(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0090069
(43) 공개일자 2024년06월21일

(51) 국제특허분류(Int. Cl.)
G06V 40/20 (2022.01) *G06V 10/30* (2022.01)
G06V 10/44 (2022.01) *G06V 10/82* (2022.01)
(52) CPC특허분류
G06T 2207/20044 (2013.01)
(21) 출원번호 10-2022-0173809
(22) 출원일자 2022년12월13일
심사청구일자 없음

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
강석주
경기도 고양시 일산서구 후곡로 36, 403동 1003호 (일산3동, 후곡마을4단지아파트)
김현성
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
김기남
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(74) 대리인
유미특허법인

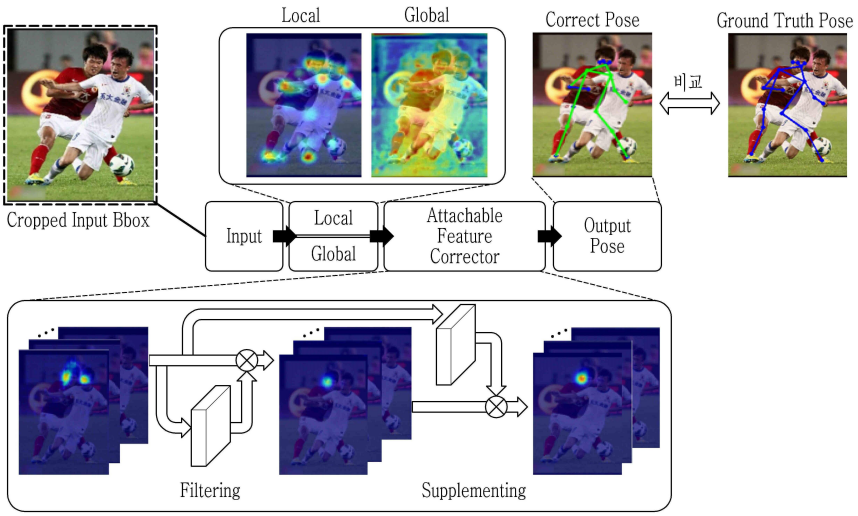
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **인공 신경망을 이용한 인간 자세 추정 장치 및 그 방법**

(57) 요 약

본 개시에 따르면, 인간 자세 추정 장치는 적어도 한 명의 인간이 포함된 입력 이미지로부터 개별 신체 부위들에 대한 지역 특징과 전체 신체에 대한 전역 특징을 추출하는 특징 추출부, 그리고 상기 입력 이미지에서 타겟 인간과 관련된 키포인트가 포함된 히트맵을 출력하기 위해, 상기 지역 특징과 상기 전역 특징을 결합한 입력 특징에 대해 불필요한 영역을 삭제하고, 필수적인 영역을 추가하는 특징 보정부를 포함할 수 있다.

대 표 도



이 발명을 지원한 국가연구개발사업
과제고유번호 1711152894
과제번호 2021-0-02308-001
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 인공지능산업원천기술개발(R&D)
연구과제명 지역 사회, 생활 문제 해결을 위한 멀티모달 환경에서의 엣지 컴퓨터 기반 심층신경망 기술 연구
기 여 율 1/2
과제수행기관명 서강대학교 산학협력단
연구기간 2022.01.01~2022.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 22PQWO-C153369-04
과제번호 22PQWO-C153369-04
부처명 다부처 (행정안전부,산업통상자원부,과학기술정보통신부,국토교통부)
과제관리(전문)기관명 국토교통과학기술진흥원
연구사업명 스마트 도로조명 플랫폼 개발 및 실증연구
연구과제명 스마트 도로조명 활용 도시재난안전관리 연계 기술 개발
기 여 율 1/2
과제수행기관명 서강대학교 산학협력단
연구기간 2022.01.01~2022.12.31

명 세 서

청구범위

청구항 1

적어도 한 명의 인간이 포함된 입력 이미지로부터 개별 신체 부위들에 대한 지역 특징과 전체 신체에 대한 전역 특징을 추출하는 특징 추출부, 그리고

상기 입력 이미지에서 타겟 인간과 관련된 키포인트가 포함된 히트맵을 출력하기 위해, 상기 지역 특징과 상기 전역 특징을 결합한 입력 특징에 대해 불필요한 영역을 삭제하고, 필수적인 영역을 추가하는 특징 보정부

를 포함하는, 인간 자세 추정 장치.

