



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0101959
(43) 공개일자 2018년09월14일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01) G06K 9/00 (2006.01)
G06K 9/48 (2006.01) G06K 9/62 (2006.01)
G06N 3/08 (2006.01) G06N 99/00 (2010.01)

(52) CPC특허분류
G06F 17/30784 (2013.01)
G06K 9/00744 (2013.01)

(21) 출원번호 10-2017-0028561

(22) 출원일자 2017년03월06일

심사청구일자 2017년03월06일

(71) 출원인

서강대학교산학협력단

서울특별시 마포구 백범로 35 (신수동, 서강대학교)

(72) 발명자

양지훈

서울특별시 은평구 백련산로2길 19(응암제2동, 백련산 힐스테이트 1차 105동 603호)

이정현

서울특별시 마포구 백범로 35 서강대학교 컴퓨터공학과

(74) 대리인

이지연

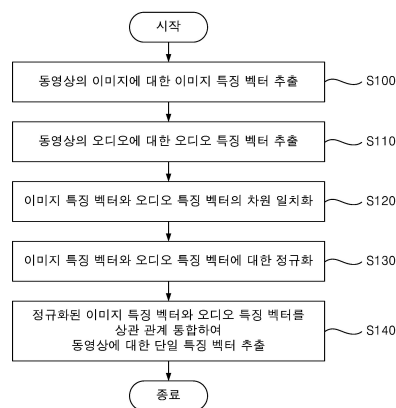
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 멀티 모달의 상관관계를 이용한 동영상 특징 벡터 추출 방법 및 시스템

(57) 요약

본 발명은 동영상의 특징 벡터 추출 방법 및 시스템에 관한 것이다. 상기 동영상의 특징 벡터 추출 방법 및 시스템은, 동영상으로부터 이미지 및 오디오를 추출하고, 상기 이미지에 대한 p 차원의 이미지 특징 벡터를 추출하고 상기 오디오에 대한 q 차원의 오디오 특징 벡터를 추출하고, 이들의 차원을 d 차원으로 일치시키고 단위 벡터로 각각 정규화시킨 후, 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대하여 상관 관계 통합(correlation pooling)을 수행하여 상기 동영상을 대표하는 단일의 특징 벡터를 추출하는 것을 특징으로 한다.

대표도 - 도1



(52) CPC특허분류

G06K 9/481 (2013.01)

G06K 9/6267 (2013.01)

G06N 3/084 (2013.01)

G06N 99/005 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 201639008

부처명 미래창조과학부

연구관리전문기관 정보통신기술진흥센터

연구사업명 정보통신방송 기술개발사업

연구과제명 퍼스널 미디어가 연결공유결합하여 재구성 가능케 하는 복합 모달리티 기반 미디어 응용 프레임워크 개발

기 여 율 50/100

주관기관 한국전자통신연구원

연구기간 2016.03.01 ~ 2017.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 201639001

부처명 미래창조과학부

연구관리전문기관 정보통신기술진흥센터

연구사업명 대학ICT연구센터육성사업

연구과제명 글로벌 핀테크 벤처 기업 및 창의적 인재 육성을 위한 스마트 핀테크 플랫폼 기술개발

기 여 율 50/100

주관기관 서강대학교 산학협력단

연구기간 2016.01.01 ~ 2016.12.31

명세서

청구범위

청구항 1

이미지와 오디오로 이루어지는 동영상의 특징 벡터 추출 방법에 있어서,

- (a) 상기 동영상의 이미지에 대한 이미지 특징 벡터를 추출하는 단계;
- (b) 상기 동영상의 오디오에 대한 오디오 특징 벡터를 추출하는 단계;
- (c) 상기 이미지 특징 벡터 및 상기 오디오 특징 벡터를 각각 단위 벡터를 이용하여 정규화시키는 단계;
- (d) 정규화된 이미지 특징 벡터와 정규화된 오디오 특징 벡터를 통합하여 상기 동영상에 대한 특징 벡터를 생성하는 단계;

를 구비하여 동영상을 대표하는 단일의 특징 벡터를 추출하는 동영상의 특징 벡터 추출 방법.

청구항 2

제1항에 있어서, 상기 (d) 단계는, 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 상관 계수를 추출하고, 상기 상관 계수를 가중값으로 이용하여 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터를 상관 관계 통합하여 상기 동영상에 대한 특징 벡터를 생성하는 것을 특징으로 하는 동영상의 특징 벡터 추출 방법.

청구항 3

제2항에 있어서, 상기 상관 계수는 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 피어슨 상관 계수인 것을 특징으로 하는 동영상의 특징 벡터 추출 방법.

청구항 4

제2항에 있어서, 상기 (d) 단계는 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대하여 평균 통합을 하여 상기 동영상에 대한 특징 벡터를 생성하되, 상기 상관 계수를 가중값으로 사용하는 것을 특징으로 하는 동영상의 특징 벡터 추출 방법.

청구항 5

제1항에 있어서, 상기 (c) 단계는

단일 계층 신경망을 사용하여 상기 (a) 단계에서 추출된 이미지 특징 벡터의 차원과 상기 (b) 단계에서 추출된 오디오 특징 벡터의 차원을 일치시키고, 서로 일치된 차원을 갖는 이미지 특징 벡터와 오디오 특징 벡터를 단위 벡터를 이용하여 정규화시키는 것을 특징으로 하며,

상기 단위 벡터는 이미지 특징 벡터와 오디오 특징 벡터의 속성은 그대로 유지하면서 크기가 1인 벡터인 것을 특징으로 하는 동영상의 특징 벡터 추출 방법.

청구항 6

제1항 내지 제5항 중 어느 한 항에 따른 동영상의 특징 벡터 추출 방법에 의해 추출된 동영상을 대표하는 단일의 특징 벡터를 이용하여 동영상에 대한 이벤트를 분류하는 것을 특징으로 하는 동영상 분류 방법.

청구항 7

동영상으로부터 이미지와 오디오를 각각 추출하는 이미지/오디오 추출 모듈;

상기 이미지/오디오 추출 모듈로부터 추출된 이미지에 대한 이미지 특징 벡터를 추출하는 이미지 특징 벡터 추출 모듈;

상기 이미지/오디오 추출 모듈로부터 추출된 오디오에 대한 오디오 특징 벡터를 추출하는 오디오 특징 벡터 추

출 모듈;

단일 계층 신경망을 이용하여 상기 이미지 특징 벡터 추출 모듈에 의해 추출된 이미지 특징 벡터의 차원과 상기 오디오 특징 벡터 추출 모듈에 의해 추출된 오디오 특징 벡터의 차원을 서로 일치시키는 차원 일치 모듈;

상기 차원 일치 모듈에 의해 차원이 일치된 이미지 특징 벡터 및 오디오 특징 벡터를 단위 벡터를 이용하여 각각 정규화시키는 정규화 모듈;

정규화된 이미지 특징 벡터와 오디오 특징 벡터를 통합시켜 동영상상을 대표하는 하나의 특징 벡터를 추출하는 벡터 통합 모듈;

을 구비하여, 동영상상을 대표하는 단일을 특징 벡터를 추출하여 제공하는 동영상상의 특징 벡터 추출 시스템.

청구항 8

제7항에 있어서, 상기 벡터 통합 모듈은, 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 상관 계수를 추출하고, 상기 상관 계수를 가중값으로 이용하여 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터를 상관 관계 통합하여 상기 동영상상에 대한 특징 벡터를 생성하는 것을 특징으로 하는 동영상상의 특징 벡터 추출 시스템.

청구항 9

제8항에 있어서, 상기 상관 계수는 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 피어슨 상관 계수인 것을 특징으로 하는 동영상상의 특징 벡터 추출 시스템.

청구항 10

제8항에 있어서, 상기 벡터 통합 모듈은, 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대하여 평균 통합을 하여 상기 동영상상에 대한 특징 벡터를 생성하되, 상기 상관 계수를 가중값으로 사용하는 것을 특징으로 하는 동영상상의 특징 벡터 추출 시스템.

