ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

J Intell Inform Syst 2015 March: 21(1): 83~98 http://dx.doi.org/10.13088/jiis.2015.21.1.83

효과적인 인터랙티브 비디오 저작을 위한 얼굴영역 기반의 어노테이션 방법

윤의녕
인하대학교 컴퓨터정보공학부
(entymos@hotmail.com)

가명현
인하대학교 컴퓨터정보공학부
(gagaman7777@eslab.inha.ac.kr)

조근식
인하대학교 컴퓨터정보공학부
(gsjo@inha.ac.kr)

TV를 보면서 방송에 관련된 정보를 검색하려는 많은 시청자들은 정보 검색을 위해 주로 포털 사이트를 이용하고 있으며, 무분별한 정보 속에서 원하는 정보를 찾기 위해 많은 시간을 소비하고 있다. 이와 같은 문제를 해결하기 위한 연구로써, 인터랙티브 비디오에 대한 연구가 활발하게 진행되고 있다. 인터랙티브 비디오는 일반적인 비디오에 추가 정보를 갖는 클릭 가능한 객체, 영역, 또는 핫스팟을 동시에 제공하여 사용자와 상호작용이 가능한 비디오를 말한다. 클릭 가능한 객체를 제공하는 인터랙티브 비디오를 저작하기 위해서는 첫째, 증강 객체를 생성하고, 둘째, 어노테이터가 비디오 위에 클릭 가능한 객체의 영역과 객체가 등장할 시간을 지정하고, 셋째, 객체를 클릭할 때 사용자에게 제공할 추가 정보를 지정하는 과정을 인터랙티브 비디오 저작 도구를 이용하여 수행한다. 그러나 기존의 저작 도구를 이용하여 인터랙티브 비디오를 저작할 때, 객체의 영역과 등장할 시간을 지정하는데 많은 시간을 소비하고 있다. 본 논문에서는 이와 같은 문제를 해결하기 위해 유사한 샷들의 모임인 샷 시퀀스의 모든 샷에서 얼굴 영역을 검출한 샷 시퀀스 메타데이터 모델과 객체의 어노테이션 결과를 저장할 인터랙티브 오브젝트 메타데이터 모델, 그리고 어노테이션 후 발생될 수 있는 부정확한 객체의 위치 문제를 보완할 사용자 피드백 모델을 적용한 얼굴영역을 기반으로 하는 새로운 형태의 어노테이션 방법을 제안한다. 마지막으로 제안한 어노테이션 방법의 성능을 검증하기 위해서 인터랙티브 비디오 저작 시스템을 구현하여 기존의 저작 도구들과 저작 시간을 비교하였고, 사용자 평가를 진행 하였다. 비교 분석 결과 평균 저작 시간이 다른 저작 도구에 비해 2배 감소하였고, 사용자 평가 결과 약 10% 더 유용한다고 평가 되었다.

주제어 : 인터랙티브 비디오, 저작 도구, 어노테이션, 샷 시퀀스 정렬

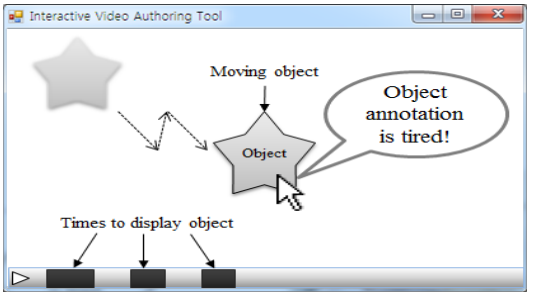
논문접수일 : 2014년 11월 26일 논문수정일 : 2014년 12월 26일 게재확정일 : 2015년 1월 2일
투고유형 : 국문급행 교신저자 : 조근식

1. 서론

최근 TV 시청 중 태블릿과 스마트폰 사용자의 84%가 쇼핑을 하거나, 방송과 관련된 정보를 검색하기 위해 자신의 모바일 디바이스를 세컨드 스크린(Second Screen)으로 사용하는 것으로 조사되었다(Nielsen, 2014). 그러나 현재 대부분의 세컨드 스크린 사용자들은 주로 포털 사이트에서 방송과 관련된 정보를 검색하고 있으며, 무분별한 정보 속에서 원하는 정보를 찾기 위해 많은 시간을 소비하고 있다. 이러한 정보 검색 방법은 정보를 짧은 시간 안에 파악하고 소비하는 나우이즘 (Now-ism)을 원하는 소비자들의 요구를 만족시켜주지 못하고 있다. 이와 같은 문제를 해결하기 위한 연구로써, 인터랙티브 비디오에 대한 연구가 활발히 진행되고 있다(Lin, 2013; Miller et al., 2011).

인터랙티브 비디오는 일반적인 비디오에 클릭

가능한 객체(Clickable Object), 영역(Area), 또는 핫스팟(Hotspot)을 동시에 제공하여 사용자와 상호작용이 가능한 비디오를 말한다. 클릭 가능한 객체를 제공하는 인터랙티브 비디오를 저작하기 위해서는 어노테이터(Annotator)가 인터랙티브 비디오 저작 도구를 이용하여 비디오 위에 어노테이션 될 객체의 영역과, 객체를 클릭할 때 보여줄 정보를 지정 하는 과정이 필요하다. 그러나 <Figure 1>과 같이 기존의 저작 도구를 사용하여 객체를 수동으로 어노테이션 할 때 객체가 등장할 시간과, 객체가 증강될 위치를 지정하는 일련의 반복되는 작업으로 인해 많은 시간을 소비하고 있다(Lee et al., 2014; Yoon et al., 2014).



〈Figure 1〉 Limitation of Manual Annotation

본 논문에서는 이러한 문제를 해결하기 위해서 얼굴 영역을 기반으로 하는 새로운 형태의 객체 어노테이션 방법을 제안한다. 이를 통해 인터랙티브 비디오 시청자가 관심 있어할만한 드라마나 영화의 등장인물이 입고 있는 옷이나, 가방, 또는 엑세서리 등에 관련된 추가 정보를 쉽고, 효율적으로 어노테이션 하고자 한다. 얼굴 영역을 기반으로 하는 이유는 영상에서 등장인물의 얼굴을 기준으로 등장인물이 입고 있는 옷이나 가방, 엑세서리 등의 상대 위치가 일반적으로 크게 변하지 않기 때문이다.