청구항 2

제1항에서,

상기 특징 보정부는,

복수개의 잔차 블록으로 구성된 잔차 신경망으로서, 상기 잔차 신경망을 통해 상기 입력 특징에서 불필요한 영역을 삭제한 필터링 특징을 출력하는 특징 필터링 모듈

을 포함하는, 인간 자세 추정 장치.

청구항 3

제2항에서,

상기 특징 필터링 모듈은,

순차적으로 동작하는 상기 복수개의 잔차 블록과 시그모이드 함수(sigmoid function)를 이용하여 상기 입력 특징에서 불필요한 영역을 제거한 필터링 특징을 출력하는, 인간 자세 추정 장치.

청구항 4

제3항에서,

상기 특징 보정부는,

상기 복수개의 잔차 블록을 이용하여, 상기 입력 특징에서 필요한 영역을 추가한 보충 특징을 추출하는 특징 보충 모듈

을 더 포함하는, 인간 자세 추정 장치.

청구항 5

제4항에서,

상기 복수개의 잔차 블록은,

컨벌루션 레이어(Convolution Layer), 그리고 어텐션 모듈인 NORM, RELU를 포함하는, 인간 자세 추정 장치.

청구항 6

제4항에서,

상기 특징 보정부는,

상기 입력 특징에 상기 필터링 특징을 요소별 곱셈 연산(element-wise multiplication)하고, 요소별 곱셈 연산한 결과에 상기 보충 특징을 요소별 합산 연산(element-wise summation)하는 연산부

를 포함하는, 인간 자세 추정 장치.

청구항 7

제6항에서,

상기 연산부의 연산 결과를 컨벌루션 레이어에 통과시켜 상기 히트맵을 추출하는 히트맵 추출부

를 더 포함하는, 인간 자세 추정 장치.

청구항 8

제1항에서,

상기 특징 추출부는,

트랜스포머 기반 신경망을 이용하여, 상기 입력 이미지로부터 전체 신체에 대한 전역 특징을 추출하는 전역 특징 추출부

를 포함하는, 인간 자세 추정 장치.

청구항 9

제8항에서,

상기 특징 추출부는,

컨벌루션(convolution) 신경망을 이용하여, 상기 입력 이미지로부터 이미지 특징을 추출하고, 컨벌루션 연산을 통해 상기 이미지 특징으로부터 타겟 인간의 키포인트와 관련된 지역 특징을 추출하는 지역 특징 추출부

를 더 포함하는, 인간 자세 추정 장치.

청구항 10

제9항에서,

상기 전역 특징 추출부는,

상기 이미지 특징과 위치 인코딩을 입력으로 사용하여, 상기 이미지 특징을 전역적으로 이해하기 위한 인코딩을 수행하고, 인코딩 결과를 사용하여 입력 임베딩을 출력 임베딩으로 변환하는 디코딩을 수행하며, 디코딩 결과에 중선형보간법(Bilinear interpolation)을 적용하여 상기 전역 특징을 추출하는, 인간 자세 추정 장치.

청구항 11

제9항에서,

상기 지역 특징 추출부는,

상기 컨벌루션 신경망을 이용하여, 상기 입력 이미지로부터 이미지 특징 및 고해상도 특징을 추출하고,

상기 특징 보정부는,

상기 고해상도 특징, 상기 지역 특징 및 상기 전역 특징을 결합한 상기 입력 특징을 이용하는, 인간 자세 추정 장치.

