(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0006472
(43) 공개일자 2025년01월13일

(51) 국제특허분류(Int. Cl.)
G06V 20/40 (2022.01) *G06V 10/80* (2022.01)
G06V 10/94 (2022.01)
(52) CPC특허분류
G06V 20/47 (2022.01)
G06V 10/806 (2023.08)
(21) 출원번호 10-2023-0086289
(22) 출원일자 2023년07월04일
심사청구일자 2023년07월04일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
백성욱
서울특별시 강남구 영동대로 22, 813동 1303호(일원동, 디에이치자이 개포)
이미영
서울특별시 강남구 학동로82길 6 (삼성동, 삼성리치빌)
(뒷면에 계속)
(74) 대리인
특허법인위더피플

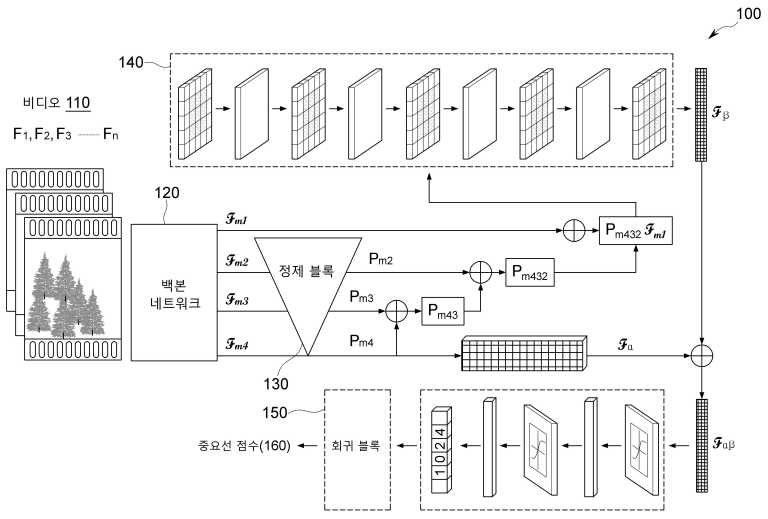
전체 청구항 수 : 총 21 항

(54) 발명의 명칭 **멀티 스케일 특징 기반의 비디오 요약 방법 및 장치**

(57) 요 약

본 발명은 멀티 스케일 특징 기반의 비디오 요약 방법에 관한 것이다. 방법은, 복수의 프레임을 포함하는 비디오를 수신하는 단계, 복수의 프레임을 백본 네트워크에 제공하여 복수의 프레임에 포함된 각각의 프레임에 대응하는 제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성하는 단계, 생성된 제2 특징맵, 제3 특징맵 및 제4 특징맵을 정제 블록에 제공하여 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 생성하는 단계, 제1 특징맵, 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 기초로 특징 퓨전을 수행하여 최종 특징을 생성하는 단계, 생성된 최종 특징을 회귀 블록에 제공하여 각각의 프레임에 대응하는 중요성 점수를 산출하는 단계 및 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 수신된 비디오에 대한 비디오 요약을 수행하는 단계를 포함한다.

대 표 도



(52) CPC특허분류
G06V 10/95 (2022.01)
(72) 발명자
하비브 칸
서울특별시 광진구 능동로 209 세종대학교 대양AI 센터 411호
사미 울라 칸
서울특별시 광진구 능동로 209 세종대학교 대양AI 센터 411호
줄피카 아마드 칸
서울특별시 광진구 능동로 209 세종대학교 대양AI 센터 411호
윤상일
경기도 시흥시 은행로 188-15, 2동 108호(대야동, 삼보아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711188556
과제번호	2023R1A2C1005788
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	커넥티드 비전 환경에서 범죄/사고 예방을 위한 이상행동상황인식 기술
기 여 율	1/1
과제수행기관명	세종대학교
연구기간	2023.03.01 ~ 2024.02.29

명 세 서

청구범위

청구항 1

적어도 하나의 프로세서에 의해 수행되는 멀티 스케일 특징 기반의 비디오 요약 방법으로서,

복수의 프레임을 포함하는 비디오를 수신하는 단계;

상기 복수의 프레임을 백본 네트워크(backbone network)에 제공하여 상기 복수의 프레임에 포함된 각각의 프레임에 대응하는 제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성하는 단계;

상기 생성된 제2 특징맵, 제3 특징맵 및 제4 특징맵을 정제 블록(refinement block)에 제공하여 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 생성하는 단계;

상기 제1 특징맵, 상기 정제된 제2 특징맵, 상기 정제된 제3 특징맵 및 상기 정제된 제4 특징맵을 기초로 특징 퓨전(feature fusion)을 수행하여 최종 특징을 생성하는 단계;

상기 생성된 최종 특징을 회귀 블록(regressor block)에 제공하여 상기 각각의 프레임에 대응하는 중요성 점수를 산출하는 단계; 및

상기 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 상기 수신된 비디오에 대한 비디오 요약을 수행하는 단계;

를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 2

제1항에 있어서,

상기 백본 네트워크는 ViT(vision transformer) 기반의 백본 네트워크를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 3

제1항에 있어서,

상기 각각의 프레임에 대응하는 제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성하는 단계는,

상기 복수의 프레임을 상기 백본 네트워크의 제1 스테이지에 제공하여 상기 각각의 프레임에 대응하는 상기 제1 특징맵을 생성하는 단계;

상기 제1 특징맵을 상기 백본 네트워크의 제2 스테이지에 제공하여 상기 제2 특징맵을 생성하는 단계;

상기 제1 특징맵 및 상기 제2 특징맵을 상기 백본 네트워크의 제2 스테이지에 제공하여 상기 제3 특징맵을 생성하는 단계; 및

상기 제1 특징맵, 상기 제2 특징맵 및 상기 제3 특징맵을 상기 백본 네트워크의 제3 스테이지에 제공하여 상기 제4 특징맵을 생성하는 단계;

를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 4

제1항에 있어서,

상기 정제 블록은 피라미드형 정제 블록(pyramidal refinement block)을 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 5

제1항에 있어서,

상기 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 생성하는 단계는,

상기 제2 특징맵을 상기 정제 블록의 제2 스테이지의 3개의 DASPP(dense atrous spatial pyramid pooling) 모듈과 하나의 MHSA(multi-headed self-attention) 모듈에 제공하여 상기 정제된 제2 특징맵을 생성하는 단계;

상기 제3 특징맵을 상기 정제 블록의 제3 스테이지의 2개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 상기 정제된 제3 특징맵을 생성하는 단계; 및

상기 제4 특징맵을 상기 정제 블록의 제4 스테이지의 하나의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 상기 정제된 제4 특징맵을 생성하는 단계;

를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 6

제1항에 있어서,

상기 특징 퓨전을 수행하여 최종 특징을 생성하는 단계는,

상기 정제된 제3 특징맵과 상기 정제된 제4 특징맵에 대한 특징 퓨전을 수행하여 제1 퓨전값을 생성하는 단계;

상기 제1 퓨전값과 상기 정제된 제2 특징맵에 대한 특징 퓨전을 수행하여 제2 퓨전값을 생성하는 단계;

상기 제2 퓨전값과 상기 제1 특징맵에 대한 특징 퓨전을 수행하여 제3 퓨전값을 생성하는 단계; 및

상기 제3 퓨전값 및 상기 제4 특징맵에 평탄화(flattening)를 수행하여 상기 최종 특징을 생성하는 단계;

를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 7

제1항에 있어서,

상기 회귀 블록은 콘볼루션(convolution), 특징 정규화(feature normalization) 및 디콘볼루션(deconvolution)을 위한 복수의 레이어를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 8

제1항에 있어서,

상기 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 상기 수신된 비디오에 대한 비디오 요약을 수행하는 단계는,

상기 각각의 프레임에 대응하는 중요성 점수를 이용하여 상기 비디오를 구성하는 각각의 샷(shot)에 대응하는 요약 점수를 산출하는 단계; 및

상기 각각의 샷에 대응하는 요약 점수를 기초로 상기 비디오에 대한 비디오 요약을 수행하는 단계;

를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 9

제8항에 있어서,

상기 각각의 샷에 대응하는 요약 점수를 산출하는 단계는,

상기 각각의 샷에 포함된 프레임들의 중요성 점수의 평균을 기초로 상기 각각의 샷에 대응하는 요약 점수를 산출하는 단계;

를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 10

제8항에 있어서,

상기 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 상기 수신된 비디오에 대한 비디오 요약을 수행하는 단계는,

사전 결정된 비율 이내의 길이를 갖으면서 상기 요약 점수의 합이 최대가 되도록, 상기 비디오를 구성하는 복수의 샷 중 적어도 일부의 샷을 키샷(keyshot)으로 추출하여 비디오 요약을 수행하는 단계;

를 포함하는 멀티 스케일 특징 기반의 비디오 요약 방법.

청구항 11

제1항 내지 제10항 중 어느 한 항에 따른 방법을 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램.

