(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0011707
(43) 공개일자 2021년02월02일

(51) 국제특허분류(Int. Cl.)
G06K 9/00 (2006.01) *G06K 9/48* (2006.01)
G06K 9/62 (2006.01) *G06N 3/08* (2006.01)
(52) CPC특허분류
G06K 9/00718 (2013.01)
G06K 9/481 (2013.01)
(21) 출원번호 10-2019-0088936
(22) 출원일자 2019년07월23일
심사청구일자 없음

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
낭종호
서울특별시 강남구 언주로146길 18 (논현2동, 동현아파트) 1동 206호
신광수
서울특별시 관악구 신림로14길 27-9, 401호
(74) 대리인
이지연

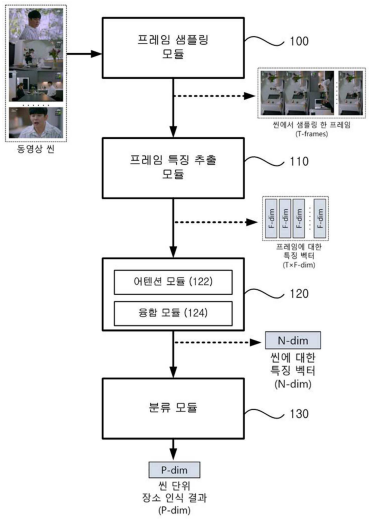
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **CNN을 기반으로 한 동영상의 씬 단위 장소 분류 방법 및 이를 이용한 장소 분류 장치**

(57) 요 약

본 발명은 동영상의 씬 단위 장소 분류 방법 및 장치에 관한 것이다. 상기 동영상의 씬 단위 장소 분류 방법은, 동영상의 각 씬에 대하여, 이진 분류기를 사용하여 각 씬을 구성하는 프레임들로부터 사전 설정된 개수의 프레임들을 샘플링하는 프레임 샘플링 단계; 상기 프레임 샘플링 모듈에 의해 샘플링된 프레임들에 대한 특징 벡터들을 추출하는 특징 추출 단계; 각 씬으로부터 샘플링된 프레임들에 대하여 가중치를 부여하는 어텐션 맵을 생성하고, 이를 이용하여 각 프레임에 대해 강조된 특징벡터를 생성하고, 상기 프레임들에 대한 강조된 특징 벡터들을 융합하여 씬 단위 특징 벡터를 생성하는 씬 단위 특징 벡터 생성 단계; 씬에 대한 씬 단위 특징 벡터를 이용하여 해당 씬에 대한 장소를 분류하는 분류 단계; 를 구비하여, 동영상의 각 씬에 대한 장소를 인식하여 분류한다.

대 표 도 - 도3



(52) CPC특허분류
G06K 9/627 (2013.01)
G06N 3/08 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2017-0-1781
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	혁신성장동력프로젝트(인공지능) 사업
연구과제명	(3세부) 비디오 이해를 위한 데이터 수집 및 보정 자동화 시스템 개발
기 여 율	1/1
과제수행기관명	㈜ 코난테크놀로지
연구기간	2019.01.01 ~ 2019.12.31

명 세 서

청구범위

청구항 1

사전 설정된 프로그램을 실행하는 마이크로프로세서에 의해, 동영상의 각 씬에 대한 장소를 분류하도록 구성된 동영상의 씬단위 장소 분류 방법에 있어서,

동영상의 각 씬에 대하여, 각 씬을 구성하는 프레임들로부터 사전 설정된 개수의 프레임들을 샘플링하는 프레임 샘플링 단계;

상기 프레임 샘플링 모듈에 의해 샘플링된 프레임들에 대한 특징 벡터들을 추출하는 특징 추출 단계;

각 씬으로부터 샘플링된 프레임들로부터 상기 특징 추출 모듈에 의해 추출된 프레임들에 대한 특징 벡터들을 융합하여 씬에 대한 특징 벡터를 생성하는 씬 단위 특징 벡터 생성 단계;

씬에 대한 씬 단위 특징 벡터를 이용하여 해당 씬에 대한 장소를 분류하는 분류 단계;

를 구비하여, 동영상의 각 씬에 대한 장소를 인식하여 분류하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 방법.

청구항 2

제1항에 있어서, 상기 프레임 샘플링 단계는,

사전 학습에 의해 구성된 이진 분류기를 사용하여 장소 포함 확률이 50% 이상인 프레임들을 후보 프레임으로 1차 추출하고,

각 씬으로부터 1차 추출된 후보 프레임들의 개수가 사전 설정된 개수보다 많으면, 장소 포함 확률이 높은 순서에 따라 사전 설정된 개수만큼 프레임들을 최종 샘플링하고,

각 씬으로부터 1차 추출된 후보 프레임들의 개수가 사전 설정된 개수보다 작으면, 상기 후보 프레임들을 반복적으로 추출하여 사전 설정된 개수만큼 프레임들을 최종 샘플링하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 방법.

청구항 3

제1항에 있어서, 상기 특징 추출 단계는,

사전 학습에 의해 구성된 CNN(Convolution Neural Network)을 기반으로 한 분류기를 사용하여, 각 프레임에 대한 장소 인식을 위한 특징을 추출하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 방법.

청구항 4

제1항에 있어서, 상기 씬 단위 특징 벡터 생성 단계는,

각 프레임에 대한 특징을 분석하여 각 특징에 대한 가중치를 부여하고 프레임에 대한 강조된 특징 벡터를 생성하는 어텐션 단계; 및

하나의 씬으로부터 샘플링된 프레임들에 대하여 상기 강조된 특징 벡터들을 융합하여 씬 단위 특징 벡터를 생성하는 융합 단계;를 구비하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 방법.

청구항 5

제4항에 있어서, 상기 어텐션 단계는,

각 프레임들에 대한 특징 벡터들을 구성하는 특징을 이용하여 어텐션 맵(Attention Map)을 생성하고,

어텐션 맵을 구성하는 각 특징에 대한 가중치를 조절하고,

상기 가중치가 조절된 어텐션 맵과 입력된 프레임의 특징 벡터들을 곱하여, 일부 특징들이 강조된 특징 벡터를

생성하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 방법.

청구항 6

제4항에 있어서, 상기 융합 단계는,

풀링 방법, 합성곱 방법, 클러스터링 방법 중 하나를 사용하여, 사전 설정된 개수의 프레임들에 대한 강조된 특징 벡터들을 융합하여 씬 단위 특징 벡터를 생성하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 방법.

청구항 7

마이크로프로세서를 구비하여 동영상의 각 씬에 대한 장소를 분류하고 인식하는 동영상의 씬단위 장소 분류 장치에 있어서,

동영상의 각 씬에 대하여, 각 씬을 구성하는 프레임들로부터 사전 설정된 개수의 프레임들을 샘플링하는 프레임 샘플링 모듈;

상기 프레임 샘플링 모듈에 의해 샘플링된 프레임들에 대한 장소 인식을 위한 특징 벡터들을 추출하는 특징 추출 모듈;

각 씬으로부터 샘플링된 프레임들로부터 상기 특징 추출 모듈에 의해 추출된 프레임들에 대한 특징 벡터들을 융합하여 씬에 대한 특징 벡터를 생성하는 씬 단위 특징 벡터 생성 모듈;

씬에 대한 씬 단위 특징 벡터를 이용하여 해당 씬에 대한 장소를 분류하는 분류 모듈;

을 구비하여, 동영상의 각 씬에 대한 장소를 인식하여 분류하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 장치.

