다채널 영상 기반 비디오 스위칭 자동화 설계 및 구현

김순철, 오혜주

한국전자통신연구원(ETRI)
choulsim@etri.re.kr, feeler@etri.re.kr

Implementation of video switching automation from multi cameras

Soon-Choul Kim, Hye-ju Oh
Electronics and Telecommunications Research Institute

요 약

최근 미디어 제작 워크플로우에 인공지능 기술을 적용하고자 하는 관심이 확산되고 있으며, 기존 비디오 스위칭에서의 기계적인 기술 적용 이외에 딥러닝 기반 영상분석을 통한 소프트웨어 중심 제작 자동화에 대한 연구개발들이 활발하다. 본 논문은 다채널 카메라 영상에서 만들어진 비디오 소스를 기반으로 화자의 발화구간을 실시간 추출하고 인물 샷 중심으로 다양한 프로그램영상(PGM) 출력을 자동화하는 AI 비디오 스위칭 기술의 설계 및 구현 결과를 설명한다.

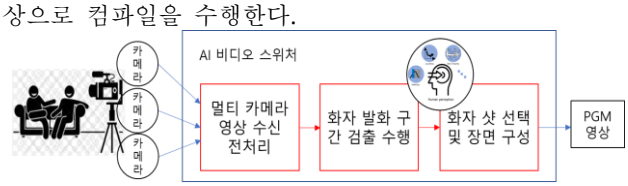
1. 서 론

최근 미디어 제작 환경은 다양한 AI 기술 도입과 클라우드 및 네트워크 인프라 고도화를 통해 입력/편집/배포 워크플로우 향상과 제작자의 업무 효율화가 빠르게 진행되고 있다. 비디오 스위칭 시스템은 복수의 영상 소스를 입력으로 하드웨어 장치 혹은 소프트웨어 응용처리를 통해 TV 방송이나 OTT 플랫폼으로 송출되는 프로그램 영상(Program, PGM)을 심리스(seamless)하게 처리할 수 있도록 기능을 제공하는 제작 장치로서, 화자 중심의 비디오 스위칭 자동화 기술은 토크쇼/회의 녹화 등에서 특정 화자가 말할 때 자동으로 해당 화자를 촬영하는 카메라로 전환(스위칭)될 수 있도록 한다. 일반적으로 주제작자(PD)가 다수의 입력 영상(Preview, PRV)을 보면서 녹화 또는 실시간 중계를 위해 조작/편집하거나 화자 마이크의 발성 감지를 통해 화자의 카메라로 스위칭을 자동화 처리하고 있다. 이는 제작자나 장치의 일시적인 오류와 다수가 동시에 말할 때와 같은 비디오 스위칭의 혼선이 발생할 수 있으며, 무엇보다도 1 인/소규모 제작자에게는 시간적/공간적/비용적 한계가 존재한다. 딥러닝 기반의 자동화된 비디오 스위칭 기술은 마이크 오류나 음성 인식 한계가 있을 때도 보완적으로 작동해 안정적인 스위칭을 제공할 수 있도록 한다[1][2][3].

본 논문은 다채널의 영상 기반으로 화자 중심의 자동화된 비디오 스위칭 기술 설계 및 구현에 관한 것으로서, 각 영상 채널에 대해 화자 구간 검출과 발화 확률값의 병렬 처리 과정을 통해 실시간 자동 편집 속도의 성능을 향상시켰다. 또한, 영상 채널 속 화자 결정 과정에서 화자의 발화구간이 너무 짧거나 발화구간 추출이 모호한 경우 화면 전환(카메라 스위칭)이 빈번히 발생되는 오류에 대해 자연스러운 화면 구성이 이뤄질 수 있도록 하는 것을 특징으로 한다.