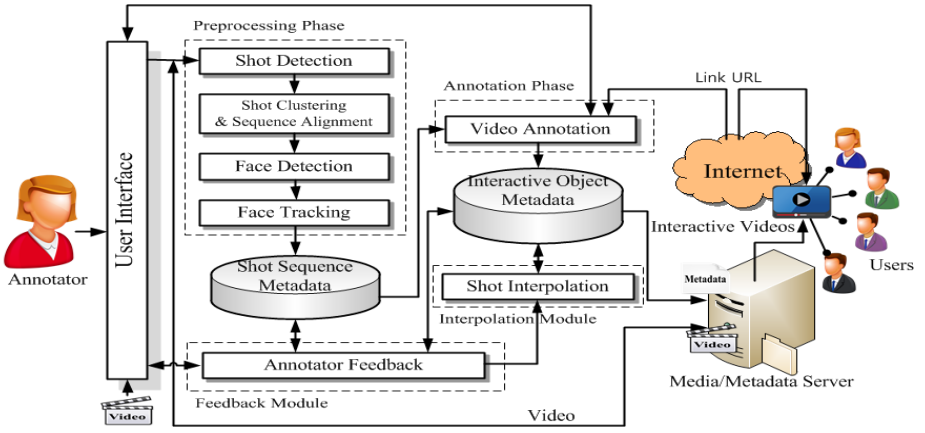
이를 위해 본 논문의 2장에서는 기존의 인터랙티브 비디오 어노테이션 도구들을 설명한다. 3장에서는 효과적인 인터랙티브 비디오 저작을 위한 얼굴영역 기반의 어노테이션 방법을 이용한 저작 시스템의 구조와 새로운 형태의 객체 어노테이션 방법을 위한 전처리 과정과, 어노테이션 과정을 설명한다. 또한, 전처리 과정 또는 어노테이션 과정 후 발생할 수 있는 잘못 정렬된 샷의 문제, 객체의 부정확한 위치 문제들을 보완하기 위한 피드백 모델과, 피드백 후 발생 할 수 있는 샷 사이에 객체가 어노테이션 되지 않은 샷이 있는 경우를 위한 보간 모듈에 대해 설명 한다. 4장에서는 기존의 저작 방법과 제안하는 어노테이션 방법의 성능을 비교하고, 인터랙티브 비디오 저작 경험이 있는 사용자들을 대상으로 한 사용자 평가 결과를 분석한다. 마지막으로 5장에서는 결론을 맺고 향후 연구 방향을 제시한다.

2. 관련 연구

2.1. 인터랙티브 비디오 저작 도구

인터랙티브 비디오 저작 도구는 기존의 비디오 위에 추가 정보를 갖는 객체를 어노테이션 하는 다양한 기능을 제공한다. 기존의 인터랙티브 비디오 저작 도구들로써 Popcorn Maker(Mozilla, 2014), Zentrick(Zentrick, 2014)와 같이 수동으로 객체의 위치 및 크기를 지정하는 방법을 사용하는 저작 도구와, wireWAX(wireWAX, 2014) 와 같이 비전 기반의 객체 검출 기법을 이용한 어노테이션 기법을 사용하는 저작 도구 등이 있다.

Popcorn Maker는 Mozilla에서 만든 인터랙티브 비디오 저작 도구로써 객체를 레이어의 형태



〈Figure 2〉 Interactive Video Authoring System Architecture

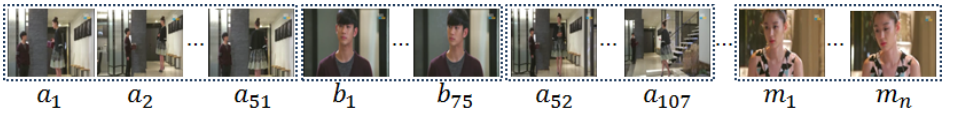
로 특정 위치에 어노테이션하며, 객체가 화면에 증강될 시작 시간과 종료 시간을 레이어(Layer)에 있는 막대의 크기를 조정하거나 위치를 조정하여 설정한다. 객체가 나타나는 시간을 조작하는 방법은 매우 간단하지만, 영상을 확인하며 모든 객체의 등장 시간을 각각 지정해야 하기 때문에 저작 시간이 오래 걸린다. 또한 한번 사용된 객체는 재사용하지 못하는 단점이 있으며, 위치가 고정된 객체를 어노테이션 할 수 있지만, 움직이는 객체를 어노테이션 하는 기능은 지원하지 않는다. 하지만 수동으로 객체를 어노테이션 함으로써 정확도가 높아지는 장점이 있다.

Zentrick은 Popcorn Maker 유사한 인터페이스를 갖고 있다. 그러나 Popcorn Maker와는 달리 움직이는 객체를 어노테이션 할 수 있으며, 한번 사용된 객체를 재사용하기 위해 증강된 객체를 새로운 시간대로 복사하여 사용할 수 있다. 그러나 여전히 객체의 등장시간과 객체의 위치를 수동으로 지정하는데 많은 시간을 소비해야 한다.

wireWAX는 어노테이터가 원하는 객체의 영역을 선택하여 해당 영역과 유사한 영역을 영상 내의 모든 프레임에서 검출할 수 있다. 어노테이터는 검출된 객체의 영역을 선택하고 해당 위치에 추가 정보를 지정 할 수 있으며, 사용된 객체의 정보를 재사용할 수 있다. 또한 검출된 객체에 쉽게 레이블링(Labeling)하기 위해 모든 유사한 객체를 수동으로 정렬하고 한 번에 레이블링 할 수 있는 기능이 제공된다. 하지만 객체를 검출하는데 많은 시간이 소비되며, 비디오의 크기가 커질수록 수동으로 정렬해야 할 객체가 많아져 어노테이션 시간이 오래 걸리는 단점이 있다.

이와 같이 기존의 인터랙티브 비디오 저작 도구들의 단점인 저작 시간을 줄이기 위해 수동 어노테이션 방법의 장점인 정확도와 비전 기반의 어노테이션 방법의 장점인 편의성을 결합한 새로운 형태의 어노테이션 방법에 대한 연구가 필요하다.

3. 효과적인 인터랙티브 비디오 저작을 위한 얼굴영역 기반의 어노테이션 방법



〈Figure 3〉 Input Frames

3.1 인터랙티브 비디오 저작 시스템 구조

제안하는 인터랙티브 비디오 저작 시스템의 구조는 <Figure 2>와 같다. 어노테이터가 유저 인터페이스를 통해 객체를 어노테이션 하려는 비디오를 서버에 업로드 하거나, 서버에 있는 비디오를 불러오면, 저작 시스템에서는 자동으로 전처리 단계를 수행한다.