청구항 8

제7항에 있어서, 상기 프레임 샘플링 모듈은,

이진 분류기를 사용하여 장소 포함 확률이 50% 이상인 프레임들을 추출하고,

각 씬으로부터 추출된 프레임이 사전 설정된 개수보다 많으면, 장소 포함 확률이 높은 순서에 따라 사전 설정된 개수의 프레임들을 샘플링하고,

각 씬으로부터 추출된 프레임이 사전 설정된 개수보다 작으면, 추출된 프레임들을 반복적으로 추출하여 사전 설정된 개수의 프레임들을 샘플링하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 장치.

청구항 9

제7항에 있어서, 상기 씬 단위 특징 벡터 생성 모듈은,

프레임에 대한 특징 벡터들에 대하여 각 특징에 대한 가중치를 조정하는 어텐션 모듈; 및

하나의 씬으로부터 샘플링된 프레임들에 대하여 어텐션 모듈에 의해 강조된 특징 벡터들을 융합하여 씬 단위 특징 벡터를 생성하는 융합 모듈;을 구비하고,

는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 장치.

청구항 10

제9항에 있어서, 상기 어텐션 모듈은,

각 프레임들에 대한 특징 벡터들을 구성하는 특징을 이용하여 어텐션 맵(Attention Map)을 생성하고,

학습을 통해 어텐션 맵을 구성하는 특징들에 대한 가중치를 조절하고,

상기 가중치가 조절된 어텐션 맵과 입력된 프레임의 특징 벡터를 곱하여, 일부 특징들이 강조된 특징 벡터를 생성하는 것을 특징으로 하는 동영상의 씬 단위 장소 분류 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 CNN을 기반으로 한 동영상의 씬 단위 장소 분류 방법에 관한 것으로서, 더욱 구체적으로는, 동영상의 각 씬으로부터 동일한 개수의 프레임을 샘플링하고, 샘플링된 프레임들로부터 특징 벡터를 추출하고 이들을 융합시켜 씬 단위 특징 벡터를 생성하고, 씬 단위 특징 벡터를 이용하여 해당 씬의 장소를 보다 정확하게 인식하고 분류하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 스마트폰과 통신 기술의 발달은 동영상에 대한 수요 및 공급을 폭발적으로 증가시켰다. 이러한 동영상 데이터의 증가와 함께 이를 자동으로 처리하고자 하는 수요 또한 늘어났다. 이에 따라, 동영상 내 등장 인물, 장소, 객체, 음성 등을 통해 동영상이 가지고 있는 의미를 분석하고 이를 자동으로 분류하고자 하는 다양한 시도가 있다. 특히 최근 딥러닝 기술의 발달을 통해, 동영상 내 이미지에 대한 장소 인식 연구의 경우 다양한 시도를 통해 높은 성능을 내고 있다.

[0003] 동영상은 프레임(Frame), 샷(Shot) 그리고 씬(Scene) 단위로 구별할 수 있다. 도 1은 동영상에서의 프레임(frame), 샷(Shot), 씬(Scene)에 대한 분류방법을 예시적으로 도시한 것이다. 도 1을 참조하면, 동영상은 카메라를 통해 일정 시간을 두고 연속적으로 취득된 이미지들을 모아 제작되는데, '프레임(frame)'은 동영상에서의 각 이미지들을 의미한다. '샷(Shot)'은 전술한 프레임에 대해 카메라의 녹화가 시작되어 종료될 때까지 연속적으로 촬영한 구간을 의미한다. '씬(Scene)'은 동일한 장소에서 촬영된 샷들의 집합을 의미한다. 이러한 동영상 구분에서는 샷 내 프레임 간에는 유사한 특성을 가지고 있으며, 하나의 씬 내의 샷 간에는 동일한 인물과 동일한 장소에서 이루어지는 장면들이 모여 있음을 알 수 있다. 예를 들면 <그림 1>에서 Shot 1에서 Shot 5까지는 동일한 장소에서 한 남자와 한 여자를 대상으로 촬영되었으며 그 촬영 구도만 변화함을 알 수 있다. 그러므로 이러한 샷들을 모아 Scene 1을 구성할 수 있다. Shot 6의 경우 이전의 샷과는 다르게 다른 장소 및 다른 인물이 나오며, 이야기 또한 이어지지 않는다. 따라서 Shot 5부터는 Scene 1 과 연관이 없으므로 Scene 2로 구분할 수 있다.

[0004] 일반적으로, 동영상의 한 씬에서 장소가 변하는 경우는 매우 드물다. 한 샷에서 인물들의 움직임에 따라 장소가 변하거나 또는 한 씬 에서 인물들 간 전화통화 등의 특별한 경우가 아니라면 한 씬은 주로 동일한 장소에서 이루어진다. 하지만 동영상에서 씬 단위로 장소를 인식하는데 몇 가지 어려움이 있다. 첫 번째로 씬 내부에서 장소를 알 수 없는 샷이 등장하는 문제가 있다. 예를 들어, Scene 1에서 Shot 1, Shot 2 그리고 Shot 5는 장소를 추출하기 쉽지만, Shot 3과 Shot 4는 인물이 크게 나타나므로 그 장소를 쉽게 추정하기 어렵다. 비슷하게 Scene 2의 Shot 7은 특정 물체가 크게 디스플레이되므로 장소를 인식할 수 없다. 두 번째로 동영상 별 씬 길이가 서로 다르며 장소를 추정할 수 있는 샷이 나오는 위치가 다르며, 그 샷의 길이 또한 다르다. 따라서 단순한 알고리즘을 사용하여 장소가 잘 나타나는 샷이나 프레임을 추출하는 것이 불가능하다. 마지막으로, 동영상에서는 같은 장소임에도 제작자의 의도에 따라 다양한 위치에서 다르게 촬영이 이루어진다. 예를 들어 Scene 3에서 동일한 장소에 대해 정면으로 촬영한 Shot 9와 우측 상단에서 촬영한 Shot 10이 연속으로 나오는 것을 볼 수 있다. 따라서 기존의 이미지에 대한 장소 인식기를 그대로 사용하는 경우 그 성능이 좋지 못하다. 이러한 문제점들로 인해 동영상에서 씬 단위로 장소를 인식하는 것은 어려움을 가지고 있으며, 이를 해결하기 위한 새로운 방법의 필요성이 존재한다.

[0005] 장소 인식과 더불어 전통적인 컴퓨터 비전의 다양한 문제를 해결하고자 다양한 연구들이 진행되어왔다. 도 2의 (a)와 (b)는 각각 컴퓨터 비전에서의 전통적인 분석 방법과 딥러닝 기반의 분석 방법의 구조를 도시한 블록도이다.

[0006] 도 2를 참조하면, 컴퓨터 비전(Computer Vision) 기술은 전통적인 방법과 최근 연구가 활발히 진행되고 있는 딥러닝 기반 방법으로 나눌 수 있다. 전통적인 방법은, 도 2의 (a)에 도시된 바와 같이, 입력 데이터에 대해 미리 정해놓은 기법을 사용하여 특징을 추출한다. 이미지는 데이터 셋에 미리 정해놓은 Hand-craft 기반의 방법을 사용하여 특징을 추출하고, 이러한 특징에 따라 분류기를 학습하여 최종 결과를 얻는 과정을 진행한다. 한편, 딥러닝 기반의 연구 방법은, 도 2의 (b)에 도시된 바와 같이, 특징을 추출하는 부분부터 학습을 진행하므로 데이터에 맞춘 특징 추출 방법을 스스로 학습하고, 이런 특징들에 대한 분류기 또한 함께 학습하여 최종 결과를 얻게 된다.