청구항 11

제7항 내지 제10항 중 어느 한 항에 따른 동영상상의 특징 벡터 추출 시스템에 의해 추출된 동영상상을 대표하는 단일의 특징 벡터를 이용하여 동영상상에 대한 이벤트를 분류하는 것을 특징으로 하는 동영상 분류 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 동영상 특징 벡터 추출 방법 및 시스템에 관한 것으로서, 더욱 구체적으로는 멀티 모달의 상관 관계를 이용하여 동영상상을 대표하는 단일의 특징 벡터를 추출하는 방법 및 시스템에 관한 것이다.

배경 기술

[0002] 다양한 기계학습 알고리즘들이 연구되고 성능이 향상됨에 따라 인공지능의 최종 목표인 인간 수준의 인공지능 실현을 위해서는 인간 수준의 기계학습 기술을 개발하기 위해 다양한 연구가 진행 중이다. 인간은 데이터를 받아들일 때 촉각, 시각, 후각, 미각, 청각 다섯 개의 감각을 통해서 받아들인다. 이에 인간 수준이란 인간과 같이 다양한 종류의 데이터를 받아들이고 학습할 수 있어야 하며 새로운 데이터에 대해서도 적절한 판단이 진행될 수 있어야 한다. 지금 대부분의 기계학습은 한 종류만의 제한된 입력을 받는 단일 모달리티(uni-modality)로 구성되어 있다.

[0003] 단일 모달리티의 경우는 인간을 뛰어넘는 경우가 있다. 하지만 실제 인간이 사는 환경에서는 단 하나만의 정제된 모달리티가 입력되지 않으며 다양한 종류의 멀티 모달리티(multi-modality)를 인식하고 이를 종합하여 판단을 진행해야 한다. 실생활에 더욱 인간과 같은 인공지능 적용을 위해서는 다양한 종류의 멀티 모달리티를 인식하고 학습할 수 있어야 한다. 이에 더욱 사람과 같이 다양한 종류의 데이터를 같이 받아들임으로써 유연한 학습이 가능하고 더 나은 성능을 얻을 수 있도록 하는 멀티 모달리티 기반의 기계학습 연구들이 진행되고 있다.

- [0004] 멀티 모달리티 기반의 기계학습 연구로 대표되는 것은 동영상 분류 시스템이다. 동영상은 이미지, 오디오, 텍스트 등 다양한 종류의 모달리티를 가지는 데이터라고 할 수 있기 때문이다. 하지만 대부분의 동영상 분류 시스템은 이미지 하나만을 가지고 동영상을 분류하는 경우가 많다. 물론 동영상에서 시각적인 정보가 대부분을 차지하지만, 동영상에서의 모든 이미지가 동영상의 주제에 맞지는 않을 것이다. 동영상에서 주제와는 무관한 이미지가 출력될 때 음성이나 텍스트 등 다른 모달리티가 동영상의 주제와 관련있다면, 이미지 대신 이들을 활용하여 동영상의 주제를 분류할 수 있을 것이다. 따라서 이미지 정보뿐 아니라 오디오와 텍스트 정보를 같이 활용하는 멀티 모달리티 기반의 기계학습을 통하여 동영상 분류 성능이 향상될 수 있을 것이다.
- [0005] 기존의 동영상 분류 시스템은 대부분 이미지만을 사용한다. 그러므로 기계학습에서 이미지 분류에 대표적으로 사용되는 알고리즘인 CNNs (Convolutional Neural Networks)를 통하여 이미지를 분류하고 그 결과를 동영상 분류에 활용한다. 동영상 하나의 길이가 T 초라고 할 때, 1초마다 이미지 하나를 추출하면 동영상 하나로부터 T 개의 이미지가 나온다. 학습 동영상으로부터 이미지를 추출하여 각각의 이미지를 CNNs를 통하여 동영상에 해당하는 이벤트 클래스로 추가 학습(fine tuning)을 진행한다. 이렇게 추가 학습된 CNNs를 바탕으로 새로운 동영상에서 추출한 T 개의 이미지를 입력하면 T 개의 이미지에 대해 각각 이벤트 클래스가 분류되게 된다. 이로부터 한 동영상 내에서 가장 높은 빈도로 분류된 이벤트 클래스를 해당 동영상의 이벤트 클래스라고 할 수 있다. 이는 투표 형식으로 진행되는 동영상 분류 시스템이다. 더 나아가 이미지로부터 다양한 특징 벡터 추출과 분류 알고리즘을 활용하여 가장 많은 투표를 받은 이벤트 클래스로 분류하는 앙상블(ensemble) 기법을 적용한 이벤트 분류 시스템이 있다.
- [0006] 투표 형식과 유사하지만 다양한 방법으로 접근할 수 있는 동영상 분류 기법으로는 통합(pooling) 방법이 있다. 이는 CNNs에서와 같이 이미지에서 대표 특징을 추출하기 위하여 사용하는 통합 기법을 적용한 것으로, 동영상으로부터 추출된 여러 개의 이미지를 대표하는 하나의 특징 벡터를 구하는 방법이다. 동영상에서 추출된 T 개의 이미지에 대해 CNNs를 통하여 T 개의 특징 벡터 행렬을 구하고 이를 하나의 대표 벡터로 변환한 뒤, 이를 분류에 사용한다. 통합의 종류는 평균 통합(average pooling), 최댓값 통합(max pooling), 지역적 선택 통합(local pooling), 추가 학습 진행 통합(slow, late pooling) 등 다양한 종류의 통합 방법이 있으며 평균 통합을 제외하고 가장 높은 성능을 내고 있는 것은 최댓값 통합이다.
- [0007] 투표 방식과 통합 방식을 사용하는 경우 동영상의 길이에 상관없이 동영상을 대표하는 특징 벡터를 빠르고 쉽게 구할 수 있다. 하지만 동영상이 시계열 데이터임에 주목하고 이를 반영하여 동영상 이벤트를 분류하는 대표적인 연구로 3D CNNs 와 LSTM based 동영상 분류 시스템이 있다. 3D CNNs 동영상 분류 시스템은 2D 이미지에 동영상 길이를 포함한 3D 상태로 변환한 뒤 이를 CNNs를 통하여 학습하는 방식으로 학습 단계에서 시공간의 정보를 포함할 수 있다. LSTM based 동영상 분류 시스템은 시계열 데이터를 처리하는 데 있어 대표적인 기계학습 알고리즘인 RNNs(Recurrent Neural Networks) 기반의 LSTM(Long Short Term Memory)을 사용한다. 이는 CNNs를 통해 추출된 이미지들의 특징 벡터를 바탕으로 다시 LSTM을 통해 시계열 정보를 학습하는 방법으로 동영상의 시계열 정보를 효과적으로 학습할 수 있다.
- [0008] 그러나 3D CNNs 동영상 분류 시스템은 고정된 값을 입력으로 받아들이므로 동영상의 길이가 유동적인 경우에는 활용하기가 어렵다. 또한, LSTM based 경우 CNNs 기반의 이미지 분류 알고리즘 위에 LSTM 알고리즘을 적용하는 방법으로 두 종류의 네트워크가 존재한다. 따라서 파라미터(parameter)의 개수가 많아져 학습에 필요한 학습 데이터와 시간이 늘어났으며 학습 단계에서 두 종류의 네트워크 간 파라미터 공유가 어렵다.

선행기술문헌

특허문헌

- [0009] (특허문헌 0001) 한국등록특허공보 제 10-0792016호
(특허문헌 0002) 한국공개특허공보 제 10-2007-0107628호

발명의 내용

해결하려는 과제

- [0010] 전술한 문제점을 해결하기 위한 본 발명의 목적은 멀티 모달의 상관 관계 통합(correlation pooling)을 이용하

여 동영상을 대표하는 단일의 특징 벡터를 생성할 수 있는 동영상에 대한 특징 벡터 추출 방법 및 시스템을 제공하는 것이다.