청구항 12

컴퓨팅 장치로서,

통신 모듈;

메모리; 및

상기 메모리와 연결되고, 상기 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서

를 포함하고,

상기 적어도 하나의 프로그램은,

복수의 프레임을 포함하는 비디오를 수신하고,

상기 복수의 프레임을 백본 네트워크에 제공하여 상기 복수의 프레임에 포함된 각각의 프레임에 대응하는 제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성하고,

상기 생성된 제2 특징맵, 제3 특징맵 및 제4 특징맵을 정제 블록에 제공하여 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 생성하고,

상기 제1 특징맵, 상기 정제된 제2 특징맵, 상기 정제된 제3 특징맵 및 상기 정제된 제4 특징맵을 기초로 특징 퓨전을 수행하여 최종 특징을 생성하고,

상기 생성된 최종 특징을 회귀 블록에 제공하여 상기 각각의 프레임에 대응하는 중요성 점수를 산출하고,

상기 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 상기 수신된 비디오에 대한 비디오 요약을 수행하기 위한 명령어들을 포함하는 컴퓨팅 장치.

청구항 13

제12항에 있어서,

상기 백본 네트워크는 ViT 기반의 백본 네트워크를 포함하는 컴퓨팅 장치.

청구항 14

제12항에 있어서,

상기 적어도 하나의 프로그램은,

상기 복수의 프레임을 상기 백본 네트워크의 제1 스테이지에 제공하여 상기 각각의 프레임에 대응하는 상기 제1 특징맵을 생성하고,

상기 제1 특징맵을 상기 백본 네트워크의 제2 스테이지에 제공하여 상기 제2 특징맵을 생성하고,

상기 제1 특징맵 및 상기 제2 특징맵을 상기 백본 네트워크의 제2 스테이지에 제공하여 상기 제3 특징맵을 생성하고,

상기 제1 특징맵, 상기 제2 특징맵 및 상기 제3 특징맵을 상기 백본 네트워크의 제3 스테이지에 제공하여 상기 제4 특징맵을 생성하기 위한 명령어들을 더 포함하는 컴퓨팅 장치.

청구항 15

제12항에 있어서,

상기 정제 블록은 피라미드형 정제 블록을 포함하는 컴퓨팅 장치.

청구항 16

제12항에 있어서,

상기 적어도 하나의 프로그램은,

상기 제2 특징맵을 상기 정제 블록의 제2 스테이지의 3개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 상기 정제된 제2 특징맵을 생성하고,

상기 제3 특징맵을 상기 정제 블록의 제3 스테이지의 2개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 상기 정제된 제3 특징맵을 생성하고,

상기 제4 특징맵을 상기 정제 블록의 제4 스테이지의 하나의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 상기 정제된 제4 특징맵을 생성하기 위한 명령어들을 더 포함하는 컴퓨팅 장치.

청구항 17

제12항에 있어서,

상기 적어도 하나의 프로그램은,

상기 정제된 제3 특징맵과 상기 정제된 제4 특징맵에 대한 특징 퓨전을 수행하여 제1 퓨전값을 생성하고,

상기 제1 퓨전값과 상기 정제된 제2 특징맵에 대한 특징 퓨전을 수행하여 제2 퓨전값을 생성하고,

상기 제2 퓨전값과 상기 제1 특징맵에 대한 특징 퓨전을 수행하여 제3 퓨전값을 생성하고,

상기 제3 퓨전값 및 상기 제4 특징맵에 평탄화를 수행하여 상기 최종 특징을 생성하기 위한 명령어들을 더 포함

하는 컴퓨팅 장치.

청구항 18

제12항에 있어서,

상기 회귀 블록은 콘볼루션, 특징 정규화 및 디콘볼루션을 위한 복수의 레이어를 포함하는 컴퓨팅 장치.

청구항 19

제12항에 있어서,

상기 적어도 하나의 프로그램은,

상기 각각의 프레임에 대응하는 중요성 점수를 이용하여 상기 비디오를 구성하는 각각의 샷에 대응하는 요약 점수를 산출하고,

상기 각각의 샷에 대응하는 요약 점수를 기초로 상기 비디오에 대한 비디오 요약을 수행하기 위한 명령어들을 더 포함하는 컴퓨팅 장치.

청구항 20

제19항에 있어서,

상기 적어도 하나의 프로그램은,

상기 각각의 샷에 포함된 프레임들의 중요성 점수의 평균을 기초로 상기 각각의 샷에 대응하는 요약 점수를 산출하기 위한 명령어들을 더 포함하는 컴퓨팅 장치.

청구항 21

제19항에 있어서,

상기 적어도 하나의 프로그램은,

사전 결정된 비율 이내의 길이를 갖으면서 상기 요약 점수의 합이 최대가 되도록, 상기 비디오를 구성하는 복수의 샷 중 적어도 일부의 샷을 키샷으로 추출하여 비디오 요약을 수행하기 위한 명령어들을 더 포함하는 컴퓨팅 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 멀티 스케일 특징 기반의 비디오 요약 방법 및 장치에 관한 것으로, 비디오를 대표하는 최적의 프레임을 추출하여 비디오 요약을 생성하기 위한 멀티 스케일 특징 기반의 비디오 요약 방법 및 장치에 관한 것이다.

배 경 기 술

[0003] 최근 보안을 위한 감시 카메라 등의 급격한 확장으로 대량의 비디오 콘텐츠가 매일 생성되고 있다. 예를 들어, 공항, 교통 시스템, 병원 및 기타 공공 장소에는 수많은 감시 카메라가 설치되어 있으며, 비디오 모니터링 시스템의 관리자는 감시 카메라에서 촬영된 비디오를 지속적으로 모니터링하고 있다. 그러나, 감시 카메라에 의해 촬영된 비디오는 실시간으로 관찰하기 어려운 대량의 시각적 데이터를 포함하고 있어, 관리자가 해당 비디오를

모두 모니터링하는 것은 어려운 문제가 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 상기와 같은 문제점을 해결하기 위한 멀티 스케일 특징 기반의 비디오 요약 방법, 컴퓨터 판독 가능 매체에 저장된 컴퓨터 프로그램, 컴퓨터 프로그램이 저장된 컴퓨터 판독 가능 매체 및 장치(시스템)를 제공한다.

과제의 해결 수단

[0007] 본 발명은 방법, 장치(시스템), 컴퓨터 판독 가능 매체에 저장된 컴퓨터 프로그램 또는 컴퓨터 프로그램이 저장된 컴퓨터 판독 가능 매체를 포함한 다양한 방식으로 구현될 수 있다.

[0008] 본 발명의 일 실시예에 따르면, 적어도 하나의 프로세서에 의해 수행되는 멀티 스케일 특징 기반의 비디오 요약 방법은, 복수의 프레임을 포함하는 비디오를 수신하는 단계, 복수의 프레임을 백본 네트워크에 제공하여 복수의 프레임에 포함된 각각의 프레임에 대응하는 제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성하는 단계, 생성된 제2 특징맵, 제3 특징맵 및 제4 특징맵을 정제 블록에 제공하여 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 생성하는 단계, 제1 특징맵, 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 기초로 특징 퓨전을 수행하여 최종 특징을 생성하는 단계, 생성된 최종 특징을 회귀 블록에 제공하여 각각의 프레임에 대응하는 중요성 점수를 산출하는 단계 및 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 수신된 비디오에 대한 비디오 요약을 수행하는 단계를 포함한다.

[0009] 본 발명의 일 실시예에 따르면, 백본 네트워크는 ViT(vision transformer) 기반의 백본 네트워크를 포함한다.

[0010] 본 발명의 일 실시예에 따르면, 각각의 프레임에 대응하는 제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성하는 단계는, 복수의 프레임을 백본 네트워크의 제1 스테이지에 제공하여 각각의 프레임에 대응하는 제1 특징맵을 생성하는 단계, 제1 특징맵을 백본 네트워크의 제2 스테이지에 제공하여 제2 특징맵을 생성하는 단계, 제1 특징맵 및 제2 특징맵을 백본 네트워크의 제2 스테이지에 제공하여 제3 특징맵을 생성하는 단계 및 제1 특징맵, 제2 특징맵 및 제3 특징맵을 백본 네트워크의 제3 스테이지에 제공하여 제4 특징맵을 생성하는 단계를 포함한다.

[0011] 본 발명의 일 실시예에 따르면, 정제 블록은 피라미드형 정제 블록을 포함한다.

[0012] 본 발명의 일 실시예에 따르면, 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 생성하는 단계는, 제2 특징맵을 정제 블록의 제2 스테이지의 3개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제2 특징맵을 생성하는 단계, 제3 특징맵을 정제 블록의 제3 스테이지의 2개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제3 특징맵을 생성하는 단계 및 제4 특징맵을 정제 블록의 제4 스테이지의 하나의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제4 특징맵을 생성하는 단계를 포함한다.

[0013] 본 발명의 일 실시예에 따르면, 특징 퓨전을 수행하여 최종 특징을 생성하는 단계는, 정제된 제3 특징맵과 정제된 제4 특징맵에 대한 특징 퓨전을 수행하여 제1 퓨전값을 생성하는 단계, 제1 퓨전값과 정제된 제2 특징맵에 대한 특징 퓨전을 수행하여 제2 퓨전값을 생성하는 단계, 제2 퓨전값과 제1 특징맵에 대한 특징 퓨전을 수행하여 제3 퓨전값을 생성하는 단계 및 제3 퓨전값 및 제4 특징맵에 평탄화를 수행하여 최종 특징을 생성하는 단계를 포함한다.