[0007] 전통적인 방법의 이미지에 대한 장소 인식 연구는 HoG나 SIFT 방법을 사용한 Hand-craft 기반 특징을 추출한 후, SVM(Support Vector Machine)등의 분류기를 사용하는 방식으로 분류를 진행했으며 성능을 높이기 위한 다양한 시도가 있어왔다. 전통적인 방법의 이미지에 대한 장소 인식은 주로 이미지를 패치 단위로 나누어 인식하거나, 물체 등의 부가 정보를 통해 장소를 인식하고자 노력하였다. 하지만 이러한 방법들은 장소 인식을 하는데 몇 가지 문제가 있다. 먼저, 이미지를 패치 단위로 나누어 인식하기 때문에, 검출하고자 하는 장소의 특성에 따라 그 성능이 크게 차이가 나게 된다. 예를 들어 사람이 많이 등장하는 장소들은 패치 단위로 나누게 되면, 장소의 특성의 차이가 크게 나타나지 않게 되어 분류를 진행하기 어렵게 된다. 한편, 소개한 물체 등의 부가 정보를 통해 장소 인식하는 경우에는, 해당 장소에 대한 특별한 물체가 자주 등장하지 않게 되면, 그 성능이 크게 떨어진다. 예를 들어, 자연 풍경과 관련된 장소를 인식하려는 경우, 해당 장소만이 가지는 특별한 물체가 없어 그 성능이 크게 떨어질 수 있다. 또한, 이러한 방법들은 장소에 대한 레이블의 수가 늘어남에 따라 유사한 장소가 생기게 되고, 그 결과 장소 인식 성능이 크게 떨어지는 한계점을 가지게 된다. 즉, 유사한 장소에 대해 구분하기 위한 뚜렷한 Hand-craft 기반의 특징을 추출하지 못하는 경우가 있어 그 성능을 보장하기 매우 어렵다는 문제점을 가지고 있다.

[0008] 전술한 문제점은 CNN(Convolutional Neural Network)을 사용한 딥러닝 기반의 방식이 도입되면서 그 성능이 크게 향상되었다. 전통적인 방식과 달리, 딥러닝 기반의 방식은 데이터셋에 맞는 특징을 추출하고 분류기와 함께 학습하는 구조로 되어있다. 따라서 장소 추출에 적합한 특징을 추출할 수 있고, 이에 따른 성능 향상을 기대할 수 있다. CNN을 기반으로 한 VGG, GoogLeNet 그리고 ResNet 등의 다양한 모델로 이미지에 대한 장소 인식을 진행하고자 하는 연구가 활발하게 진행되었다. 이와 함께 Places365 데이터셋이 공개되면서 그 연구가 활발하게 진행되고 있음을 알 수 있다.

[0009] 지금까지 동영상에 대한 장소 인식에 관한 연구는 그리 활발하게 진행되지 못하였다. 이미지에 대한 장소 인식에 대한 성능이 높지 않았으며, 장소 인식을 위한 동영상에 대한 데이터셋이 부족하여 그 연구가 활발히 이루어지지 않았다. 전통적인 방법의 동영상에 대한 장소 인식에 관한 연구는 프레임을 대상으로 이미지에 대한 방식을 사용하여 각각의 프레임에 대해 패치 단위로 나누어 특징을 추출한 후, 이러한 프레임 단위 특징에 대해 누적한 결과에 대해 분류기를 통과하여 장소를 추출하는 방식이 주로 사용되었다. 전통적인 이미지에 대한 장소 인식 방법처럼 비디오 프레임에 대해 패치를 나누어 특징을 추출한 후, 해당 특징들을 프레임 단위 및 이전 프레임들의 정보와 합친 후 최종적으로 분류기를 거쳐 해당 동영상에 대한 장소를 인식하는 형태로 전체 구조를 구성하게 된다.

[0010] 이러한 방식 또한 동영상에 대한 분류에 딥러닝이 활용되면서 큰 변화를 겪게 되었다. 딥러닝을 활용한 동영상 내 행동 분석의 모델을 차용하면서 많은 시도들이 진행되어왔다. 프레임 단위에 대해 이미지 CNN 기반 모델을 활용하여 특징을 추출한 뒤, 뭉치는 작업을 통해 샷 또는 씬 단위 특징을 생성하고 해당 특징에 대해 분류기를 학습하여 장소를 추정하는 방법들이 다양하게 제안되었다.

선행기술문헌

특허문헌

[0011] (특허문헌 0001) 한국등록특허공보 제 10-1826669호

(특허문헌 0002) 한국등록특허공보 제 10-1963558호

(특허문헌 0003) 한국공개특허공보 제 10-2018-0018031호

발명의 내용

해결하려는 과제

[0012] 전술한 문제점들을 해결하기 위한 본 발명의 목적은 어텐션 방법을 사용한 CNN 기반으로 하여 방송 동영상의 씬 단위 장소 분류 방법 및 장치를 제공하는 것이다.

과제의 해결 수단

[0013] 전술한 기술적 과제를 달성하기 위한 본 발명의 제1 특징에 따른 동영상의 씬 단위 장소 분류 방법은, 사전 설

정된 프로그램을 실행하는 마이크로프로세서에 의해, 동영상의 각 씬에 대한 장소를 분류하도록 구성된 동영상의 씬단위 장소 분류 방법에 관한 것으로서, 동영상의 각 씬에 대하여, 각 씬을 구성하는 프레임들로부터 사전 설정된 개수의 프레임들을 샘플링하는 프레임 샘플링 단계; 상기 프레임 샘플링 모듈에 의해 샘플링된 프레임들에 대한 특징 벡터들을 추출하는 특징 추출 단계; 각 씬으로부터 샘플링된 프레임들로부터 상기 특징 추출 모듈에 의해 추출된 프레임들에 대한 특징 벡터들을 융합하여 씬에 대한 특징 벡터를 생성하는 씬 단위 특징 벡터 생성 단계; 씬에 대한 씬 단위 특징 벡터를 이용하여 해당 씬에 대한 장소를 분류하는 분류 단계; 를 구비하여, 동영상의 각 씬에 대한 장소를 인식하여 분류한다.

[0014] 전술한 제1 특징에 따른 동영상의 씬 단위 장소 분류 방법에 있어서, 상기 프레임 샘플링 단계는, 사전 학습에 의해 구성된 이진 분류기를 사용하여 장소 포함 확률이 50% 이상인 프레임들을 후보 프레임으로 1차 추출하고, 각 씬으로부터 1차 추출된 후보 프레임들의 개수가 사전 설정된 개수보다 많으면, 장소 포함 확률이 높은 순서에 따라 사전 설정된 개수만큼 프레임들을 최종 샘플링하고, 각 씬으로부터 1차 추출된 후보 프레임들의 개수가 사전 설정된 개수보다 작으면, 상기 후보 프레임들을 반복적으로 추출하여 사전 설정된 개수만큼 프레임들을 최종 샘플링하는 것이 바람직하다.