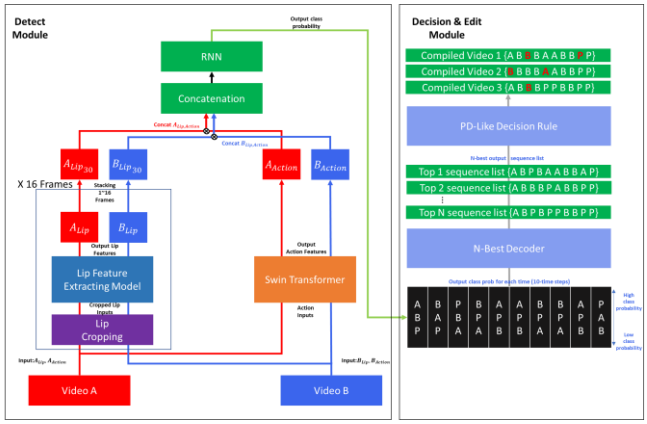
II. 딥러닝 기반 비디오 스위칭 자동화 모델 설계/구현

[그림 1]은 본 논문에서 적용한 화자 중심의 비디오 스위칭 자동화 기술 개념도를 나타내며, 주요하게 3 단계의 과정들로 구성된다. 멀티 카메라 영상 수신 전처리 단계는 다채널 영상의 디코딩과 각 비디오 프레임으로부터 얼굴의 3 차원 좌표 생성 및 입술 관심영역을 검출하는 병렬처리 과정이다. 화자 발화 구간 검출 수행 단계는 비디오 프레임 단위로 화자 구간(수~수십 프레임)의 발화 여부를 확률값(0..1)으로 산출하는 과정이다. 마지막으로 화자 샷 선택 및 장면 구성 단계는 이전 단계에서의 확률값들을 토대로 최적의 화자 샷들을 구성하여 PGM 영상으로 컴파일을 수행한다.



<그림 1> 화자중심 비디오 스위칭 자동화 개념도

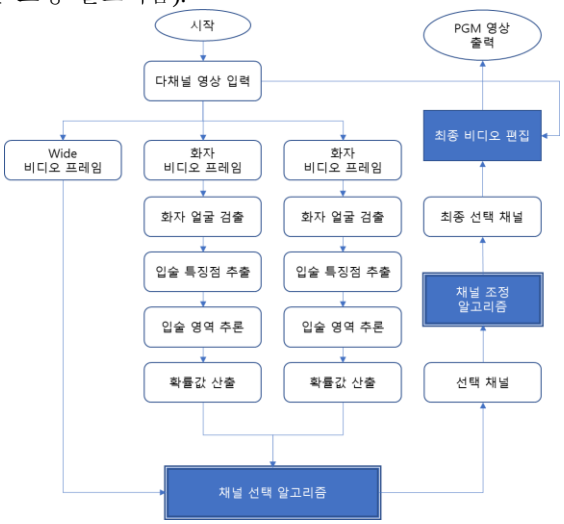
비디오 스위칭 자동화 모델은 입술모양에 기반한 화자 구간 검출 외에 제작자의 다양한 의도/규칙을 내포한 PGM 영상들을 내보낼 수 있도록 화자 음성 혹은/그리고 행동(손짓과 같은 움직임) 인식을 통한 멀티모달 융합 모델로 확장될 수 있도록 유연한 설계를 갖는다<그림 2>. 설계된 모델에 기반한 비디오 스위처는 화자 비디오 프레임 확률값과 비디오 스위칭 민감도 시간을 조절하여 화자 중심의 화면 전환 시 빠른 또는 느린 템포의 비디오 스위칭이 가변적으로 발생하는 PGM 출력을 구성할 수 있도록 구현되었다.



<그림 2> 화자 입술과 행동 융합 모델을 기반으로 한 비디오 스위칭 자동화 아키텍처 설계

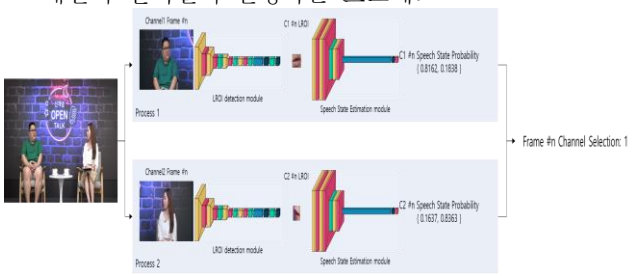
본 논문에서 제시된 비디오 스위칭 자동화 모델에 기반하여 2 명의 화자 채널(A, B)과 전체(wide) 화자 채널(W)을 포함한 총 3 개의 영상 입력 채널(A,B,W)을 사용하

여 <그림 3>과 같은 절차로 구현하였다. A, B 채널의 화자 구간을 검출하여 확률값에 근거하여 A 또는 B 또는 W 채널을 선택해서 PGM 채널의 시퀀스를 버퍼에 임시 구성하도록 하였다(채널 선택 알고리즘). 임시 버퍼에 저장된 선택 채널들의 시퀀스에는 확률값 오차에 따른 비정상 프레임들이 존재할 수 있으며 이로 인한 짧은 시간 내 잦은 비디오 스위칭 발생으로 시청에 불편감을 초래한다. 임시 버퍼에 저장된 비디오 프레임 시퀀스 중에 비정상 프레임들을 찾아 정상 프레임들로 대체하는 과정이 요구된다. 즉, 최종 PGM 영상 생성 시 이전에 나열된 프레임 집합들 중 비정상 프레임 집합이 있는지 여부를 판단하고, 비정상 프레임 집합이 있는 경우 비정상 프레임 집합을 정상 프레임 집합으로 대체할 수 있도록 한다(채널 조정 알고리즘).