청구항 12

적어도 하나의 프로세서에 의해 동작하는 인간 자세 추정 장치의 동작 방법으로서,

적어도 한 명의 인간이 포함된 입력 이미지로부터 개별 신체 부위들에 대한 지역 특징과 전체 신체에 대한 전역 특징을 추출하는 단계,

상기 지역 특징과 상기 전역 특징을 결합한 입력 특징에 대해 불필요한 영역을 삭제하고, 필수적인 영역을 추가하는 특징 보정을 수행하는 단계, 그리고

히트맵 생성 네트워크를 통해, 상기 특징 보정이 수행된 결과로부터 타겟 인간과 관련된 키포인트가 포함된 히트맵을 출력하는 단계

를 포함하는, 방법.

청구항 13

제12항에서,

상기 추출하는 단계는,

컨벌루션(convolution) 신경망을 이용하여, 상기 입력 이미지로부터 이미지 특징 및 고해상도 특징을 추출하는 단계,

컨벌루션 연산을 통해 상기 이미지 특징으로부터 타겟 인간의 키포인트와 관련된 지역 특징을 추출하는 단계, 그리고

트랜스포머 기반 신경망을 이용하여, 상기 이미지 특징으로부터 전체 신체에 대한 전역 특징을 추출하는 단계를 포함하고,

상기 입력 특징은,

컨벌루션 블록을 통해 상기 고해상도 특징, 상기 지역 특징 및 상기 전역 특징이 결합되어 생성되는, 방법.

청구항 14

제13항에서,

상기 전역 특징을 추출하는 단계는,

상기 이미지 특징과 위치 인코딩을 입력으로 사용하여, 상기 이미지 특징을 전역적으로 이해하기 위한 인코딩을 수행하는 단계,

인코딩 결과를 사용하여 입력 임베딩을 출력 임베딩으로 변환하는 디코딩을 수행하는 단계, 그리고

디코딩 결과에 중선형보간법(Bilinear interpolation)을 적용하여 상기 전역 특징을 추출하는 단계

를 포함하는, 방법.

청구항 15

제13항에서,

상기 특징 보정을 수행하는 단계는,

순차적으로 동작하는 복수개의 잔차 블록과 시그모이드 함수(sigmoid function)를 이용하여, 상기 입력 특징에서 불필요한 영역을 제거한 필터링 특징을 출력하는 단계,

상기 복수개의 잔차 블록을 이용하여, 상기 입력 특징에서 필수영역을 추가한 보충 특징을 출력하는 단계, 그리고

상기 입력 특징에 상기 필터링 특징을 요소별 곱셈 연산(element-wise multiplication)하고, 요소별 곱셈 연산한 결과에 상기 보충 특징을 요소별 합산 연산(element-wise summation)하는 단계

를 포함하는, 방법.

청구항 16

제15항에서,

상기 히트맵을 출력하는 단계는,

상기 입력 특징에 요소별 곱셈 연산된 필터링 특징 및 요소별 합산 연산된 보충 특징을 컨벌루션 레이어에 통과시켜 상기 히트맵을 추출하는, 방법.

청구항 17

인간 자세 추정 프로그램이 저장된 메모리, 그리고

상기 인간 자세 추정 프로그램을 실행하는 적어도 하나의 프로세서를 포함하고,

상기 인간 자세 추정 프로그램은,

컨벌루션(convolution) 신경망을 이용하여, 적어도 한 명의 인간이 포함된 입력 이미지로부터 개별 신체 부위들에 대한 지역 특징을 추출하고,

트랜스포머 기반 신경망을 이용하여, 상기 입력 이미지로부터 전체 신체에 대한 전역 특징을 추출하며,

상기 지역 특징과 상기 전역 특징을 결합한 입력 특징을 이용하여 상기 입력 이미지로부터 타겟 인간과 관련된 키포인트가 포함된 히트맵을 생성하는 명령어들(Instruction)을 포함하는, 인간 자세 추정 장치.