[0015] 전술한 제1 특징에 따른 동영상의 씬 단위 장소 분류 방법에 있어서, 상기 특징 추출 단계는, 사전 학습에 의해 구성된 CNN(Convolution Neural Network)을 기반으로 한 분류기를 사용하여, 각 프레임에 대한 장소 인식을 위한 특징을 추출하는 것이 바람직하다.

[0016] 전술한 제1 특징에 따른 동영상의 씬 단위 장소 분류 방법에 있어서, 상기 씬 단위 특징 벡터 생성 단계는, 각 프레임에 대한 특징을 분석하여 각 특징에 대한 가중치를 부여하고 프레임에 대한 강조된 특징 벡터를 생성하는 어텐션 단계; 및 하나의 씬으로부터 샘플링된 프레임들에 대하여 상기 강조된 특징 벡터들을 융합하여 씬 단위 특징 벡터를 생성하는 융합 단계;를 구비하고,

[0017] 상기 어텐션 단계는, 각 프레임들에 대한 특징 벡터들을 구성하는 특징을 이용하여 어텐션 맵(Attention Map)을 생성하고, 어텐션 맵을 구성하는 각 특징들에 대한 가중치를 조절하고, 상기 가중치가 조절된 어텐션 맵과 입력된 프레임의 특징 벡터를 곱하여 프레임이 강조된 특징 벡터를 생성하며,

[0018] 상기 융합 단계는, 풀링 방법, 합성곱 방법, 클러스터링 방법 중 하나를 사용하여, 사전 설정된 개수의 프레임들에 대한 강조된 특징 벡터들을 융합하여 씬 단위 특징 벡터를 생성하는 것이 바람직하다.

발명의 효과

[0019] 본 발명에 따른 씬 단위 장소 분류 장치 및 방법에 있어서, 이진 분류기를 사용하여 씬에 대해 동일한 개수의 프레임을 샘플링하고 씬 단위 특징 벡터를 생성함으로써, 씬의 길이에 상관없이 장소 인식이 가능해질 뿐만 아니라 씬의 장소 등장 위치에 강건한 장소 인식이 가능해진다.

[0020] 본 발명에 따른 씬 단위 장소 분류 장치 및 방법에 있어서, 이진 분류기를 사용하여 각 씬으로부터 사전 설정된 동일한 개수의 프레임들을 샘플링함으로써, 장소가 주로 나타나는 프레임을 위주로 샘플링되어, 장소 인식에 대한 대한 성능이 향상된다.

[0021] 또한, 본 발명에 따른 씬 단위 장소 분류 장치 및 방법에 있어서, 어텐션 방법을 사용함으로서, 샘플링 단계에서 장소가 포함되어 있지 않은 프레임들이 제대로 걸러지지 못하더라도 어텐션 맵을 통해 추가적으로 필터링되어, 장소 인식에 대한 정확도가 향상된다.

도면의 간단한 설명

[0022] 도 1은 동영상에서의 프레임(frame), 샷(Shot), 씬(Scene)에 대한 분류방법을 예시적으로 도시한 것이다.

도 2의 (a)와 (b)는 각각 컴퓨터 비전에서의 전통적인 분석 방법과 딥러닝 기반의 분석 방법의 구조를 도시한 블록도이다.

도 3은 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치의 구성을 개략적으로 도시한 블록도이다.

도 4는 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 프레임 샘플링 모듈(100)의 동작을 설명하기 위하여 도시한 구성도이다.

도 5는 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 특징 추출 모듈(110)의

동작을 설명하기 위하여 도시한 구성도이다.

도 6은 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 씬 단위 특징 벡터 생성 모듈(120)의 어텐션 모듈(122)의 동작을 설명하기 위하여 도시한 구성도이다.

도 7의 (a), (b), (c)는 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 씬 단위 특징 벡터 생성 모듈(120)의 융합 모듈(124)의 융합 방법들을 설명하기 위하여 도시한 구성도들이다.

도 8은 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 분류 모듈(130)의 동작을 설명하기 위하여 도시한 구성도들이다.

발명을 실시하기 위한 구체적인 내용

[0023] 본 발명에 따른 동영상의 씬 단위 장소 분류 장치 및 방법은, 동영상의 각 씬에 대하여, 장소 포함 확률이 높은 프레임들을 사전 설정 개수만큼 샘플링하고, 학습을 통해 샘플링된 프레임들에 대한 특징 벡터를 생성하고, 학습을 통해 프레임들에 대한 특징 벡터들의 특징을 강조시킨 후 융합하여 씬 단위 특징 벡터를 생성하고, 상기 씬 단위 특징 벡터에 대한 전결합 레이터를 사용한 후 각 장소의 레이블에 대한 확률값을 추출하고, 이를 이용하여 각 씬에 대한 장소를 인식하는 것을 특징으로 한다.

[0024] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치 및 분류 방법에 대하여 구체적으로 설명한다.

[0025] 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치는 씬 단위 장소 분류 방법을 구현하기 위한 프로그램이 저장된 마이크로프로세서를 포함하는 장치 또는 마이크로컴퓨터 등으로 구성될 수 있다. 이하, 본 발명의 바람직한 실시예에 따라 상기 프로그램에 의해 구현되는 동영상의 씬단위 장소 분류 방법에 대하여 설명한다.

[0026] 도 3은 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치의 구성을 개략적으로 도시한 블록도이다.

[0027] 도 3을 참조하면, 본 발명에 따른 동영상의 씬 단위 장소 분류 장치(1)는 프레임 샘플링 모듈(100), 프레임 특징 추출 모듈(110), 씬단위 특징벡터 생성 모듈(120), 분류 모듈(130)을 포함한다.

[0028] 상기 프레임 샘플링 모듈(100)은, 동영상의 각 씬에 대하여, 각 씬을 구성하는 프레임들로부터 사전 설정된 개수의 프레임들을 샘플링한다. 상기 프레임 특징 추출 모듈(110)은 상기 프레임 샘플링 모듈에 의해 샘플링된 프레임들에 대한 특징 벡터들을 추출한다. 상기 씬단위 특징벡터 생성 모듈(120)은 각 씬으로부터 샘플링된 프레임들로부터 상기 특징 추출 모듈에 의해 추출된 프레임들에 대한 특징 벡터들을 융합하여 씬에 대한 특징 벡터를 생성한다. 상기 분류 모듈(130) 씬에 대한 씬 단위 특징 벡터를 이용하여 해당 씬에 대한 장소를 분류한다. 전술한 구성들에 의하여, 본 발명에 따른 동영상의 씬 단위 장소 분류 장치는 우수한 성능으로 동영상의 각 씬에 대한 장소를 인식하여 분류할 수 있게 된다. 이하, 각 모듈들의 동작에 대하여 보다 구체적으로 설명한다.

[0029] 상기 프레임 샘플링 모듈(100)은 이진 분류기를 사용하여 장소가 잘 나타나는 이미지를 분류하고, 분류된 이미지들 중 사전 설정된 개수만큼 샘플링하는 것을 특징으로 한다.