[0011] 본 발명의 다른 목적은 전술한 방법을 적용하여, 동영상으로부터 추출이 가능한 이미지와 오디오 두 개의 멀티모달리티를 사용하여 추출된 동영상에 대한 단일의 특징 벡터를 이용하여 동영상에 대한 이벤트를 분류하여 동영상 이벤트 분류 성능을 향상시킬 수 있는 동영상 분류 시스템 및 방법을 제공하는 것이다.

과제의 해결 수단

[0012] 전술한 기술적 과제를 달성하기 위한 본 발명의 제1 특징에 따른 이미지와 오디오로 이루어지는 동영상의 특징 벡터 추출 방법은, (a) 상기 동영상의 이미지에 대한 이미지 특징 벡터를 추출하는 단계; (b) 상기 동영상의 오디오에 대한 오디오 특징 벡터를 추출하는 단계; (c) 상기 이미지 특징 벡터 및 상기 오디오 특징 벡터를 각각 단위 벡터를 이용하여 정규화시키는 단계; (d) 정규화된 이미지 특징 벡터와 정규화된 오디오 특징 벡터를 통합하여 상기 동영상에 대한 특징 벡터를 생성하는 단계; 를 구비하여 동영상을 대표하는 단일의 특징 벡터를 추출한다.

[0013] 전술한 제1 특징에 따른 동영상의 특징 벡터 추출 방법에 있어서, 상기 (d) 단계는, 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 상관 계수를 추출하고, 상기 상관 계수를 가중값으로 이용하여 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터를 상관 관계 통합하여 상기 동영상에 대한 특징 벡터를 생성하는 것이 바람직하다.

[0014] 전술한 제1 특징에 따른 동영상의 특징 벡터 추출 방법에 있어서, 상기 상관 계수는 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 피어슨 상관 계수인 것이 바람직하다.

[0015] 전술한 제1 특징에 따른 동영상의 특징 벡터 추출 방법에 있어서, 상기 (d) 단계는 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대하여 평균 통합을 하여 상기 동영상에 대한 특징 벡터를 생성하되, 상기 상관 계수를 가중값으로 사용하는 것이 바람직하다.

[0016] 전술한 제1 특징에 따른 동영상의 특징 벡터 추출 방법에 있어서, 상기 (c) 단계는 단일 계층 신경망을 사용하여 상기 (a) 단계에서 추출된 이미지 특징 벡터의 차원과 상기 (b) 단계에서 추출된 오디오 특징 벡터의 차원을 일치시키고, 서로 일치된 차원을 갖는 이미지 특징 벡터와 오디오 특징 벡터를 단위 벡터를 이용하여 정규화시키는 것을 특징으로 하며, 상기 단위 벡터는 이미지 특징 벡터와 오디오 특징 벡터의 속성은 그대로 유지하면서 크기가 1인 벡터인 것이 바람직하다.

[0017] 본 발명의 제2 특징에 따른 동영상 분류 방법은, 전술한 제1 특징에 따른 동영상의 특징 벡터 추출 방법에 의해 추출된 동영상을 대표하는 단일의 특징 벡터를 이용하여 동영상에 대한 이벤트를 분류하는 것을 특징으로 한다.

[0018] 본 발명의 제3 특징에 따른 동영상의 특징 벡터 추출 시스템은, 동영상으로부터 이미지와 오디오를 각각 추출하는 이미지/오디오 추출 모듈; 상기 이미지/오디오 추출 모듈로부터 추출된 이미지에 대한 이미지 특징 벡터를 추출하는 이미지 특징 벡터 추출 모듈; 상기 이미지/오디오 추출 모듈로부터 추출된 오디오에 대한 오디오 특징 벡터를 추출하는 오디오 특징 벡터 추출 모듈; 단일 계층 신경망을 이용하여 상기 이미지 특징 벡터 추출 모듈에 의해 추출된 이미지 특징 벡터의 차원과 상기 오디오 특징 벡터 추출 모듈에 의해 추출된 오디오 특징 벡터의 차원을 서로 일치시키는 차원 일치 모듈; 상기 차원 일치 모듈에 의해 차원이 일치된 이미지 특징 벡터 및 오디오 특징 벡터를 단위 벡터를 이용하여 각각 정규화시키는 정규화 모듈; 정규화된 이미지 특징 벡터와 오디오 특징 벡터를 통합시켜 동영상을 대표하는 하나의 특징 벡터를 추출하는 벡터 통합 모듈;을 구비하여, 동영상을 대표하는 단일을 특징 벡터를 추출하여 제공한다.

[0019] 전술한 제3 특징에 따른 동영상의 특징 벡터 추출 시스템에 있어서, 상기 벡터 통합 모듈은, 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 상관 계수를 추출하고, 상기 상관 계수를 가중값으로 이용하여 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터를 상관 관계 통합하여 상기 동영상에 대한 특징 벡터를 생성하는 것이 바람직하다.

[0020] 전술한 제3 특징에 따른 동영상의 특징 벡터 추출 시스템에 있어서, 상기 상관 계수는 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 피어슨 상관 계수인 것이 바람직하다.

[0021] 전술한 제3 특징에 따른 동영상의 특징 벡터 추출 시스템에 있어서, 상기 벡터 통합 모듈은, 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대하여 평균 통합을 하여 상기 동영상에 대한 특징 벡터를 생성하되, 상

기 상관 계수를 가중값으로 사용하는 것이 바람직하다.

[0022] 본 발명의 제4 특징에 따른 동영상 분류 시스템은, 전술한 제3 특징에 따른 동영상의 특징 벡터 추출 시스템에 의해 추출된 동영상을 대표하는 단일의 특징 벡터를 이용하여 동영상에 대한 이벤트를 분류하는 것을 특징으로 한다.

발명의 효과

[0023] 본 발명에 따른 동영상의 특징 벡터 추출 방법 및 시스템은 동영상의 이미지 특징 벡터와 오디오 특징 벡터를 추출하고, 이들의 차원을 일치시키고 단위 벡터로 정규화시킨 후, 상관 관계 통합(correlation pooling)을 이용하여 동영상을 대표하는 단일의 특징 벡터를 추출할 수 있다.

[0024] 또한, 이렇게 추출된 동영상 특징 벡터를 활용함으로써 보다 효율적으로 동영상에 대한 이벤트를 분류할 수 있게 된다.

[0025] 본 발명에 따른 특징 벡터 추출 방법에서는, 인공지능이 더욱 인간과 같은 학습을 할 수 있도록 단일 모달리티가 아닌 동영상으로부터 추출된 이미지와 오디오를 같이 학습하는 멀티 모달 학습을 시도하였다. 이에 각 모달리티를 단위 벡터로 정규화하는 방식을 제안함으로써, 본 발명은 멀티 모달의 통합 단계에서 효과적으로 통합할 수 있다. 그리고 이를 통해 하나의 동영상으로부터 얻을 수 있는 여러 종류의 특징 벡터를 효율적으로 통합하여 하나의 대표 특징 벡터를 구할 수 있다.

[0026] 또한, 본 발명에 따른 특징 벡터 추출 방법에서는, 멀티 모달리티를 사용할 때, 제안한 상관관계 통합이 기존의 통합 방법을 사용하는 것보다 성능이 향상됨을 확인할 수 있었다. 이를 통해 인공지능이 더욱 인간과 같아질 수 있도록 멀티미디어 학습이 가능함을 확인할 수 있었다.

도면의 간단한 설명

[0027] 도 1은 본 발명의 바람직한 실시예에 따른 동영상의 특징 벡터 추출 방법을 순차적으로 도시한 흐름도이다.

도 2는 피어슨 상관계수에 따른 선형관계를 설명하는 그래프이다.

도 3은 AlexNet의 구조도이다.

도 4는 GoogLeNet의 구조도이다.

