다중 인터페이스 비디오 주석 도구의 설계 및 구현

최기석, 오두희°, 낭종호

서강대학교 컴퓨터공학과

gschoe@sogang.ac.kr, doohllo7@sogang.ac.kr, jhnang@sogang.ac.kr

Design and Implementation of a Multi-Interface Video Annotation Tool

Giseok Choe, Doohee Oh°, Jongho Nang

Department of Computer Science and Engineering. Sogang University.

요 약

콘텐츠의 내용에 맞는 의미 주석이 삽입되어 있는 경우 대화형 서비스 같은 향상된 서비스를 제공해 주거나 새로운 분석이 가능해진다. 그러나 현재 의미 주석의 자동 생성은 일부 제한적인 분야에서 사용되고 있으며, 사람이 직접 입력하는 주석이 필요하다. 또한 비디오를 분석하여 의미 주석을 자동으로 생성하는 연구에서도 학습 데이터나 정답 데이터를 필요로 하기 때문에 많은 연구에서 목적에 맞는 도구를 개발하여 데이터를 생성해 왔다.

최근의 연구는 빅데이터와 딥러닝의 영향으로 더 많은 학습 데이터를 사용하는 경향이 있어 때문에 비디오 주석 도구의 효율성이 중요해졌다. 본 연구에서는 빠른 주석 작업을 위한 다중 인터페이스 비디오 주석 도구를 제안하고 설계 및 구현하였다.

1. 서론

Youtube를 중심으로 인터넷 비디오 스트리밍은 주류미디어의 하나로 자리잡았다.[1] 국내에서도 네이버 TV캐스트, Daum tv팟, 아프리카TV 등의 많은 비디오 스트리밍 서비스가 제공되고 있다. 이전에는 상상하기 어려운 양의 콘텐츠가 SNS(Social Network Service)를 통하여 유통되고 있으며, 콘텐츠의 종류와 형태는 TV방송, 음악, 영화, 광고 등의 기존 형식 뿐만 아니라 UGC(User Generated Contents)로 불리는 형태까지 다양한 형태를 가지고 있다.[2, 3] 인터넷 비디오 데이터는 단순한 소비에서 그치지 않고, 전통적인 비디오 연구를 하던 분야에서부터 사회학, 의학 등의 다양한 분야에서 분석을 위한 데이터로 활용되어 연구되는 것을 볼 수 있다.

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 연구결과로 수행되었음. [R0126-15-1112, 퍼스널 미디어가 연결 공유 결합하여 재구성 가능케 하는 복합 모달리티 기반 미디어 응용 프레임워크 개발]

자동으로 생성한 데이터를 기반으로 서비스하기 위해 불완전한 부분을 채우거나, 실험에 필요한 학습/정답 데이터를 모을 필요가 있다. 각각의 연구에서는 용도에 맞는 자체 개발한 주석 도구를 사용하였다.

최근의 연구는 빅데이터와 딥러닝의 영향으로 더 많은 학습 데이터를 사용하는 경향이 있다. 반면 비디오 주석 작업은 사람이 직접 주석을 달아야 하기 때문에 많은 시간과 노력이 필요하기 때문에 빠른 주석 작업의 중요성이 높아졌다.

본 연구에서는 빠른 주석 작업을 위한 비디오 주석 도구를 제안하고 설계 및 구현하였다.

2. 관련연구

비디오에 의미 주석을 추가하려는 시도는 오래 전부터 있었다[4]. 콘텐츠 기술 표준인 MPEG-7[5]이 채택된 이후에는 MPEG-7을 사용하거나 흡사한 형태의 자료구조를 가진 의미 주석 구조를 사용하였다.

콘텐츠 기반 검색(Content-based Search)이나 이미지 분석을 위한 실험 도구로 주석 도구가 소개되거나[6-8], 주석 도구 자체를 연구하는 형태[9-11]로 소개되었다.