[0030] 도 4는 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 프레임 샘플링 모듈(100)의 동작을 설명하기 위하여 도시한 구성도이다. 도 4를 참조하면, 상기 프레임 샘플링 모듈(100)은, 사전 학습에 의해 구성된 이진 분류기를 사용하여 해당 프레임의 이미지에 대하여 장소 포함 여부에 대한 확률값을 구하고, 해당 프레임의 이미지가 장소 포함 확률이 50% 이상인 프레임들을 1차 추출하여 후보 프레임들을 샘플링한다. 여기서, 프레임에 대한 장소 포함 확률이 50% 이상인 경우, 해당 프레임은 장소를 파악하기 쉬운 이미지이며, 그 이하인 경우 해당 프레임은 장소를 파악하기 어려운 이미지이다. 다음, 각 씬에 대하여 상기 후보 프레임들로부터 고정된 개수(T개)의 프레임들을 최종 샘플링한다. 이때, 각 씬으로부터 추출된 후보 프레임들의 개수가 사전 설정된 개수보다 많으면, 장소 포함 확률이 높은 순서에 따라 사전 설정된 개수의 프레임들을 최종 샘플링한다. 반면에, 각 씬으로부터 추출된 후보 프레임들의 개수가 사전 설정된 개수보다 작으면, 추출된 후보 프레임들을 반복적으로 추출하여 사전 설정된 개수의 프레임들을 샘플링함으로써, 장소를 파악하기 어려운 이미지는 샘플링되지 못하도록 하는 것이 바람직하다.

[0031] 상기 이진 분류기는 프레임 샘플링시에 해당 프레임의 이미지에 장소 포함 여부를 판단하기 위한 분류기로서, 장소에 대해 인식이 가능한 이미지와 장소에 대한 분류가 불가능한 이미지에 대하여 학습을 진행함으로써 구성

된다. 이렇게 학습에 의해 구성된 이진 분류기는, 한 장의 이미지가 입력되면 해당 이미지에 대한 장소가 포함될 확률을 계산하여 출력하게 되는데, 장소가 잘 드러나는 이미지에 대하여는 확률값이 1에 가까운 값을 출력하고 장소가 잘 드러나지 않는 이미지에 대하여는 확률값이 0에 가까운 값을 출력하게 된다. 따라서, 본 발명에 따른 프레임 샘플링 모듈은 이진 분류기에 의해 장소 포함 확률값이 0.5이상인 프레임들, 즉 장소 포함 확률이 50% 이상인 프레임들을 1차 샘플링하게 된다.

[0032] 상기 특징 추출 모듈은, ResNet-50 기반의 네트워크에 대하여 학습용 데이터셋을 이용하여 주어진 이미지에 대해 장소를 인식하기 위한 특징을 추출하도록 사전 학습시킴으로써, 프레임 단위 특징 추출에 가장 적합한 네트워크를 사용하게 된다. 따라서, 상기 특징 추출 모듈(110)은, 각 프레임의 이미지로부터, CNN(Convolution Neural Network)을 기반으로 한 분류기를 사용하여, 각 프레임에 대한 장소 인식을 위한 특징을 추출하게 된다. 상기 CNN은 장소 인식을 위한 최적의 특징을 뽑아 분류에 사용하기 때문에, 이러한 점을 활용하여 동영상 콘텐츠의 프레임 단위 특징을 추출하기가 용이하다.

[0033] 도 5는 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 특징 추출 모듈(110)의 동작을 설명하기 위하여 도시한 구성도이다. 도 5를 참조하면, 본 발명의 일실시형태에 따른 특징 추출 모듈은 Places365 데이터셋에 대하여 학습된 ResNet-50 기반 네트워크를 사용하여 전결합 레이어 직전에서의 2,048 차원의 벡터를 프레임 한 장에 대한 특징으로 사용하게 된다.

[0034] 상기 씬 단위 특징 벡터 생성 모듈(120)은 사전 설정된 개수(T)개의 프레임들에서 추출된 특징 벡터들을 융합하여 단일의 씬 단위 특징 벡터를 생성하게 된다. 상기 씬 단위 특징 벡터 생성 모듈(120)은, 프레임에 대한 특징 벡터들에 대하여 각 특징에 따른 비율을 조정하는 어텐션 모듈(122); 및 하나의 씬으로부터 샘플링된 프레임들에 대하여 상기 강조된 특징 벡터들을 융합하여 씬 단위 특징 벡터를 생성하는 융합 모듈(124);을 구비한다.

[0035] 상기 어텐션 모듈(122)은, 프레임별 특징을 분석하여 각 특징에 따른 비율을 조정함으로써, 장소 추출에 더 적합한 특징을 강조하여 분류 성능을 높이기 위한 것이다. 상기 어텐션 모듈은, 씬 내에서 장소가 잘 나타나는 특징을 가진 프레임에 대해서는 높은 가중치를 주어 강조하고, 인물, 물체, 영상 편집 등에 의해 장소가 잘 나타나지 않는 특징을 가진 프레임에 대해서는 낮은 가중치를 주어 반영 비율을 줄임으로써, 프레임 샘플링 단계에서 걸러지지 않은 프레임에 대해 한번 더 필터링하는 효과를 가지게 된다.

[0036] 도 6은 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 씬 단위 특징 벡터 생성 모듈(120)의 어텐션 모듈(122)의 동작을 설명하기 위하여 도시한 구성도이다. 도 6을 참조하면, 전술한 기능을 위하여 상기 어텐션 모듈은, 임시적인 어텐션 맵(Attention Map)을 생성하고, 각 프레임들에 대한 특징 벡터들을 구성하는 특징을 분석하여 어텐션 맵(Attention Map)에 대한 가중치를 조절하고, 상기 가중치가 조절된 어텐션 맵과 입력된 프레임의 특징 벡터를 곱하여, 장소가 잘 나타내는 프레임들이 강조된 특징 벡터들을 생성하게 된다.

[0037] 상기 어텐션 모듈의 동작을 보다 구체적으로 설명하면, 먼저 T개의 F차원을 갖는 특징에 대하여 각 특징 별로 CNN을 사용하여 각각의 F 차원을 1차원으로 축소하며, 이 과정을 통해 씬에서 추출한 T×F의 특징은 T×1로 간단히 표시된다. 이렇게 나온 T×1 차원의 벡터에 대하여 Softmax를 진행하여 T×1 차원에 대하여 각각의 가중치를 부여하게 된다. 이러한 가중치를 어텐션 맵(Attention Map)이라고 한다. 이러한 과정을 통해 얻게 된 T×1 차원의 어텐션 맵을 T×F 차원인 프레임에 대한 특징에 대해 요소별 곱셈을 진행하게 되고, 이를 통해 각 특징별 분석을 통해 가중치를 부여하여 장소 인식이 효과적으로 진행되도록 구성하게 된다 .

[0038] 상기 융합 모듈(124)은 상기 어텐션 모듈에 의해 강조된 프레임 단위의 특징 벡터들에 대하여 풀링 방법, 합성곱 방법, 클러스터링 방법 중 하나를 사용하여, 사전 설정된 개수의 프레임들에 대한 강조된 특징 벡터들을 융합하여 씬 단위 특징 벡터를 생성한다.

[0039] 도 7은 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 씬 단위 특징 벡터 생성 모듈(120)의 융합 모듈(124)의 융합 방법들을 설명하기 위하여 도시한 구성도들이다. 도 7의 (a)를 참조하면, 풀링 방법은 평균값과 최대값에 대한 풀링으로 각각 진행하는 것으로서, 프레임 단위로 추출된 T×F 차원인 특징들에 대하여 T를 축으로 평균값과 최대값 풀링(Pooling)을 사용하여 씬에 대한 특징벡터를 계산하게 된다.