도 5는 본 발명의 바람직한 실시예에 따른 동영상에 대한 특징 벡터 추출 시스템을 전체적으로 도시한 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0028] 본 발명에 따른 동영상의 특징 벡터 추출 방법 및 시스템은 동영상의 이미지 특징 벡터와 오디오 특징 벡터를 추출하고, 이들의 차원을 일치시키고 단위 벡터로 정규화시킨 후, 상관 관계 통합(correlation pooling)을 이용하여 동영상을 대표하는 단일의 특징 벡터를 추출하는 것을 특징으로 한다.

[0029] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 동영상의 특징 벡터 추출 방법 및 시스템에 대하여 구체적으로 설명한다.

[0030] < 동영상의 특징 벡터 추출 방법 >

[0031] 먼저, 본 발명의 바람직한 실시예에 따른 동영상의 특징 벡터 추출 방법에 대하여 구체적으로 설명한다. 도 1은 본 발명의 바람직한 실시예에 따른 동영상의 특징 벡터 추출 방법을 순차적으로 도시한 흐름도이다.

[0032] 도 1을 참조하면, 본 발명에 따른 동영상의 특징 벡터 추출 방법은, 이미지와 오디오로 이루어지는 동영상을 대표하는 단일의 특징 벡터를 추출하는 방법으로서, 상기 동영상의 이미지에 대한 이미지 특징 벡터를 추출하는 단계(단계 100), 상기 동영상의 오디오에 대한 오디오 특징 벡터를 추출하는 단계(단계 110), 상기 이미지 특징 벡터 및 상기 오디오 특징 벡터의 차원을 일치시키는 단계(단계 120), 차원을 일치시킨 상기 이미지 특징 벡터 및 상기 오디오 특징 벡터를 각각 단위 벡터로 정규화시키는 단계(단계 130), 및 정규화된 이미지 특징 벡터와 정규화된 오디오 특징 벡터를 통합하여 상기 동영상에 대한 특징 벡터를 생성하는 단계(단계 140)를 구비하여 동영상을 대표하는 단일의 특징 벡터를 추출한다. 이하, 전술한 각 단계들에 대하여 구체적으로 설명한다.

[0033] 먼저, 상기 동영상의 이미지에 대한 이미지 특징 벡터를 추출하는 단계(단계 100)에 대하여 구체적으로 설명한

다.

[0034] 동영상의 프레임(Frame)은 1초에 출력되는 이미지의 개수를 의미한다. 따라서 동영상 프레임이 f FSP(Frames Per Second)이고 동영상의 길이가 T 초이면, 동영상에서 추출되는 이미지의 개수는 모두 $f \times T$ 개이다. 동영상의 길이가 길어지고 프레임의 수가 커질수록 학습에 사용되는 이미지는 많아지게 되므로 학습하는데 시간이 많이 걸린다. 또한, 1초에 f 개의 다른 이미지가 지나가더라도 사람의 눈은 모든 이미지를 인식하기 어려우며 동영상에서 1초 동안 출력되는 이미지는 급변하기 보다는 연속된 동작이나 장면이므로 대부분의 비슷할 것이다. 그러므로 본 발명에 따른 방법에서는 매 1초마다 이미지를 추출하여 총 T 개의 이미지를 사용한다.

[0035] 길이가 T 인 동영상의 t 초에 해당하는 이미지를 x_t^{img} 라 하면, 추출된 이미지들의 집합 $\mathbf{X}^{img}$ 는 수학적 식 1과 같다.

수학적 식 1

$$\mathbf{X}^{img} = [x_t^{img} \mid t = 1, \dots, T]$$

[0036]

[0037] 두 개의 CNNs(Convolutional Neural Networks) 알고리즘을 사용하여 이미지의 특징 벡터를 추출한다. CNNs는 기계학습에서 이미지 분류에 대표적으로 사용되는 알고리즘으로서, 여러 개의 합성곱 계층(convolutional layer)과 통합 계층(pooling layer)들로 이루어져 있다. 합성곱 계층은 일종의 필터 형태로 다량의 학습 가능한 가중치(weight)를 합성곱 연산을 통해 입력으로부터 특징을 추출해낸다. 통합 계층은 합성곱 계층으로부터 추출된 특징의 차원을 축소하는 역할을 한다. 마지막으로 추출된 특징을 두 층의 신경망 계층을 통해 분류한다. 이에 t 초에 해당하는 이미지는 수학적 식 2와 같이 CNNs 알고리즘을 통해 p 차원의 특징 벡터로 변환된다.

수학적 식 2

$$\mathbf{v}_t^{img} = CNNs(x_t^{img}) \in \mathbb{R}^p$$

[0038]

[0039] 이를 T 개의 이미지에 대해 모두 수행하면 수학적 식 3과 같이 동영상으로부터 $T \times p$ 차원의 특징 벡터 행렬을 구할 수 있다.

수학적 식 3

$$\mathbf{V}^{img} = [\mathbf{v}_t^{img} \mid t = 1, \dots, T] \in \mathbb{R}^{T \times p}$$

[0040]

[0041] 전술한 CNNs 알고리즘들 중 하나는 AlexNet 으로서, AlexNet(Krizhevsky, A., Sutskever, I., and Hinton, E. Imagenet classification with deep convolutional neural networks. *Proceedings of Advances in Neural Information Processing Systems*, pp.1097-1105, 2012.)은 2012년 영상 분석 관련 대회인 ILSVRC(ImageNet Large-Scale Visual Recognition Challenge)에서 혁신적인 구조로 가장 높은 성능을 낸 알고리즘이다. 도 3은 AlexNet의 구조도이다. 도 3과 같은 구조로 5단계의 합성곱 계층 및 통합 계층들과 3개의 fc 계층(fully connected layers)들로 이루어져 있다. 본 발명에서는 1000개의 클래스를 가진 ILSVRC 2012 데이터로 사전에 학습된(pre-trained) AlexNet 모델을 사용하여 fc 계층 중 두 번째 단계인 fc7 계층으로부터 4096차원의 특징 벡터를 추출하였다.

[0042] 전술한 CNNs 알고리즘들 중 다른 하나는 Inception-v3으로서, Inception-v3(Szegedy, C., Vanhoucke, V.,

Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.2818-2826, 2016)는 Google에서 개발된 CNNs 기반 이미지 분류 알고리즘인 GoogLeNet(Inception)(Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... and Rabinovich, A. Going deeper with convolutions. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.1-9, 2015.)의 세 번째 버전이다. 도 4는 GoogLeNet 의 구조도이다. GoogLeNet은 도 4와 같은 구조로 합성곱 계층 및 통합 계층을 넓고 깊게 사용하는 Inception이라는 모듈 9개를 사용한다. 또한 중간에 두 개의 보조 분류기를 사용하여 구조가 깊어질수록 발생하는 문제를 해결하였다. Inception-v3는 Inception 모듈을 보다 효율적으로 넓고 깊게 사용하도록 개선된 방법이다. 본 발명에서는 AlexNet과 마찬가지로 ILSVRC 2012 데이터로 사전에 학습된 Inception-v3를 사용하여 마지막 분류기의 평균 통합을 수행하는 pool3 계층으로부터 2048차원의 특징 벡터를 추출하였다.

[0043] 다음, 상기 동영상의 오디오에 대한 오디오 특징 벡터를 추출하는 단계(단계 110)에 대하여 구체적으로 설명한다.

[0044] 이미지와는 다르게 오디오는 시계열 데이터로 표현이 된다. 따라서 동영상으로부터 오디오를 추출할 때, 이미지를 추출하는 t 초에서 ± 0.5 초 구간의 오디오를 추출한다. 이에 길이가 T 인 동영상에서 구한 오디오를

x_t^{aud} 라 하면, 추출된 오디오들의 집합 $\mathbf{X}^{aud}$ 는 수학적 4와 같다.