전처리 단계의 샷 검출(Shot Detection) 단계에서는 비디오의 모든 프레임을 추출하여 샷을 생성한다. 샷 클러스터링 & 시퀀스 정렬(Shot Clustering & Sequence Alignment) 단계 에서는 유사한 샷을 묶고, 시간 순으로 정렬한 샷 시퀀스들을 생성한다. 얼굴 검출(Face Detection)과 얼굴 트래킹(Face Tracking) 단계에서는 샷 시퀀스(Shot Sequence)들의 모든 샷의 키프레임에서 등장인물의 얼굴을 검출한다. 전처리 과정이 끝나면 결과를 샷 시퀀스 메타데이터에 저장한다. 마지막으로 피드백 모듈(Feedback Module)에서는 어노테이터의 피드백을 통해 저장된 샷 시퀀스 메타데이터에서 잘못 인식된 얼굴의 영역 선택하여 재설정하거나, 샷 시퀀스를 재정렬 한다.

어노테이션 단계에서는 어노테이터가 전처리 단계로부터 얻은 샷 시퀀스 메타데이터 중 얼굴이 인식된 한 개의 샷의 키프레임을 선택하여 얼굴의 상대 위치에 클릭 가능한 객체를 어노테이션하고, 어노테이터가 관련 정보를 갖는 링크 URL을 가져와 객체에 지정해 준다. 그러면 저작 도구는 자동으로 샷 시퀀스의 모든 키프레임에 등장한 동일한 등장인물 얼굴의 상대위치에 앞에서 어노테이션 한 객체의 정보를 이용하여 어노테이션하며, 그 결과를 인터랙티브 오브젝트 메타데이터에 저장한다. 생성된 메타데이터 정보는 피드백 모듈과 보간 모듈(Interpolation Module)을 통해 최종 수정되어 어노테이션 된 객체들의 위치 정확도를 향상 시킨다.

피드백이 완료되면 어노테이터는 인터랙티브 오브젝트 메타데이터(Interactive Object Metadata)를 서버에 저장한다. 저장된 메타데이터는 사용자들이 이용할 인터랙티브 비디오 플레이어에서 비디오 위에 객체를 증강하기 위해 불려와 사용된다.

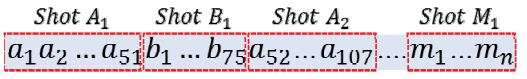
3.2. 전처리 과정

전처리 과정은 어노테이터가 컨텐츠를 쉽게 찾을 수 있도록 비디오의 내용을 샷 시퀀스로 나누어 각 샷의 대표 이미지를 썸네일(Thumbnail)로 만든다.

샷 검출 단계에서는 샷을 검출하기에 앞서 <Figure 3>과 같이 입력된 비디오로부터 프레임들을 불러온다. <Figure 3>에 있는 각각의 프레임들($a_1, b_1, \cdots, m_n$)에서 $a, b, \cdots, m$은 비디오 영상의 서로 다른 샷들을 명시적으로 표현한 것이다. 그리고 $1, 1, n$은 유사한 샷들을 모두 연결하였을 때, 그 샷 시퀀스에 속한 프레임들의 순서를 말한다. 여기서 샷은 한 대의 카메라가 영상을 찍기 시작하면서 부터, 끝날 때까지의 프레임

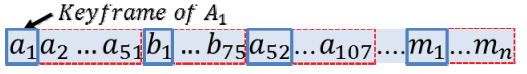
들의 모임을 말한다(Rui et al., 1999). 또한 a_1과 a_{52}는 유사한 샷에 속한 프레임으로 같은 알파벳을 사용하였다.

프레임을 불러온 후 프레임들로부터 샷의 경계를 검출하는 방법으로 널리 사용되는 칼라 히스토그램 기반의 샷 경계 검출 방법(Lienhart, 1998)을 이용하여 샷을 <Figure 4>와 같이 검출한다.



〈Figure 4〉 Detected Shots

이후 샷을 정렬하고 등장인물의 얼굴을 검출하기 위해 샷의 대표 프레임인 키프레임을 지정한다. 키프레임은 <Figure 5>와 같이 검출된 샷들의 첫 번째 프레임을 사용한다.

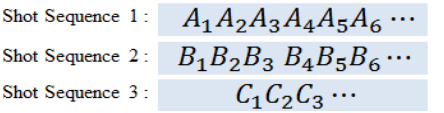


〈Figure 5〉 Keyframe Selection

샷 클러스터링 & 시퀀스 정렬 단계에서는 모든 샷들의 유사도를 구한 뒤 그룹화 하여 시간 순서대로 정렬한다(Chasanis et al., 2009). 유사도를 구하기 위해 키프레임들($a_1, b_1, a_{52}, \cdots, m_1$)의 히스토그램($H_{a_1}, H_{b_2}, H_{a_{52}}, \cdots, H_{m_1}$)을 각각 구한다. 그리고 구한 히스토그램들을 이용하여 자기 자신을 제외한 모든 키프레임 간의 유사도를 바타차야 계수(Bhattacharyya Coefficient) 식 (1)을 통해 구한다. 바타차야 계수는 일반적으로 두 개의 확률 분포 사이의 유사도을 측정하기 위해 사용된다(Lee et al., 2011).

$$BC(H_i H_j) = \sum_{x \in X} \sqrt{H_i(x) H_j(x)} \quad (1)$$

식 (1)에서는 $a_1 < i, j < m_1$인 i번째 히스토그램 H_i와 j번째 히스토그램 H_j에 속한 색상 분포를 이용하여 두 히스토그램 간의 유사도를 구한다. 그 다음 유사도가 0.8 이상인 샷들은 그룹화 하여 시간 순으로 정렬한다. 샷을 유사한 샷으로 이루어진 시퀀스로 만드는 이유는 특정 등장인물이 등장하는 샷들을 찾기 위해 소비되는 시간을 줄일 수 있기 때문이다. <Figure 6>은 유사한 샷들을 정렬한 샷 시퀀스이다.