[0040] 도 7의 (b)를 참조하면, 합성곱 방법은 T 차원 축에 대하여 Convolution layer를 더 쌓아 특징을 압축시킴으로써, 씬에 대한 특징 벡터를 만드는 것이다. 상기 합성곱 방법은 전술한 풀링 방법과 유사한 방법을 사용하되, 학습이 가능하도록 합성곱 레이어를 구성하여 T×F 차원인 프레임에 대한 특징 벡터에 대해 T를 축으로 하여 합성곱 레이어를 쌓았으며, 이 결과로 나오는 F 차원의 벡터를 씬에 대한 특징 벡터로 사용하게 된다.

[0041] 도 7의 (c)를 참조하면 클러스터링 방법은 학습을 통해 클러스터링을 진행하는 NetVLAD 방식을 사용하여 특징들의 분포를 통해 씬에 대한 특징벡터를 구하는 것이다. 이는 T×F 차원인 특징들의 분포를 보고 해당 특징들을 가장 잘 나타낼 수 있는 대표 특징을 계산하여 K 차원의 벡터로 나타내고, 해당 차원에 대해 전결합 레이어를 사용하여 필요한 정보만을 남기고 추상화하여 N 차원의 벡터를 씬에 대한 특징 벡터로 사용하게 된다.

[0042] 상기 분류 모듈(130)은, 분류기로서 SVM, 의사 결정 트리(Decision Tree), 전결합 레이어(Fully Connected Layer) 등을 사용할 수 있으며, 본 명세서에서는 특히 전결합 레이어를 사용하여 각 레이블에 대한 분류를 진행하는 것을 설명한다. 상기 전결합 레이어는 이전 레이어들의 모든 출력이 입력으로 구성되어 값을 계산하는 방식으로 구성된다.

[0043] 도 8은 본 발명의 바람직한 실시예에 따른 동영상의 씬 단위 장소 분류 장치에 있어서, 분류 모듈(130)의 동작을 설명하기 위하여 도시한 구성도들이다. 도 8을 참조하면, 본 발명에 따른 분류 모듈은 씬 단위 특징 벡터에 대하여 전결합 레이어를 세 개를 쌓아 씬에 대한 특징 벡터 N 차원을 씬 단위 장소 인식 결과인 P차원 벡터로 구성함으로써, 전결합 레이어를 통한 분류 성능을 높이게 된다. 상기 분류 모듈은 전결합 레이어후, 소프트맥스(Softmax)를 사용하여 사전 설정된 개수의 각 장소의 레이블에 대한 확률값을 추출하게 된다. 상기 분류 모듈은 전술한 과정을 통해 얻게 된 확률값이 가장 큰 레이블이 해당 씬에 대한 장소로 분류하여 인식하고, 인식된 장소 레이블을 출력한다.

[0044] 이상에서 본 발명에 대하여 그 바람직한 실시예를 중심으로 설명하였으나, 이는 단지 예시일 뿐 본 발명을 한정하는 것이 아니며, 본 발명이 속하는 분야의 통상의 지식을 가진 자라면 본 발명의 본질적인 특성을 벗어나지 않는 범위에서 이상에 예시되지 않은 여러 가지의 변형과 응용이 가능함을 알 수 있을 것이다. 그리고, 이러한 변형과 응용에 관계된 차이점들은 첨부된 청구 범위에서 규정하는 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0045] 1 : 씬 단위 장소 분류 장치