청구항 18

제17항에서,

상기 인간 자세 추정 프로그램은,

상기 지역 특징과 상기 전역 특징을 결합한 입력 특징에 대해 불필요한 영역을 삭제하고, 필수적인 영역을 추가하는 특징 보정을 수행하고, 특징 보정된 입력 이미지로부터 상기 히트맵을 생성하는 명령어들을 추가로 포함하는, 인간 자세 추정 장치.

청구항 19

제18항에서,

상기 인간 자세 추정 프로그램은,

순차적으로 동작하는 복수개의 잔차 블록과 시그모이드 함수(sigmoid function)를 이용하여 상기 특징 보정을 수행하는 명령어들을 추가로 포함하는, 인간 자세 추정 장치.

청구항 20

제18항에서,

상기 인간 자세 추정 프로그램은,

상기 컨벌루션 신경망을 이용하여 추출한 고해상도 특징을 상기 입력 특징에 추가로 포함시키는 명령어들을 추가로 포함하는, 인간 자세 추정 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공 신경망을 이용한 인간 자세 추정 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0002] 인공 지능을 이용해 이미지 내 인간의 위치를 검출하거나 영역을 표시하는 등의 기술이 활발히 연구되고 있다. 이러한 기술의 연장선으로서 인간 자세 추정 기술은 이미지 내 인간의 위치를 알아내는 것뿐만 아니라 그 인간의 주요 관절의 위치가 어디인지 알아내는 기술을 의미한다.

[0003] 종래에 알려진 기술로서, 두가지의 인간 자세 추정 기술이 있다. 첫번째 방식은 우선 인간의 위치를 검출한 다음에 해당 인간에 대한 이미지만 입력으로 해서 그 인간의 관절 좌표를 얻는 방식이다. 두번째 방식은 이미지 전체 내에서 주요 관절의 위치가 될 만한 곳을 전부 추려낸 다음에 같은 인간의 관절이라고 추정되는 것들끼리 연결하는 방식이다.

[0004] 첫번째 방식은 인간 위치 검출 후에 관절 위치 검출을 진행하기 때문에 다른 인간의 관절들과 엮일 가능성이 다른 방식에 비해 낮지만, 여러 인간의 관절을 한 인간의 관절로 착각하여 추정하는 문제는 여전히 존재한다.

[0005] 애초에 여러 인간이 겹쳐있는 경우에는 한 인간의 위치를 추정하여 그 이미지만 자른다 하더라도 그 안에 여러 인간이 존재할 수 있고, 위치를 추정하는 네트워크의 성능이 부족하여 실제 인간이 존재하는 영역보다 넓게 존재한다고 추정한다면 그 안에는 다른 인간이 포함될 수도 있다. 이를 해결하기 위해서 일반적으로 주요 관절의

위치를 찾은 후에, 입력으로 들어갔던 이미지와 출력으로 나온 관절 좌표 모두를 다시 입력으로 하여 해당 관절 좌표를 수정하는 방법이 있다. 또한, 이러한 방법과 다르게 관절의 위치를 연결한 그래프 구조를 사용해서 기존의 방식들이 추정한 위치를 수정하는 방법도 있다. 이 방법은 기존 방식대로 추정해 얻은 각 관절 위치 주변에서 새롭게 후보 위치들을 고른 후에 이 위치들을 이용해서 새로운 그래프들을 만들어내고 이들을 정답 그래프와 비교하면서 가장 그럴듯한 그래프를 고르는 학습을 진행한다.