〈Figure 6〉 Aligned Shot Sequences

얼굴 검출 단계에서는 샷 시퀀스 생성 후 미리 학습된 Haar-like 특징을 이용해서 샷 시퀀스의 제일 앞에 있는 모든 키프레임에 등장하는 인물들의 얼굴을 검출 한다(Viola and Jones, 2001; Swiki, 2014). 그리고 얼굴 트래킹 단계에서 샷 시퀀스 중 하나의 시퀀스를 선택하여 첫 키프레임에서 인식된 얼굴의 개수만큼 모든 시퀀스 내의 모든 프레임들에서 각각 얼굴을 트래킹한다. 이때 각각의 얼굴에 숫자를 지정하여 동일한 인물의 얼굴임을 표시 한다. 얼굴을 트래킹하기 위한 방법으로 트래킹에 우수하다고 알려진 루카스-카나데 옵티칼 플로우 방법(Lucas-Kanade Optical Flow Method)을 사용하였다(Lucas and Kanade, 1981). 얼굴이 검출된 키프레임을 포함한 샷 시퀀스들의 정보는 <Figure 7>과 같이 샷 시퀀스 메타데이터에 저장한다.

- Detected Shots (List)
  - Keyframe Number (Integer)
  - Current Sequence Number (Integer)
  - Keyframe Image Bitmap (Image)
  - Detected Face Areas (List)
    - Face Number (Integer)
    - X Coordinate (Integer)
    - Y Coordinate (Integer)
    - Width (Integer)
    - Height (Integer)

〈Figure 7〉 Shot Sequence Metadata

3.3 어노테이션 과정

Shot Sequence 1 : $[A_1^*]A_2^*A_3^*A_4^*A_5A_6\cdots$

[A*] : selected key frame
A* : key frame with extracted face
A : key frame with no extracted face

〈Figure 8〉 Selected Shot Sequence

어노테이터는 객체를 어노테이션 하기 전에 전처리 단계에서 생성한 샷 시퀀스들 중 객체를 어노테이션 하려는 하나의 샷 시퀀스를 선택한다. 그리고 선택된 샷 시퀀스 중 <Figure 8>와 같이 얼굴이 검출된 하나의 키프레임 $[A_1^*]$을 선택한다.

비디오 어노테이션 단계에서는 <Figure 9>와 같이 등장인물이 입고 있는 옷, 또는 장신구의 추가정보를 갖는 객체를 어노테이션 하기 위해, 어노테이터가 클릭 가능한 객체($ObjX$)의 영역(위치, 크기)을 정하고, 객체에 추가적인 상품 정보(이미지, 위치, 상품 구매 웹페이지 링크 등)를 지정 한다. 그러면 저작 시스템은 얼굴($Face$)영역의 외곽과 객체($ObjX$)의 중점 사이의 최단거리를 맨해튼 거리(Manhattan Distance)를 이용하여 구하는데 이를 상대거리 $X_{RelativeDis}$로 사용한다.

〈Figure 9〉 Face Area based Object Annotation Method

제안하는 시스템은 저작 시간을 줄이기 위해 <Figure 8>과 같이 어노테이터가 직접 어노테이션 한 키프레임($[A_1^*]$) 이외에 얼굴이 검출된 나머지 키프레임들(A_2^*, A_3^*, A_4^*)에 자동으로 어노테이션 한다. 이를 위해 앞에서 구한 $X_{RelativeDis}$의 너비($X_{RelativeDis}^{width}$), 높이($X_{RelativeDis}^{height}$)와 구하고자 하는 키프레임에 등장하는 인물의 얼굴 위치(F^x, F^y), 반지름(F_{Radius})을 사용해 A_2^*, A_3^*, A_4^*키프레임들에서의 객체가 위치할 x, y좌표를 각각 식 (2), 식 (3)을 이용하여 구한다. 그리고 앞에서 어노테이션 했던 객체의 정보를 그대로 가져와 어노테이션 할 객체에 지정한다. 식 (2), 식 (3)에서 t_n은 n번째 키프레임의 프레임 순서를 말한다.

$$O^x(t_n) = F^x(t_n) + X_{RelativeDis}^{width} + F_{Radius}(t_n) \quad (2)$$

$$O^y(t_n) = F^y(t_n) + X_{RelativeDis}^{height} + F_{Radius}(t_n) \quad (3)$$

객체 어노테이션이 완료되면, 저작 시스템은 자동으로 <Figure 10>과 같이 인터랙티브 오브젝트 메타데이터를 생성한다.

- Object Name (String)
- Object Image URL (String)
- Product Information URL (String)
- Start Time (Integer)
- End Time (Integer)
- Object Area Information
  - X Coordinate (Integer)
  - Y Coordinate (Integer)
  - Width (Integer)
  - Height (Integer)

〈Figure 10〉 Interactive Object Metadata

만약 객체가 제대로 어노테이션이 되지 않았다면, 어노테이터는 피드백 통해 객체의 위치, 등장할 시간을 수정한다. 객체의 어노테이션이 성공적으로 이루어졌다면 어노테이터는 생성된 인터랙티브 오브젝트 메타데이터를 서버에 업로드 한다. 인터랙티브 오브젝트 메타데이터는 객체의 이름, 유저 인터페이스에서 보여줄 객체 의 이미지, 객체를 클릭했을 때 보여줄 제품 정보 링크 URL, 어노테이터가 지정한 객체의 등장 시간(Start Time, End Time), 그리고 객체의 크기, 위치로 구성되어 있다. 만약 동일한 객체가 여러 번 사용되는 경우 각각의 객체 정보를 메타데이터에 추가한다.

3.4. 피드백

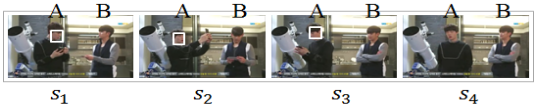
전처리, 어노테이션 단계 결과 샷이 잘못 정렬되거나, 객체가 부정확한 위치에 어노테이션 될 수 있다. 이러한 문제들로 인해 어노테이터는 인터랙티브 비디오를 저작하는데 더 많은 시간을 소비하게 된다. 그러므로 이러한 문제들을 효율적으로 해결할 수 있는 방안이 필요하다. 이를 위해 본 논문에서는 피드백 모듈을 추가하고 어노테이터가 인터랙티브 비디오를 저작 중 메타데이터 수정이 필요할 경우를 위한 피드백 모델을 제안한다. 샷 시퀀스 메타데이터와 인터랙티브 오브젝트 메타데이터는 각각 서로 다른 피드백 절차를 밟는다.

3.4.1 샷 시퀀스 메타데이터 피드백

샷 시퀀스 메타데이터에서는 크게 두 가지의 문제가 발생될 수 있다.