100 : 프레임 샘플링 모듈

110 : 프레임 특징 추출 모듈

120 : 씬 단위 특징벡터 생성 모듈

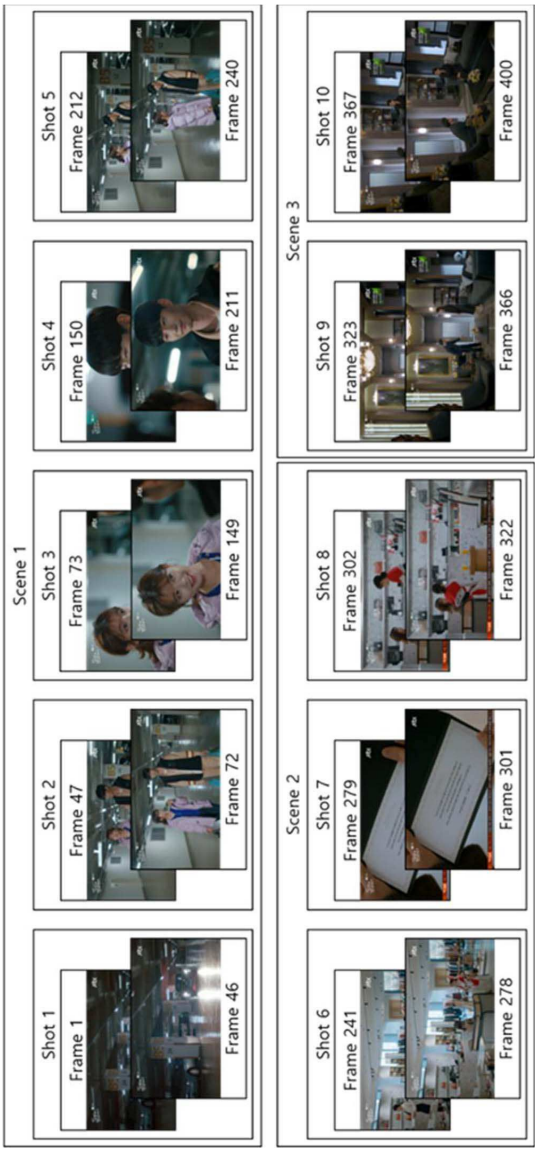
122 : 어텐션 모듈

124 : 융합 모듈

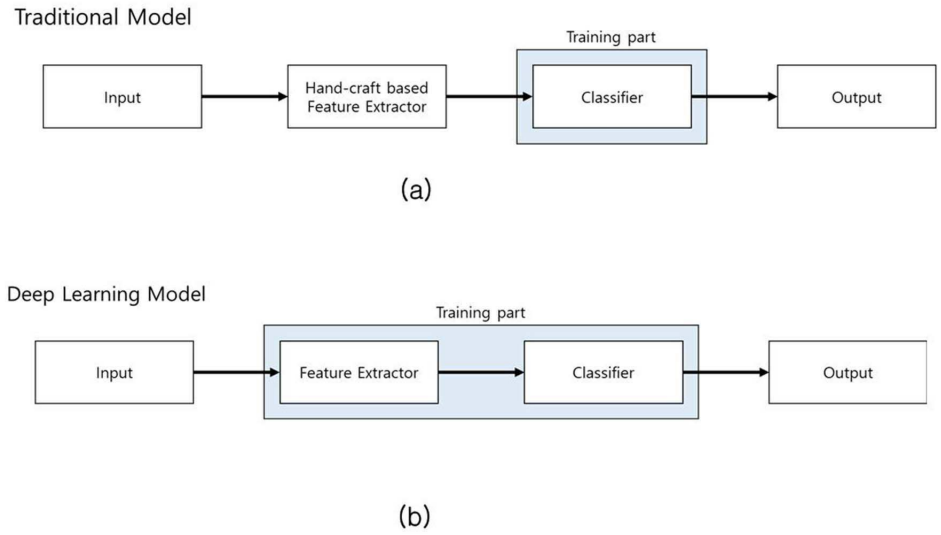
130 : 분류 모듈

도면

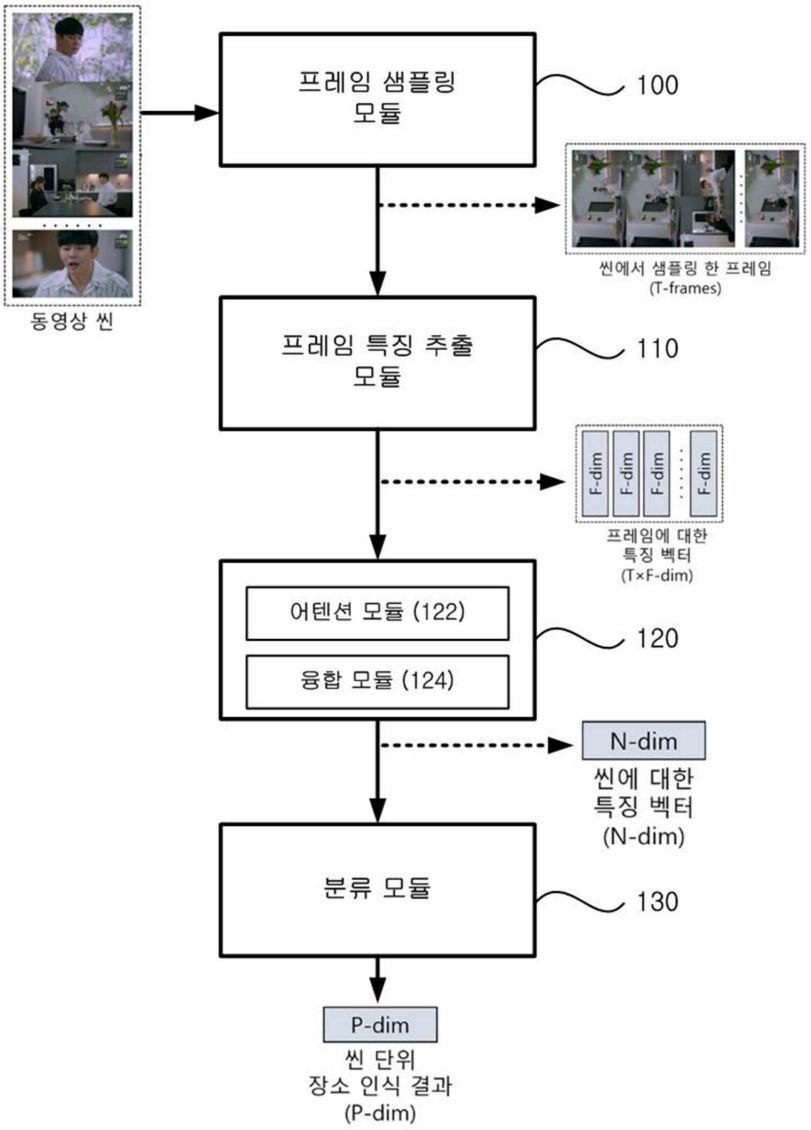
도면1


Scene 1
Shot 1
Frame 1
Frame 46
Shot 2
Frame 47
Frame 72
Shot 3
Frame 73
Frame 149
Shot 4
Frame 150
Frame 211
Shot 5
Frame 212
Frame 240
Scene 2
Shot 6
Frame 241
Frame 278
Shot 7
Frame 279
Frame 301
Shot 8
Frame 302
Frame 322
Scene 3
Shot 9
Frame 323
Frame 366
Shot 10
Frame 367
Frame 400

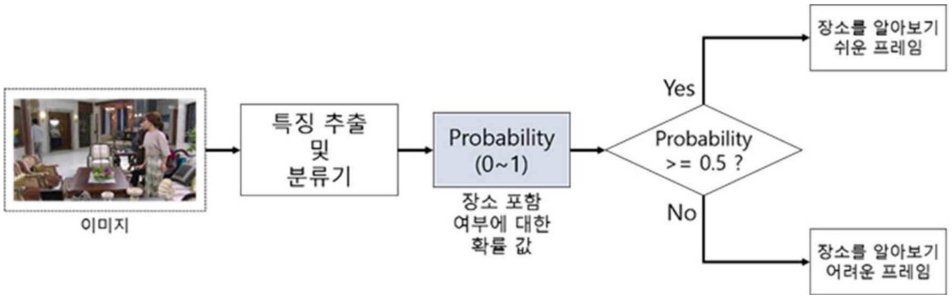
도면2


Traditional Model
Input
Hand-craft based Feature Extractor
Training part
Classifier
Output
(a)
Deep Learning Model
Input
Training part
Feature Extractor
Classifier
Output
(b)

도면3


동영상 씬
프레임 샘플링 모듈
100
씬에서 샘플링 한 프레임 (T-frames)
프레임 특징 추출 모듈
110
F-dim
프레임에 대한 특징 벡터 (T×F-dim)
어텐션 모듈 (122)
융합 모듈 (124)
120
N-dim
씬에 대한 특징 벡터 (N-dim)
분류 모듈
130
P-dim
씬 단위 장소 인식 결과 (P-dim)

도면4


이미지
특징 추출
및
분류기
Probability
(0~1)
장소 포함
여부에 대한
확률 값
Probability
>= 0.5 ?
Yes
No
장소를 알아보기
쉬운 프레임
장소를 알아보기
어려운 프레임

도면5

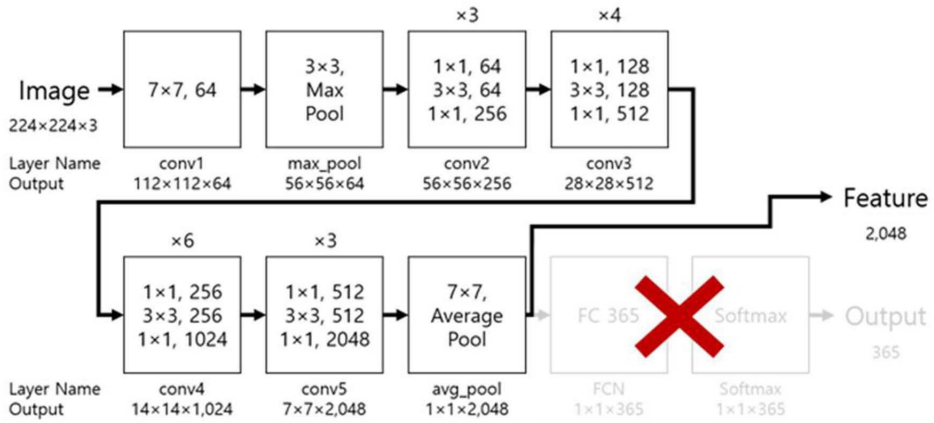
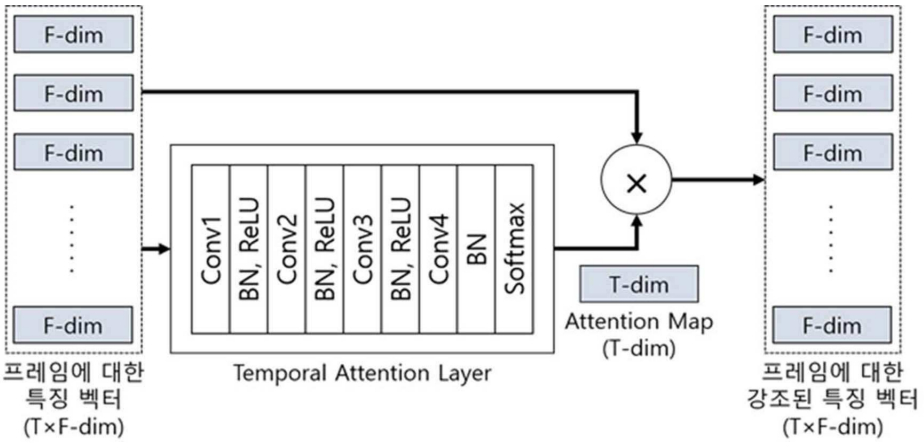
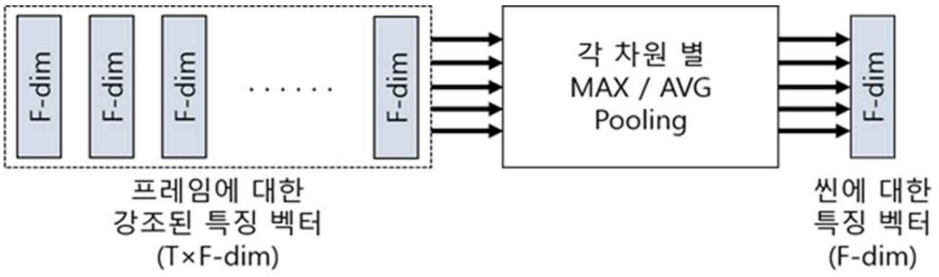