[0014] 본 발명의 일 실시예에 따르면, 회귀 블록은 콘볼루션, 특징 정규화 및 디콘볼루션을 위한 복수의 레이어를 포함한다.

[0015] 본 발명의 일 실시예에 따르면, 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 수신된 비디오에 대한 비디오 요약을 수행하는 단계는, 각각의 프레임에 대응하는 중요성 점수를 이용하여 비디오를 구성하는 각각의 샷에 대응하는 요약 점수를 산출하는 단계 및 각각의 샷에 대응하는 요약 점수를 기초로 비디오에 대한 비디오 요약을 수행하는 단계를 더 포함한다.

[0016] 본 발명의 일 실시예에 따르면, 각각의 샷에 대응하는 요약 점수를 산출하는 단계는, 각각의 샷에 포함된 프레임들의 중요성 점수의 평균을 기초로 각각의 샷에 대응하는 요약 점수를 산출하는 단계를 포함한다.

[0017] 본 발명의 일 실시예에 따르면, 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 수신된 비디오에 대한 비디오 요약을 수행하는 단계는, 사전 결정된 비율 이내의 길이를 갖으면서 요약 점수의 합이 최대가 되도록, 비디오를 구성하는 복수의 샷 중 적어도 일부의 샷을 키샷으로 추출하여 비디오 요약을 수행하는 단계를 포함한다.

[0018] 본 발명의 일 실시예에 따른 상술된 방법을 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램이 제공된다.

[0019] 본 발명의 일 실시예에 따른 컴퓨팅 장치는, 통신 모듈, 메모리 및 메모리와 연결되고, 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서를 포함한다. 적어도 하나의 프로그램은, 복수의 프레임을 포함하는 비디오를 수신하고, 복수의 프레임을 백본 네트워크에 제공하여 복수의 프레임에 포함된 각각의 프레임에 대응하는 제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성하고, 생성된 제2 특징맵, 제3 특징맵 및 제4 특징맵을 정제 블록에 제공하여 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 생성하고, 제1 특징맵, 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 기초로 특징 퓨전을 수행하여 최종 특징을 생성하고, 생성된 최종 특징을 회귀 블록에 제공하여 각각의 프레임에 대응하는 중요성 점수를 산출하고, 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 수신된 비디오에 대한 비디오 요약을 수행하기 위한 명령어들을 포함한다.

[0020] 본 발명의 일 실시예에 따르면, 적어도 하나의 프로그램은, 복수의 프레임을 백본 네트워크의 제1 스테이지에 제공하여 각각의 프레임에 대응하는 제1 특징맵을 생성하고, 제1 특징맵을 백본 네트워크의 제2 스테이지에 제공하여 제2 특징맵을 생성하고, 제1 특징맵 및 제2 특징맵을 백본 네트워크의 제2 스테이지에 제공하여 제3 특징맵을 생성하고, 제1 특징맵, 제2 특징맵 및 제3 특징맵을 백본 네트워크의 제3 스테이지에 제공하여 제4 특징맵을 생성하기 위한 명령어들을 더 포함한다.

[0021] 본 발명의 일 실시예에 따르면, 적어도 하나의 프로그램은, 제2 특징맵을 정제 블록의 제2 스테이지의 3개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제2 특징맵을 생성하고, 제3 특징맵을 정제 블록의 제3 스테이지의 2개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제3 특징맵을 생성하고, 제4 특징맵을 정제 블록의 제4 스테이지의 하나의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제4 특징맵을 생성하기 위한 명령어들을 더 포함한다.

[0022] 본 발명의 일 실시예에 따르면, 적어도 하나의 프로그램은, 정제된 제3 특징맵과 정제된 제4 특징맵에 대한 특징 퓨전을 수행하여 제1 퓨전값을 생성하고, 제1 퓨전값과 정제된 제2 특징맵에 대한 특징 퓨전을 수행하여 제2 퓨전값을 생성하고, 제2 퓨전값과 제1 특징맵에 대한 특징 퓨전을 수행하여 제3 퓨전값을 생성하고, 제3 퓨전값 및 정제된 제4 특징맵에 평탄화를 수행하여 최종 특징을 생성하기 위한 명령어들을 더 포함한다.

[0023] 본 발명의 일 실시예에 따르면, 적어도 하나의 프로그램은, 각각의 프레임에 대응하는 중요성 점수를 이용하여 비디오를 구성하는 각각의 샷에 대응하는 요약 점수를 산출하고, 각각의 샷에 대응하는 요약 점수를 기초로 비디오에 대한 비디오 요약을 수행하기 위한 명령어들을 더 포함한다.

[0024] 본 발명의 일 실시예에 따르면, 적어도 하나의 프로그램은, 각각의 샷에 포함된 프레임들의 중요성 점수의 평균을 기초로 각각의 샷에 대응하는 요약 점수를 산출하기 위한 명령어들을 더 포함한다.

[0025] 본 발명의 일 실시예에 따르면, 적어도 하나의 프로그램은, 사전 결정된 비율 이내의 길이를 갖으면서 요약 점수의 합이 최대가 되도록, 비디오를 구성하는 복수의 샷 중 적어도 일부의 샷을 키샷으로 추출하여 비디오 요약을 수행하기 위한 명령어들을 더 포함한다.

발명의 효과

[0027] 본 발명의 다양한 실시예에서 인공신경망 모델을 통해 피라미드형 특징맵 생성, 특징 정제 및 특징 퓨전의 과정을 거쳐 프레임의 초기 레벨의 정보와 전역 특징 정보를 효율적으로 포함하는 최종 특징이 생성될 수 있으며, 해당 최종 특징을 이용하여 각각의 프레임과 비디오 사이의 연관성을 효과적으로 판단하여 비디오를 대표하는 비디오 요약이 수행될 수 있다.

[0028] 본 발명의 효과는 이상에서 언급한 효과로 제한되지 않으며, 언급되지 않은 다른 효과들은 청구범위의 기재로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자("통상의 기술자"라 함)에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0030] 본 발명의 실시예들은, 이하 설명하는 첨부 도면들을 참조하여 설명될 것이며, 여기서 유사한 참조 번호는 유사한 요소들을 나타내지만, 이에 한정되지는 않는다.

도 1은 본 발명의 일 실시예에 따른 인공신경망 모델의 구조를 나타내는 예시적인 도면이다.

도 2는 본 발명의 일 실시예에 따른 회귀 블록의 구조를 나타내는 예시적인 도면이다.

도 3은 본 발명의 일 실시예에 따른 멀티 스케일 특징 기반의 비디오 요약 방법의 예시를 나타내는 도면이다.

도 4는 본 발명의 일 실시예에 따른 컴퓨팅 장치의 내부 구성을 나타내는 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 이하, 본 발명의 실시를 위한 구체적인 내용을 첨부된 도면을 참조하여 상세히 설명한다. 다만, 이하의 설명에서는 본 발명의 요지를 불필요하게 흐릴 우려가 있는 경우, 널리 알려진 기능이나 구성에 관한 구체적 설명은 생략하기로 한다.

[0032] 첨부된 도면에서, 동일하거나 대응하는 구성요소에는 동일한 참조부호가 부여되어 있다. 또한, 이하의 실시예들의 설명에 있어서, 동일하거나 대응되는 구성요소를 중복하여 기술하는 것이 생략될 수 있다. 그러나, 구성요소에 관한 기술이 생략되어도, 그러한 구성요소가 어떤 실시예에 포함되지 않는 것으로 의도되지는 않는다.

[0033] 개시된 실시예의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명이 완전하도록 하고, 본 발명이 통상의 기술자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것일 뿐이다.

[0034] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 개시된 실시예에 대해 구체적으로 설명하기로 한다. 본 명세서에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 관련 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서, 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0035] 본 명세서에서의 단수의 표현은 문맥상 명백하게 단수인 것으로 특정하지 않는 한, 복수의 표현을 포함한다. 또한, 복수의 표현은 문맥상 명백하게 복수인 것으로 특정하지 않는 한, 단수의 표현을 포함한다. 명세서 전체에서 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다.

[0036] 본 발명에서, "포함하다", "포함하는" 등의 용어는 특징들, 단계들, 동작들, 요소들 및/또는 구성 요소들이 존재하는 것을 나타낼 수 있으나, 이러한 용어가 하나 이상의 다른 기능들, 단계들, 동작들, 요소들, 구성 요소들 및/또는 이들의 조합이 추가되는 것을 배제하지는 않는다.

[0037] 본 발명에서, 특정 구성 요소가 임의의 다른 구성 요소에 "결합", "조합", "연결" 되거나, "반응" 하는 것으로 언급된 경우, 특정 구성 요소는 다른 구성 요소에 직접 결합, 조합 및/또는 연결되거나, 반응할 수 있으나, 이에 한정되지 않는다. 예를 들어, 특정 구성 요소와 다른 구성 요소 사이에 하나 이상의 중간 구성 요소가 존재할 수 있다. 또한, 본 발명에서 "및/또는"은 열거된 하나 이상의 항목의 각각 또는 하나 이상의 항목의 적어도 일부의 조합을 포함할 수 있다.