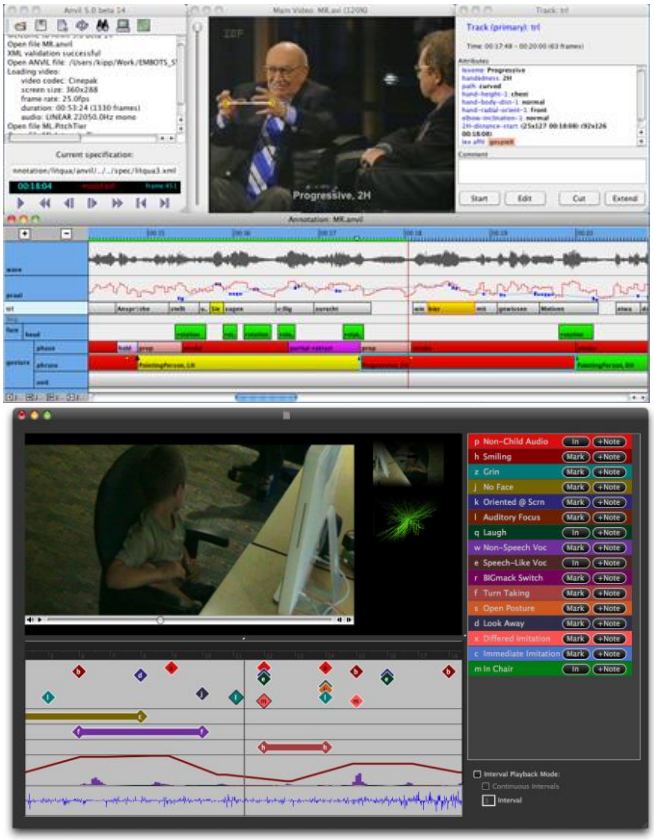

그림 1 주석 도구의 사용자 인터페이스 Anvil[9](상), VCode and VData[10](하)

그림1은 기존 연구에서 개발한 Anvil, Vcode and VData의 구동 화면이다. 용도에 따라서 세부적인 부분에 차이가 보인다. 하지만 타임라인, 모니터, 의미 주석을 위한 인터페이스는 공통적으로 존재한다.

3. 설계 및 구현

도구를 설계할 때 빠른 주석을 위해서 다음 두 가지를 고려하였다.

- 복잡한 인터페이스 제거 및 단순화
- 직관적인 인터페이스
- 빠른 인터페이스

기존의 주석 도구들에서 공통적으로 발견되는 인터페이스는 타임라인인데, 사용하기에 복잡한 인터페이스이다. 타임라인 인터페이스를 제거하고, 기존의 도구에서 주석 구간을 프레임 단위로 제어하던 부분을 영상의 샷단위 제어로 대체하였다. 그리고 그림2와 같은 형태의 샷 경계 검출기를 별도로 제작하였다.

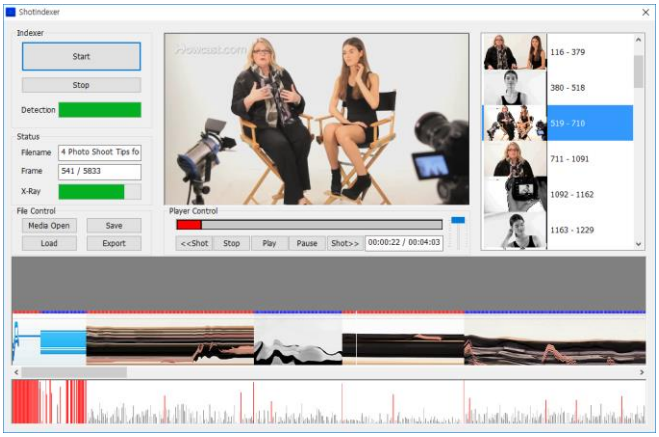

그림 2 구현한 샷 경계 검출기

비디오 구조를 정의하는 방법 중에서 샷을 검출하고, 검출된 샷을 묶어서 씬으로 정의하는 방법을 고려하였다[12]. 샷은 물리적인 단위로서 모호하지 않기때문에 주석 단계에서 수정이 필요하지 않기 때문에 별도의 도구에서 자동 검출하였다.

인터페이스가 복잡해질 사용자의 작업 속도가 느려진 점을 발견하였다. 빠른 인터페이스를 위해 키보드 단축키를 중심으로한 인터페이스를 그림3의 모습으로 설계하였고, 직관적인 인터페이스를 위해 마우스 중심의 인터페이스를 그림4의 모습으로 설계하였다. 상황에 따라서 두 가지 인터페이스를 전환하면서 작업 할 수 있도록 구성하였다.

인터페이스를 단순화 하기 위해 기본 주석 단위를 샷으로 제한하였기 때문에 부분 이미지 객체 및 사운드를 필요에 따라 별도로 선택 가능하도록 그림5의 모습으로 설계하였다.