〈Figure 11〉 Example of Wrong Sorted Shot Problem

<Figure 11>에서 s_3는 잘못 정렬된 샷의 키프레임 이다. 이를 위해 어노테이터는 피드백을 통해 잘못 정렬된 샷을 제거하거나, 필요에 따라 샷을 추가한다.



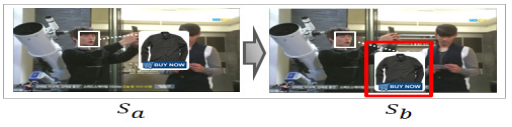
〈Figure 12〉 Example of Face Detection Failures

<Figure 12>는 얼굴 트래킹이 정상적으로 안 된 경우와 얼굴이 검출되지 않은 경우를 보여준다. <Figure 12>에서 A는 얼굴이 s_3까지는 정상적으로 검출되었지만 s_4에서 얼굴이 검출되지 않았다. 이를 해결하기 위해 어노테이터는 s_4의 이미지를 확인하고, s_3와 유사한 샷이라고 판단될 때, 피드백을 통해 임의로 얼굴 영역을 지정한다. <Figure 12>에서 B는 s_1에서 얼굴이 인식되지 않아 s_2이후 트래킹이 안 된 경우를 보여준다. 이를 해결하기 위해 어노테이터가 임의로 모

든 키프레임에 얼굴 영역을 지정한다. 만약 샷 시퀀스에 있는 이미지가 많은 경우 샷 시퀀스의 첫 번째 키프레임 이미지인 s_1과 마지막 키프레임 이미지인 s_4에만 얼굴 영역을 지정하고 보간한다.

3.4.2 인터랙티브 오브젝트 메타데이터 피드백

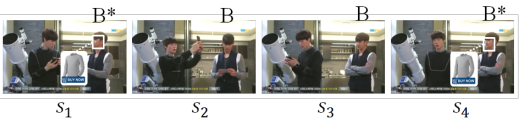
<Figure 13>와 같이 영상이 재생되며 샷 내부의 등장인물들이 움직이다가 증강된 객체에 의해 s_a 에서 다른 등장인물이 가려지는 문제가 발생할 수 있다. 이를 위해 시청자의 시청을 방해하지 않는 선에서 어노테이터가 임의로 메타데이터에서 객체의 위치정보를 수정하여 <Figure 13>의 s_b와 같이 업데이트한다.



〈Figure 13〉 Example of Wrong Position Problem

3.5 샷 보간 모듈

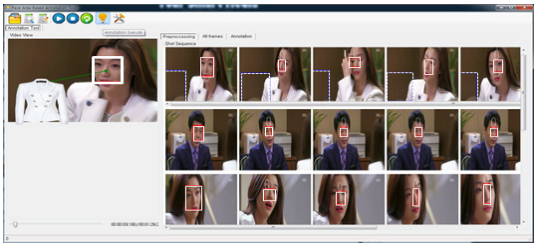
샷 보간 모듈 에서는 피드백에 의해 <Figure 14>와 같이 객체가 어노테이션 된 샷들(s_1, s_4)이 있을 때, s_1, s_4사이에 있는 객체가 어노테이션 되지 않은 샷들(s_2, s_3)에 어노테이션 될 객체의 위치를 추정한다. 보간 방법은 선형 보간법(Linear Interpolation)을 이용하며, s_1, s_4에 있는 등장인물의 얼굴 영역을 이용해 s_2와 s_3에 있는 등장인물의 얼굴 영역을 보간법을 이용해 구한다. 그리고 그 얼굴 영역의 상대 위치에 객체를 어노테이션 한다.



〈Figure 14〉 Shots with Face Detection Failures

4. 구현 및 실험

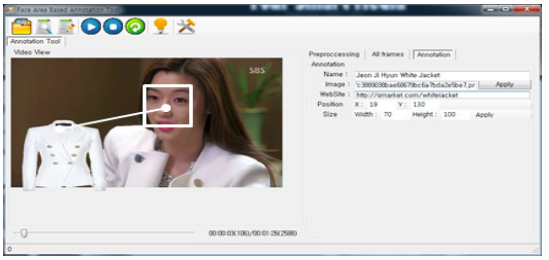
4.1 사용자 인터페이스



〈Figure 15〉 User Interface

<Figure 15>는 제안한 인터랙티브 비디오 저작 시스템의 인터페이스이다. 어노테이터는 인터페이스를 통해 비디오를 불러오거나, 재생, 정지 등의 비디오 컨트롤을 할 수 있다. 어노테이터는 정렬된 샷 시퀀스에서 첫 번째 샷의 키프레임을 선택하여 객체를 어노테이션 할 수 있는데, 이를 위해 마우스를 이용하여 얼굴 영역으로부터 상대거리에 객체의 위치와 크기를 지정한 후 객체의 추가 정보를 <Figure 16>과 같이 입력한다. 어노테이션 후 어노테이션 한 객체를 이용 나머지 샷 시퀀스의 모든 키프레임에 어노테이션 한다.

객체가 위치할 영역과 객체의 추가 정보가 지정 되면, 샷 시퀀스에 있는 나머지 샷의 키프레임에도 검출된 얼굴의 상대거리에 객체를 어노



〈Figure 16〉 Additional Information to the Object

테이션하며, 메타데이터에 저장한다. 어노테이터는 인터페이스에 있는 피드백을 버튼을 클릭하여 잘못 정렬된 샷 시퀀스를 재정렬 할 수 있으며, 어노테이션 된 결과인 인터랙티브 오브젝트 메타데이터를 객체의 추가정보 입력 창을 통해 수정할 수 있다. 또한 보간 버튼을 클릭하여 첫 번째 샷의 키프레임과 마지막 샷 키프레임 사이에 어노테이션 되지 않은 객체의 위치를 추정할 수 있다. 마지막으로 어노테이션 된 결과를 메타데이터로 저장하고, 그 메타데이터를 불러와 프리뷰(Preview)의 형태로 재생할 수 있다.

4.2 어노테이션 성능 평가

제안한 어노테이션 방법의 성능을 평가하기 위해 인터랙티브 비디오를 저작 할 때 소요되는 어노테이션 시간을 측정 하였으며, 이를 위해 <Table 1>의 대화 씬이 주로 나오는 5개의 드라마 비디오 클립(Video Clip)을 사용하였다.