수학적 4

$$\mathbf{X}^{aud} = [x_t^{aud} | t = 1, \dots, T]$$

[0045]

[0046] 본 발명에 따른 동영상 특징 벡터 추출 방법에서는, 오디오로부터 특징 벡터를 추출하기 음성신호처리 분야에서 음성의 특성을 표현하기 위해 대표적으로 사용되는 방식인 MFCC (Mel Frequency Cepstral coefficient)(Sahidullah, M., and Saha, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Communication*, Vol. 54, No.4, pp.543-565, 2012.)를 사용한다. 이에 t ± 0.5 초에 해당하는 오디오를 수학적 5와 같이 MFCC를 통해 q 차원의 오디오 특징 벡터로 변환한다.

수학적 5

$$\mathbf{v}_t^{aud} = MFCC(x_t^{aud}) \in \mathbb{R}^q$$

[0047]

[0048] 이를 T 개의 오디오에 대해 모두 수행하면 수학적 6과 같이 동영상으로부터 $T \times q$ 차원의 오디오 특징 벡터 행렬을 구할 수 있다.

수학적 6

$$\mathbf{V}^{aud} = [\mathbf{v}_t^{aud} | t = 1, \dots, T] \in \mathbb{R}^{T \times q}$$

[0049]

[0050] 다음, 상기 이미지 특징 벡터 및 상기 오디오 특징 벡터의 차원을 일치시키는 단계(단계 120) 및 전술한 단계에서 차원을 일치시킨 상기 이미지 특징 벡터 및 상기 오디오 특징 벡터를 각각 단위 벡터로 정규화시키는 단계(단계 130)에 대하여 구체적으로 설명한다.

[0051] 단계 100 및 단계 110에서 구해진 이미지 특징 벡터와 오디오 특징 벡터의 차원은 각각 p 와 q 이므로, 이들은 일반적인 통합 기법 적용이 어렵다. 따라서, 이들의 차원을 동일하게 해줄 필요가 있으며 통합 단계에서 둘 간의 상관관계가 반영될 수 있어야 한다. 이에 단일 계층 신경망(single layer neural network)을 사용하여 d 차원으로 일치시켜야 한다. 신경망의 활성화 함수는 수학식 7과 같이 딥 러닝 알고리즘에서 대표적으로 사용되는 ReLU(Rectified Linear Unit)(Nair, V., and Hinton, E. Rectified linear units improve restricted boltzmann machines. *Proceedings of International Conference on Machine Learning*, pp.807-814, 2010.)를 사용할 수 있다.

수학식 7

$$f(x) = \max(0, x)$$

[0052]

[0053] 본 발명에 따른 특징 벡터 추출 방법은, 멀티 모달리티의 특징 벡터들을 하나로 통합하는 단계를 거쳐 동영상을 대표하는 새로운 특징 벡터를 추출하고자 한다. 하지만, 수학식 7에서 확인할 수 있듯이 신경망의 활성화 함수인 ReLU의 결과값은 무한대로 치솟을 수 있다. 이는 통합 단계에서 특징 벡터들의 분포 차이가 발생할 수 있으며 그 차이가 크다면 분포가 작은 모달리티의 특징이 제대로 반영되지 않을 것이다. 이에 수학식 8과 같이 단위 벡터로의 정규화 수행함으로써, 전술한 문제를 해결할 수 있으며 두 모달리티의 상관관계를 효율적으로 구할 수 있게 된다.

수학식 8

$$\hat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$$

[0054]

[0055] 본 발명에 있어서, 단일 계층의 신경망을 통해 추출된 차원의 특징 벡터를 다음과 같이 벡터의 속성을 그대로 유지하면서 크기가 1인 단위 벡터로 정규화를 수행한다. 이미지와 오디오의 신경망을 $\phi_{img}(\cdot)$, $\phi_{aud}(\cdot)$ 라 하면 수학식 9 및 수학식 10과 같이 차원을 일치시켜 차원이 d 인 이미지 특징 벡터($\mathbf{v}'^{img}$)와 오디오 특징 벡터($\mathbf{v}'^{aud}$)를 구할 수 있다.

수학식 9

$$\mathbf{v}'_t^{img} = \phi_{img}(\mathbf{v}_t^{img}) \in \mathbb{R}^d$$

$$\mathbf{V}'^{img} = [\mathbf{v}'_t^{img} | t = 1, \dots, T] \in \mathbb{R}^{T \times d}$$

[0056]

수학식 10

$$\mathbf{v}'_{t^{aud}} = \phi_{aud}(\mathbf{v}_t^{aud}) \in \mathbb{R}^d$$

$$\mathbf{V}'^{aud} = [\mathbf{v}'_t^{aud} | t = 1, \dots, T] \in \mathbb{R}^{T \times d}$$

[0057]

[0058] 단위 벡터를 구하여 최종적으로 구해진 이미지 특징 벡터와 오디오 특징 벡터는 수학식 11 및 수학식 12와 같다.

수학식 11

$$\widehat{\mathbf{v}}_t^{img} = \mathbf{v}'_t^{img}$$

$$\widehat{\mathbf{V}}^{img} = \mathbf{V}'^{img}$$

[0059]

수학식 12

$$\widehat{\mathbf{v}}_t^{aud} = \mathbf{v}'_t^{aud}$$

$$\widehat{\mathbf{V}}^{aud} = \mathbf{V}'^{aud}$$

[0060]

[0061] 다음, 정규화된 이미지 특징 벡터와 정규화된 오디오 특징 벡터를 통합하여 상기 동영상에 대한 특징 벡터를 생성하는 단계(단계 140)에 대하여 구체적으로 설명한다.

[0062] 최종적으로 구해진 이미지 특징 벡터와 오디오 특징 벡터들을 통해 동영상을 대표하는 하나의 특징 벡터(u)를 생성한다.

[0063] 상관 관계 분석(correlation analysis)은 두 개의 변수 사이에 선형적 관계를 파악하기 위한 분석 방법이다. 이때 두개의 변수 사이의 선형적 관계를 상관 계수(correlation coefficient)라 하는데, 이를 구하기 위해서는 보편적으로 피어슨 상관 계수(Pearson correlation coefficient)를 사용한다. 피어슨 상관 계수는 두 변수가 변하는 정도인 공분산(covariance)에 각 변수가 변하는 정도인 표준편차 쌍의 곱을 나누어 구할 수 있으며, -1에서 1 사이의 값을 가진다. 도 2는 피어슨 상관계수에 따른 선형관계를 설명하는 그래프이다. 도 2를 참조하면, 2개의 변수들의 분포가 유사하면 계수가 1에 가까운 양의 상관 관계를 가지고, 분포가 반대이면 계수가 -1에 가까운 음의 상관 관계를 가진다. 0에 가까울수록 선형 관계가 거의 없다고 할 수 있다.

[0064] 상관 계수를 활용하는 경우는 상관 계수를 통하여 여러 모델의 분포를 분산시켜 앙상블시 개별 네트워크를 특화하는 방법이나 오토 인코더를 통해 멀티 모달을 학습할 때 사용하는 상관관계 신경망(correlation neural networks)이 있다. 이 상관관계 신경망은 두 개의 멀티 모달리티가 입력으로 사용되며 오토 인코더를 통하여 각각의 특징 벡터를 학습할 때, 피어슨 상관계수를 목적 함수(object function)를 최소화하는 과정에서 일정 가중치로 사용한다. 이로 인하여 두 개의 멀티 모달리티를 입력하여 학습할 때 두 개의 모달리티가 유사한 상관관계를 가지게끔 유도할 수 있다.

[0065] 본 발명에서는 유연한 길이의 동영상을 활용할 수 있고 효과적으로 두 모달리티 간의 상관관계를 반영하기 위해

통합 방법을 활용한다. 통합 방법 중 대표적으로 사용되는 최댓값 통합은 수학식 13과 같이 표현이 가능하다.