[0038] 본 발명에서, "제1", "제2" 등의 용어는 특정 구성 요소를 다른 구성 요소와 구별하기 위해 사용되는 것으로, 이러한 용어에 의해 상술된 구성 요소가 제한되진 않는다. 예를 들어, "제1" 구성 요소는 "제2" 구성 요소와 동일하거나 유사한 형태의 요소일 수 있다.

[0039] 본 발명에서, "비디오 요약(video summarization)"은 비디오를 구성하는 복수의 프레임 중 해당 비디오를 대표하는 프레임들을 추출하는 작업을 지칭할 수 있다.

[0040] 본 발명에서, "특징 정제(feature reinforcement)"는 특징맵을 강조하거나 향상시켜 더욱 대표성을 갖도록 만드는 작업을 지칭할 수 있다. 또한, 피라미드형 특징 정제는 점진적으로 특징맵을 정제하여, 다양한 레벨에서 수집된 특성을 병합하여 출력을 형성하는 것을 지칭할 수 있다.

[0041] 본 발명에서, "MHSA(multi-headed self-attention) 모듈"은 특성 표현을 처리하기 위해 여러 개의 병렬적인 셀프 어텐션 매커니즘(self-attention mechanism)을 계산하여 여러 개의 어텐션 벡터를 생성하기 위한 모듈을 지칭할 수 있다. 또한, 본 발명에서, "DASPP(dense atrous spatial pyramid pooling) 모듈"은 다중 스케일의 특성을 고려하여 지역적 문맥 정보(local context information) 및 전역적 문맥 정보(global context information)를 추출하기 위한 모듈을 지칭할 수 있다.

[0042] 본 발명에서, "평탄화(flattening)"는 앞단에서 추출된 특징맵을 전결합층에 전달하기 위해 1차원 자료로 변환시키는 것을 지칭할 수 있다.

[0043] 본 발명에서, "샷(shot)"은 비디오를 구성하는 복수의 프레임 중 적어도 일부를 포함하는 프레임 시퀀스(frame sequence)를 나타낼 수 있다.

[0044] 도 1은 본 발명의 일 실시예에 따른 인공신경망 모델(100)의 구조를 나타내는 예시적인 도면이다. 도시된 것과 같이, 인공신경망 모델(100)은 비디오 요약을 수행하기 위한 모델로서, 백본 네트워크(backbone network)(120), 정제 블록(refinement block)(130), 회귀 블록(regressor block)(150) 등을 포함하도록 구성될 수 있다. 예를 들어, 인공신경망 모델(100)은 비디오를 구성하는 각각의 프레임($F_1, F_2, F_3, \cdots, F_n$)에 대한 점수를 산출하고, 산출된 점수를 기초로 비디오를 대표하는 프레임을 추출하여 비디오 요약을 수행할 수 있다.

[0045] 일 실시예에 따르면, 인공신경망 모델(100)의 백본 네트워크(120) 상에 복수의 프레임을 포함하는 비디오(110)가 제공될 수 있다. 여기서, 백본 네트워크(120)는 비디오를 구성하는 각각의 프레임 이미지로부터 모서리(edge), 색상(color), 모양(shape) 등의 특징을 추출하여 특징맵(feature map)을 생성하도록 구성될 수 있다. 즉, 복수의 프레임을 포함하는 비디오(110)를 수신하는 경우, 백본 네트워크(120)는 복수의 프레임에 포함된 각각의 프레임에 대응하는 특징맵을 생성할 수 있다.

[0046] 일 실시예에 따르면, 백본 네트워크(120)는 프레임으로부터 멀티 스케일 특징(multi-scale feature)을 추출하도록 구성될 수 있다. 예를 들어, 백본 네트워크(120)는 이미지 태스크를 트랜스포머 모델로 구현한 ViT(vision transformer) 기반의 모델로서, 피라미드형의 복수의 스테이지(stage)에 대응하는 복수의 특징맵을 생성하도록 구성될 수 있다. 즉, 백본 네트워크(120)는 복수의 프레임에 포함된 각각의 프레임에 대응하는 제1 특징맵($\mathcal{F}_{m1}$), 제2 특징맵($\mathcal{F}_{m2}$), 제3 특징맵($\mathcal{F}_{m3}$) 및 제4 특징맵($\mathcal{F}_{m4}$)을 생성할 수 있다.

[0047] 일 실시예에 따르면, 백본 네트워크(120)는 제1 스테이지에 각각의 프레임을 제공하여 각각의 프레임에 대응하는 제1 특징맵($\mathcal{F}_{m1}$)을 생성할 수 있다. 예를 들어, 제1 특징맵($\mathcal{F}_{m1}$)은 다음의 수학식 1에 의해 생성될 수 있다.

수학식 1

[0049] $$\mathcal{F}_{m1} = E(TB_1(\mathcal{F}_{ith}RGB))$$

[0051] 여기서, E는 백본 네트워크(120)의 인코더(encoder)를 지칭하고, $\mathcal{F}_{ith}RGB$는 i번째 입력된 프레임을 지칭할 수 있다. 또한, TB_1는 백본 네트워크(120)의 제1 스테이지를 지칭할 수 있다. 즉, 비디오(110)의 프레임이 백본 네트워크(120)에 입력되는 경우 제1 특징맵($\mathcal{F}_{m1}$)이 먼저 생성될 수 있다.

[0052] 또한, 백본 네트워크(120)는 이와 같이 제1 특징맵($\mathcal{F}_{m1}$)이 생성되는 경우 제1 특징맵($\mathcal{F}_{m1}$)을 제2 스테이지에 제공하여 제2 특징맵($\mathcal{F}_{m2}$)을 생성할 수 있다. 이 경우, 제2 특징맵($\mathcal{F}_{m2}$)은 제1 특징맵($\mathcal{F}_{m1}$)에 관한 정보를 포함할 수 있다. 예를 들어, 제2 특징맵($\mathcal{F}_{m2}$)은 다음의 수학식 2에 의해 생성될 수 있다.

수학식 2

[0054] $$\mathcal{F}_{m2} = E(TB_2(\mathcal{F}_{m1}))$$

[0056] 여기서, TB_2는 백본 네트워크(120)의 제2 스테이지를 지칭할 수 있다. 추가적으로, 백본 네트워크(120)는 제1 특징맵($\mathcal{F}_{m1}$) 및 제2 특징맵($\mathcal{F}_{m2}$)을 제3 스테이지에 제공하여 제3 특징맵($\mathcal{F}_{m3}$)을 생성할 수 있다. 이 경우, 제3 특징맵($\mathcal{F}_{m3}$)은 제1 특징맵($\mathcal{F}_{m1}$) 및 제2 특징맵($\mathcal{F}_{m2}$)에 관한 정보를 포함할 수 있다. 예를 들어, 제3 특징맵($\mathcal{F}_{m3}$)은 다음의 수학식 3에 의해 생성될 수 있다.

수학식 3

[0058] $$\mathcal{F}_{m3} = E(TB_3(\mathcal{F}_{m1} \oplus \mathcal{F}_{m2}))$$

[0060] 여기서, TB_3는 백본 네트워크(120)의 제3 스테이지를 지칭할 수 있다. 마찬가지로, 백본 네트워크(120)는 제1 특징맵($\mathcal{F}_{m1}$), 제2 특징맵($\mathcal{F}_{m2}$) 및 제3 특징맵($\mathcal{F}_{m3}$)을 제4 스테이지에 제공하여 제4 특징맵($\mathcal{F}_{m4}$)을 생성할 수 있다. 이 경우, 제4 특징맵($\mathcal{F}_{m4}$)은 제1 특징맵($\mathcal{F}_{m1}$), 제2 특징맵($\mathcal{F}_{m2}$) 및 제3 특징맵($\mathcal{F}_{m3}$)에 관한 정보를 포함할 수 있다. 예를 들어, 제4 특징맵($\mathcal{F}_{m4}$)은 다음의 수학식 4에 의해 생성될 수 있다.

수학식 4

[0062] $$\mathcal{F}_{m4} = E(TB_4(\mathcal{F}_{m1} \oplus \mathcal{F}_{m2} \oplus \mathcal{F}_{m3}))$$

[0064] 여기서, TB_4는 백본 네트워크(120)의 제4 스테이지를 지칭할 수 있다. 상술된 과정에 의해, 피라미드형으로 복수의 특징맵이 생성될 수 있으며, 이 경우, 제1 특징맵($\mathcal{F}_{m1}$)은 프레임의 초기 레벨(initial level)의 정보를 포함할 수 있으며, 깊이가 깊어질수록 멀티 레벨의 전역 특징(global feature) 정보를 포함할 수 있다. 이와 같이 생성된 복수의 특징맵을 이용하는 경우, 비디오 프레임에 대한 상당한 정보를 활용할 수 있어, 정보의 손실 없이 해당 비디오(110)를 대표하는 프레임들을 추출할 수 있다.