[0006] 그러나, 이러한 방법들은 두 번의 네트워크를 통과하거나 여러 그래프를 고르는 과정 등에 의해서 계산양이 많아진다거나 연산 시간이 길어진다는 단점이 존재한다. 또한, 여러 인간의 관절 좌표를 한 인간의 관절 좌표로 연결한다는 문제들이 존재한다. 이는 인간 위치 추정기의 성능이 좋지 않거나 애초 입력 이미지에 인간들이 겹쳐서 존재하기 때문에 발생하는 문제들이다. 이를 해결하기 위해서 두 번의 관절 위치 추정기를 통과하거나 아니면 그래프 구조를 사용해서 연결 관계를 학습하는 방법 등이 적용됐지만 그 수행 시간이나 연산 양에 있어서 단점을 보인다.

발명의 내용

해결하려는 과제

[0007] 본 개시는, 지역 특징 뿐만 아니라 전역 특징을 고려하고, 지역 특징과 전역 특징으로 구성된 특징 맵(Feature Map)에서 불필요한 영역은 필터링하고 필요한 영역은 보충하여 지역 특징과 전역 특징을 조화시켜 최종적인 타겟 인간의 자세를 추정하는 장치 및 그 방법을 제공하는 것이다.

과제의 해결 수단

[0008] 본 발명의 한 특징에 따르면, 인간 자세 추정 장치는 적어도 한 명의 인간이 포함된 입력 이미지로부터 개별 신체 부위들에 대한 지역 특징과 전체 신체에 대한 전역 특징을 추출하는 특징 추출부, 그리고 상기 입력 이미지에서 타겟 인간과 관련된 키포인트가 포함된 히트맵을 출력하기 위해, 상기 지역 특징과 상기 전역 특징을 결합한 입력 특징에 대해 불필요한 영역을 삭제하고, 필수적인 영역을 추가하는 특징 보정부를 포함한다.

[0009] 상기 특징 보정부는, 복수개의 잔차 블록으로 구성된 잔차 신경망으로서, 상기 잔차 신경망을 통해 상기 입력 특징에서 불필요한 영역을 삭제한 필터링 특징을 출력하는 특징 필터링 모듈을 포함할 수 있다.

[0010] 상기 특징 필터링 모듈은, 순차적으로 동작하는 상기 복수개의 잔차 블록과 시그모이드 함수(sigmoid function)를 이용하여 상기 입력 특징에서 불필요한 영역을 제거한 필터링 특징을 출력할 수 있다.

[0011] 상기 특징 보정부는, 상기 복수개의 잔차 블록을 이용하여, 상기 입력 특징에서 필요한 영역을 추가한 보충 특징을 추출하는 특징 보충 모듈을 더 포함할 수 있다.

[0012] 상기 복수개의 잔차 블록은, 컨벌루션 레이어(Convolution Layer), 그리고 어텐션 모듈인 NORM, RELU를 포함할 수 있다.

[0013] 상기 특징 보정부는, 상기 입력 특징에 상기 필터링 특징을 요소별 곱셈 연산(element-wise multiplication)하고, 요소별 곱셈 연산한 결과에 상기 보충 특징을 요소별 합산 연산(element-wise summation)하는 연산부를 포함할 수 있다.

[0014] 상기 인간 자세 추정 장치는 상기 연산부의 연산 결과를 컨벌루션 레이어에 통과시켜 상기 히트맵을 추출하는 히트맵 추출부를 더 포함할 수 있다.

[0015] 상기 특징 추출부는, 트랜스포머 기반 신경망을 이용하여, 상기 입력 이미지로부터 전체 신체에 대한 전역 특징을 추출하는 전역 특징 추출부를 포함할 수 있다.

[0016] 상기 특징 추출부는, 컨벌루션(convolution) 신경망을 이용하여, 상기 입력 이미지로부터 이미지 특징을 추출하고, 컨벌루션 연산을 통해 상기 이미지 특징으로부터 타겟 인간의 키포인트와 관련된 지역 특징을 추출하는 지역 특징 추출부를 더 포함할 수 있다.