수학식 13

$$\mathbf{u} = [u_1, \dots, u_N] \in \mathbb{R}^d$$

$$(\text{단, } u_i = \max(V_{1i}, \dots, V_{Ti}))$$

[0066]

[0067]

최댓값 통합은 특징 벡터 행렬(V)의 각 열에 대해서 최댓값을 찾는 방식이다. 따라서 멀티 모달리티로 이미지와 오디오 특징 벡터가 동시에 입력되면 두 특징 벡터의 쌍을 유지할 의미가 없어진다. 따라서, 본 발명의 바람직한 실시예에 따른 특징 벡터 추출 방법에서는 이미지와 오디오의 특징 벡터가 입력되었을 때 두 특징벡터의 쌍을 유지하면서 상관관계가 반영될 수 있도록 수학식 14와 같은 평균 통합을 사용한다.

수학식 14

$$\mathbf{u} = \frac{1}{T} \sum_{t=1}^T \mathbf{v}_t \in \mathbb{R}^d$$

[0068]

[0069]

서로간의 상관관계를 반영하기 위해 피어슨 상관계수를 사용한다. 변수 X 와 Y 사이의 피어슨 상관계수 $corr(X, Y)$ 은 수학식 15와 같다.

수학식 15

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

[0070]

[0071]

본 발명의 바람직한 실시예에 따른 특징 벡터 추출 방법에서 사용하는 상관관계 통합은 평균 통합을 수행할 때, 시간 t 에서의 이미지와 오디오의 특징 벡터 사이의 상관계수를 일정 가중치로 반영하는 방법이다. 피어슨 상관계수가 1에 가까울수록 두 변수 사이의 분포가 유사한데 이는 동영상으로부터 추출된 시간 t 에서의 이미지와 오디오가 모두 동영상을 대표하는 특징 벡터라 가정할 수 있다. 반면 피어슨 상관 계수가 -1에 가까워진다면 둘 중 하나만이 동영상을 대표하는 특징 벡터라 가정할 수 있다.

[0072]

따라서 평균 통합에서 매 시간 t 에서의 이미지와 오디오 간의 상관계수를 구하여 두 모달리티가 모두 동영상을 대표할 가능성이 높으면, 동영상의 특징 벡터($\mathbf{u}$)에 더 큰 영향을 줄 수 있도록 한다. 또한, 상관 계수의 범위는 -1에서 1 사이므로 단위 벡터로 정규화된 특징 벡터를 입력으로 사용하면 보다 효율적으로 상관관계를 반영

할 수 있을 것이다. 이에 수학적 식 16과 같이 이전 단계에서 단위 벡터로 정규화를 진행한 이미지와 오디오의 최

중 특징 벡터 행렬 $\hat{\mathbf{V}}^{img}$ 와 $\hat{\mathbf{V}}^{aud}$ 를 사용하는 상관관계 통합한다.

수학적 식 16

$$\mathbf{u} = [u_1, \dots, u_N] \in \mathbb{R}^d$$

$$(\text{단, } u_i = \frac{1}{T} \sum_{t=1}^T \{ (\hat{\mathbf{v}}_{ti}^{img} + \hat{\mathbf{v}}_{ti}^{aud} - (1 - \text{corr}(\hat{\mathbf{v}}_t^{img}, \hat{\mathbf{v}}_t^{aud})) \})$$

[0073]

[0074] 한편, 전술한 본 발명에 따른 동영상에 대한 특징 벡터 추출 방법을 통해 상관관계 통합으로 구해진 동영상의 특징 벡터($\mathbf{u}$)를 SVM(Support Vector Machine)을 통하여 이벤트 클래스에 대한 확률 분포를 구하고 이를 소프트맥스(soft-max)를 통하여 동영상에 대한 이벤트를 선택함으로써, 동영상을 분류할 수 있게 된다.

[0075] < 동영상에 대한 특징 벡터 추출 시스템 >

[0076] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 동영상에 대한 특징 벡터 추출 시스템에 대하여 구체적으로 설명한다.

[0077] 도 5는 본 발명의 바람직한 실시예에 따른 동영상에 대한 특징 벡터 추출 시스템을 전체적으로 도시한 블록도이다. 도 5를 참조하면, 본 발명에 따른 동영상에 대한 특징 벡터 추출 시스템(30)은 이미지/오디오 추출 모듈(300), 이미지 특징 벡터 추출 모듈(310), 오디오 특징 벡터 추출 모듈(312), 차원 일치 모듈(320), 정규화 모듈(330), 벡터 통합 모듈(340)을 구비하여, 이미지와 오디오로 이루어지는 동영상을 대표하는 단일의 특징 벡터를 추출하여 제공한다.

[0078] 상기 이미지/오디오 추출 모듈(300)은 동영상으로부터 이미지와 오디오를 각각 추출한다.

[0079] 상기 이미지 특징 벡터 추출 모듈(310)은 상기 이미지/오디오 추출 모듈로부터 추출된 이미지에 대한 이미지 특징 벡터를 추출하여 제공한다. 본 발명의 바람직한 실시예에 따른 특징 벡터 추출 시스템은 두 개의 CNNs 알고리즘을 사용하여 각 이미지에 대하여 p 차원의 특징 벡터로 변환시키면, 길이가 T인 동영상에 대하여 수학적 식 3과 같은 $T \times p$ 차원의 이미지 특징 벡터 행렬을 추출하여 제공하게 된다.

[0080] 상기 오디오 특징 벡터 추출 모듈(312)은, 상기 이미지/오디오 추출 모듈로부터 추출된 오디오에 대한 오디오 특징 벡터를 추출하여 제공한다. 본 발명의 바람직한 실시예에 따른 특징 벡터 추출 시스템은 오디오로부터 특징 벡터를 추출하기 위하여 음성 신호처리 분야에서 음성의 특성을 표현하기 위하여 대표적으로 사용되는 방식인 MFCC(Mel Frequency Cepstral Coefficient)를 사용한다. 이에 의하여, 각 오디오에 대하여 q 차원의 특징 벡터로 변환시키면, 동영상으로부터 수학적 식 6과 같은 $T \times q$ 차원의 오디오 특징 벡터 행렬을 추출하여 제공하게 된다.

[0081] 상기 차원 일치 모듈(320)은, 단일 계층 신경망을 이용하여 상기 이미지 특징 벡터 추출 모듈에 의해 추출된 이미지 특징 벡터의 차원과 상기 오디오 특징 벡터 추출 모듈에 의해 추출된 오디오 특징 벡터의 차원을 하나의 차원으로 일치시킨다. 따라서, 상기 차원 일치 모듈(320)에 의해 d 차원의 이미지 특징 벡터 및 오디오 특징 벡터를 구하여 제공하게 된다.

[0082] 상기 정규화 모듈(330)은, 상기 차원 일치 모듈에 의해 차원이 일치되어 모두 d 차원의 이미지 특징 벡터 및 오디오 특징 벡터를 단위 벡터를 이용하여 각각 정규화시켜 제공하게 된다.

[0083] 상기 벡터 통합 모듈(340)은 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 상관 계수를 추출하고, 상기 상관 계수를 가중값으로 이용하여 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터를 상관관계 통합하여 동영상을 대표하는 하나의 특징 벡터($\mathbf{u}$)를 추출하여 제공한다. 상기 상관 계수는 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대한 피어슨 상관 계수인 것이 바람직하다.