[0065] 일 실시예에 따르면, 생성된 제2 특징맵($\mathcal{F}_{m2}$), 제3 특징맵($\mathcal{F}_{m3}$) 및 제4 특징맵($\mathcal{F}_{m4}$)에 대한 특징 정제

(feature reinforcement)가 수행될 수 있다. 예를 들어, 제2 특징맵($\mathcal{F}_{m2}$), 제3 특징맵($\mathcal{F}_{m3}$) 및 제4 특징맵($\mathcal{F}_{m4}$)이 정제 블록(130)에 제공되는 경우, 정제된 제2 특징맵(P_{m2}), 정제된 제3 특징맵(P_{m3}) 및 정제된 제4 특징맵(P_{m4})이 생성될 수 있다. 도시된 것과 같이, 정제 블록(130)은 피라미드형 정제 블록(pyramidal refinement block)을 포함할 수 있다.

[0066] 일 실시예에 따르면, 정제 블록(130)은 3개의 스테이지를 포함하도록 구성될 수 있다. 즉, 제1 특징맵($\mathcal{F}_{m1}$)은 프레임의 초기 수준의 정보를 포함하는 것이어서 특징 퓨전(feature fusion) 시까지 유지하고, 나머지 특징맵들은 특징 정제를 위해 정제 블록에 제공될 수 있다. 정제 블록은 DASPP(dense atrous spatial pyramid pooling) 모듈 및 MHSA(multi-headed self-attention) 모듈로 구성될 수 있으며, 스테이지마다 상이한 개수의 모듈들을 이용할 수 있다.

[0067] 일 실시예에 따르면, 정제 블록은 제2 스테이지의 3개의 DASPP 모듈과 하나의 MHSA 모듈을 이용하여 제2 특징맵($\mathcal{F}_{m2}$)을 기초로 정제된 제2 특징맵(P_{m2})을 생성할 수 있다. 또한, 정제 블록은 제3 스테이지의 2개의 DASPP 모듈과 하나의 MHSA 모듈을 이용하여 제3 특징맵($\mathcal{F}_{m3}$)을 기초로 정제된 제3 특징맵(P_{m3})을 생성할 수 있다. 마지막으로, 정제 블록은 제4 스테이지의 하나의 DASPP 모듈과 하나의 MHSA 모듈을 이용하여 제4 특징맵($\mathcal{F}_{m4}$)을 기초로 정제된 제4 특징맵(P_{m4})을 생성할 수 있다. 이와 같이 특징 정제가 수행되는 경우, 디코딩 중 복구할 수 없는 정보의 손실이 최소화될 수 있다.

[0068] 일 실시예에 따르면, 생성되거나 정제된 특징맵에 대한 특징 퓨전이 수행될 수 있다. 먼저, 정제된 제3 특징맵(P_{m3})과 정제된 제4 특징맵(P_{m4})에 대한 특징 퓨전이 수행되어 제1 퓨전값(P_{m43})이 생성될 수 있다. 예를 들어, 제1 퓨전값(P_{m43})은 다음의 수학식 5에 의해 생성될 수 있다.

수학식 5

[0070] $$P_{m43} = \sigma(P_{m4} \oplus P_{m3})$$

[0072] 여기서, σ는 잔차 채널 어텐션(residual channel attention)을 지칭할 수 있다. 또한, 제1 퓨전값(P_{m43})과 정제된 제2 특징맵(P_{m2})에 대한 특징 퓨전이 수행되어 제2 퓨전값(P_{m432})이 생성될 수 있다. 예를 들어, 제2 퓨전값(P_{m432})은 다음의 수학식 6에 의해 생성될 수 있다.

수학식 6

[0074] $$P_{m432} = \sigma(\Omega(P_{m43} \oplus P_{m2}))$$

[0076] 여기서, Ω는 업 샘플링(up-sampling) 동작을 지칭할 수 있다. 그리고 나서, 제2 퓨전값(P_{m432})과 제1 특징맵($\mathcal{F}_{m1}$)에 대한 특징 퓨전이 수행되어 제3 퓨전값($P_{m432}\mathcal{F}_{m1}$)이 생성될 수 있다. 예를 들어, 제3 퓨전값

($P_{m432}\mathcal{F}_{m1}$)은 다음의 수학식 7에 의해 생성될 수 있다.

수학식 7

[0078] $$P_{m432}\mathcal{F}_{m1} = \sigma(\Omega(P_{m432} \oplus \mathcal{F}_{m1}))$$

[0080] 이와 같이 초기 레벨의 정보를 포함하는 제1 특징맵($\mathcal{F}_{m1}$)을 가장 마지막에 특징 퓨전하여 이용함으로써 초기 레벨의 정보의 손실을 최소화하면서도 프레임을 대표하는 특징을 효과적으로 결정할 수 있다.

[0081] 상술된 것과 같이, 특징 퓨전이 수행된 후, 제3 퓨전값($P_{m432}\mathcal{F}_{m1}$) 및 제4 특징맵($\mathcal{F}_{m4}$)에 평탄화(flattening)가 수행되어 최종 특징이 생성될 수 있다. 일 실시예에 따르면, 제3 퓨전값($P_{m432}\mathcal{F}_{m1}$)은 콘볼루션 헤드(convolution head)(140) 및 평탄화 레이어를 통해 평탄화될 수 있다. 예를 들어, 제3 퓨전값($P_{m432}\mathcal{F}_{m1}$)은 다음의 수학식 8에 의해 평탄화될 수 있다.

수학식 8

[0083] $$P_{m432}\mathcal{F}_{m1} \rightarrow C_H + F_{L1} = \mathcal{F}_{\beta}$$

[0085] 여기서, C_H는 콘볼루션 헤드(140)를 지칭하는 것으로, 3개의 1차원 콘볼루션 레이어(convolution layer)와 2개의 맥스 풀링 레이어(max pooling layer)로 구성될 수 있다. 또한, F_{L1}는 평탄화 레이어를 지칭할 수 있으며, $\mathcal{F}_{\beta}$는 평탄화된 제3 퓨전값을 지칭할 수 있다.

[0086] 추가적으로, 제4 특징맵($\mathcal{F}_{m4}$)은 평탄화 레이어를 통해 평탄화될 수 있다. 예를 들어, 제4 특징맵($\mathcal{F}_{m4}$)은 다음의 수학식 9에 의해 평탄화될 수 있다.

수학식 9

[0088] $$\mathcal{F}_{m4} \rightarrow F_{L2} = \mathcal{F}_{\alpha}$$

[0090] 여기서, F_{L2}는 평탄화 레이어를 지칭할 수 있으며, $\mathcal{F}_{\alpha}$는 평탄화된 제4 특징맵을 지칭할 수 있다. 이와 같이 평탄화가 수행된 후, 다음의 수학식 10과 같이 $\mathcal{F}_{\alpha}$ 및 $\mathcal{F}_{\beta}$에 대한 특징 퓨전이 수행될 수 있다.

수학식 10

[0092] $$\mathcal{F}_{\alpha\beta} = \mathfrak{N}_{h\times w}(B_c(\mathcal{F}_\alpha \oplus \mathcal{F}_\beta))$$

[0094] 여기서, B_c는 밸런싱 콘볼루션 레이어(balancing convolution layer)를 지칭할 수 있으며, h는 패치의 높이를 나타내고, w는 패치의 너비를 나타낼 수 있다. 이와 같이 생성된 $\mathcal{F}_{\alpha\beta}$가 ReLU 및 완전 연결 레이어(fully connected layer)를 포함하는 완전 연결 블록(fully connected block)에 제공되는 경우, 1024 차원의 최종 특징이 생성될 수 있다.

[0095] 이와 같이 생성된 최종 특징이 회귀 블록(150)에 제공되는 경우, 각각의 프레임에 대응하는 중요성 점수(160)가 산출될 수 있다. 여기서, 중요성 점수(160)는 해당 프레임과 비디오(110)의 연관성이 높을수록 높게 산출될 수 있다. 이 경우, 산출된 각각의 프레임에 대응하는 중요성 점수(160)를 이용하여 수신된 비디오(110)에 대한 비디오 요약이 수행될 수 있다.

[0096] 일 실시예에 따르면, 각각의 프레임에 대한 중요성 점수가 산출된 후, 각각의 프레임에 대응하는 중요성 점수를 이용하여 비디오(110)를 구성하는 각각의 샷(shot)에 대응하는 요약 점수가 산출될 수 있다. 또한, 각각의 샷에 대응하는 요약 점수를 기초로 비디오에 대한 비디오 요약이 수행될 수 있다. 예를 들어, 샷은 특정 개수의 프레임들을 포함하는 프레임 시퀀스로 구성될 수 있으며, 샷의 요약 점수는 해당 샷에 포함된 각각의 프레임들의 중요성 점수의 평균을 기초로 산출될 수 있다. 평균 점수가 산출되면, 사전 결정된 비율(예를 들어, 비디오의 전체 길이의 15%) 이내의 길이를 갖으면서 요약 점수의 합이 최대가 되도록, 비디오(110)를 구성하는 복수의 샷 중 적어도 일부의 샷이 키샷(keyshot)으로 추출되어 비디오 요약이 수행될 수 있다.

