(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0077330
(43) 공개일자 2023년06월01일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2023.01) *G06N 3/08* (2023.01)
G06N 5/04 (2023.01)
(52) CPC특허분류
G06N 3/049 (2023.01)
G06N 3/082 (2023.01)
(21) 출원번호 10-2021-0164497
(22) 출원일자 2021년11월25일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
조민수
경상북도 포항시 남구 효성로 55, 105동 2502호 (효자동, 효자풍림아이원아파트)
권희승
경상북도 포항시 북구 새천년대로1076번길 38, 305동 1102호 (두호동, 창포아이파크3차아파트)
(뒷면에 계속)
(74) 대리인
특허법인 무한

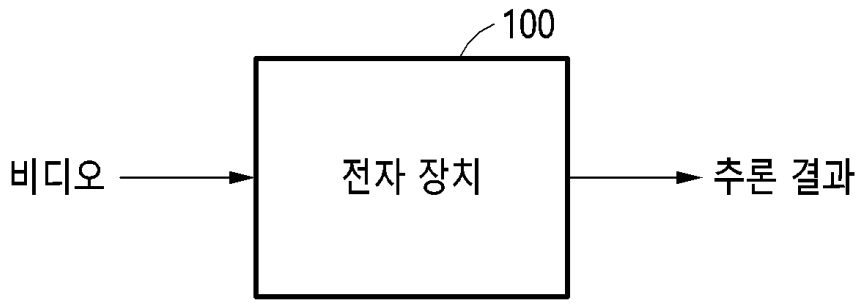
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **시공간 자기-유사도를 이용하는 전자 장치 및 그 동작 방법**

(57) 요 약

시공간 자기-유사도를 이용하는 전자 장치 및 그 동작 방법이 개시된다. 전자 장치는 프로세서 및 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리를 포함하고, 적어도 하나의 명령어가 프로세서에서 실행되면, 프로세서는 입력 비디오에 대응하는 비디오 특징 맵의 각 위치에 대해 하나의 공간 자기-유사도 맵 및 복수의 공간 교차-유사도 맵들을 포함하는 STSS 텐서를 시간 오프셋 및 공간 오프셋에 기초하여 결정하고, STSS 텐서의 각 위치에 대해 공간 오프셋에 관한 차원을 감소시키고, 시간 오프셋에 관한 차원을 유지시킴으로써, STSS 텐서로부터 STSS 특징 벡터들을 결정하고, STSS 특징 벡터들의 각 위치에 대한 시간 오프셋에 관한 차원을 통합함으로써, STSS 특징 맵을 결정하고, STSS 특징 맵을 비디오 특징 맵에 더한 결과에 기초하여 입력 비디오에 대한 추론을 수행한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/084 (2023.01)
G06N 5/04 (2023.01)
(72) 발명자
김만진
경상북도 포항시 남구 지곡로 83, 6동 401호 (지곡동, 포스빌)
곽수하
경상북도 포항시 남구 지곡로 155, 4동 201호 (지곡동, 교수아파트)

명 세 서

청구범위

청구항 1

전자 장치에 있어서,

프로세서; 및

상기 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리

를 포함하고,

상기 적어도 하나의 명령어가 상기 프로세서에서 실행되면, 상기 프로세서는

입력 비디오에 대응하는 비디오 특징 맵의 각 위치에 대해 하나의 공간 자기-유사도(spatial self-similarity) 맵 및 복수의 공간 교차-유사도(spatial cross-similarity) 맵들을 포함하는 STSS(spatio-temporal self-similarity) 텐서를 시간 오프셋 및 공간 오프셋에 기초하여 결정하고,

상기 STSS 텐서의 각 위치에 대해 상기 공간 오프셋에 관한 차원을 감소시키고, 상기 시간 오프셋에 관한 차원을 유지시킴으로써, 상기 STSS 텐서로부터 STSS 특징 벡터들을 결정하고,

상기 STSS 특징 벡터들의 각 위치에 대한 상기 시간 오프셋에 관한 차원을 통합함으로써, STSS 특징 맵을 결정하고,

상기 STSS 특징 맵을 상기 비디오 특징 맵에 더한 결과에 기초하여 상기 입력 비디오에 대한 추론을 수행하는,

전자 장치.

청구항 2

제1항에 있어서,

상기 프로세서는

상기 STSS 텐서의 각 위치마다 상기 시간 오프셋에 따라 시간 축으로 인접하는 이웃 프레임들에 대한 복수의 모션 특징들을 나타내는 상기 STSS 특징 벡터들을 결정하는,

전자 장치.

청구항 3

제1항에 있어서,

상기 프로세서는

상기 STSS 텐서에 상기 공간 오프셋과 상기 시간 오프셋에 관한 복수의 3차원 컨볼루션 레이어들을 적용함으로써, 상기 STSS 특징 벡터들을 결정하는,

전자 장치.

청구항 4

제1항에 있어서,

상기 프로세서는

상기 STSS 텐서의 각 위치에 대한 상기 공간 오프셋에 대한 차원을 평탄화(flatten)한 후 MLP(multi-layer perceptron)을 적용함으로써, 상기 STSS 특징 벡터들을 결정하는,

전자 장치.

청구항 5

제1항에 있어서,

상기 프로세서는

상기 STSS 텐서의 각 위치에 대한 상기 공간 오프셋 및 상기 시간 오프셋에 대한 차원들에 soft-argmax를 적용함으로써, 상기 STSS 텐서의 각 위치에 대해 복수의 2차원 모션 특징들을 포함하는 상기 STSS 특징 벡터들을 결정하는,

전자 장치.

청구항 6

제1항에 있어서,

상기 프로세서는

상기 STSS 특징 벡터들에 상기 공간 오프셋 및 상기 시간 오프셋에 대한 차원들 방향으로 인지장(receptive field)이 넓어지는 컨볼루션 레이어들을 적용한 결과를 상기 시간 오프셋에 대한 차원으로 평탄화한 후 1x1x1 컨볼루션 레이어를 적용함으로써, 상기 STSS 특징 맵을 결정하는,

전자 장치.

청구항 7

제1항에 있어서,

상기 프로세서는

상기 STSS 특징 벡터들에 MLP를 적용한 결과를 상기 시간 오프셋에 대한 차원으로 평탄화한 후 1x1x1 컨볼루션 레이어를 적용함으로써, 상기 STSS 특징 맵을 결정하는,

전자 장치.

청구항 8

제1항에 있어서,

상기 STSS 특징 맵은

상기 비디오 특징 맵과 동일한 크기를 가지는,

전자 장치.

청구항 9

제1항에 있어서,

상기 프로세서는

상기 STSS 특징 맵 및 상기 비디오 특징 맵을 요소별 덧셈(element-wise addition)으로 더하는,

전자 장치.

청구항 10

제1항에 있어서,

상기 복수의 공간 교차-유사도 맵은

상기 시간 오프셋에 기초하여 선택된 시간 축으로 인접하는 프레임들로부터 정방향(forward), 역방향(backward), 단기(short-term) 및 장기(long-term)에 대한 모션 정보를 포함하는,

전자 장치.

청구항 11

제1항에 있어서,

상기 입력 비디오에 대한 추론은

상기 입력 비디오에 나타난 행동 및/또는 제스처에 대한 인식 및/또는 분류를 포함하는,

전자 장치.

청구항 12

전자 장치의 동작 방법에 있어서,

입력 비디오에 대응하는 비디오 특징 맵의 각 위치에 대해 하나의 공간 자기-유사도 맵 및 복수의 공간 교차-유사도 맵들을 포함하는 STSS 텐서를 시간 오프셋 및 공간 오프셋에 기초하여 결정하는 동작;

상기 STSS 텐서의 각 위치에 대해 상기 공간 오프셋에 관한 차원을 감소시키고, 상기 시간 오프셋에 관한 차원을 유지시킴으로써, 상기 STSS 텐서로부터 STSS 특징 벡터들을 결정하는 동작;

상기 STSS 특징 벡터들의 각 위치에 대한 상기 시간 오프셋에 관한 차원을 통합함으로써, STSS 특징 맵을 결정하는 동작; 및

상기 STSS 특징 맵을 상기 비디오 특징 맵에 더한 결과에 기초하여 상기 입력 비디오에 대한 추론을 수행하는 동작

을 포함하는

전자 장치의 동작 방법.

청구항 13

제12항에 있어서,

상기 STSS 특징 벡터들을 결정하는 동작은

상기 STSS 텐서의 각 위치마다 상기 시간 오프셋에 따라 시간 축으로 인접하는 이웃 프레임들에 대한 복수의 모션 특징들을 나타내는 상기 STSS 특징 벡터들을 결정하는,

전자 장치의 동작 방법.

청구항 14

제12항에 있어서,

상기 STSS 특징 벡터들을 결정하는 동작은

상기 STSS 텐서에 상기 공간 오프셋과 상기 시간 오프셋에 관한 복수의 3차원 컨볼루션 레이어들을 적용함으로써, 상기 STSS 특징 벡터들을 결정하는,

전자 장치의 동작 방법.

청구항 15

제12항에 있어서,

상기 STSS 특징 벡터들을 결정하는 동작은

상기 STSS 텐서의 각 위치에 대한 상기 공간 오프셋에 대한 차원을 평탄화한 후 MLP을 적용함으로써, 상기 STSS 특징 벡터들을 결정하는,

전자 장치의 동작 방법.

