상황정보의 XML 파일 내용

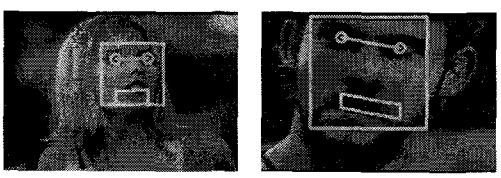


그림 11. 화자인식이 잘못된 경우

5. 결 론

본 논문은 동영상에서 여러 대화에 대한 상황정보

를 추출하여 자동으로 어노테이션 하는 방법론을 제안하였고, 구현 및 실험을 통해 이용 가능성을 증명하였다. 제안된 방법론의 구조는 영상 처리부, Timed Text 처리부, 대화 상황 생성부로 구성된다. 본 논문에서는 화자 인식을 위해 프레임 간 히스토그램 비교 분석방법을 사용하였고, 대화 상황의 핵심 요소인 동작형태 판단을 위한 다양한 규칙을 제안하였다. 이러한 규칙은 화자와 청자 간의 대화 형태를 규정한 것으로 등장인물수와 화자와 청자와 자막유무를 통해 동작 형태를 결정하도록 이루어져 있다. 또한 대화마다 상황정보를 어노테이션하기 위해 XML로 저장하는 형식을 제안하고 본 논문에서 제안한 동영상 어노테이션 방법론을 구현 실험하였다. 제안된 방법론 이용 한다면 영화 사이트나 UCC 사이트들의 많은 동영상뿐만 아니라 인터넷 상에 존재하는 무수한 동영상들을 자동으로 어노테이션 하는 것이 가능하다.

대부분 동영상의 줄거리 전개의 중심축은 등장인물이며 등장인물 간의 대화 상황을 자동으로 어노테이션 하는 것은 동영상의 줄거리 전개와 관련된 중요한 정보를 대화 단위로 확보할 수 있다는 것을 의미한다. 기존의 연구들이 영상에 있는 개체들의 단순 정보만을 어노테이션의 대상으로 삼은데 반해 본 연구는 등장인물 간의 상황을 함께 어노테이션 하여 대화 장면에 대한 폭넓은 정보를 담을 수 있게 하였다. 상황정보를 어노테이션 할 경우 개체들 간의 관계나 행동에 대한 검색이 가능해 보다 풍부한 표현의 검색이 가능해 진다. 또한 축약의 경우도 마찬가지로 장면 선택 시에 상황정보를 유용한 정보로 활용할 수 있다.

본 논문은 한 숏(Shot) 내에서만 화자와 얼굴간의 동작관계를 추론하였다. 하지만 상황적으로 사람이 있음에도 뒷모습이나 가려진 얼굴로 인해 잘못된 상황 정보를 추출할 수도 있다. 그리고 화자인식을 위해 사용한 입의 위치가 얼굴이 갖는 개인별 특징 및 크기와 각도에 따라 변화할 수 있다. 따라서 향후에 숏과 숏 사이의 얼굴들 간의 연계를 통해 검출되지 않은 사람을 인식하거나 영상으로부터 입의 특징을 찾아 얼굴의 각도와 관계없이 입의 위치를 추출하여 상황정보의 정확성을 높이는 연구가 필요하며, 또한 제안된 방법론으로 어노테이션 된 동영상을 검색하기 위한 연구가 추가적으로 필요할 것으로 판단된다.

참 고 문 헌

[1] D. Yamamoto and K. Nagao, "iVAS: Web-based Video Annotation System and its Applications," *In Proceedings of the 3rd International Semantic Web Conference, Demonstration Session,* 2004.

[2] 박주현, 낭종호, "MPEG-7시각 정보 기술자와 텍스트 정보를 이용한 내용 기반 웹 이미지 검색 시스템," 한국정보과학회 학술발표 논문집 한국정보과학회 2006. 한국컴퓨터종합학술 대회 논문집(A), pp. 232-234, 2006.

[3] 김태희, 이응희, 정동석, "MPEG 압축 영역에서 내용 곡선을 이용한 Video 요약 기법," 한국통신학회논문지, 제27권, 제10A호, pp. 1021-1028, 2002.

[4] 임동혁, 이석룡, 정진완, "비디오 검색과 시각적 요약을 위한 장면 기반 계층적 브라우징 기법," 정보과학회논문지: 데이터베이스, 제28권, 제2호, pp. 181-187, 2001.

[5] 오형철, 최종호, "에지 투영 및 방향성 벡터를 이용한 차량번호판 인식 알고리즘," 한국정보기술학회논문지, 제7권, 제1호, pp. 83-92, 2009.

[6] J.R. Cózar, N. Guil, J.M. González-Linares, E.L. Zapata, and E. Izquierdox, "Logo type detection to support semantic-based video annotation," *Signal Processing: Image Communication*, Vol. 22, Issues 7-8, pp. 669-679, 2006.

[7] J. Assfalg, M. Bertini, C. Colombo, A.D. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *Computer Vision and Image Understanding*, Vol. 92, Issues 2-3, pp. 285-305, 2003.

[8] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in TV video," *Image and Vision Computing, InPress, Corrected Proof*, May 2008.

[9] S. Park, K. Oh, H. Kim, and G. Jo, "Automatic Subtitles Localization through Speaker

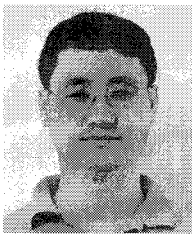
Identification in Multimedia System," *Semantic Computing and Applications, IWSCA'08 IEEE International Workshop on,* pp. 166-172, 2008.

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

[11] L. Chen, G. Chen, C. Xu, J. March, and S. Benford, "EmoPlayer: A media player for video clips with affective annotations," *Interacting with Computers,* Vol.20, pp. 17-28, 2008.

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

[13] 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, pp. 180-186, Jan. 2009.



박 승 보

1995년 2월 인하대학교 전기공학과 공학사
1997년 2월 인하대학교 전기공학과 공학석사
1996년 12월~2002년 5월 대우전자 품질경영연구소, 디지털 TV 연구소 연구원
2003년 9월~현재 인하대학교 정보공학과 박사과정
관심분야 : 영상정보 표현, 얼굴인식, USN

김 유 원

1987년 2월 경희대학교 공학사
2003년 8월 인하대학교 공학석사
2003년 9월~현재 인하대학교 정보공학과 박사과정
관심분야 : 디지털 방송, 컴퓨터비전, 멀티미디어 프로세싱

조 근 식

1982년 3월 인하대학교 전자계산학과 공학사
1991년 City University of New York Computer Science 공학박사
1992년 3월~현재 인하대학교 컴퓨터정보공학과 교수
관심분야 : 인공지능, Semantic Web, 지능형 에이전트 시스템