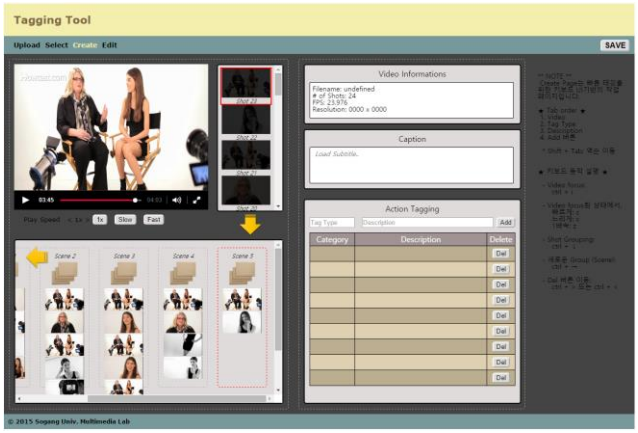

그림 3 키보드 중심의 인터페이스

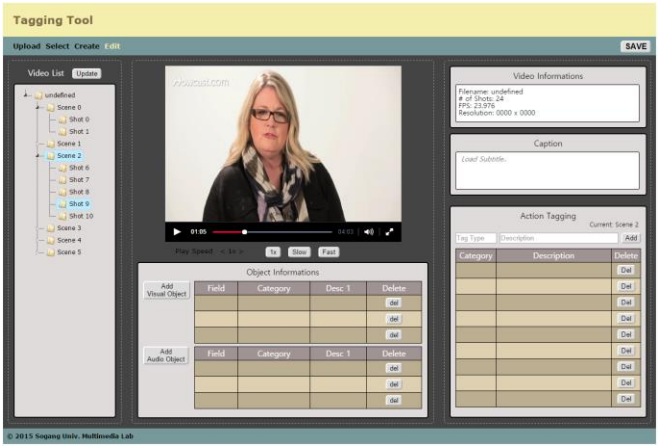

그림 4 마우스 중심의 인터페이스

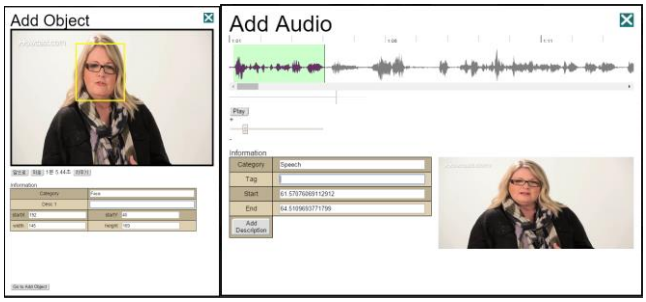

그림 5 이미지 영역 및 사운드 선택 인터페이스

4. 구현 및 구동 환경

샷 경계 검출 도구는 Windows Application으로 구현하였다.

주석 도구는 HTML5의 웹 기반 응용으로 구현하였으며 Javascript와 PHP를 사용하였다. 사용자는 HTML5를 지원하는 인터넷 브라우저로 서버에 접근하면 사용할 수 있는 형태이다.

5. 평가 및 향후 과제

의미 주석 작업은 콘텐츠를 이해해야 하므로 최소콘텐츠 길이만큼 작업 시간이 걸린다. 기존의 도구는 그 복잡성으로 인해 콘텐츠 길이의 약 4~5배의 시간이 소요되었으나 제안한 도구는 콘텐츠 길이의 약 3배 정도의 시간으로 작업 시간이 단축되었음을 확인하였다.

향후 과제로는 많은 사람이 동시에 주석 작업을 할수 있는 소셜 주석 기능을 추가한다면 대규모 주석 데이터를 생성하기에 좀더 수월할 것이라 생각한다.

참고문헌

[1] J. Burgess, and J. Green, *YouTube: Online video and participatory culture*: John Wiley & Sons, 2013.

[2] G. Vickery, and S. Wunsch-Vincent, *Participative web and user-created content: Web 2.0 wikis and social networking*: Organization for Economic Cooperation and Development (OECD), 2007.

[3] J. Krumm, N. Davies, and C. Narayanaswami, "User-generated content," *IEEE Pervasive Computing*, no. 4, pp. 10-11, 2008.

[4] B. L. Harrison, and R. M. Baecker, "Designing video annotation and analysis systems." pp. 157-166.

[5] M. Abdel-Mottaleb, N. Dimitrova, L. Agnihori, S. Dagtas, S. Jeannin, S. Krishnamachari, T. McGee, and G. Vaithilingam, "MPEG-7: a content description standard beyond compression." pp. 770-777.

[6] M. Lux, W. Klieber, J. Becker, K. Tochtermann, H. Mayer, H. Neuschmied, and W. Haas, "XML and MPEG-7 for Interactive Annotation and Retrieval using Semantic Meta-data," *J. UCS,* vol. 8, no. 10, pp. 965-984, 2002.

[7] H. Sack, and J. Waitelonis, "Integrating social tagging and document annotation for content-based search in multimedia data."

[8] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision,* vol. 101, no. 1, pp. 184-204, 2013.

[9] M. Kipp, "Anvil: The video annotation research tool," *Handbook of Corpus Phonology. Oxford University Press, Oxford (to appear, 2011)*, 2010.

[10] J. Hagedorn, J. Hailpern, and K. G. Karahalios, "VCode and VData: illustrating a new framework for supporting the video annotation workflow." pp. 317-321.

[11] X. Mu, "Towards effective video annotation: An approach to automatically link notes with video content," *Computers & Education,* vol. 55, no. 4, pp. 1752-1763, 2010.

[12] T. Lin, and H.-J. Zhang, "Automatic video scene extraction by shot grouping." pp. 39-42.