청구항 16

제12항에 있어서,

상기 STSS 특징 벡터들을 결정하는 동작은

상기 STSS 텐서의 각 위치에 대한 상기 공간 오프셋 및 상기 시간 오프셋에 대한 차원들에 soft-argmax를 적용함으로써, 상기 STSS 텐서의 각 위치에 대해 복수의 2차원 모션 특징들을 포함하는 상기 STSS 특징 벡터들을 결정하는,

전자 장치의 동작 방법.

청구항 17

제12항에 있어서,

상기 STSS 특징 맵을 결정하는 동작은

상기 STSS 특징 벡터들에 상기 공간 오프셋 및 상기 시간 오프셋에 대한 차원들 방향으로 인지장(receptive field)이 넓어지는 컨볼루션 레이어들을 적용한 결과를 상기 시간 오프셋에 대한 차원으로 평탄화한 후 1x1x1 컨볼루션 레이어를 적용함으로써, 상기 STSS 특징 맵을 결정하는,

전자 장치의 동작 방법.

청구항 18

제12항에 있어서,

상기 STSS 특징 맵은 상기 비디오 특징 맵과 동일한 크기를 가지고,

상기 STSS 특징 맵 및 상기 비디오 특징 맵은 요소별 덧셈으로 서로 더해지는,

전자 장치의 동작 방법.

청구항 19

제12항에 있어서,

상기 복수의 공간 교차-유사도 맵은

상기 시간 오프셋에 기초하여 선택된 시간 축으로 인접하는 프레임들로부터 정방향, 역방향, 단기 및 장기에 대한 모션 정보를 포함하는,

전자 장치의 동작 방법.

청구항 20

제12항 내지 제20항 중에서 어느 한 항의 방법을 실행하는 컴퓨터 프로그램을 저장하는 컴퓨터 판독가능 기록매체.

발명의 설명

기 술 분 야

[0001] 아래의 개시는 시공간 자기-유사도를 이용하는 전자 장치 및 그 동작 방법에 관한 것이다.

배 경 기 술

[0003] 비디오 내의 3차원 시공간 정보를 학습하는 데에는 3차원 컨볼루션 뉴럴 네트워크(3D CNN)가 주로 쓰이는데, 3차원 컨볼루션 뉴럴 네트워크는 비디오의 여러 프레임 내의 시공간 정보를 학습하기 위해 기존 이미지에서 사용되는 이차원의 컨볼루션 뉴럴 네트워크를 시간 축으로 확장한 것일 수 있다. 또한, 비디오 내의 모션 정보를 학습하기 위해서, 옵티컬 플로우(optical flow)가 사용될 수도 있다. 다만, 3차원 컨볼루션이나 옵티컬 플로우는 그 연산량이 상당하기에 연산 효율성을 높일 수 있는 다양한 연구가 수행되고 있다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0005] 일 실시예에 따른 전자 장치는 프로세서 및 상기 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리를 포함하고, 상기 적어도 하나의 명령어가 상기 프로세서에서 실행되면, 상기 프로세서는 입력 비디오에 대응하는 비디오 특징 맵의 각 위치에 대해 하나의 공간 자기-유사도(spatial self-similarity) 맵 및 복수의 공간 교차-유사도(spatial cross-similarity) 맵들을 포함하는 STSS(spatio-temporal self-similarity) 텐서를 시간 오프셋 및 공간 오프셋에 기초하여 결정하고, 상기 STSS 텐서의 각 위치에 대해 상기 공간 오프셋에 관한 차원을 감소시키고, 상기 시간 오프셋에 관한 차원을 유지시킴으로써, 상기 STSS 텐서로부터 STSS 특징 벡터들을 결정하고, 상기 STSS 특징 벡터들의 각 위치에 대한 상기 시간 오프셋에 관한 차원을 통합함으로써, STSS 특징 맵을 결정하고, 상기 STSS 특징 맵을 상기 비디오 특징 맵에 더한 결과에 기초하여 상기 입력 비디오에 대한 추론을 수행한다.

[0006] 상기 프로세서는 상기 STSS 텐서의 각 위치마다 상기 시간 오프셋에 따라 시간 축으로 인접하는 이웃 프레임들에 대한 복수의 모션 특징들을 나타내는 상기 STSS 특징 벡터들을 결정할 수 있다.

[0007] 상기 프로세서는 상기 STSS 텐서에 상기 공간 오프셋과 상기 시간 오프셋에 관한 복수의 3차원 컨볼루션 레이어들을 적용함으로써, 상기 STSS 특징 벡터들을 결정할 수 있다.

[0008] 상기 프로세서는 상기 STSS 텐서의 각 위치에 대한 상기 공간 오프셋에 대한 차원을 평탄화(flatten)한 후

MLP(multi-layer perceptron)을 적용함으로써, 상기 STSS 특징 벡터들을 결정할 수 있다.

[0009] 상기 프로세서는 상기 STSS 텐서의 각 위치에 대한 상기 공간 오프셋 및 상기 시간 오프셋에 대한 차원들에 soft-argmax를 적용함으로써, 상기 STSS 텐서의 각 위치에 대해 복수의 2차원 모션 특징들을 포함하는 상기 STSS 특징 벡터들을 결정할 수 있다.

[0010] 상기 프로세서는 상기 STSS 특징 벡터들에 상기 공간 오프셋 및 상기 시간 오프셋에 대한 차원들 방향으로 인지장(receptive field)이 넓어지는 컨볼루션 레이어들을 적용한 결과를 상기 시간 오프셋에 대한 차원으로 평탄화한 후 1x1x1 컨볼루션 레이어를 적용함으로써, 상기 STSS 특징 맵을 결정할 수 있다.

[0011] 상기 프로세서는 상기 STSS 특징 벡터들에 MLP를 적용한 결과를 상기 시간 오프셋에 대한 차원으로 평탄화한 후 1x1x1 컨볼루션 레이어를 적용함으로써, 상기 STSS 특징 맵을 결정할 수 있다.

[0012] 상기 STSS 특징 맵은 상기 비디오 특징 맵과 동일한 크기를 가질 수 있다.

[0013] 상기 프로세서는 상기 STSS 특징 맵 및 상기 비디오 특징 맵을 요소별 덧셈(element-wise addition)으로 더할 수 있다.

[0014] 상기 복수의 공간 교차-유사도 맵은 상기 시간 오프셋에 기초하여 선택된 시간 축으로 인접하는 프레임들로부터 정방향(forward), 역방향(backward), 단기(short-term) 및 장기(long-term)에 대한 모션 정보를 포함할 수 있다.

[0015] 상기 입력 비디오에 대한 추론은 상기 입력 비디오에 나타난 행동 및/또는 제스처에 대한 인식 및/또는 분류를 포함할 수 있다.

[0016] 일 실시예에 따른 전자 장치의 동작 방법은 입력 비디오에 대응하는 비디오 특징 맵의 각 위치에 대해 하나의 공간 자기-유사도 맵 및 복수의 공간 교차-유사도 맵들을 포함하는 STSS 텐서를 시간 오프셋 및 공간 오프셋에 기초하여 결정하는 동작, 상기 STSS 텐서의 각 위치에 대해 상기 공간 오프셋에 관한 차원을 감소시키고, 상기 시간 오프셋에 관한 차원을 유지시킴으로써, 상기 STSS 텐서로부터 STSS 특징 벡터들을 결정하는 동작, 상기 STSS 특징 벡터들의 각 위치에 대한 상기 시간 오프셋에 관한 차원을 통합함으로써, STSS 특징 맵을 결정하는 동작 및 상기 STSS 특징 맵을 상기 비디오 특징 맵에 더한 결과에 기초하여 상기 입력 비디오에 대한 추론을 수행하는 동작을 포함한다.

도면의 간단한 설명

[0018] 도 1 내지 도 2는 일 실시예에 따른 전자 장치의 동작을 설명하기 위한 도면이다.

도 3 및 도 4는 일 실시예에 따른 SELFY 블록의 동작을 설명하기 위한 도면이다.

도 5 내지 도 7은 일 실시예에 따른 STSS 텐서를 결정하는 동작을 설명하기 위한 도면이다.

도 8 내지 도 10은 일 실시예에 따른 STSS 특징 벡터들을 결정하는 동작을 설명하기 위한 도면이다.

도 11 및 도 12는 일 실시예에 따른 STSS 특징 맵을 결정하는 동작을 설명하기 위한 도면이다.

도 13은 일 실시예에 따른 전자 장치의 동작 방법을 나타낸 도면이다.

도 14는 일 실시예에 따른 전자 장치를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0019] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0020] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0021] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어

있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0022] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0023] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0024] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.

[0026] 도 1 내지 도 2는 일 실시예에 따른 전자 장치의 동작을 설명하기 위한 도면이다.

[0027] 도 1를 참조하면, 전자 장치(100)는 입력된 비디오에 대해 추론을 수행하고 그 결과를 출력할 수 있다.