〈Table 1〉 Video Clip Dataset

Video ID	Genre	Duration
V1	Drama	01:32
V2	Drama	04:24
V3	Drama	03:37
V4	Drama	04:46
V5	Drama	04:23

비디오에 객체를 어노테이션 하기 위해 제안하는 어노테이션 방법과 다른 방법을 사용하는 저작 도구인 wireWAX, Popcorn Maker, Zentrick을 이용하였으며, 영상 안에 등장하는 특정 인물이 입고 있는 옷에 추가 정보를 갖는 객체를 어노테이션 하는 시간을 <Table 2>와 같이 각각 측정하였다.

〈Table 2〉 Comparison of Annotation Time

Video ID	Proposed Method	wireWAX	Popcorn Maker	Zentrick
V1	1:30	3:46	4:59	4:05
V2	4:01	4:47	3:22	3:29
V3	1:12	5:10	2:13	3:20
V4	1:44	6:30	2:59	2:29
V5	38	3:29	2:48	2:02

<Table 2>의 실험 결과 제안한 방법을 이용한 평균 어노테이션 시간인 1분 49초와 다른 저작 도구를 이용한 어노테이션 평균인 3분 38초를 통해 어노테이션 성능이 2배 향상된 것을 확인할 수 있었다. 다른 저작 도구는 어노테이션 하려는 특정 등장인물이 입고 있는 옷의 등장 시간과 위치를 찾기 위해 영상의 내용을 확인하는데 많은 시간을 소요하였다. 반면 제안한 방법은 유사한 샷을 묶어 썸네일의 형태로 제공하기 때문에 쉽게 특정 등장인물을 찾고 그 인물이 입고 있는 옷의 위치 및 시간을 찾을 수 있어서 짧은 시간 안에 객체를 어노테이션 할 수 있었던 것으로 판단된다.

비디오 V2의 경우 제안한 방법의 어노테이션 시간이 수동 어노테이션 방법을 사용하는 Popcorn Maker와 Zentrick보다 오래 걸린 것을 확인 할 수 있는데, 그 이유는 제안한 저작 시스템의 전처리 단계에서 실제 샷의 개수인 6개가

아닌 20개의 샷을 잘못 검출하였고, 등장인물 얼굴의 영역이 잘못 인식되어 어노테이터가 피드백을 통해 수동으로 다시 보정하는데 많은 시간이 소비하였기 때문이다. V2와 같이 어노테이션 시간이 오래 걸리는 경우, Edge Change Ratio (Lienhart, 1998)와 같은 향상된 샷 경계 검출 기법을 이용하거나 Modified Census Transform 기반의 얼굴 검출 방법(Froba and Ernst, 2004)과 같이 성능이 우수한 얼굴 검출 방법을 사용하여 에러를 최소화함으로써 피드백 시간을 줄일 수 있을 것으로 예상된다.

wireWAX의 경우 객체를 검출하고 추적하는 기법을 사용하기 때문에 저작 시간이 짧을 것으로 예상했으나, 특정 객체(등장인물의 옷)를 비디오 내에서 제대로 검출하지 못해 많은 시간을 피드백 하는데 소비하였다.

4.3 시스템 평가

제안한 어노테이션 방법을 이용한 저작 시스템의 유용성을 평가하기 위해 인터랙티브 비디오 저작 경험이 있는 사용자 19명을 대상으로 설문을 진행 하였다. 본 논문에서 사용한 설문 방법은 IBM의 Computer System Usability Questionnaire(Lewis, 1995)이며, 컴퓨터 시스템에 대한 유용성을 평가하는 방법이다. 이 방법은 사용자가 미리 구성된 19개의 문항들을 이용하여 시스템을 점수로 평가하며, 최저 1점부터 최대 7점까지 평가가 가능하다.

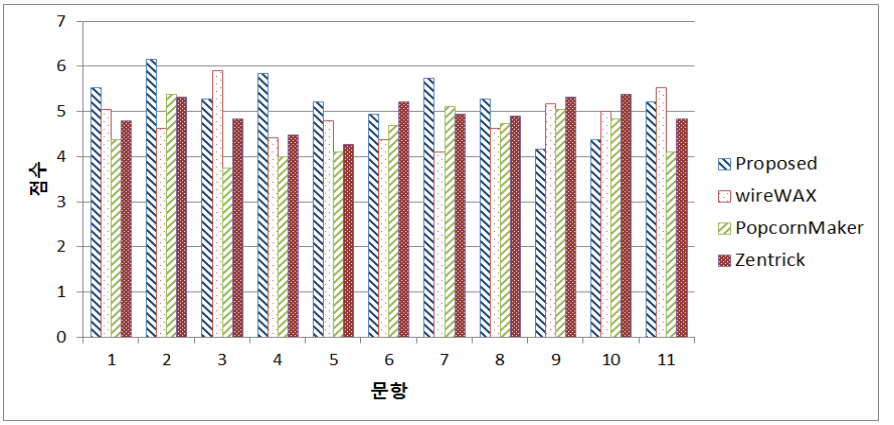
본 논문에서는 제안한 시스템을 평가하기 위해 <Table 3>과 같이 CSUQ의 19개 문항 중 11개 문항을 사용하였다. 사용하지 않은 나머지 문항들은 정보의 질(Information Quality)을 평가하는 문항이므로 제안한 시스템을 평가하기에 부적절하여 제외하였으며, 19번째 문항은 시스템의 전체적인 만족도를 묻는 문항인데, 전체 문항의 평균으로 대체할 수 있어 제외하였다.

〈Table 3〉 CSUQ

Score Name	Question ID	Questions
SYSUSE	Q1	Overall, I am satisfied with how easy it is to use this system
	Q2	It is simple to use this system
	Q3	I can effectively complete my work using this system
	Q4	I am able to complete my work quickly using this system
	Q5	I am able to efficiently complete my work using this system
	Q6	I feel comfortable using this system
	Q7	It was easy to learn to use system
	Q8	I believe I became productive quickly using this system
INTERQUAL	Q16	The interface of this system is pleasant
	Q17	I like using the interface of this system
	Q18	This system has all the functions and capabilities I expect it to have

제안한 어노테이션 방법을 사용하는 저작 시스템의 평가에 참여하는 사용자 19명은 제안한 인터랙티브 비디오 저작 시스템과, popcorn Maker, Zentrick, wireWAX를 이용하여 각각의 저작 도구에서 인터랙티브 비디오를 저작해보고 선택된 11개의 문항에 점수를 매겼다.