[0097] 이와 같은 구성에 의해, 인공신경망 모델(100)을 통해 피라미드형 특징맵 생성, 특징 정제 및 특징 퓨전의 과정을 거쳐 프레임의 초기 레벨의 정보와 전역 특징 정보를 효율적으로 포함하는 최종 특징이 생성될 수 있으며, 해당 최종 특징을 이용하여 각각의 프레임과 비디오(110) 사이의 연관성을 효과적으로 판단하여 비디오(110)를 대표하는 비디오 요약이 수행될 수 있다.

[0098] 도 2는 본 발명의 일 실시예에 따른 회귀 블록(150)의 구조를 나타내는 예시적인 도면이다. 일 실시예에 따르면, 회귀 블록(150)은 콘볼루션 연산을 사용하여 모든 프레임을 동시에 처리할 수 있다. 예를 들어, 회귀 블록(150)은 콘볼루션(convolution), 특징 정규화(feature normalization) 및 디콘볼루션(deconvolution)을 위한 복수의 레이어를 포함할 수 있다.

[0099] 도시된 예에서, 콘볼루션 레이어는 1차원 콘볼루션(1D Con1, 1D Con2), 배치 정규화(BN), ReLU 활성화 함수 및 드랍 아웃(drop out)을 포함하도록 구성될 수 있다. 이러한 콘볼루션 레이어의 후단에 복원을 수행하기 위한 디콘볼루션 레이어(DeCon1, DeCon2)가 배열될 수 있으며, 소프트맥스(softmax)를 통해 중요성 점수(160)가 산출될 수 있다.

[0100] 도 3은 본 발명의 일 실시예에 따른 멀티 스케일 특징 기반의 비디오 요약 방법(300)의 예시를 나타내는 도면이다. 멀티 스케일 특징 기반의 비디오 요약 방법(300)은 적어도 하나의 프로세서(예: 컴퓨팅 장치의 적어도 하나의 프로세서)에 의해 수행될 수 있다. 멀티 스케일 특징 기반의 비디오 요약 방법(300)은 프로세서가 복수의 프레임을 포함하는 비디오를 수신함으로써 개시될 수 있다(S310).

[0101] 이 경우, 프로세서는 복수의 프레임을 백본 네트워크에 제공하여 복수의 프레임에 포함된 각각의 프레임에 대응하는 제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성할 수 있다(S320). 여기서, 백본 네트워크는 ViT(vision transformer) 기반의 백본 네트워크를 포함할 수 있다. 일 실시예에 따르면, 프로세서는 복수의 프레임을 백본 네트워크의 제1 스테이지에 제공하여 각각의 프레임에 대응하는 제1 특징맵을 생성하고, 제1 특징맵을 백본 네트워크의 제2 스테이지에 제공하여 제2 특징맵을 생성할 수 있다. 또한, 프로세서는 제1 특징맵 및 제2 특징맵을 백본 네트워크의 제2 스테이지에 제공하여 제3 특징맵을 생성하고, 제1 특징맵, 제2 특징맵 및 제3 특징맵을 백본 네트워크의 제3 스테이지에 제공하여 제4 특징맵을 생성할 수 있다.

[0102] 프로세서는 생성된 제2 특징맵, 제3 특징맵 및 제4 특징맵을 정제 블록에 제공하여 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 생성할 수 있다(S330). 여기서, 정제 블록은 피라미드형 정제 블록

(pyramidal refinement block)을 포함할 수 있다. 예를 들어, 프로세서는 제2 특징맵을 정제 블록의 제2 스테이지의 3개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제2 특징맵을 생성하고, 제3 특징맵을 정제 블록의 제3 스테이지의 2개의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제3 특징맵을 생성하고, 제4 특징맵을 정제 블록의 제4 스테이지의 하나의 DASPP 모듈과 하나의 MHSA 모듈에 제공하여 정제된 제4 특징맵을 생성할 수 있다.

[0103] 프로세서는 제1 특징맵, 정제된 제2 특징맵, 정제된 제3 특징맵 및 정제된 제4 특징맵을 기초로 특징 퓨전을 수행하여 최종 특징을 생성할 수 있다(S340). 예를 들어, 프로세서는 정제된 제3 특징맵과 정제된 제4 특징맵에 대한 특징 퓨전을 수행하여 제1 퓨전값을 생성하고, 제1 퓨전값과 정제된 제2 특징맵에 대한 특징 퓨전을 수행하여 제2 퓨전값을 생성할 수 있다. 또한, 프로세서는 제2 퓨전값과 제1 특징맵에 대한 특징 퓨전을 수행하여 제3 퓨전값을 생성하고, 제3 퓨전값 및 제4 특징맵에 평탄화(flattening)를 수행하여 최종 특징을 생성할 수 있다.

[0104] 프로세서는 생성된 최종 특징을 회귀 블록에 제공하여 각각의 프레임에 대응하는 중요성 점수를 산출할 수 있다(S350). 여기서, 회귀 블록은 콘볼루션(convolution), 특징 정규화(feature normalization) 및 디콘볼루션(deconvolution)을 위한 복수의 레이어를 포함할 수 있다.

[0105] 그리고 나서, 프로세서는 산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여 수신된 디오에 대한 비디오 요약을 수행할 수 있다(S360). 일 실시예에 따르면, 프로세서는 각각의 프레임에 대응하는 중요성 점수를 이용하여 비디오를 구성하는 각각의 샷(shot)에 대응하는 요약 점수를 산출하고, 각각의 샷에 대응하는 요약 점수를 기초로 비디오에 대한 비디오 요약을 수행할 수 있다. 예를 들어, 프로세서는 각각의 샷에 포함된 프레임들의 중요성 점수의 평균을 기초로 각각의 샷에 대응하는 요약 점수를 산출할 수 있으며, 사전 결정된 비율 이내의 길이를 갖으면서 요약 점수의 합이 최대가 되도록, 비디오를 구성하는 복수의 샷 중 적어도 일부의 샷을 키샷(keyshot)으로 추출하여 비디오 요약을 수행할 수 있다.

[0106] 도 4는 본 발명의 일 실시예에 따른 컴퓨팅 장치(400)의 내부 구성을 나타내는 블록도이다. 컴퓨팅 장치(400)는 메모리(410), 프로세서(420), 통신 모듈(430) 및 입출력 인터페이스(440)를 포함할 수 있으며, 도 4에 도시된 바와 같이, 컴퓨팅 장치(400)는 통신 모듈(430)을 이용하여 네트워크를 통해 정보 및/또는 데이터를 통신할 수 있도록 구성될 수 있다.

[0107] 메모리(410)는 비-일시적인 임의의 컴퓨터 판독 가능한 기록매체를 포함할 수 있다. 일 실시예에 따르면, 메모리(410)는 RAM(random access memory), ROM(read only memory), 디스크 드라이브, SSD(solid state drive), 플래시 메모리(flash memory) 등과 같은 비소멸성 대용량 저장 장치(permanent mass storage device)를 포함할 수 있다. 다른 예로서, ROM, SSD, 플래시 메모리, 디스크 드라이브 등과 같은 비소멸성 대용량 저장 장치는 메모리와는 구분되는 별도의 영구 저장 장치로서 컴퓨팅 장치(400)에 포함될 수 있다. 또한, 메모리(410)에는 운영체제와 적어도 하나의 프로그램 코드가 저장될 수 있다.

[0108] 이러한 소프트웨어 구성요소들은 메모리(410)와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 이러한 컴퓨팅 장치(400)에 직접 연결가능한 기록 매체를 포함할 수 있는데, 예를 들어, 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 예로서, 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록매체가 아닌 통신 모듈(430)을 통해 메모리(410)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 애플리케이션의 설치 파일을 배포하는 파일 배포 시스템이 통신 모듈(430)을 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 메모리(410)에 로딩될 수 있다.

[0109] 프로세서(420)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(410) 또는 통신 모듈(430)에 의해 다른 사용자 단말(미도시) 또는 다른 외부 시스템으로 제공될 수 있다.

[0110] 통신 모듈(430)은 네트워크를 통해 사용자 단말(미도시)과 컴퓨팅 장치(400)가 서로 통신하기 위한 구성 또는 기능을 제공할 수 있으며, 컴퓨팅 장치(400)가 외부 시스템(일례로 별도의 클라우드 시스템 등)과 통신하기 위한 구성 또는 기능을 제공할 수 있다. 일례로, 컴퓨팅 장치(400)의 프로세서(420)의 제어에 따라 제공되는 제어 신호, 명령, 데이터 등이 통신 모듈(430)과 네트워크를 거쳐 사용자 단말 및/또는 외부 시스템의 통신 모듈을 통해 사용자 단말 및/또는 외부 시스템으로 전송될 수 있다.

[0111] 또한, 컴퓨팅 장치(400)의 입출력 인터페이스(440)는 컴퓨팅 장치(400)와 연결되거나 컴퓨팅 장치(400)가 포함

할 수 있는 입력 또는 출력을 위한 장치(미도시)와의 인터페이스를 위한 수단일 수 있다. 도 4에서는 입출력 인터페이스(440)가 프로세서(420)와 별도로 구성된 요소로서 도시되었으나, 이에 한정되지 않으며, 입출력 인터페이스(440)가 프로세서(420)에 포함되도록 구성될 수 있다. 컴퓨팅 장치(400)는 도 4의 구성요소들보다 더 많은 구성요소들을 포함할 수 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다.