[0028] 전자 장치(100)는 컴퓨터 비전 및 기계 학습 기술을 이용한 비디오 행동 인식(video action recognition)을 수행할 수 있다. 비디오 행동 인식은 비디오에 촬영된 사람의 행동을 분석하기 위한 기술로, 예를 들어, 5~10초 정도의 짧은 영상 구간에 나타난 사람의 행동을 인식하고 분석하여 미리 정해진 동작 클래스들 중 하나로 분류할 수 있다. 이후에 상세히 설명하겠으나, 전자 장치(100)는 시공간 자기-유사도(spatio-temporal self-similarity; STSS)를 이용하여 비디오 행동 인식을 수행할 수 있다. 시공간 자기-유사도는 그들 사이의 유사도를 시간 및 공간 차원에서 계산하여 개별 이미지 특징의 관계 구조(relational structure)를 나타낼 수 있다. 전자 장치(100)에서 수행되는 추론 동작이 전술한 예에 한정되지는 않으며, 이외에도 이미지 분류, 비디오 행동 검출, 비디오 객체 트랙킹, 비디오 객체 세그멘테이션 등을 포함할 수 있다.

[0029] 전자 장치(100)는 휴대폰, 스마트 폰, 태블릿, 전자북 장치, 랩탑, 퍼스널 컴퓨터, 또는 서버와 같은 다양한 컴퓨팅 장치, 스마트 시계, 스마트 안경, HMD(Head-Mounted Display), 또는 스마트 의류와 같은 다양한 웨어러블 기기, 스마트 스피커, 스마트 TV, 또는 스마트 냉장고와 같은 다양한 가전장치, 스마트 자동차, 스마트 키오스크, IoT(Internet of Things) 기기, WAD(Walking Assist Device), 드론, 또는 로봇을 포함할 수 있다.

[0030] 또한, 전자 장치(100)는 다양한 비디오 인식 모델에 결합하는 형태로 호환되어, 비디오 인식과 관련된 다양한 장치에 적용될 수 있다. 예를 들어, 전자 장치(100)는 비디오 검색 장치, 비디오 감시 장치, 제스처 인식 장치, 인간-로봇 상호 작용 장치, 자율 주행 및 스포츠 분석 장치 등에 적용될 수 있다.

[0032] 도 2를 참조하면, 전자 장치는 비디오 내 모션 정보를 추출하는 신경 모듈(neural module)인 SELFY 블록(231)을 이용하여 비디오 추론을 수행할 수 있다.

[0033] 전자 장치는 입력 비디오(210)에 컨볼루션 특징 추출(convolution feature extraction)(220)을 수행함으로써 비디오 특징 맵 V을 결정할 수 있다. 모션 특징 추출(motion feature extraction)(230)을 수행하는 네트워크 중간에 SELFY 블록(231)이 삽입되어, 비디오 특징 맵 V로부터 STSS 특징 맵(spatio-temporal self-similarity feature map) Z가 결정될 수 있다. SELFY 블록(231)은 비디오 내 모션 정보를 학습하여 STSS 특징 맵을 결정하는 신경 모듈로서, STSS 변환, STSS 특징 추출, STSS 특징 통합을 순차적으로 수행할 수 있다. SELFY 블록(231)은 비디오 추론을 수행하는 뉴럴 아키텍처에 쉽게 삽입될 수 있고, 추가 감독 없이 종단 간 학습이 가능할 수 있다. SELFY 블록(231)에 대해서는 이후 도면들을 참조하여 상세히 설명한다. STSS 특징 맵 Z은 비디오 특징 맵 V에 더해져서, 다운스트림(downstream) 비디오 처리 뉴럴 네트워크에 통합될 수 있다. STSS 특징 맵 Z 및 비디오 특징 맵 V은 T x H x W x C의 4차원 특징 맵일 수 있다. 전자 장치는 STSS 특징 맵 Z과 비디오 특징 맵 V 간 덧셈 결과에 컨볼루션 특징 추출(240)을 수행하고, 그 결과를 이용하여 최종 행동 추정(250)을 수행할 수 있다.

[0034] 도 3 및 도 4는 일 실시예에 따른 SELFY 블록의 동작을 설명하기 위한 도면이다.

[0035] 도 3을 참조하면, SELFY 블록(300)은 비디오 특징 맵 V(310)를 입력 받아 자기-유사도 변환을 수행하여 STSS 텐서 S(320)로 변환할 수 있다. 또한, SELFY 블록(300)은 특징 추출 동작에 기반하여 STSS 텐서 S(320)로부터 STSS 특징 벡터들 F(330)를 추출할 수 있다. 그리고, SELFY 블록(300)은 특징 통합 동작에 기반하여 STSS 특징 벡터들 F(330)로부터 STSS 특징 맵 Z(340)을 생성할 수 있다. STSS 특징 맵 Z(340)은 요소별 덧셈(element-wise addition)에 의해 비디오 특징 맵 V(310)에 융합됨으로써, SELFY 블록(300)이 잔차 블록(residual block)으로 동작하게 할 수 있다.

[0036] SELFY 블록(300)은 시공간 자기-유사도(STSS) 특징을 학습하여 비디오 처리 인공 신경망의 비디오 표현 능력을 향상시키는 인공 신경 모듈일 수 있다. 자기-유사도(self-similarity)는 각 로컬 영역을 공간적 이웃(spatial neighbors)에 대한 유사도로 표현하여, 내부 구조(intra-structures)를 효과적으로 포착하는 이미지에 대한 관계형 설명자(relational descriptor)일 수 있다. STSS는 외형 특징을 관계형 값(relational values)으로 변환함으로써 학습자(learner)가 공간과 시간의 구조적 패턴을 더 잘 인식할 수 있도록 할 수 있다.

[0037] 도 4를 참조하면, 비디오 특징 맵 V로부터 STSS 특징이 추출되는 동작을 설명하기 위한 예시가 도시된다.

[0038] STSS는 쿼리(410)로 표현되는 각 위치를 공간 및 시간에서 이웃과의 유사도로 설명될 수 있다. STSS를 통해, 동일 프레임에 있는 이웃에 대해서는 공간적 자기-유사도 맵(spatial self-similarity map)이 추출되고, 다른 프레임에 있는 이웃(예: 단기 이웃, 장기 이웃)에 대해서는 움직임 가능성 맵(motion likelihood map)이 추출될 수 있다. STSS는 움직임에 대한 일반화되고 원시적 시각(generalized, far-sighted view)을 가질 수 있다. STSS를 통해 추가 감독 없이 풍부한 모션 표현이 추출될 수 있다.

[0039] 프레임 시퀀스, 즉 비디오가 주어지면, STSS는 공간과 시간에서 이웃과 유사한 각 로컬 영역을 나타낼 수 있다. 외형 특징(appearance features)이 관계형 값(relational values)으로 변환되어, 학습자가 시공간의 구조적 패턴을 더 잘 인식할 수 있도록 할 수 있다. SELFY 블록은 STSS의 전체 볼륨을 활용하여 효과적인 모션 표현을 추출할 수 있다. 공간적, 시간적 측면에서 이웃의 충분한 볼륨으로 비디오에서 장기적인 상호 작용(long-term interaction)과 빠른 움직임(fast motion)을 효과적으로 포착하여 강력한 행동 인식을 수행할 수 있다.

[0040] 도 5 내지 도 7은 일 실시예에 따른 STSS 텐서를 결정하는 동작을 설명하기 위한 도면이다.

[0041] 도 5를 참조하면, 공간 자기-유사도(spatial self-similarity)(510)와 공간 교차-유사도(spatial cross-similarity)(520)를 설명하기 위한 예시가 도시된다.

[0042] 이미지 특징 맵 $\mathbf{I} \in \mathbb{R}^{X \times Y \times C}$ 가 주어지면, I의 자기-유사도 변환(spatial self-similarity transformation)으로 4차원 텐서 $\mathbf{S} \in \mathbb{R}^{X \times Y \times U \times V}$ 가 생성되며, 공간 자기-유사도(510)는 아래의 수학식1로 표현될 수 있다.

수학식 1

[0043] $$\mathbf{S}_{x,y,u,v} = \mathrm{sim}(\mathbf{I}_{x,y}, \mathbf{I}_{x+u,y+v})$$

[0044] 여기서 sim(· , ·)은 유사도 함수(예: 코사인 유사도(cosine similarity))일 수 있다. (x,y)는 쿼리 좌표(query coordinate)이고, (u,v)는 쿼리 좌표로부터의 공간 오프셋(spatial offset)일 수 있다. 지역성(locality)을 부과하기 위해, 오프셋은 이웃(neighborhood) $(u, v) \in [-d_U, d_U] \times [-d_V, d_V]$ 으로 제한될 수 있다. 따라서, U = $2d_U$ +1 및 V = $2d_V$ +1이 각각 될 수 있다. C차원 형태의 특징 $I_{x,y}$를 UV차원 상대적 특징(relational feature) $S_{x,y}$로 변환하여, 외형의 변화를 억제하고 영상의 공간 구조(spatial structures)를 나타낼 수 있다.

[0045] 자기-유사도 변환은 두 개의 서로 다른 특징 맵($\mathbf{I}, \mathbf{I}' \in \mathbb{R}^{X \times Y \times C}$)에 걸친 공간 교차-유사도(520)(또는, 상관 관계(correlation))과 밀접하게 관련될 수 있으며, 아래의 수학식 2로 표현될 수 있다.