[0017] 상기 전역 특징 추출부는, 상기 이미지 특징과 위치 인코딩을 입력으로 사용하여, 상기 이미지 특징을 전역적으로 이해하기 위한 인코딩을 수행하고, 인코딩 결과를 사용하여 입력 임베딩을 출력 임베딩으로 변환하는 디코딩을 수행하며, 디코딩 결과에 중선형보간법(Bilinear interpolation)을 적용하여 상기 전역 특징을 추출할 수 있다.

[0018] 상기 지역 특징 추출부는, 상기 컨벌루션 신경망을 이용하여, 상기 입력 이미지로부터 이미지 특징 및 고해상도 특징을 추출하고, 상기 특징 보정부는, 상기 고해상도 특징, 상기 지역 특징 및 상기 전역 특징을 결합한 상기 입력 특징을 이용할 수 있다.

[0019] 다른 특징에 따르면, 적어도 하나의 프로세서에 의해 동작하는 인간 자세 추정 장치의 동작 방법으로서, 적어도 한 명의 인간이 포함된 입력 이미지로부터 개별 신체 부위들에 대한 지역 특징과 전체 신체에 대한 전역 특징을 추출하는 단계, 상기 지역 특징과 상기 전역 특징을 결합한 입력 특징에 대해 불필요한 영역을 삭제하고, 필수적인 영역을 추가하는 특징 보정을 수행하는 단계, 그리고 히트맵 생성 네트워크를 통해, 상기 특징 보정이 수행된 결과로부터 타겟 인간과 관련된 키포인트가 포함된 히트맵을 출력하는 단계를 포함한다.

[0020] 상기 추출하는 단계는, 컨벌루션(convolution) 신경망을 이용하여, 상기 입력 이미지로부터 이미지 특징 및 고해상도 특징을 추출하는 단계, 컨벌루션 연산을 통해 상기 이미지 특징으로부터 타겟 인간의 키포인트와 관련된 지역 특징을 추출하는 단계, 그리고 트랜스포머 기반 신경망을 이용하여, 상기 이미지 특징으로부터 전체 신체에 대한 전역 특징을 추출하는 단계를 포함하고, 상기 입력 특징은, 컨벌루션 블록을 통해 상기 고해상도 특징, 상기 지역 특징 및 상기 전역 특징이 결합되어 생성될 수 있다.

[0021] 상기 전역 특징을 추출하는 단계는, 상기 이미지 특징과 위치 인코딩을 입력으로 사용하여, 상기 이미지 특징을 전역적으로 이해하기 위한 인코딩을 수행하는 단계, 인코딩 결과를 사용하여 입력 임베딩을 출력 임베딩으로 변환하는 디코딩을 수행하는 단계, 그리고 디코딩 결과에 중선형보간법(Bilinear interpolation)을 적용하여 상기 전역 특징을 추출하는 단계를 포함할 수 있다.

[0022] 상기 특징 보정을 수행하는 단계는, 순차적으로 동작하는 복수개의 잔차 블록과 시그모이드 함수(sigmoid function)를 이용하여, 상기 입력 특징에서 불필요한 영역을 제거한 필터링 특징을 출력하는 단계, 상기 복수개의 잔차 블록을 이용하여, 상기 입력 특징에서 필수영역을 추가한 보충 특징을 출력하는 단계, 그리고 상기 입력 특징에 상기 필터링 특징을 요소별 곱셈 연산(element-wise multiplication)하고, 요소별 곱셈 연산한 결과에 상기 보충 특징을 요소별 합산 연산(element-wise summation)하는 단계를 포함할 수 있다.

[0023] 상기 히트맵을 출력하는 단계는, 상기 입력 특징에 요소별 곱셈 연산된 필터링 특징 및 요소별 합산 연산된 보충 특징을 컨벌루션 레이어에 통과시켜 상기 히트맵을 추출할 수 있다.