- [0084] 한편, 상기 벡터 통합 모듈은, 상기 정규화된 이미지 특징 벡터와 오디오 특징 벡터에 대하여 평균 통합을 하여 상기 동영상에 대한 특징 벡터를 생성하되, 상기 상관 계수를 가중값으로 사용하여 평균 통합하는 것이 바람직하다.
- [0085] 전술한 구성을 갖는 본 발명의 바람직한 실시예에 따른 동영상에 대한 특징 벡터 추출 시스템은 이미지와 오디오를 갖는 동영상으로부터 이미지와 오디오를 추출하고, 이들 각각 CNNs 알고리즘과 MFCC를 통해 각각 p 차원의 이미지 특징 벡터와 q 차원의 오디오 특징 벡터를 추출하고, 단일 계층 신경망을 통해 이들의 차원을 일치시키고 단위 벡터로 정규화시킨 후, 상관관계 통합을 사용하여 상기 동영상을 대표하는 단일의 특징 벡터(u)를 생성하여 제공하게 된다.
- [0086] 한편, 본 발명에 따른 동영상 분류 시스템은 전술한 본 발명에 따른 동영상에 대한 특징 벡터 추출 시스템을 이용하여 동영상에 대한 단일의 특징 벡터(u)를 추출하고, SVM(Support Vector Machine)을 통하여 이벤트 클래스에 대한 확률 분포를 구하고, 이를 소프트맥스(Soft-max)를 통하여 동영상에 대한 이벤트를 선택함으로써, 동영상을 분류할 수 있도록 한다.
- [0087] 전술한 본 발명에 따른 동영상에 대한 특징 벡터 생성 방법을 검증하기 위하여, 표 1과 같이 다양한 조건에서의 비교 실험을 수행하였다. 다만, 멀티 모달의 상관 관계를 반영하기 위해 제안한 상관 관계 통합은 정규화를 수행하지 않을 경우, 상관 계수의 영향이 거의 없으므로 단위 벡터로 정규화를 진행한 경우에만 실험을 수행하였다. 이미지의 경우 AlexNet과 Inception-v3로 추출한 특징 벡터에 따른 성능을 추가로 비교해 본다.

표 1

대분류	중분류	소분류
모달리티 종류	정규화 유무	통합 방법
오디오	없음	최댓값 통합
		평균 통합
	단위 벡터 정규화	최댓값 통합
		평균 통합
이미지	없음	최댓값 통합
		평균 통합
	단위 벡터 정규화	최댓값 통합
		평균 통합
오디오 + 이미지	없음	최댓값 통합
		평균 통합
	단위 벡터 정규화	최댓값 통합
		평균 통합
		상관관계 통합

- [0088]
- [0089] 실험 데이터로는 멀티 모달리티를 사용하기 위하여 YLI-MED 데이터를 사용하였으며, 이는 멀티미디어 이벤트 인식 연구(MED, Multimedia Event Detection)에 사용되는 데이터로 YFCC100M(Yahoo Flickr Creative Commons 100 Million)으로부터 추출된 이미지와 오디오가 포함된 1823개의 동영상 데이터이다. 동영상의 평균 길이는 약 43초이며 10개의 이벤트 클래스로 구성되어 있다. 1000개의 학습 데이터와 823개의 실험 데이터로 구성되어 있으며 각 클래스에 대한 자세한 정보는 표 2와 같다.

표 2

Label	Event name	Training	Test	Total
Ev101	Birthday Party	100	137	237
Ev102	Flash Mob	100	50	150
Ev103	Getting a Vehicle Unstuck	100	41	141
Ev104	Parade	100	130	230
Ev105	Person Attempting a Board Trick	100	94	194
Ev106	Person Grooming an Animal	100	39	139
Ev107	Person Hand-Feeding an Animal	100	20	220
Ev108	Person Landing a Fish	100	43	143
Ev109	Wedding Ceremony	100	119	219
Ev110	Working on a Woodworking Project	100	150	150
Total		1000	823	1823

[0090]

[0091]

본 발명에 대한 추가 실험을 위하여 YouTube로부터 이미지와 오디오가 포함된 1369개의 동영상 데이터를 자체 수집하여 구성하였다. 동영상의 평균 길이는 약 54초이며 총 16개의 이벤트 클래스로 구성되어 있다. 906개의 학습 데이터와 463개의 실험 데이터로 이루어져 있으며 각 클래스에 대한 정보는 표 3과 같다.

표 3

Label	Event name	Training	Test	Total
Ev101	First Birthday Party	64	33	97
Ev102	Camping	28	15	43
Ev103	Climbing	18	10	28
Ev104	Amusement Park	63	33	96
Ev105	Pet	64	32	96
Ev106	Entrance Ceremony	22	11	33
Ev107	Birthday Party	56	29	85
Ev108	Wedding	64	32	96
Ev109	Cute Festival	66	33	99
Ev110	Bicycle Stunt	66	34	100
Ev111	Graduation Ceremony	65	33	98
Ev112	Fishing	66	34	100
Ev113	Toddle	66	34	100
Ev114	Golf	66	33	99
Ev115	Ramen Cook	66	34	100
Ev116	Makeup	66	33	99
Total		906	463	1369

[0092]

[0093] 앞서 정의한 실험 조건에 따라 YLI-MED와 YouTube 데이터로 실험을 수행해본 결과는 각각 표 4, 표 5와 같다.

표 4

YLI-MED 데이터				
대분류	중분류	소분류	CNNs 알고리즘에 따른 분류 정확도	
모달리티 종류	정규화 유무	통합 방법	AlexNet	Inception-v3
오디오 (MFCC)	없음	최대값 통합	31.13%	
		평균 통합	28.53%	
	단위 벡터 정규화	최대값 통합	34.54%	
		평균 통합	33.92%	
이미지	없음	최대값 통합	62.37%	77.98%
		평균 통합	62.58%	78.10%
	단위 벡터 정규화	최대값 통합	66.21%	78.72%
		평균 통합	66.96%	79.59%
오디오 (MFCC) + 이미지	없음	최대값 통합	66.21%	76.22%
		평균 통합	66.96%	77.22%
	단위 벡터 정규화	최대값 통합	68.33%	78.97%
		평균 통합	69.46%	80.73%
		상관관계 통합	70.09%	81.35%

[0094]

표 5

YouTube 데이터				
대분류	중분류	소분류	CNNs 알고리즘에 따른 분류 정확도	
모달리티 종류	정규화 유무	통합 방법	AlexNet	Inception-v3
오디오 (MFCC)	없음	최댓값 통합	27.00%	
		평균 통합	28.08%	
	단위 벡터 정규화	최댓값 통합	31.10%	
		평균 통합	31.97%	
이미지	없음	최댓값 통합	81.43%	89.63%
		평균 통합	80.99%	90.48%
	단위 벡터 정규화	최댓값 통합	83.59%	91.79%
		평균 통합	84.23%	92.44%
오디오 (MFCC) + 이미지	없음	최댓값 통합	63.72%	50.76%
		평균 통합	79.05%	78.83%
	단위 벡터 정규화	최댓값 통합	84.67%	91.45%
		평균 통합	85.75%	92.66%
		상관관계 통합	85.53%	93.09%

[0095]

[0096]

실험 결과 이미지의 특징 벡터를 추출하는 CNNs 알고리즘에 따라 성능의 차이가 큰 것을 확인할 수 있었다. 이는 Inception-v3가 가장 최신의 이미지 분류 알고리즘으로 AlexNet보다 성능이 향상되었기 때문이다. 이미지에 대해 두 가지의 알고리즘을 사용한 이유는 CNNs 알고리즘의 성능 비교가 아닌 특징 벡터를 추출하는 다양한 알고리즘 사용하여 실험해 봄으로써 본 발명에서 제안하는 방법들의 성능을 검증하기 위함이다.

[0097]

최종 실험 결과 표 4와 표 5에서 확인할 수 있듯이 제안한 단위 벡터 정규화를 사용할 경우 모든 통합 방법에 대해 성능의 향상이 있었다. 또한, 제안한 상관관계 통합이 전체에 대해 가장 높은 성능을 얻을 수 있었다.

[0098]

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 해석되어

야 할 것이다.