사용자 평가 결과 <Figure 17>과 같이 시스템 유용성에 대한 문항인 Q1～Q8의 점수는 제안한 저작 시스템이 다른 저작 시스템에 비해 평균 0.8점 높은 평가를 받았다. 특히 Q2, Q4, Q7이 각각 평균 1.1점, 1.5점, 1점 더 높은 평가를 받았



〈Figure 17〉 CSUQ Results

다. 이를 통해 본 논문에서 제안한 저작 시스템을 사용하였을 때, 쉽고 빠르게 인터랙티브 비디오 저작이 가능하며, 저작 방법을 빠르게 배울 수 있다는 점에서 상대적으로 제안한 저작 시스템이 다른 저작 도구에 비해 더 유용함을 확인할 수 있었다. 그러나 Q3의 경우 wireWAX가 제안한 저작 시스템 보다 더 높은 점수를 받았는데, 이는 wireWAX가 객체를 인식하여 쉽게 어노테이션 하거나, 수동으로 객체를 어노테이션 하는 등 다양한 기능을 제공해 주고 있어 사용자가 효율적으로 인터랙티브 비디오를 저작 할 수 있기 때문인 것으로 판단된다. 하지만 어노테이션 성능 평가를 통해 빠른 시간 안에 인터랙티브 비디오를 저작하기에는 제안한 저작 도구 더 적합함을 확인할 수 있었다.

인터페이스의 질에 관한 문항 중 Q16, Q17 문항의 경우 사용자들은 제안한 저작 시스템을 wireWAX, Popcorn Maker, 그리고 Zentrick보다 평균 0.86점 낮게 평가하였다. 또한 Q16~Q18 문항의 표준편차 결과 Q1~Q8문항 보다 0.12점 높은 평균 0.99점을 나타내 제안한 시스템의 인터페이스에 관한 사용자 선호도가 유용성 평가에 비해 일관적이지 않았음을 보여 주었다. 이는 기존의 wireWAX와 같이 객체의 영역을 인식하고 그 위에 추가 정보를 지정하는 방법에 익숙한 사용자들이 얼굴 영역 기반의 객체 어노테이션 방법을 제안한 시스템의 인터페이스를 통해 쉽게 유추하지 못하였기 때문인 것으로 분석되었다. 이러한 문제를 보완하기 위해 저작 도구 위에 툴팁(Tooltip) 형태로 사용 방법을 제공하거나, 저작도구를 좀 더 직관적으로 만들면, 인터페이스의 질을 향상 시킬 수 있을 것으로 예상된다.

5. 결론 및 향후 연구

본 논문에서는 얼굴 영역의 상대거리에 객체를 어노테이션 하는 새로운 형태의 객체 어노테이션 방법을 제안하였다. 또한 제안한 어노테이션 방법을 검증하기 위해 제안한 방법을 사용하는 인터랙티브 비디오 저작 시스템을 구현하였

다. 실험은 구현된 저작 시스템과 기존의 인터랙티브 비디오 저작 도구들과의 객체 어노테이션 시간을 분석하였고, 사용자 평가를 진행하였다.

객체 어노테이션 시간 측정 결과 다른 저작 도구에 비해 제안한 어노테이션 방법의 성능이 평균 2배 더 향상된 것을 확인할 수 있었다. 또한 인터랙티브 비디오 저작 경험이 있는 사용자 19명을 대상으로 시스템 유용성 사용자 평가(CSUQ)를 진행 하였으며, 그 결과 기존의 인터랙티브 비디오 저작 방법보다 약 10% 이상 더 유용함을 확인할 수 있었다. 또한 사용자 평가를 통해 제안한 객체 어노테이션 방법이 빠르게 인터랙티브 비디오를 만드는데 효율적임을 확인하였으며, 본 저작 시스템은 사용자들이 배우기가 쉬워 비교적 저작이 어려운 다른 저작 도구에 비해 손쉽게 사용 할 수 있음을 확인하였다. 이를 통해 사용자가 드라마나 영화의 등장인물이 입고 있는 옷, 가방, 엑세서리 등에 관련된 추가 정보를 쉽고, 효율적으로 어노테이션 할 수 있는 것으로 분석 되었다.

그러나 잘못 검출된 샷과 잘못 인식된 얼굴 영역을 피드백을 통해 어노테이터가 수동으로 보정하기 때문에 어노테이션 시간이 늘어나는 단점이 있었다. 이를 개선하기 위해 향후 향상된 샷 경계 검출 방법과 얼굴 검출 방법을 사용할 필요가 있다. 또한 하나의 샷 시퀀스 안에 있는 샷의 키프레임들 간에 차이가 클 경우 얼굴 영역 트래킹이 쉽지 않은데, 이때 트래킹이 쉽도록 키프레임 사이를 보간해 주는 방법이 필요하다. 마지막으로 본 저작 시스템을 이용하여 두 개 이상의 객체를 어노테이션 하면 객체간의 오버랩(Overlap) 현상이 발생할 수 있다. 이를 해결하기 위해 향후 객체의 충돌(Collision)을 감지하고 객체의 위치를 이동하는데 등장인물을 가리지 않도록 화면에서 덜 중요한 공간을 찾아 객체를 이동시키는 방법이 필요하다.

참고문헌(References)

Chasanis, V. T., C. L. Likas, and N. P. Galatsanos, “Scene Detection in Videos Using Shot Clustering and Sequence Alignment,” *IEEE Transactions on Multimedia,* Vol.11, No.1 (2009), 89～100.

Froba, B., A. Ernst, "Face Detection with the Modified Census Transform," *Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, (2004), 91～96.

Lewis, J. R., “IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use,” *International Journal of Human-Computer Interaction*, Vol.7, No.1 (1995), 57～78.

Lienhart, R., "Comparison of automatic shot boundary detection algorithms," *Proceedings of the SPIE Conference*, Vol.3656(1998), 290～301.

Lee, K. -S., A. N. Rosli, I. A. Supandi, and G. -S. Jo, "Dynamic sampling-based interpolation algorithm for representation of clickable moving object in collaborative video annotation," *Neurocomputing,* Vol.146(2014), 291～300.

Lee, K. A., C. H. You, H. Li, T. Kinnunen, and K. C. Sim, "Using Discrete Probabilities With Bhattacharyya Measure for SVM-Based Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.19, No.4(2011), 861～870.

Lin, T. T. C., “Convergence and Regulation of Multi-Screen Television : The Singapore Experience,” *Telecommunications Policy*, Vol.37, No.8(2013), 673～685.

Lucas, B. D., T. Kanade, “An iterative image registration technique with an application to stereo vision,” *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, (1981), 674～679.