<그림 3> AI 비디오 스위칭 구현 동작

A. 채널 선택 알고리즘 구현: 각 화자 비디오 프레임의 입술 관심영역에 대한 확률값을 비교, 분석하여 어떤 채널이 선택될지 결정하는 프로세스



<그림 4> 화자 구간 검출 및 채널 선택 결과

채널 선택 알고리즘은 각 화자 비디오 프레임의 입술 관심영역에 대한 확률값을 비교, 분석하여 어떤 채널이 선택될지 결정하는 프로세스로서, (1) 각 채널의 확률 값이 일정 값 이상인 경우 중 가장 높은 확률 값을 선택, (2) 또는 높은 값을 가진 채널들의 확률값 차이가 근소할 경우(0.05 이하, 임계값) W 채널을 선택한다. <그림 4>는 두채널의 n 번째 비디오 프레임 확률값(채널 1: 0.8162, 채널 2: 0.1637)을 비교한 결과, 채널 A 의 을 PGM 영상 출력 샷으로 선택하는 결과를 나타낸 것이다.

B. 채널 조정 알고리즘 구현: 비디오 프레임에서 화자변경 정보가 너무 짧게 나타나는 경우, 이를 노이즈로 간주하고 주변의 일관된 값으로 대체함으로써 연속성과 자연스러움을 유지하는 프로세스

채널 조정 알고리즘은 (1) 채널 선택 알고리즘에 의해 선택된 채널들의 나열에서 이상값을 보이거나, (2) 일정시간(15frame, 임계값) 동안 유지되지 않는 구간을 유지하도록 선택된 채널들의 나열을 재설정하는 역할을 수행하며, 다음과 같은 동작 절차를 갖는다.

① 프레임 정렬 및 구간 나누기 : 프레임 번호를 정렬하고, 화자 값이 바뀌는 지점을 기준으로 연속된 구간(segment)을 나눔

② 짧은 구간 검출 : 각 구간의 길이를 threshold 와 비교하여 너무 짧은 구간(예: 1~30 프레임)을 보정 대상으로 판단

③ 인접 화자 값으로 대체 : 왼쪽과 오른쪽 화자 값을 확인하여, 가능한 경우 왼쪽 값을 우선적으로 사용하여 구간을 덮어씀. 이후 양쪽이 동일하면 그 값을 사용하고 왼쪽이 없으면 오른쪽을 사용

④ 일관성 유지 : 짧은 구간을 보정하여 급격한 화자 변경을 방지하고 자연스러운 전환을 유도

⑤ 결과 반환 : 보정된 화자 정보 배열을 다시 딕셔너리로 변환하여 반환

위 동작 절차에 따른 채널 조정 알고리즘 적용 결과로서 대표적인 수행 예는 다음과 같다.

- (알고리즘 적용 전)“AAAAAABAAAAAAAA”
  (알고리즘 적용 후) “AAAAAAAAAAAAAAA”
  (B: 이상값이라 판단) 스무딩, 혹은 채널 편집 품질을 매끄럽게 유지
- (알고리즘 적용 전) “AAAAAWBABBAAAAA”
  (알고리즘 적용 후) “AAAAAAAAAAAAAAA”
  (A 채널이 15 frame 동안 유지되지 않았다고 판단하여 앞 혹은 뒤의 인접 채널로 재설정)

III. 결론

본 논문은 딥러닝 기반의 영상분석을 통해 화자 중심의 비디오 스위칭 자동화 기술에 관한 것으로서, 화자 중심의 PGM 영상을 제안 알고리즘 적용을 통해 자연스럽고 매끄러운 품질 결과를 만들어 내도록 하였다. 본 논문에서 제안된 화자 검출 모델 학습 및 화자별 발성 구간 검출 정확도는 95%이상 달성하였다. 향후에는 영상과 음성을 동시에 활용한 멀티모달 기반의 실시간 비디오 스위칭 자동화 기술로 확장해서 고도화한 후에 실제 방송된 PGM 영상과 제안된 PGM 스위칭 자동화 영상과의 전문가그룹의 품질 비교(평가항목: 매끄러운 화자 전환, 장면 일관성)를 통해 AI 편집기술 활용성에 대한 평가를 진행할 계획이다.

ACKNOWLEDGMENT

“이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-01028, 클라우드-IP 기반 고품질 미디어 제작 기술 개발)”

참 고 문 헌

[1] S. Kim, et al., "Design of Cloud-based remote collaborative system for broadcasting production workflow," 2023 14th International Conference on Information and Communication Technology Convergence (ICTC), pp. 1404-1406, Oct. 2023.

[2] M. H. Wani and A. R. Faridi, "Deep Learning-Based Video Action Recognition: A Review," 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2022, pp. 243-249.

[3] C. Wright, J. Allnutt, R. Campbell, M. Evans, S. Jolly, E. Shotton, S. Lechelt, G. Phillipson, and L. Kerlin, “AI in production: Video analysis and machine learning for expanded live events coverage,” in Proceedings of the 2023 ACM International Conference on Interactive Media Experiences Workshops, 2023, pp. 77‑78.