부호의 설명

[0099]

30 : 동영상의 특징 벡터 추출 시스템

300 : 이미지/오디오 추출 모듈

310 : 이미지 특징 벡터 추출 모듈

312 : 오디오 특징 벡터 추출 모듈

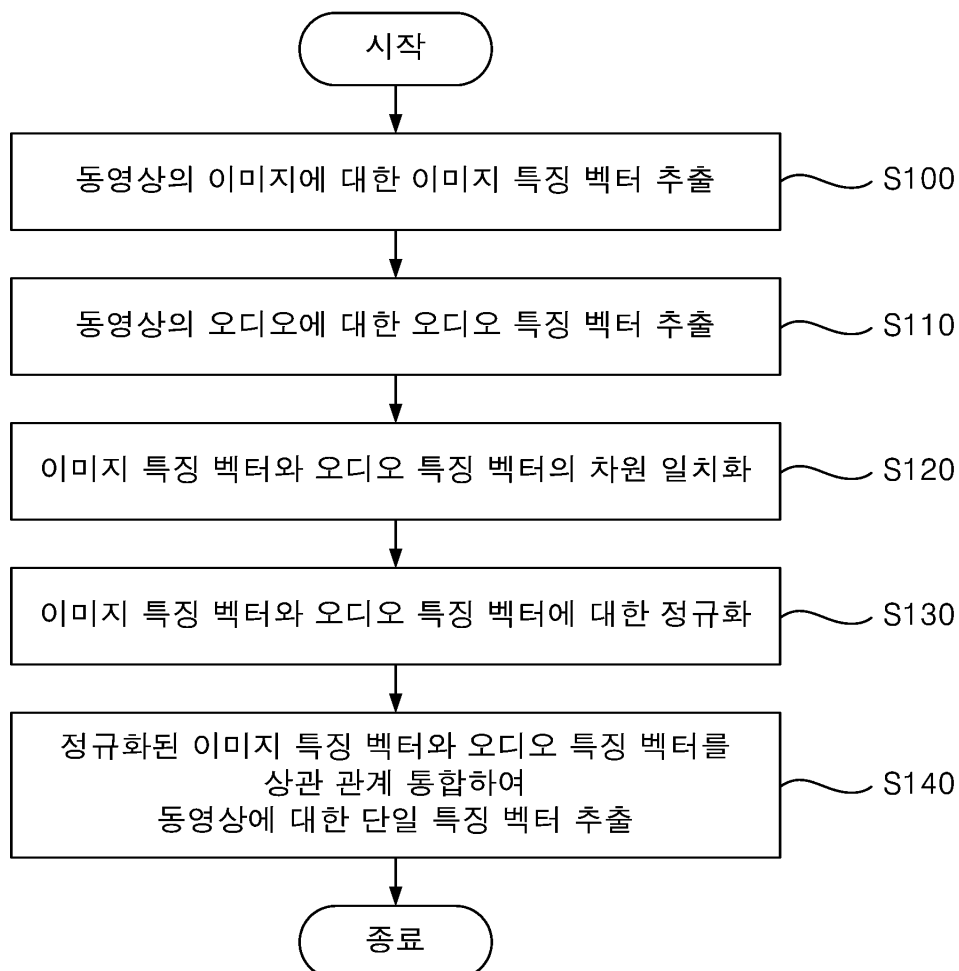
320 : 차원 일치 모듈

330 : 정규화 모듈

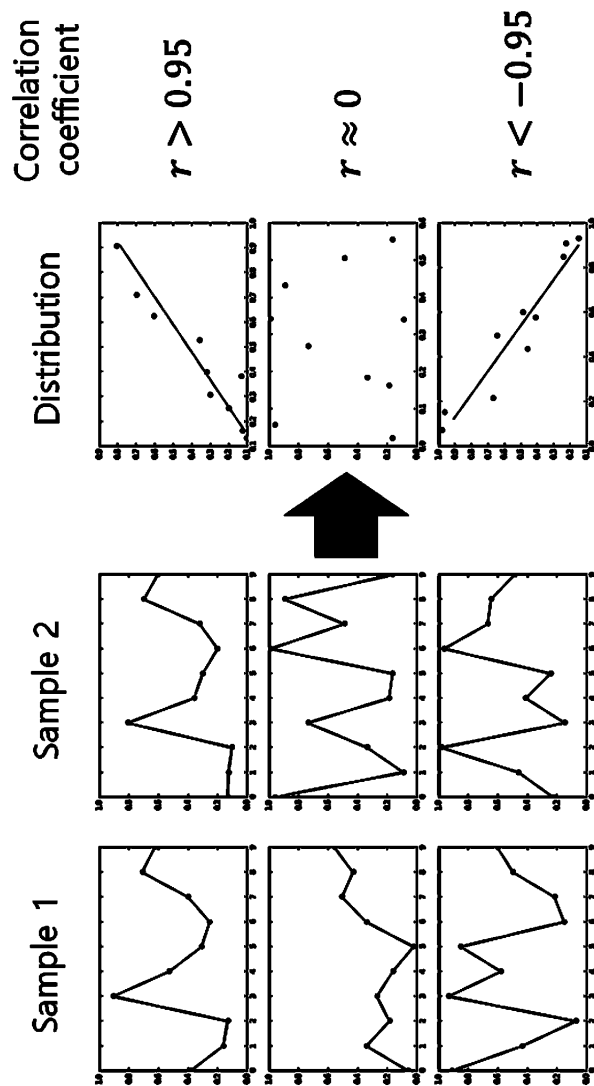
340 : 벡터 통합 모듈

도면

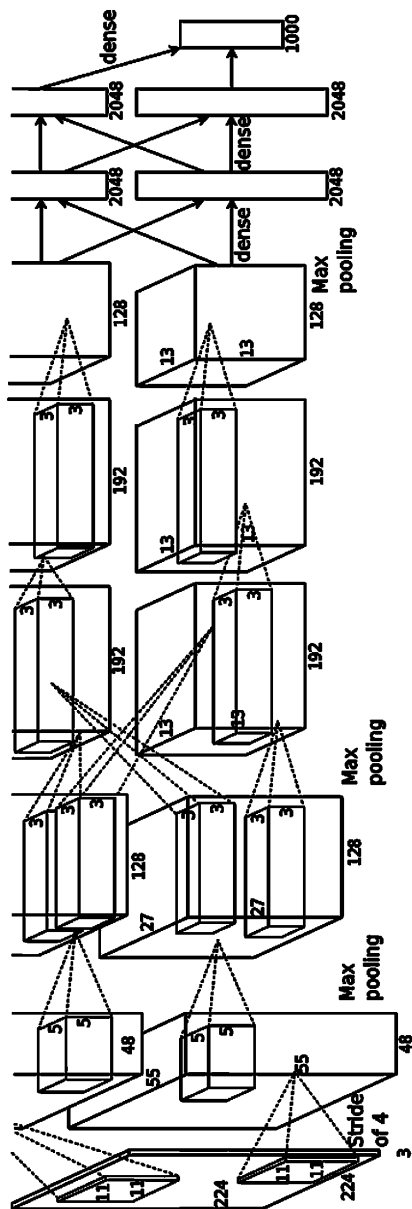
도면1



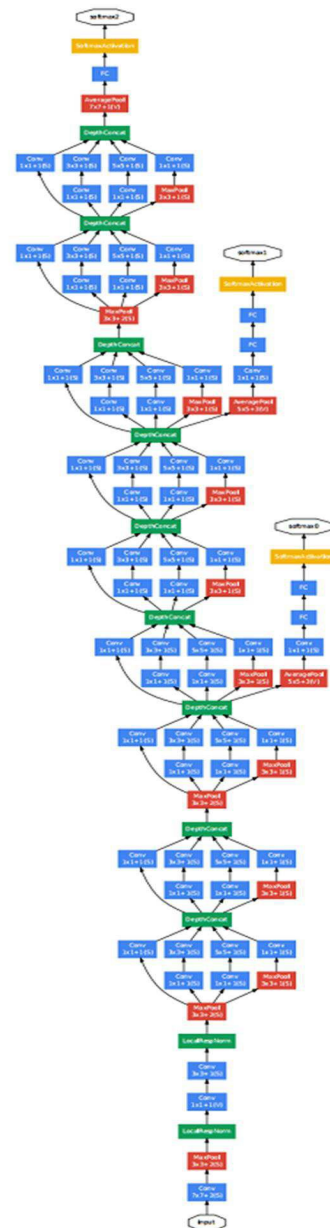
도면2



도면3



도면4



도면5