수학식 2

[0046] $$\mathbf{S}_{x,y,u,v} = \mathrm{sim}(\mathbf{I}_{x,y}, \mathbf{I}'_{x+u,y+v})$$

[0047] 이미지 특징 맵 I는 쿼리를 포함하는 프레임이고, 이미지 특징 맵 I'은 시간 이웃 프레임일 수 있다. 두 이미지의 움직이는 물체가 주어졌을 때, 교차-유사도 변환은 움직임 정보를 효과적으로 포착할 수 있다.

[0048] 도 6을 참조하면, 비디오 특징 맵 V로부터 STSS 텐서 S를 결정하는 동작을 설명하기 위한 예시가 도시된다.

[0049] SELFY 블록은 비디오 처리 뉴럴 네트워크에 삽입되어, 중간 단계의 비디오 특징 맵 V를 입력 받아 STSS 텐서 S(610)로 변환할 수 있다. 입력된 T 개의 프레임들을 가지는 비디오 특징 맵을 $\mathbf{V} \in \mathbb{R}^{T\times X\times Y\times C}$ 라고 할 때, V에 대한 STSS텐서 $\mathbf{S} \in \mathbb{R}^{T\times X\times Y\times L\times U\times V}$ 는 아래의 수학식 3으로 표현될 수 있다.

수학식 3

[0050] $$\mathbf{S}_{t,x,y,l,u,v} = \mathrm{sim}(\mathbf{V}_{t,x,y}, \mathbf{V}_{t+l,x+u,y+v})$$

[0051] 여기서, (t, x, y)는 쿼리의 좌표이고, (l, u, v)는 쿼리의 시공간 오프셋일 수 있다. sim()은 유사도 함수로서, 예를 들어, 코사인 유사도 함수일 수 있다. 공간 오프셋 u, v의 지역성 외에도 시간 오프셋 l은 시간적 이웃으로 제한될 수 있다. 따라서, 시공간 오프셋 l, u, v는 $(l,u,v) \in [-d_L, d_L] \times [-d_U, d_U] \times [-d_V, d_V]$ 범위로 제한될 수 있다. STSS 텐서 S(610)의 시공간 범위는 $L = 2d_L + 1$, $U = 2d_U + 1$, $V = 2d_V + 1$일 수 있다. 계산된 STSS tensor S는 시간 오프셋 l에 따라 하나의 공간 자기 유사도 맵과 $2d_L$개의 공간 교차-유사도 맵을 포함할 수 있다. 각각의 교차-유사도 맵은 특정 모션에 대한 가능성(likelihood) 정보를 포함할 수 있다. 예를 들어, 작은 시간 오프셋(예: l = -1 또는 +1)을 가진 부분 텐서는 인접한 프레임들에서 모션 정보를 수집하고, 큰 시간 오프셋(예: l = -2 또는 +2 등)을 가진 부분 텐서는 긴 시간 변화에서 모션 정보를 수집할 수 있다. 따라서, STSS 텐서는 정방향, 역방향, 단기, 장기 등의 모션 정보를 포함할 수 있다.

[0052] 도 7를 참조하면, 쿼리에 따라 결정된 STSS 텐서의 예시들이 도시된다. 도 7의 예시에서 첫번째 줄은 비디오에 포함된 연속적인 8개의 프레임들을 나타낼 수 있다. 예를 들어, 시공간 오프셋 (L, U, V)이 (5, 9, 9)인 경우를 가정한다.

[0053] 제1 쿼리(710)에 대해서는, 제1 쿼리(710)를 포함한 기준 프레임과 그 이웃 프레임들이 총 5개에서, 제1 쿼리(710)를 중심으로 하는 부분 이미지 (9, 9)에 대해 제1 STSS 텐서(720)가 결정될 수 있다. 제1 STSS 텐서(720)에서 밝은 부분은 제1 쿼리(710)와 높은 유사도를 나타낼 수 있다.

[0054] 제2 쿼리(730)에 대해서는, 제2 쿼리(730)를 포함한 기준 프레임과 그 이웃 프레임들이 총 5개에서, 제2 쿼리(730)를 중심으로 하는 부분 이미지 (9, 9)에 대해 제2 STSS 텐서(740)가 결정될 수 있다. 마찬가지로, 제2 STSS 텐서(740)에서 밝은 부분은 제2 쿼리(730)와 높은 유사도를 나타낼 수 있다.

[0055] 이처럼, 비디오 특징 맵의 각 위치에 대해 STSS 텐서가 결정될 수 있다.

[0056] 도 8 내지 도 10은 일 실시예에 따른 STSS 특징 벡터들을 결정하는 동작을 설명하기 위한 도면이다.

[0057] STSS 텐서 $\mathbf{S} \in \mathbb{R}^{T\times X\times Y\times L\times U\times V}$에서 각 시공간 위치(spatio-temporal position) (t,x,y) 및 시간 오프셋 l에 대한 C_F 차원 특징이 추출되어, STSS 특징 벡터들 $\mathbf{F} \in \mathbb{R}^{T\times X\times Y\times L\times C_F}$이 결정될 수 있다. 이는 공간 및 시간 오프셋에서의 변환에 기반하여 수행될 수 있다. STSS 텐서 S에서 STSS 특징 벡터들 F로 변환될 때, L 차원은 서로 다른 시간 오프셋에 걸쳐 모션 정보를 추출하기 위해 유지될 수 있다. STSS 특징 벡터들 F로의 변환은 컨볼루션, MLP, soft-argmax 중 어느 하나에 기반하여 수행될 수 있으며, 이에 대해서는 도 8 내

지 도 10을 통해 상세히 설명한다.

[0058] 도 8을 참조하면, 컨볼루션에 기반하여 STSS 텐서 S로부터 STSS 특징 벡터들 F가 결정될 수 있다.

[0059] STSS 특징 추출 동작은 앞서 계산한 $\mathbf{S} \in \mathbb{R}^{T\times X\times Y\times L\times U\times V}$ 에서 각 위치 (t, x, y) 마다 C_F차원의 STSS 특징 벡터를 추출하여 결과적으로 $\mathbf{F} \in \mathbb{R}^{T\times X\times Y\times L\times C_F}$ 형태의 텐서를 형성할 수 있다. STSS 특징 추출을 위하여, STSS 내의 구조적인 패턴을 효과적으로 학습 가능한 컨볼루션 레이어 $g(\cdot)$가 이용될 수 있다. 다시 말해, STSS 텐서 S의 (L, U, V) 볼륨에 대한 컨볼루션 커널이 학습될 수 있다. STSS 텐서 S는 $\mathbf{S} \in \mathbb{R}^{T\times X\times Y\times L\times U\times V\times 1}$ 형태의 7차원 텐서라고 할 때, $\mathbf{K}_c \in \mathbb{R}^{1\times 1\times 1\times L_n\times U_n\times V_n\times C\times C'}$ 의 다중 채널 컨볼루션 커널을 가지는 컨볼루션 레이어 $g(\cdot)$는 다음의 수학식 4처럼 표현될 수 있다.

수학식 4

[0060] $$g(\mathbf{S}) = \mathrm{ReLU}(\mathrm{Conv}(\mathbf{S}, \mathbf{K}_c))$$

[0061] 다층의 컨볼루션 레이어들에 STSS 텐서 S를 통과시키는 과정에서 점진적으로 (U, V) 차원이 줄어들면서, 결과적으로, $\mathbf{F} \in \mathbb{R}^{T\times X\times Y\times L\times C_F}$ 형태의 STSS 특징 벡터들 F가 생성될 수 있다. (U, V) 차원과 다르게, L차원은 감소되지 않고, 일정하게 유지시킴으로써 시간 정보 손실이 최소화될 수 있다.

[0062] 다시 말해, $\mathbb{R}^{T\times X\times Y\times L\times U\times V\times 1}$부터 점진적으로 다운샘플링 ($U$,$V$)하고 스트라이드(strides)가 있는 다중 컨볼루션을 통해 채널을 확장하여 최종적으로 $\mathbb{R}^{T\times X\times Y\times L\times 1\times 1\times C_F}$ 가 결정될 수 있다. 미세한 시간 해상도(fine temporal resolution)를 유지하는 것이 자세한 모션 정보를 캡처하는 데 효과적인일 수 있기 때문에, L 차원을 그대로 유지할 수 있다. STSS 텐서 S는 재구성(reshape)된 다음 S의 (l, u, v) 차원을 따라 3D 컨볼루션이 적용될 수 있다. N개의 컨볼루션 레이어들을 이용하는 STSS 특징 추출 과정은 아래와 같이 표현될 수 있다.

수학식 5

[0063] $$\mathbf{F} = (g_n \circ g_{n-1} \circ \cdots \circ g_1)(\mathbf{S})$$

[0064] 여기서, $\circ$ 는 합성 함수(composite function)를 나타낼 수 있다. 컨볼루션 기법은 이후에 설명할 MLP에 비해 STSS 텐서 내 구조적인 패턴을 학습하는 데 더 뛰어날 수 있다.

[0065] 도 9를 참조하면, MLP에 기반하여 STSS 텐서 S로부터 STSS 특징 벡터들 F가 결정될 수 있다.