Image
224×224×3
Layer Name
Output
7×7, 64
conv1
112×112×64
3×3,
Max
Pool
max_pool
56×56×64
×3
1×1, 64
3×3, 64
1×1, 256
conv2
56×56×256
×4
1×1, 128
3×3, 128
1×1, 512
conv3
28×28×512
Feature
2,048
×6
1×1, 256
3×3, 256
1×1, 1024
conv4
14×14×1,024
×3
1×1, 512
3×3, 512
1×1, 2048
conv5
7×7×2,048
7×7,
Average
Pool
avg_pool
1×1×2,048
FC 365
FCN
1×1×365
Softmax
Softmax
1×1×365
Output
365

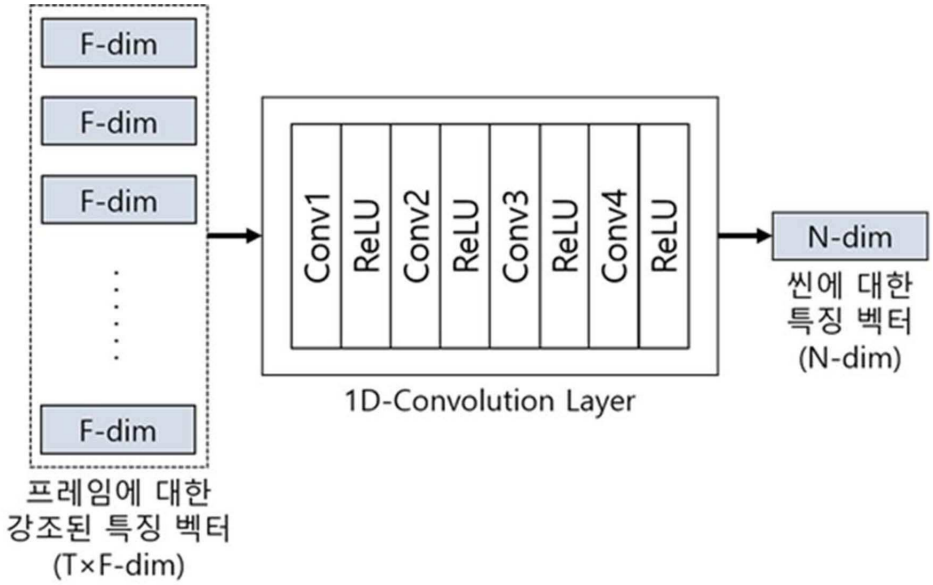
도면6


F-dim
F-dim
F-dim
F-dim
프레임에 대한
특징 벡터
(T×F-dim)
Conv1
BN, ReLU
Conv2
BN, ReLU
Conv3
BN, ReLU
Conv4
BN
Softmax
Temporal Attention Layer
T-dim
Attention Map
(T-dim)
F-dim
F-dim
F-dim
F-dim
프레임에 대한
강조된 특징 벡터
(T×F-dim)

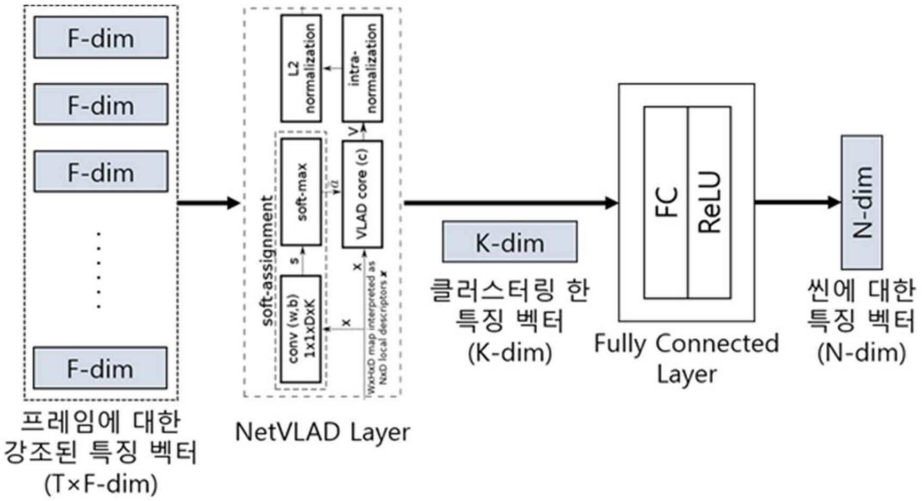
도면7a



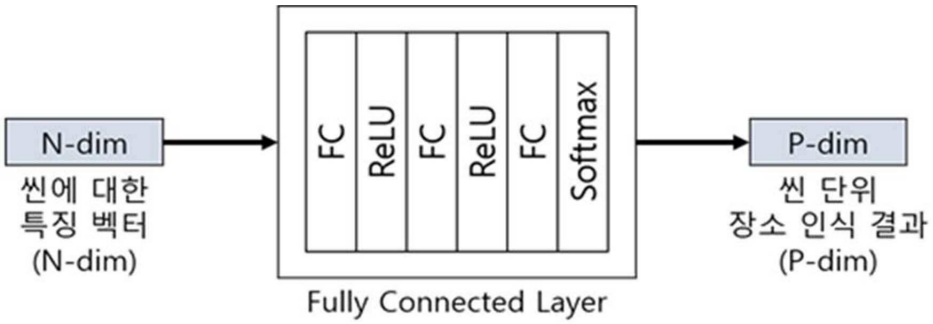
도면7b



도면7c



도면8


N-dim
씬에 대한
특징 벡터
(N-dim)
FC
ReLU
FC
ReLU
FC
Softmax
Fully Connected Layer
P-dim
씬 단위
장소 인식 결과
(P-dim)