[0112] 컴퓨팅 장치(400)의 프로세서(420)는 복수의 사용자 단말 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다.

[0113] 상술된 방법 및/또는 다양한 실시예들은, 디지털 전자 회로, 컴퓨터 하드웨어, 펌웨어, 소프트웨어 및/또는 이들의 조합으로 실현될 수 있다. 본 발명의 다양한 실시예들은 데이터 처리 장치, 예를 들어, 프로그래밍 가능한 하나 이상의 프로세서 및/또는 하나 이상의 컴퓨팅 장치에 의해 실행되거나, 컴퓨터 판독 가능한 기록 매체 및/또는 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 상술된 컴퓨터 프로그램은 컴파일된 언어 또는 해석된 언어를 포함하여 임의의 형태의 프로그래밍 언어로 작성될 수 있으며, 독립 실행형 프로그램, 모듈, 서브 루틴 등의 임의의 형태로 배포될 수 있다. 컴퓨터 프로그램은 하나의 컴퓨팅 장치, 동일한 네트워크를 통해 연결된 복수의 컴퓨팅 장치 및/또는 복수의 상이한 네트워크를 통해 연결되도록 분산된 복수의 컴퓨팅 장치를 통해 배포될 수 있다.

[0114] 상술된 방법 및/또는 다양한 실시예들은, 입력 데이터를 기초로 동작하거나 출력 데이터를 생성함으로써, 임의의 기능, 함수 등을 처리, 저장 및/또는 관리하는 하나 이상의 컴퓨터 프로그램을 실행하도록 구성된 하나 이상의 프로세서에 의해 수행될 수 있다. 예를 들어, 본 발명의 방법 및/또는 다양한 실시예는 FPGA(Field Programmable Gate Array) 또는 ASIC(Application Specific Integrated Circuit)과 같은 특수 목적 논리 회로에 의해 수행될 수 있으며, 본 발명의 방법 및/또는 실시예들을 수행하기 위한 장치 및/또는 시스템은 FPGA 또는 ASIC와 같은 특수 목적 논리 회로로서 구현될 수 있다.

[0115] 컴퓨터 프로그램을 실행하는 하나 이상의 프로세서는, 범용 목적 또는 특수 목적의 마이크로 프로세서 및/또는 임의의 종류의 디지털 컴퓨팅 장치의 하나 이상의 프로세서를 포함할 수 있다. 프로세서는 읽기 전용 메모리, 랜덤 액세스 메모리의 각각으로부터 명령 및/또는 데이터를 수신하거나, 읽기 전용 메모리와 랜덤 액세스 메모리로부터 명령 및/또는 데이터를 수신할 수 있다. 본 발명에서, 방법 및/또는 실시예들을 수행하는 컴퓨팅 장치의 구성 요소들은 명령어들을 실행하기 위한 하나 이상의 프로세서, 명령어들 및/또는 데이터를 저장하기 위한 하나 이상의 메모리 디바이스를 포함할 수 있다.

[0116] 일 실시예에 따르면, 컴퓨팅 장치는 데이터를 저장하기 위한 하나 이상의 대용량 저장 장치와 데이터를 주고받을 수 있다. 예를 들어, 컴퓨팅 장치는 자기 디스크(magnetic disc) 또는 광 디스크(optical disc)로부터 데이터를 수신하거나/수신하고, 자기 디스크 또는 광 디스크로 데이터를 전송할 수 있다. 컴퓨터 프로그램과 연관된 명령어들 및/또는 데이터를 저장하기에 적합한 컴퓨터 판독 가능한 저장 매체는, EPROM(Erasable Programmable Read-Only Memory), EEPROM(Electrically Erasable PROM), 플래시 메모리 장치 등의 반도체 메모리 장치를 포함하는 임의의 형태의 비 휘발성 메모리를 포함할 수 있으나, 이에 한정되지 않는다. 예를 들어, 컴퓨터 판독 가능한 저장 매체는 내부 하드 디스크 또는 이동식 디스크와 같은 자기 디스크, 광 자기 디스크, CD-ROM 및 DVD-ROM 디스크를 포함할 수 있다.

[0117] 사용자와의 상호 작용을 제공하기 위해, 컴퓨팅 장치는 정보를 사용자에게 제공하거나 디스플레이하기 위한 디스플레이 장치(예를 들어, CRT (Cathode Ray Tube), LCD(Liquid Crystal Display) 등) 및 사용자가 컴퓨팅 장치 상에 입력 및/또는 명령 등을 제공할 수 있는 포인팅 장치(예를 들어, 키보드, 마우스, 트랙볼 등)를 포함할 수 있으나, 이에 한정되지 않는다. 즉, 컴퓨팅 장치는 사용자와의 상호 작용을 제공하기 위한 임의의 다른 종류의 장치들을 더 포함할 수 있다. 예를 들어, 컴퓨팅 장치는 사용자와의 상호 작용을 위해, 시각적 피드백, 청각 피드백 및/또는 촉각 피드백 등을 포함하는 임의의 형태의 감각 피드백을 사용자에게 제공할 수 있다. 이에 대해, 사용자는 시각, 음성, 동작 등의 다양한 제스처를 통해 컴퓨팅 장치로 입력을 제공할 수 있다.

[0118] 본 발명에서, 다양한 실시예들은 백엔드 구성 요소(예: 데이터 서버), 미들웨어 구성 요소(예: 애플리케이션 서버) 및/또는 프론트 엔드 구성 요소를 포함하는 컴퓨팅 시스템에서 구현될 수 있다. 이 경우, 구성 요소들은 통신 네트워크와 같은 디지털 데이터 통신의 임의의 형태 또는 매체에 의해 상호 연결될 수 있다. 예를 들어, 통신 네트워크는 LAN(Local Area Network), WAN(Wide Area Network) 등을 포함할 수 있다.

[0119] 본 명세서에서 기술된 예시적인 실시예들에 기반한 컴퓨팅 장치는, 사용자 디바이스, 사용자 인터페이스(UI) 디바이스, 사용자 단말 또는 클라이언트 디바이스를 포함하여 사용자와 상호 작용하도록 구성된 하드웨어 및/또는

소프트웨어를 사용하여 구현될 수 있다. 예를 들어, 컴퓨팅 장치는 랩톱(laptop) 컴퓨터와 같은 휴대용 컴퓨팅 장치를 포함할 수 있다. 추가적으로 또는 대안적으로, 컴퓨팅 장치는, PDA(Personal Digital Assistants), 태블릿 PC, 게임 콘솔(game console), 웨어러블 디바이스(wearable device), IoT(internet of things) 디바이스, VR(virtual reality) 디바이스, AR(augmented reality) 디바이스 등을 포함할 수 있으나, 이에 한정되지 않는다. 컴퓨팅 장치는 사용자와 상호 작용하도록 구성된 다른 유형의 장치를 더 포함할 수 있다. 또한, 컴퓨팅 장치는 이동 통신 네트워크 등의 네트워크를 통한 무선 통신에 적합한 휴대용 통신 디바이스(예를 들어, 이동 전화, 스마트 전화, 무선 셀룰러 전화 등) 등을 포함할 수 있다. 컴퓨팅 장치는, 무선 주파수(RF; Radio Frequency), 마이크로파 주파수(MWF; Microwave Frequency) 및/또는 적외선 주파수(IRF; Infrared Ray Frequency)와 같은 무선 통신 기술들 및/또는 프로토콜들을 사용하여 네트워크 서버와 무선으로 통신하도록 구성될 수 있다.

[0120] 본 발명에서 특정 구조적 및 기능적 세부 사항을 포함하는 다양한 실시예들은 예시적인 것이다. 따라서, 본 발명의 실시예들은 상술된 것으로 한정되지 않으며, 여러 가지 다른 형태로 구현될 수 있다. 또한, 본 발명에서 사용된 용어는 일부 실시예를 설명하기 위한 것이며 실시예를 제한하는 것으로 해석되지 않는다. 예를 들어, 단수형 단어 및 상기는 문맥상 달리 명확하게 나타내지 않는 한 복수형도 포함하는 것으로 해석될 수 있다.

[0121] 본 발명에서, 달리 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함하여 본 명세서에서 사용되는 모든 용어는 이러한 개념이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 갖는다. 또한, 사전에 정의된 용어와 같이 일반적으로 사용되는 용어들은 관련 기술의 맥락에서의 의미와 일치하는 의미를 갖는 것으로 해석되어야 한다.

[0122] 본 명세서에서는 본 발명이 일부 실시예들과 관련하여 설명되었지만, 본 발명의 발명이 속하는 기술분야의 통상의 기술자가 이해할 수 있는 본 발명의 범위를 벗어나지 않는 범위에서 다양한 변형 및 변경이 이루어질 수 있다. 또한, 그러한 변형 및 변경은 본 명세서에 첨부된 특허청구의 범위 내에 속하는 것으로 생각되어야 한다.