[0066] STSS 특징 추출 동작은 자기-유사도 값을 특징으로 변환하는 MLP를 이용하여 수행될 수 있다. 이를 위해, 전자 장치는 STSS 텐서 S의 (U,V) 볼륨을 UV 차원 벡터로 평탄화(flatten)한 후 MLP를 적용할 수 있다. 재구성(reshape)된 텐서 $\mathbf{S} \in \mathbb{R}^{T\times X\times Y\times L\times UV}$ 에 적용되는 퍼셉트론 $f(\cdot)$는 아래의 수학식으로 표현될 수 있다.

수학식 6

[0067] $$f(\mathbf{S}) = \mathrm{ReLU}(\mathbf{S} \times_5 \mathbf{W}_\phi)$$

[0068] 여기서 x_5은 n-모드 텐서 곱(n-mode tensor product)을 나타내고, $\mathbf{W}_\phi \in \mathbb{R}^{C' \times UV}$는 퍼셉트론 파라미터이고, 출력은 $f(\mathbf{S}) \in \mathbb{R}^{T \times X \times \bar{Y} \times L \times \bar{C}'}$일 수 있다. 따라서, MLP에 기반한 특징 추출은 아래의 수학식으로 표현될 수 있다.

수학식 7

[0069] $$\mathbf{F} = (f_n \circ f_{n-1} \circ \cdots \circ f_1)(\mathbf{S})$$

[0070] 위의 수학식7로 특징 텐서 $\mathbf{F} \in \mathbb{R}^{T \times X \times Y \times L \times C_F}$가 계산될 수 있다. 이러한 방법은 변위 정보(displacement information)를 인코딩할 수 있을 뿐만 아니라, 유사도 값(similarity values)을 직접 액세스할 수 있기 때문에, 이후에 설명할 soft-argmax에 기반한 특징 추출보다 유연하고 효과적일 수 있다. MLP에 기반한 특징 추출은 모션 분포(motion distribution)를 학습하는 데 유용할 수 있다.

[0071] 도 10을 참조하면, soft-argmax에 기반하여 STSS 텐서 S로부터 STSS 특징 벡터들 F가 결정될 수 있다.

[0072] soft-argmax에 기반한 특징 추출은 공간 교차-유사도를 사용하여 명시적 변위 필드(explicit displacement fields)를 계산할 수 있다. argmax(u,v)는 유사도 값이 가장 높은 위치를 인덱싱하여 변위 필드를 추출할 수 있지만 미분이 불가능하여 뉴럴 네트워크의 역전파(back propagation) 시 그라디언트(gradient)를 전달하지 못할 수 있다. soft-argmax는 softmax 가중치로 변위 벡터(displacement vectors)를 집계(aggregate)할 수 있으므로, argmax와 달리 뉴럴 네트워크의 역전파 시 그라디언트를 전달할 수 있다. soft-argmax에 기반한 특징 추출은 아래의 수학식으로 표현될 수 있다.

수학식 8

[0073] $$\mathbf{F}_{t,x,y,l} = \sum_{u,v} \frac{\exp(\mathbf{S}_{t,x,y,l,u,v}/\tau)}{\sum_{u',v'} \exp(\mathbf{S}_{t,x,y,l,u',v'}/\tau)} [u; v]$$

[0074] 위의 수학식을 통해, STSS 특징 벡터들 $\mathbf{F} \in \mathbb{R}^{T \times X \times Y \times L \times 2}$가 생성될 수 있다. 온도 계수(temperature factor) τ는 softmax 분포를 조정할 수 있으며, 예를 들어, τ = 0.01일 수 있다.

[0075] 도 11 및 도 12는 일 실시예에 따른 STSS 특징 맵을 결정하는 동작을 설명하기 위한 도면이다.

[0076] 전자 장치는 STSS 특징 벡터들 F의 각 위치에 대한 시간 오프셋에 관한 차원을 통합함으로써, STSS특징 맵 Z을 결정할 수 있다. STSS 특징 맵 Z는 (T, X, Y, C) 볼륨을 가져 원래 스트림으로 피드백될 수 있다. STSS 특징 벡터들 F로부터 STSS 특징 맵 Z로의 변환은 컨볼루션, MLP 중 어느 하나에 기반하여 수행될 수 있으며, 이에 대해서는 도 11 및 도 12를 통해 상세히 설명한다.

[0077] 도 11을 참조하면, 컨볼루션에 기반하여 STSS 특징 벡터들 F로부터 STSS 특징 맵 Z가 결정될 수 있다.

[0078] STSS 특징 통합 동작에서는 STSS 특징 벡터들 $\mathbf{F} \in \mathbb{R}^{T \times X \times Y \times L \times C_F}$을 통합하여 STSS 특징 맵 $\mathbf{Z} \in \mathbb{R}^{T \times X \times Y \times L \times C}$이 생성될 수 있다. 이때, 컨볼루션 레이어 $h(\cdot)$가 이용될 수 있고, $\mathbf{K_i} \in \mathbb{R}^{T_n \times X_n \times Y_n \times 1 \times C_F \times C_F'}$ 형태의 다중 채널 커널을 가지는 컨볼루션 레이어는 아래와 같이 표현될 수 있다.

수학식 9

[0079] $$h(\mathbf{S}) = \mathrm{ReLU}(\mathrm{Conv}(\mathbf{F}, \mathbf{K}_i))$$

[0080] 이러한 형태의 컨볼루션 레이어 $h(\cdot)$는 (t, x, y) 차원 방향으로 인지장(receptive field)를 넓히면서 STSS 특징 벡터들 F를 효과적으로 통합할 수 있다. n층의 컨볼루션 레이어들을 통과하여 생성되는 특징 맵 $\mathbf{F}^* \in \mathbb{R}^{T\times X\times Y\times L\times C'_{\mathrm{F}}}$은 아래와 같이 계산될 수 있다.

수학식 10

[0081] $$\mathbf{F}^* = (h_n \circ h_{n-1} \circ \cdots \circ h_1)(\mathbf{F})$$

[0082] 계산된 F*의 (L, C_{F}) 행렬을 LC_{F} 차원의 벡터로 평탄화한 후 1x1x1컨볼루션 레이어를 적용하여 최종적인 STSS 특징 맵 Z가 생성될 수 있다. 이때 1x1x1 컨볼루션 레이어는 다른 시간 오프셋 정보를 통합하고, STSS 특징 맵 Z가 비디오 특징 맵 V에 더해질 수 있도록 채널 차원의 크기를 조절하는 역할을 수행할 수 있다. STSS 특징 맵 Z은 아래의 수학식으로 표현될 수 있다.

수학식 11

[0083] $$\mathbf{Z} = \mathrm{ReLU}(\mathbf{F}^* \times_4 \mathbf{W}_\theta)$$

[0084] 위의 수학식에서, $\times_4$는 n-모드 텐서 곱을 나타내고, $\mathbf{W}_\theta \in \mathbb{R}^{C\times LC^*_{\mathrm{F}}}$은 1x1x1 컨볼루션 레이어의 커널일 수 있다. 최종적으로 계산된 STSS 특징 맵 Z는 비디오 특징 맵 V에 요소별(element-wise)로 더해짐으로써, SELFY 블록이 모션 학습을 위한 잔차 블록처럼 동작할 수 있다.

[0085] 도 12를 참조하면, MLP에 기반하여 STSS 특징 벡터들 F로부터 STSS 특징 맵 Z가 결정될 수 있다. STSS 특징 통합 동작은 MLP를 이용하여 수행될 수 있다. 전자 장치는 STSS 특징 벡터들에 MLP를 적용한 결과의 (L, C_{F}) 행렬을 LC_{F} 차원의 벡터로 평탄화한 후 1x1x1컨볼루션 레이어를 적용하여 최종적인 STSS 특징 맵 Z를 생성할 수 있다. 이때 1x1x1 컨볼루션 레이어는 다른 시간 오프셋 정보를 통합하고, STSS 특징 맵 Z가 비디오 특징 맵 V에 더해질 수 있도록 채널 차원의 크기를 조절하는 역할을 수행할 수 있다. 최종적으로 계산된 STSS 특징 맵 Z는 비디오 특징 맵 V에 요소별로 더해져서, SELFY 블록이 모션 학습을 위한 잔차 블록처럼 동작할 수 있다.

[0086] 도 13은 일 실시예에 따른 전자 장치의 동작 방법을 나타낸 도면이다.

[0087] 이하 실시예에서 각 동작들은 순차적으로 수행될 수도 있으나, 반드시 순차적으로 수행되는 것은 아니다. 예를 들어, 각 동작들의 순서가 변경될 수도 있으며, 적어도 두 동작들이 병렬적으로 수행될 수도 있다. 동작들(1310~1340)은 전자 장치의 적어도 하나의 구성요소(예: 프로세서 등)에 의해 수행될 수 있다.

[0088] 동작(1310)에서, 전자 장치는 입력 비디오에 대응하는 비디오 특징 맵의 각 위치에 대해 하나의 공간 자기-유사도(spatial self-similarity) 맵 및 복수의 공간 교차-유사도(spatial cross-similarity) 맵들을 포함하는 STSS(spatio-temporal self-similarity) 텐서를 시간 오프셋 및 공간 오프셋에 기초하여 결정한다.

