하이라이트와 스토리를 통합적으로 고려한 영상 요약 기법

김은비, 전아린, 윤단비, 황기태
한성대학교

keunbi03@naver.com, wjsdkfls03@naver.com, yoondb1128@naver.com, calafk@hansung.ac.kr

A Video Summary considering Highlights and Story Together

Eunbi Kim, Ahrin Jun, Danbi Yoon, Kitae Hwang
Hansung University

요 약

기존 영상 요약 기술은 원본 영상의 중요한 부분만을 추출하는 데 초점을 맞추어 왔다. 이 문제를 해결하기 위해 본 연구는 하이라이트와 스토리의 비율을 조절할 수 있는 가중치 기반 영상 요약 알고리즘을 제안한다. 본 논문은 원본 영상의 맥락을 반영하기 위해 '다양성 기여도' 개념을 도입하였다. 그리고 누구나 쉽게 웹 브라우저만으로 활용할 수 있도록 웹 서비스로 구현하였다. 구현된 시스템을 평가한 결과, 주어진 가중치 w에 따라 하이라이트와 스토리가 적절히 배합된 영상으로 요약됨을 검증하였다. 결론적으로 본 논문에서 제시한 요약 기법은 사용자가 원하는 가중치에 따라 하이라이트와 스토리를 잘 혼합해 낼 수 있는 매우 우수한 시스템이라고 판단된다.

Ⅰ. 서 론

기존 영상 요약 연구의 대부분은 중요도를 기준으로 요약 영상을 생성한다.[1] 그러나 현실에서 중요한 장면만으로 영상을 요약하는 것은 바람직하지 않다. 예를 들어 여행, 브이로그 영상은 전반적인 내용을 고르게 담는 것이 중요하고, 영화 예고편은 하이라이트 중심이면서도 스토리 흐름을 포함해야 한다.

이에 본 연구는 사용자가 가중치를 활용하여 원하는 하이라이트와 스토리의 비율을 조절할 수 있는 새로운 영상 요약 기법을 제안하고 시스템을 구현하였다.

본 연구는 요약 영상에 스토리 전개를 반영하기 위해 '다양성 기여도(Diversity Contribution)' 개념을 도입하였다. 이는 한 장면이 요약에 포함될 때 새롭게 추가되는 정보를 수치상으로 나타낸 것으로, 다양성 기여도가 높을수록 원본 영상의 다양한 장면이 고르게 포함된 요약 영상을 생성할 수 있다.

최근에는 스토리를 반영하기 위해 영상의 시각적 특성을 텍스트로 변환해 활용하는 LLM 기반 요약 기법도 제안되고 있으나, 이러한 접근은 추가적인 시간과 비용이 필요하다. 반면 본 연구는 시각적 특징 벡터를 활용함으로써 처리 시간을 줄일 수 있다.

또한, 실험을 통해 요약 가중치에 따라 요약 영상이 하이라이트 중심에서는 중요 장면을, 스토리 중심에서는 원본의 맥락을 더 충실히 반영함을 확인하였다.

Ⅱ. 시스템 설계 및 구현

2.1 시스템 구조

VideoSummary 시스템은 그림 1와 같이 웹 기반으로 구현되었으며, 웹 클라이언트와 웹 서버로 구성된다.

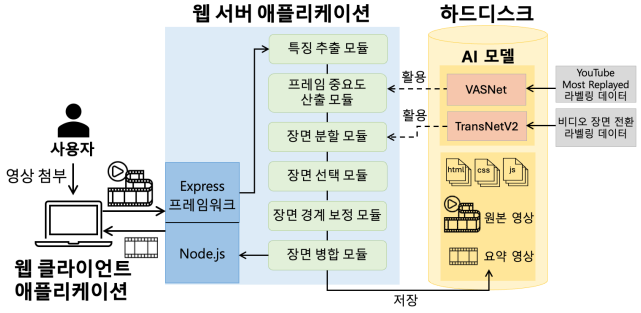


그림1. 시스템 구조도

웹 클라이언트는 사용자가 영상을 업로드하고 요약 결과를 확인할 수 있는 웹 인터페이스를 제공한다. 웹 서버는 Node.js 기반 Express 웹 프레임워크 위에서 영상 요약 파이프라인을 수행하여 요약 영상을 생성한다.

2.2 가중치에 따른 스토리-하이라이트 배합 알고리즘

이 절에서는 하이라이트 중심과 스토리 중심의 비율을 조절하여 영상을 요약하는 알고리즘을 설명한다.

먼저 원본 영상에서 1초 간격으로 대표 프레임을 추출하고, InceptionV3 모델로 각 대표 프레임마다 시각적 특징 벡터를 생성한다. 이후, 이 벡터를 입력으로 받아 Mr.HiSum 데이터셋[2]으로 학습시킨 PGL-SUM 모델이 각 대표 프레임의 중요도 점수(0-1)를 산출한다.