Miller, G., S. Fels, M. Ilich, M. M. Finke, T. Bauer, K. Wong, and S. Mueller, “An End-to-End Framework for Multi-View Video Content: Creating Multiple-Perspective Hypervideo to View On Mobile Platforms,” *Proceedings of 10th International Conference on Entertainment Computing,* (2011), 337～342.

Nielsen, *Digital Consumer Report*, 2014. Available at http://www.nielsen.com/us/en/reports.html(Accessed 13 November, 2014).

Mozilla, *Popcorn Maker*. Available at https://popcorn.webmaker.org(Accessed 13 November, 2014).

Rui, Y., T. S. Huang, and S. Mehrotra, “Constructing Table-of-Content for Videos,” *Multimedia Systems*, Vol.7, No.5(1999), 359～368.

Swiki, *Frontal Face Haar Cascade*, Available at http://alereimondo.no-ip.org/OpenCV (Accessed 13 November, 2014).

Viola, P., and M. Jones, “Rapid object detection using a boosted cascade of simple features," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2001), 511～518.

wireWAX. Available at http://wirewax.com(Accessed 13 November, 2014).

Yoon, U. N., K. S. Lee, and G. S. Jo, “Interactive Video Annotation System based on Face Area,” *Korea Computer Congress*, (2014), 755～757.

Zentrick. Available at https://www.zentrick.com (Accessed 13 November, 2014).

Abstract

Annotation Method based on Face Area for Efficient Interactive Video Authoring

Ui Nyoung Yoon* · Myeong Hyeon Ga* · Geun-Sik Jo**

Many TV viewers use mainly portal sites in order to retrieve information related to broadcast while watching TV. However retrieving information that people wanted needs a lot of time to retrieve the information because current internet presents too much information which is not required. Consequentially, this process can't satisfy users who want to consume information immediately. Interactive video is being actively investigated to solve this problem. An interactive video provides clickable objects, areas or hotspots to interact with users. When users click object on the interactive video, they can see additional information, related to video, instantly. The following shows the three basic procedures to make an interactive video using interactive video authoring tool:

(1) Create an augmented object;

(2) Set an object's area and time to be displayed on the video;

(3) Set an interactive action which is related to pages or hyperlink;

However users who use existing authoring tools such as Popcorn Maker and Zentrick spend a lot of time in step (2). If users use wireWAX then they can save sufficient time to set object's location and time to be displayed because wireWAX uses vision based annotation method. But they need to wait for time to detect and track object. Therefore, it is required to reduce the process time in step (2) using benefits of manual annotation method and vision-based annotation method effectively. This paper proposes a novel annotation method allows annotator to easily annotate based on face area. For proposing new annotation method, this paper presents two steps: pre-processing step and annotation step. The pre-processing is necessary because system detects shots for users who want to find contents of video easily. Pre-processing step is as follow: 1) Extract shots using color histogram based shot boundary detection method from frames of video; 2) Make shot clusters using similarities of shots and aligns as shot sequences; and 3) Detect and

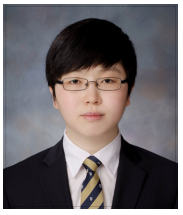
* Department of Computer Science and Information Engineering, Inha University
** Corresponding Author: Geun-Sik Jo
School of Computer Science and Information Engineering, Inha University
100 Inharo, Nam-gu, Incheon 402-751, Korea
Tel: +82-32-860-7447, Fax: +82-32-875-5863, E-mail: gsjo@inha.ac.kr

track faces from all shots of shot sequence metadata and save into the shot sequence metadata with each shot. After pre-processing, user can annotates object as follow: 1) Annotator selects a shot sequence, and then selects keyframe of shot in the shot sequence; 2) Annotator annotates objects on the relative position of the actor's face on the selected keyframe. Then same objects will be annotated automatically until the end of shot sequence which has detected face area; and 3) User assigns additional information to the annotated object. In addition, this paper designs the feedback model in order to compensate the defects which are wrong aligned shots, wrong detected faces problem and inaccurate location problem might occur after object annotation. Furthermore, users can use interpolation method to interpolate position of objects which is deleted by feedback. After feedback user can save annotated object data to the interactive object metadata. Finally, this paper shows interactive video authoring system implemented for verifying performance of proposed annotation method which uses presented models. In the experiment presents analysis of object annotation time, and user evaluation. First, result of object annotation average time shows our proposed tool is 2 times faster than existing authoring tools for object annotation. Sometimes, annotation time of proposed tool took longer than existing authoring tools, because wrong shots are detected in the pre-processing. The usefulness and convenience of the system were measured through the user evaluation which was aimed at users who have experienced in interactive video authoring system. Recruited 19 experts evaluates of 11 questions which is out of CSUQ(Computer System Usability Questionnaire). CSUQ is designed by IBM for evaluating system. Through the user evaluation, showed that proposed tool is useful for authoring interactive video than about 10% of the other interactive video authoring systems.

Key Words : Interactive Video, Authoring Tool, Annotation, Shot Sequence Alignment

Received : November 26, 2014 Revised : December 26, 2014 Accepted : January 2, 2015
Type of Submission : Fast Track Corresponding Author : Geun-Sik Jo

저 자 소 개



Ui Nyoung Yoon
Received a B.S. degree Computer and Information Engineering from Inha University, Korea, in 2013, and a M.S. degree in Information Engineering from Inha University, Korea in 2015. He is a Ph.D. Candidate in Information Engineering of Inha University, Korea. His research interests include N-Screen, Interactive Video, Semantic Web.

Myeong Hyeon Ga
Received a B.S. degree Computer and Information Engineering from Inha University, Korea, in 2012, and a M.S. degree in Information Engineering from Inha University, Korea in 2014. He is a Ph.D. Candidate in Information Engineering of Inha University, Korea. His research interests include Recommender System and Data Mining.

Geun-Sik Jo
Is a Professor in Computer and Information Engineering, Inha University, Korea. He received the B.S. degree in Computer Science from Inha University in 1982. He received the M.S. and the Ph.D. degrees in Computer Science from City University of New York in 1985 and 1991, respectively. He has been the General Chair and/or Technical Program Chair of more than 20 international conferences and workshops on artificial intelligence, knowledge management, and semantic applications. His research interests include knowledge-based scheduling, ontology, semantic Web, intelligent E-Commerce, constraint-directed scheduling, knowledge-based systems, decision support systems, and intelligent agents. He has authored and co-authored five books and 293 publications.