[0089] 동작(1320)에서, 전자 장치는 STSS 텐서의 각 위치에 대해 공간 오프셋에 관한 차원을 감소시키고, 시간 오프셋에 관한 차원을 유지시킴으로써, STSS 텐서로부터 STSS 특징 벡터들을 결정한다. 전자 장치는 STSS 텐서의 각 위치마다 시간 오프셋에 따라 시간 축으로 인접하는 이웃 프레임들에 대한 복수의 모션 특징들을 나타내는 STSS 특징 벡터들을 결정할 수 있다. 예를 들어, 전자 장치는 STSS 텐서에 공간 오프셋과 시간 오프셋에 관한 복수의 3차원 컨볼루션 레이어들을 적용함으로써, STSS 특징 벡터들을 결정할 수 있다. 이외에도 전자 장치는 MLP 또는 soft-argmax를 이용하여 STSS 특징 벡터들을 결정할 수도 있다.

[0090] 동작(1330)에서, 전자 장치는 STSS 특징 벡터들의 각 위치에 대한 시간 오프셋에 관한 차원을 통합함으로써, STSS 특징 맵을 결정한다. 예를 들어, 전자 장치는 STSS 특징 벡터들에 공간 오프셋 및 시간 오프셋에 대한 차원들 방향으로 인지장(receptive field)이 넓어지는 컨볼루션 레이어들을 적용한 결과를 시간 오프셋에 대한 차원으로 평탄화한 후 1x1x1 컨볼루션 레이어를 적용함으로써, STSS 특징 맵을 결정할 수 있다. 이외에도 전자 장치는 MLP를 이용하여 STSS 특징 맵을 결정할 수도 있다.

[0091] 동작(1340)에서, 전자 장치는 STSS 특징 맵을 비디오 특징 맵에 더한 결과에 기초하여 입력 비디오에 대한 추론을 수행한다. STSS 특징 맵은 비디오 특징 맵과 동일한 크기를 가질 수 있다. 전자 장치는 STSS 특징 맵 및 비디오 특징 맵을 요소별 덧셈으로 더할 수 있다. 입력 비디오에 대한 추론은 입력 비디오에 나타난 행동 및/또는 제스처에 대한 인식 및/또는 분류를 포함할 수 있다.

[0092] 도 13에 도시된 각 동작들에는 도 1 내지 도 12를 통해 전술한 사항들이 그대로 적용되므로, 보다 상세한 설명은 생략한다.

[0093] 도 14는 일 실시예에 따른 전자 장치를 나타낸 도면이다.

[0094] 도 14를 참조하면, 일 실시예에 따른 전자 장치(1400)는 메모리(1410) 및 프로세서(1420)를 포함한다. 메모리(1410) 및 프로세서(1420)는 버스 (bus), PCIe(Peripheral Component Interconnect Express) 및/또는 NoC(Network on a Chip) 등을 통하여 서로 통신할 수 있다.

[0095] 메모리(1410)는 컴퓨터에서 읽을 수 있는 명령어를 포함할 수 있다. 프로세서(1420)는 메모리(1410)에 저장된 명령어가 프로세서(1420)에서 실행됨에 따라 앞서 언급된 동작들을 수행할 수 있다. 메모리(1410)는 휘발성 메모리 또는 비휘발성 메모리일 수 있다.

[0096] 프로세서(1420)는 명령어들, 혹은 프로그램들을 실행하거나, 전자 장치(1400)를 제어하는 장치로서, 예를 들어, CPU(Central Processing Unit) 및/또는 GPU(Graphic Processing Unit), 가속기 등을 포함할 수 있다. 프로세서(1420)는 입력 비디오에 대응하는 비디오 특징 맵의 각 위치에 대해 하나의 공간 자기-유사도 맵 및 복수의 공간 교차-유사도 맵들을 포함하는 STSS 텐서를 시간 오프셋 및 공간 오프셋에 기초하여 결정하고, STSS 텐서의 각 위치에 대해 공간 오프셋에 관한 차원을 감소시키고, 시간 오프셋에 관한 차원을 유지시킴으로써, STSS 텐서로부터 STSS 특징 벡터들을 결정하고, STSS 특징 벡터들의 각 위치에 대한 시간 오프셋에 관한 차원을 통합함으로써, STSS 특징 맵을 결정하고, STSS 특징 맵을 비디오 특징 맵에 더한 결과에 기초하여 입력 비디오에 대한 추론을 수행한다.

[0097] 그 밖에, 전자 장치(1400)에 관해서는 상술된 동작을 처리할 수 있다.

[0098] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0099] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0100] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 저장할 수 있으며 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

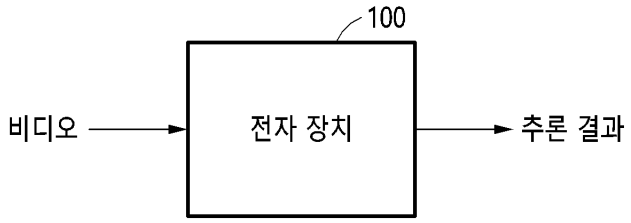
[0101] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0102] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

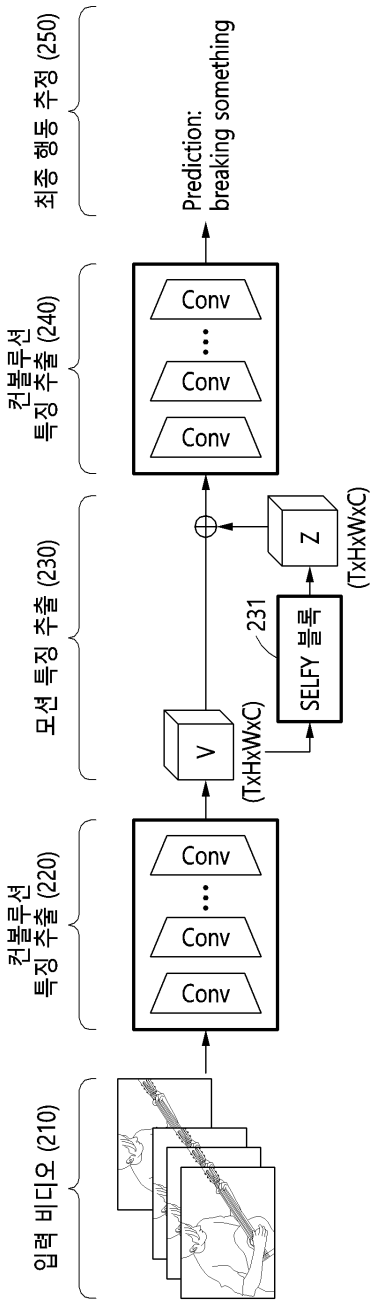
[0103] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