단순히 중요한 프레임만 연결하면 요약 영상이 부자연스럽기 때문에, TransNetV2 모델로 장면 전환점을 탐지하고 연속된 프레임을 세그먼트 단위로 분할해 요약의 기본 단위로 사용하였다.

세그먼트 i의 중요도 점수 $I(i)$는 다음과 같이 계산한다.

$$I(i) = \alpha \cdot \mu + (1-\alpha) \cdot m - \beta \cdot \sigma \ (where\ \alpha = 0.7, \beta = 0.3)$$

여기서 μ, m, σ는 세그먼트 i에 포함된 대표 프레임의 중요도 점수에 대한 평균, 최댓값, 표준편차이다.

스토리 기반의 영상을 만들기 위해서는 서로 유사성이 낮은 세그먼트들을 고르게 선택해야 한다. 이를 위해 본 연구는 다양성 기여도 개념을 도입하였다. 다양성 기여도는 특정 세그먼트가 요약 영상에 얼마나 새로운 정보를 더하는지를 정량적으로 나타낸다.

선택된 요약 세그먼트 집합을 $A = \{s_1, s_2, \cdots, s_N\}$라고 할 때, 원본 영상의 각 세그먼트 i 에 대한 커버리지는 다음과 같이 정의된다.

$$Coverage(i, A) = max_{s \in A}\big(sim(i, s)\big)$$

여기서 $sim(i,s)$ 은 세그먼트 i 와 s 간의 시각적 유사도로, 앞서 추출한 시각적 특징 벡터를 활용하여 벡터 간 코사인 유사도를 계산한 값이다. 후보 세그먼트 j가 추가될 때 커버리지 증가량은 다음과 같이 정의된다.

$$\Delta Coverage(i, j, A) = max\big(Coverage(i, A), sim(i, j)\big) - Coverage(i, A)$$

다양성 기여도 $D(j|A)$ 는 이 증가량에 세그먼트 i 의 중요도 $I(i)$를 곱한 값의 총합으로, 다음과 같이 계산된다.

$$D(j|A) = \sum_{i \in V} I(i) \cdot \Delta Coverage(i, j, A)$$

후보 세그먼트 j의 최종 가치는 중요도 $I(j)$와 다양성 기여도 $D(j|A)$의 가중합으로 다음과 같이 정의한다.

$$v(j) = (1 - w) \cdot I(j) + w \cdot D(j|A)$$

여기서 $w \in [0,1]$은 하이라이트 중심과 스토리 중심 요약 간의 균형을 조절하는 가중치이다. w가 0에 가까울수록 중요도가 높은 세그먼트가 선택되어 하이라이트 중심 요약이, 1에 가까울수록 다양성이 높은 세그먼트가 선택되어 스토리 중심 요약이 생성된다.

세그먼트 j의 최종 적합성 $f(j)$는 다음과 같이 정의한다.

$$f(j) = \frac{v(j)}{c_j}$$

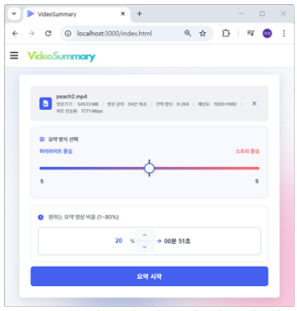
여기서 c_j는 세그먼트 j의 길이이며, $v(j)$를 c_j로 나누어 세그먼트 j의 단위 길이(초)당 가치를 산출한다.

최종 세그먼트 선택은 0-1 냅색(Knapsack)문제로 해결한다. 집합 A의 총 세그먼트 길이가 L을 넘지 않는 범위에서 $f(j)$ 값이 가장 큰 세그먼트 j를 A에 추가한다.

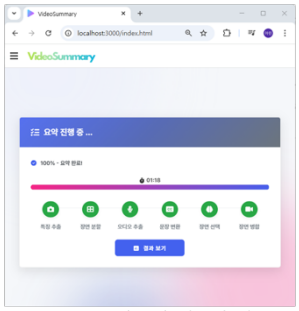
세그먼트의 경계가 음성 발화의 시작이나 종료 지점과 일치하지 않는 경우, Whisper와 Silero-VAD 모델을 사용해 탐지한 단어 단위의 발화 구간까지 세그먼트를 확장하였다. 이를 통해 음성이 끊기지 않는 자연스러운 요약 영상을 생성할 수 있다.