부호의 설명

[0124] 100: 인공신경망 모델

110: 비디오

120: 백본 네트워크

130: 정제 블록

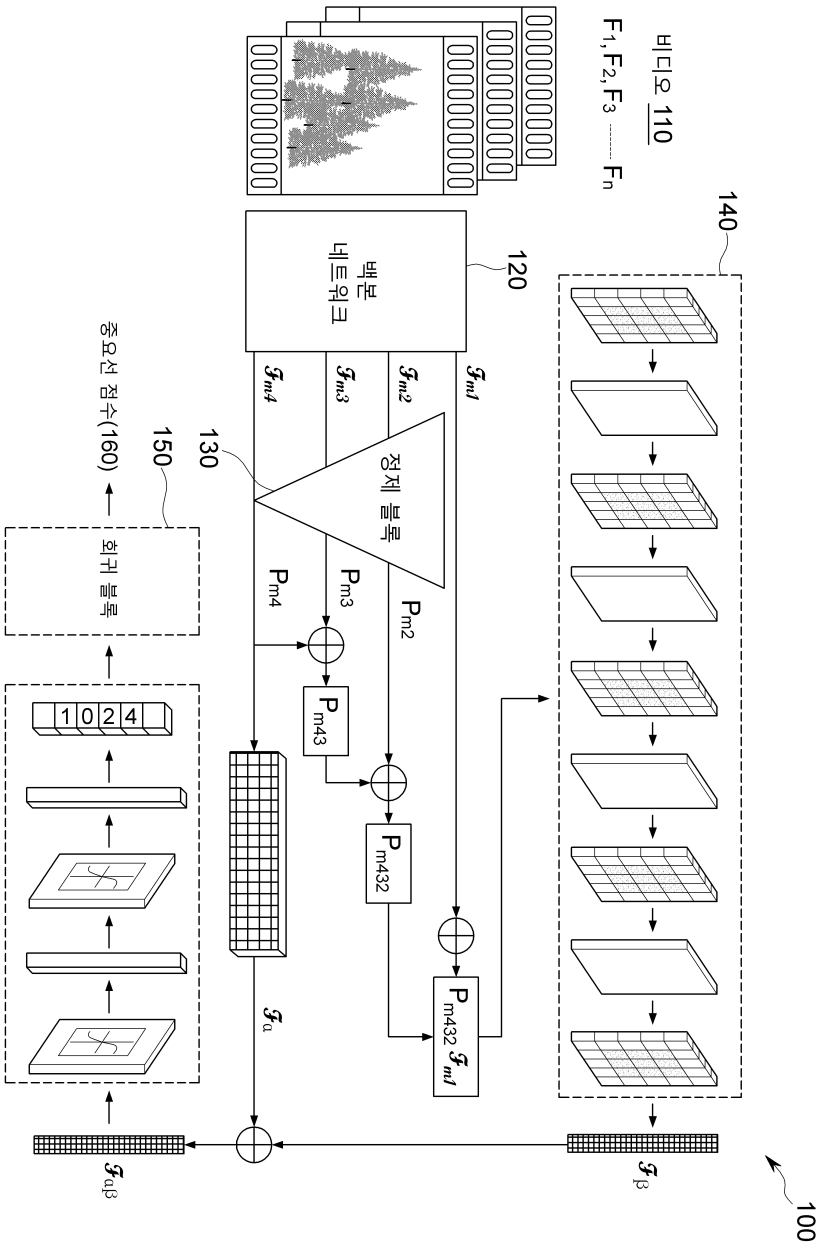
140: 콘볼루션 헤드

150: 회귀 블록

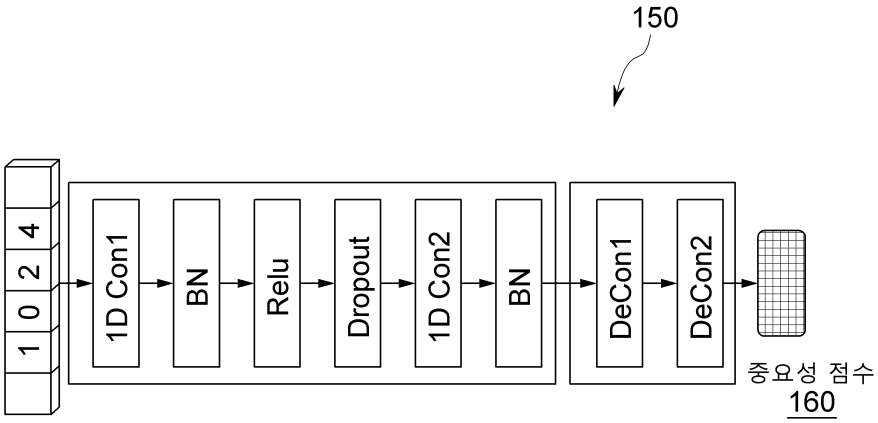
160: 중요성 점수

도면

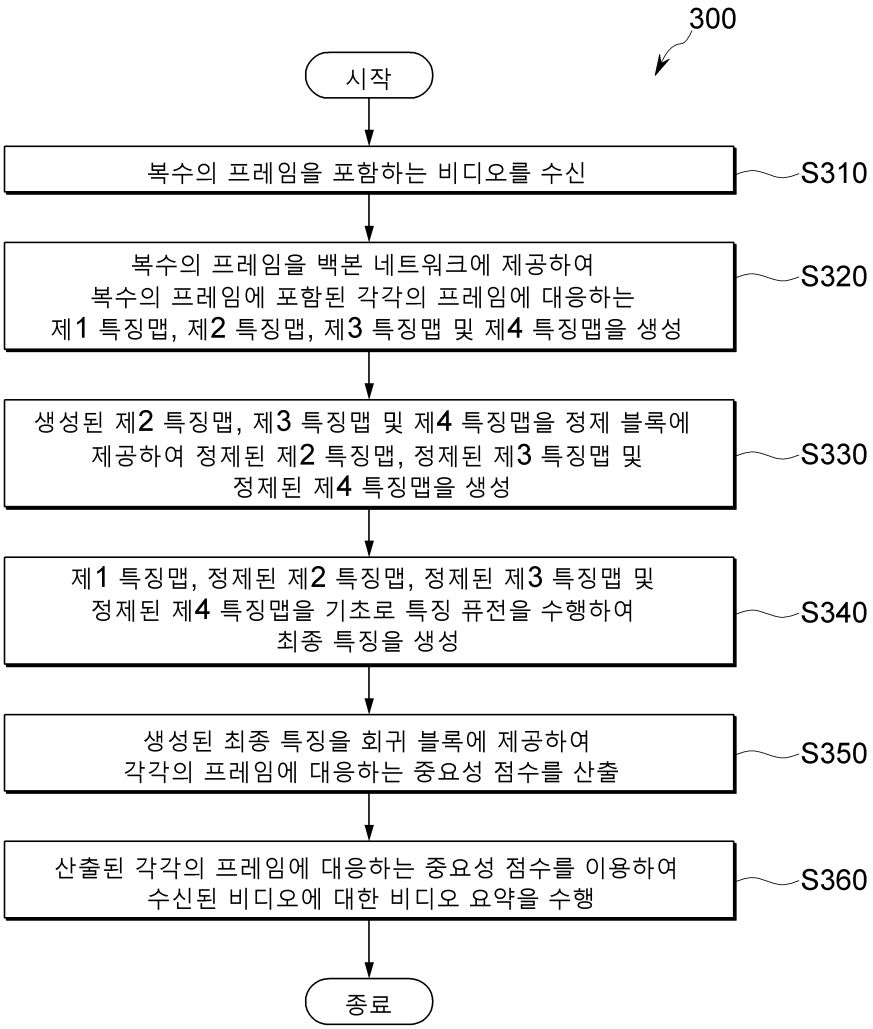
도면1


비디오 110
F1, F2, F3 ..... Fn
백본 네트워크
120
140
130
정제 블록
150
회귀 블록
중요선 점수(160)
1 0 2 4
100

도면2


150
1
0
2
4
1D Con1
BN
Relu
Dropout
1D Con2
BN
DeCon1
DeCon2
중요성 점수
160

도면3


300
시작
복수의 프레임을 포함하는 비디오를 수신
S310
복수의 프레임을 백본 네트워크에 제공하여
복수의 프레임에 포함된 각각의 프레임에 대응하는
제1 특징맵, 제2 특징맵, 제3 특징맵 및 제4 특징맵을 생성
S320
생성된 제2 특징맵, 제3 특징맵 및 제4 특징맵을 정제 블록에
제공하여 정제된 제2 특징맵, 정제된 제3 특징맵 및
정제된 제4 특징맵을 생성
S330
제1 특징맵, 정제된 제2 특징맵, 정제된 제3 특징맵 및
정제된 제4 특징맵을 기초로 특징 퓨전을 수행하여
최종 특징을 생성
S340
생성된 최종 특징을 회귀 블록에 제공하여
각각의 프레임에 대응하는 중요성 점수를 산출
S350
산출된 각각의 프레임에 대응하는 중요성 점수를 이용하여
수신된 비디오에 대한 비디오 요약을 수행
S360
종료

도면4