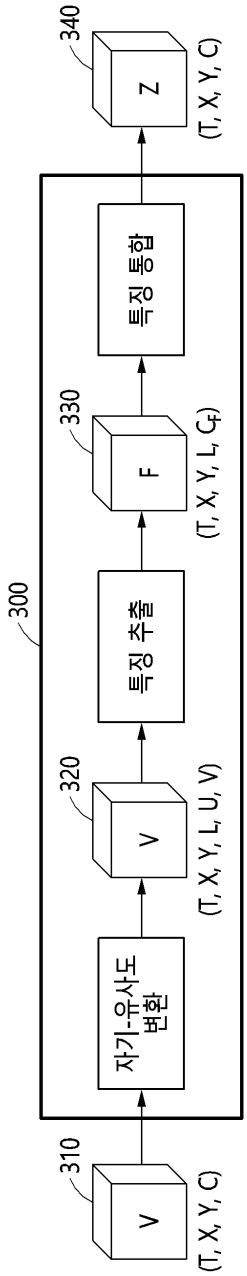
도면1



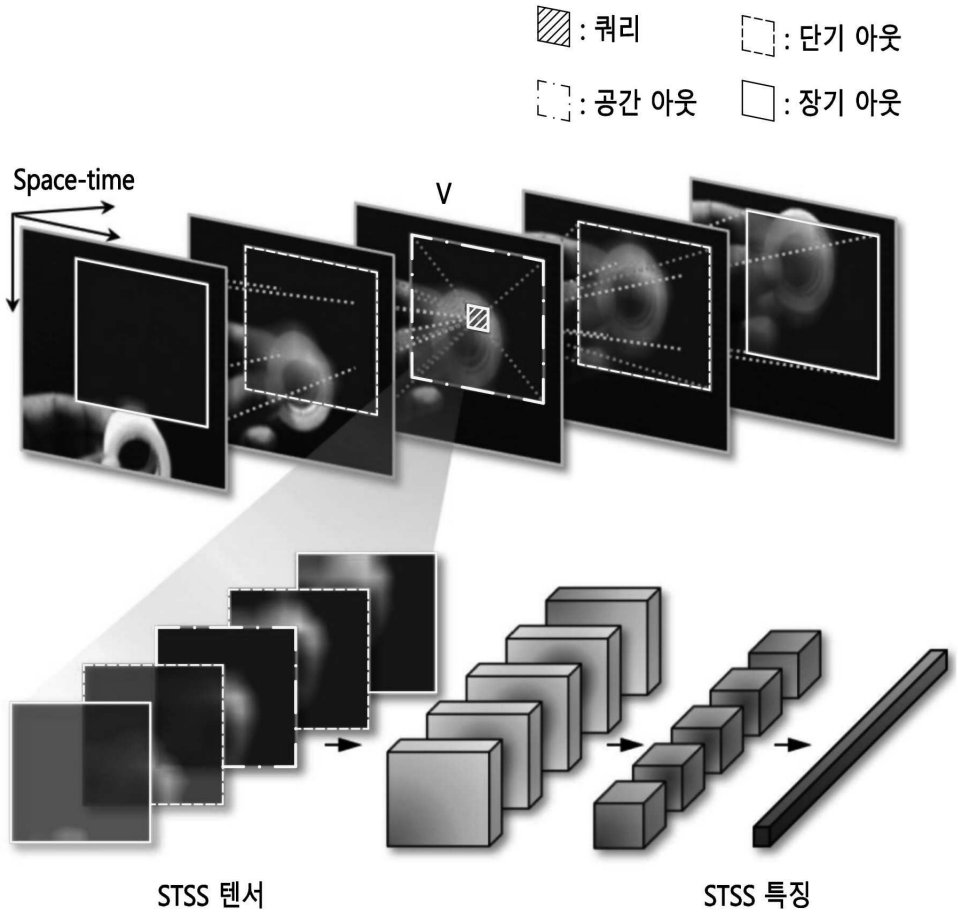
도면2


입력 비디오 (210)
컨볼루션 특징 추출 (220)
Conv
Conv
Conv
모션 특징 추출 (230)
V
(TxHxWxC)
231
SELFY 블록
Z
(TxHxWxC)
컨볼루션 특징 추출 (240)
Conv
Conv
Conv
최종 행동 추정 (250)
Prediction:
breaking something