2.3 시스템 구현

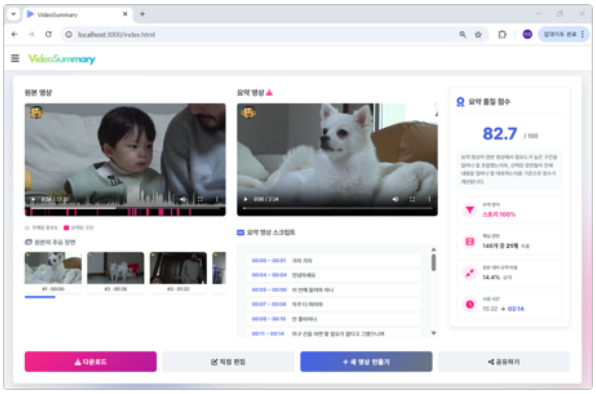
그림 2은 VideoSummary 시스템의 웹페이지 화면이다. 그림 2(a)는 사용자가 요약할 영상을 업로드하고 요약 영상의 길이와 하이라이트 중심과 스토리 중심 사이에서 요약 방식을 설정하는 화면이다. 이때 설정한 비율이 요약 시에 가중치 w로 적용된다. 그림 2(b)는 요약 진행 과정을 확인할 수 있는 화면이다. 그림 2(c)는 요약 결과 화면으로, 원본 영상과 요약 영상, 요약 품질 및 상세 정보를 확인할 수 있다.



(a) 요약 방식 설정 화면



(b) 요약 진행 화면



(c) 요약 결과 화면

그림2. VideoSummary 웹페이지 구현 화면

Ⅲ. 성능 평가

우리는 10개의 장르, 20개의 영상을 대상으로 본 논문에서 제시하고 구현한 시스템의 성능을 두 가지 지표로 평가하였다. 두 가지 지표인 세그먼트 평균 중요도 점수와 스토리 보존 점수를 평가한 결과, 그림 3과 같은 결과를 얻었다.

세그먼트 평균 중요도 점수는 원본 영상의 중요한 장면이 얼마나 잘 선택 되었는지를 나타내며, 가중치 w를 0.0에서 1.0까지 0.1 단위로 조정하며 평가하였다.

스토리 보존 점수는 Google Gemini 2.5 Pro를 사용하여 요약 영상이 원본의 맥락을 얼마나 유지하는지를 평가하였다. 이를 위해 원본의 시놉시스와 가중치 w를 0.0, 0.5, 1.0으로 설정한 요약 영상의 시놉시스를 각각 생성하고, 두 텍스트 간 일치도를 비교하였다.

w 값이 증가함에 따라 세그먼트 평균 중요도 점수가 비례적으로 낮아지고, 스토리 보존 점수는 비례적으로 높아지는 것을 확인할 수 있다.

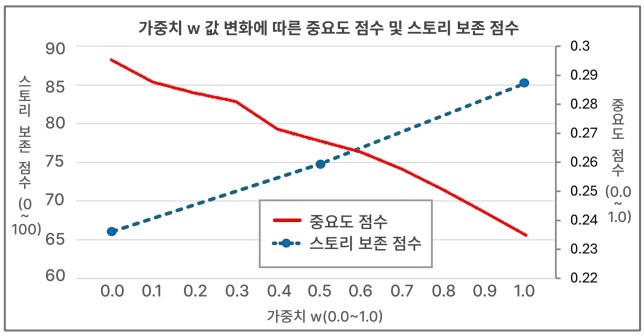


그림3. 가중치 w값에 따른 세그먼트 평균 중요도 점수 및 스토리 보존 점수 변화

Ⅳ. 결론

본 논문은 가중치 w를 통해 하이라이트 중심과 스토리 중심 비율을 조절할 수 있는 가중치 기반 영상 요약 기법을 제안하고 실제 시스템을 설계하고 구현하였다. 핵심 알고리즘으로, 스토리 맥락 반영을 위해 '다양성 기여도' 개념을 도입하였다.

본 논문에서 제안한 요약 기법을 평가한 결과, w값에 따라 세그먼트 평균 중요도 점수와 스토리 보존 점수가 상반된 경향을 보이며, 하이라이트 중심과 스토리 중심의 특성이 뚜렷하게 구분되는 것을 확인하였다.

이를 통해 가중치 w를 조절함으로써 사용자의 목적에 부합하는 요약 영상을 생성할 수 있음을 검증하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2024년도 SW중심대학사업의 결과로 수행되었음(2024-0-00049).

참 고 문 헌

[1]. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video Summarization Using Deep Neural Networks: A Survey," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1838-1863, Nov. 2021, doi: 10.1109/JPROC.2021.3117472.

[2] J. Sul, J. Han, and J. Lee, "Mr. HiSum: A large-scale dataset for video highlight detection and summarization," in Proc. Advances in Neural Information Processing Systems, vol. 36, pp. 40542-40555, 2023, doi: 10.48550/arXiv.2504.18689.