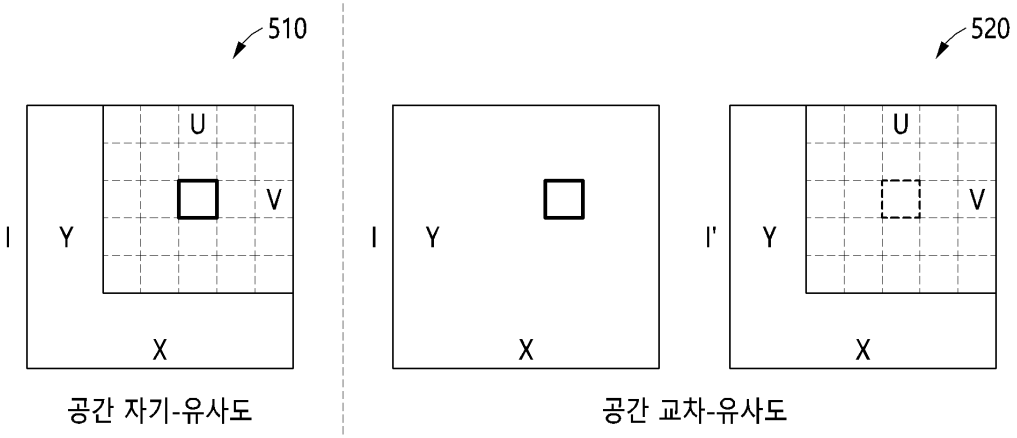
도면3


340
Z
(T, X, Y, C)
330
F
(T, X, Y, L, C_F)
특징 통합
특징 추출
320
V
(T, X, Y, L, U, V)
차기-유사도 변환
300
310
V
(T, X, Y, C)

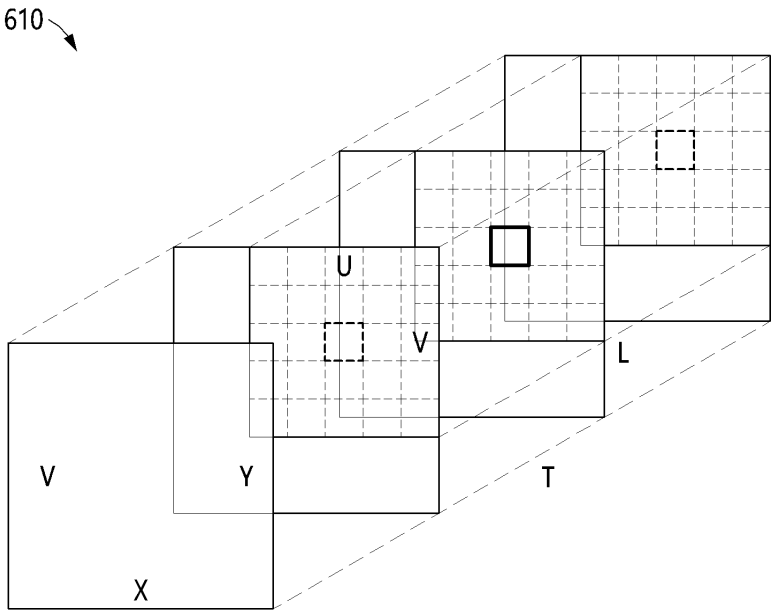
도면4


: 쿼리
: 단기 아웃
: 공간 아웃
: 장기 아웃
Space-time
V
STSS 텐서
STSS 특징

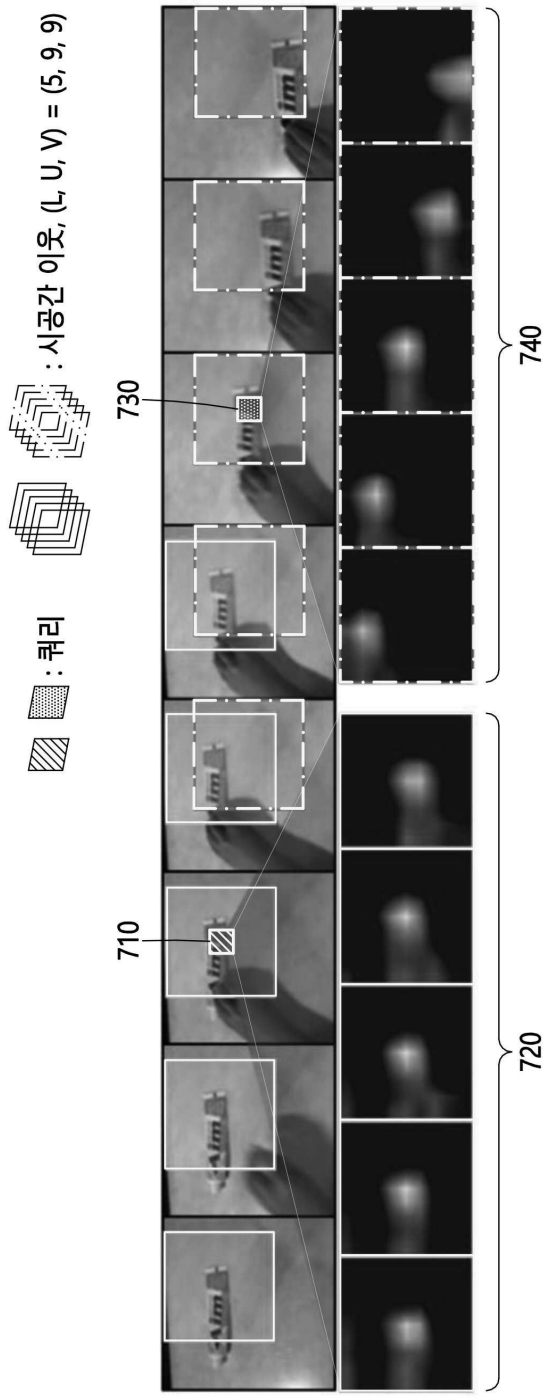
도면5


510
520
U
V
I
Y
X
I'
공간 자기-유사도
공간 교차-유사도

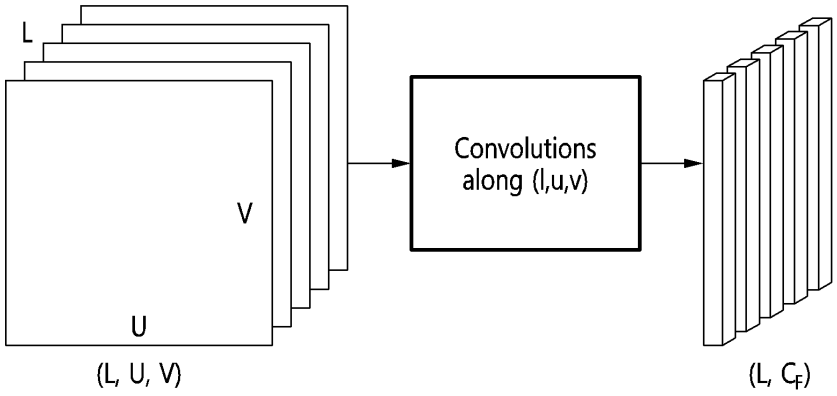
도면6



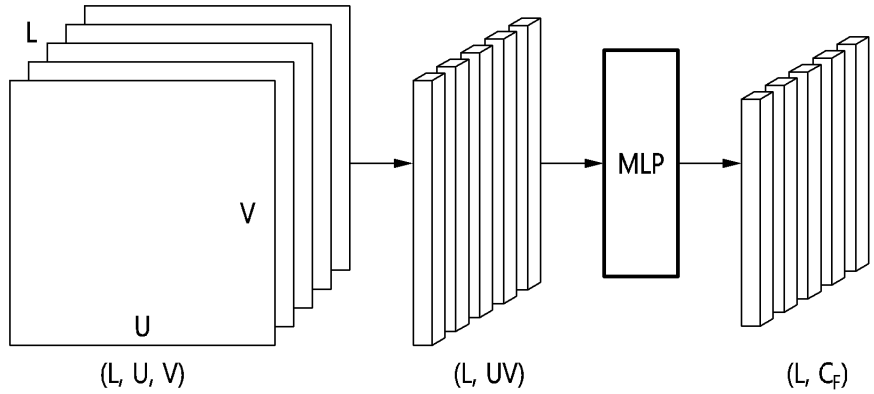
도면7



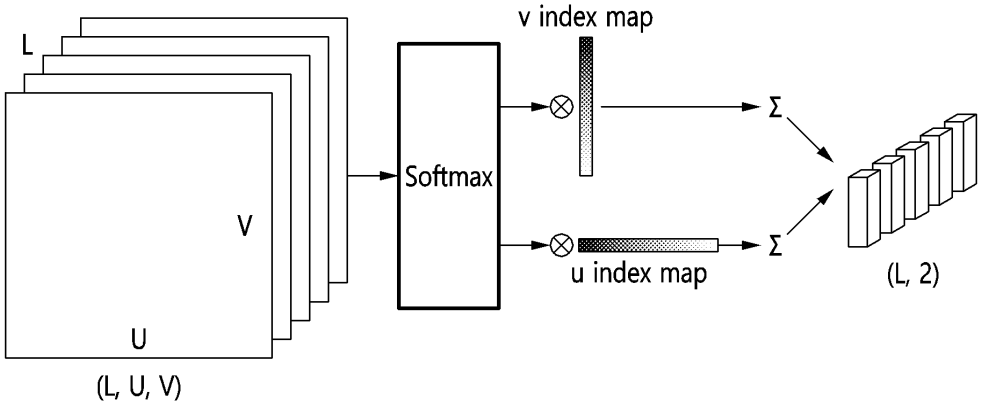
도면8


L
V
U
(L, U, V)
Convolutions
along (l,u,v)
(L, C_f)

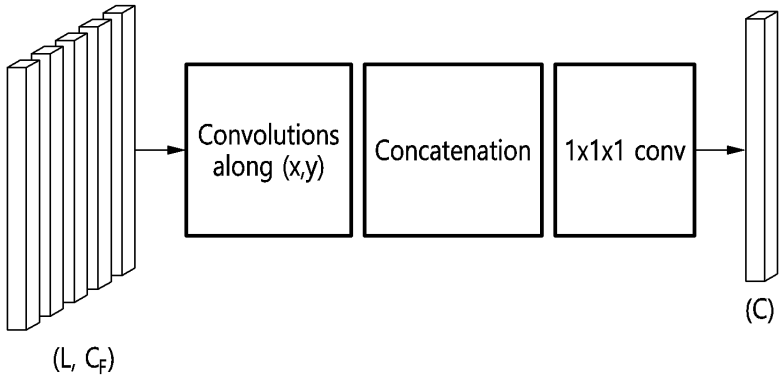
도면9


L
V
U
(L, U, V)
(L, UV)
MLP
(L, C_f)

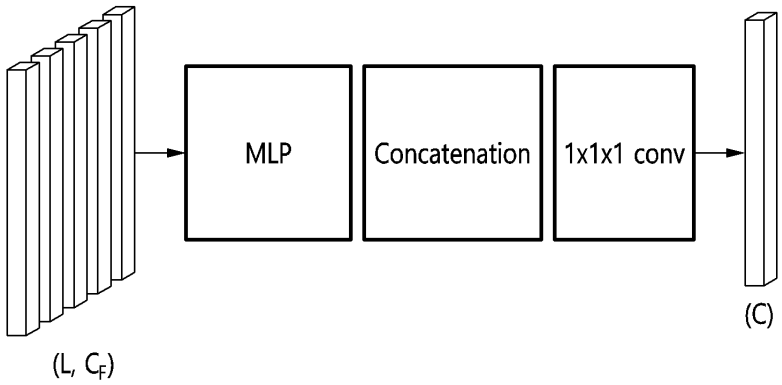
도면10


L
V
U
(L, U, V)
Softmax
v index map
u index map
Σ
Σ
(L, 2)

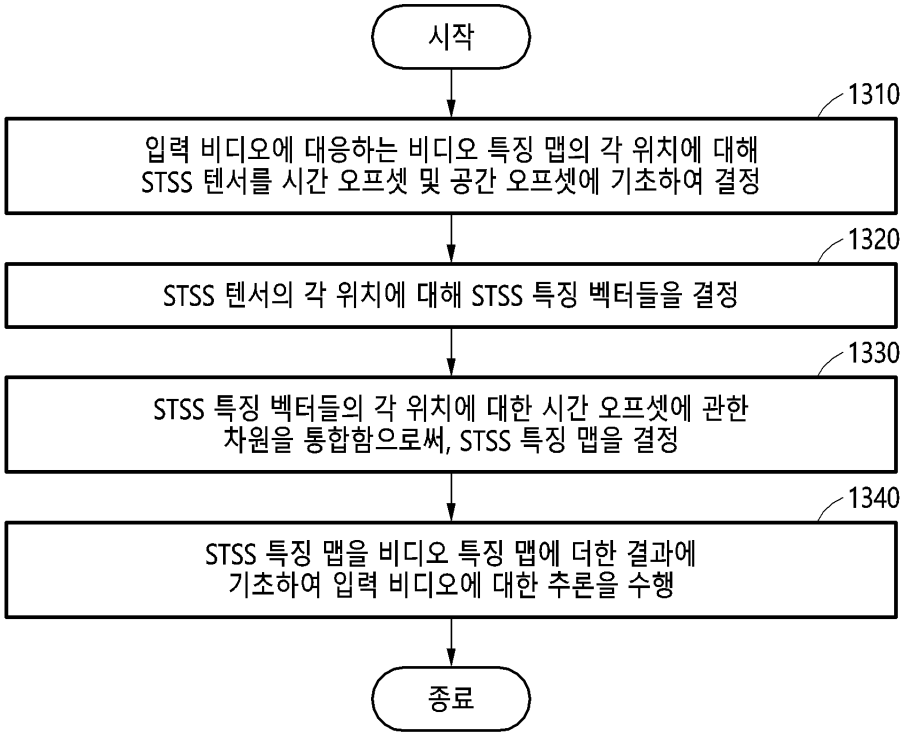
도면11


Convolutions along (x,y)
Concatenation
1x1x1 conv
(C)
(L, C_F)

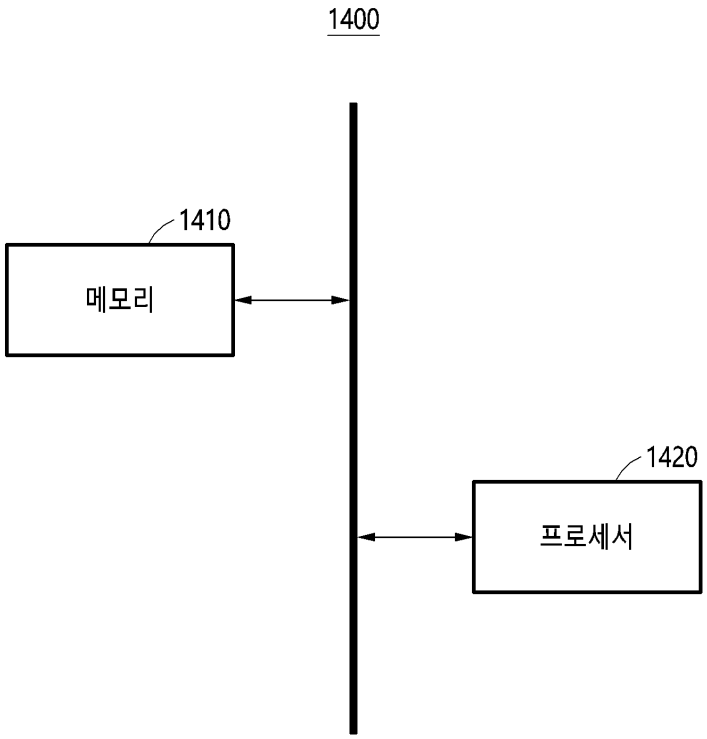
도면12


MLP
Concatenation
1x1x1 conv
(C)
(L, C_F)

도면13


시작
1310
입력 비디오에 대응하는 비디오 특징 맵의 각 위치에 대해
STSS 텐서를 시간 오프셋 및 공간 오프셋에 기초하여 결정
1320
STSS 텐서의 각 위치에 대해 STSS 특징 벡터들을 결정
1330
STSS 특징 벡터들의 각 위치에 대한 시간 오프셋에 관한
차원을 통합함으로써, STSS 특징 맵을 결정
1340
STSS 특징 맵을 비디오 특징 맵에 더한 결과에
기초하여 입력 비디오에 대한 추론을 수행
종료

도면14


1400
1410
메모리
1420
프로세서