[0024] 또 다른 특징에 따르면, 인간 자세 추정 장치는 인간 자세 추정 프로그램이 저장된 메모리, 그리고 상기 인간 자세 추정 프로그램을 실행하는 적어도 하나의 프로세서를 포함하고, 상기 인간 자세 추정 프로그램은, 컨벌루션(convolution) 신경망을 이용하여, 적어도 한 명의 인간이 포함된 입력 이미지로부터 개별 신체 부위들에 대한 지역 특징을 추출하고, 트랜스포머 기반 신경망을 이용하여, 상기 입력 이미지로부터 전체 신체에 대한 전역 특징을 추출하며, 상기 지역 특징과 상기 전역 특징을 결합한 입력 특징을 이용하여 상기 입력 이미지로부터 타겟 인간과 관련된 키포인트가 포함된 히트맵을 생성하는 명령어들(Instruction)을 포함할 수 있다.

[0025] 상기 인간 자세 추정 프로그램은, 상기 지역 특징과 상기 전역 특징을 결합한 입력 특징에 대해 불필요한 영역을 삭제하고, 필수적인 영역을 추가하는 특징 보정을 수행하고, 특징 보정된 입력 이미지로부터 상기 히트맵을 생성하는 명령어들을 추가로 포함할 수 있다.

[0026] 상기 인간 자세 추정 프로그램은, 순차적으로 동작하는 복수개의 잔차 블록과 시그모이드 함수(sigmoid function)를 이용하여 상기 특징 보정을 수행하는 명령어들을 추가로 포함할 수 있다.

[0027] 상기 인간 자세 추정 프로그램은, 상기 컨벌루션 신경망을 이용하여 추출한 고해상도 특징을 상기 입력 특징에 추가로 포함시키는 명령어들을 더 포함할 수 있다.

발명의 효과

[0028] 본 개시에 따르면, CNN 기반 네트워크를 사용하여 지역 특징을 추출하고, 트랜스포머(Transformer)를 사용하여 전역 특징을 추출하여, 지역 특징과 전역 특징을 고려하여 타겟 인간의 자세를 추정하므로, CNN 기반으로 지역 특징만을 고려하는 종래 방식의 폐색 문제를 완화할 수 있다.

[0029] 또한, 지역 특징과 전역 특징으로 구성된 특징 맵(Feature Map)에서 불필요한 영역은 필터링하고 필요한 영역은 보충하여 지역 특징과 전역 특징을 조화시켜 최종적인 자세 추정을 수행하므로, 자세 추정 성능이 크게 개선될 수 있다.

[0030] 또한, 종래에 인간의 자세 추정을 위한 지역 특징을 추출하는 CNN 기반 네트워크 구조를 유지하면서 이러한 네

트워크 구조에 연결하여 사용할 수 있어, 호환성이 좋아 구현이 용이한 장점이 있다.

도면의 간단한 설명

[0031] 도 1은 실시예에 따른 인간 자세 추정 구조를 개략적으로 나타낸다.

도 2는 종래의 방식에 따른 인간 자세 추정 구조를 나타낸다.

도 3은 실시예에 따른 인간 자세 추정 장치의 개략적인 구성을 나타낸 블록도이다.

도 4는 실시예에 따른 인간 자세 추정 장치의 세부적인 구성을 나타낸 블록도이다.

도 5는 도 4에서 FFM(121)과 FSM(122)의 잔차 블록(Residual Bolck)의 구조를 나타낸다.

도 6은 한 실시예에 따른 인간 자세 추정 장치의 동작을 설명한 순서도이다.

도 7은 실시예에 따른 컴퓨팅 장치의 하드웨어 구성을 나타낸 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0032] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0033] 설명에서, 도면 부호 및 이름은 설명의 편의를 위해 붙인 것으로서, 장치들이 반드시 도면 부호나 이름으로 한정되는 것은 아니다.

[0034] 설명에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0035] 또한, 명세서에 기재된 "…부", "…기", "…모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0036] 본 개시의 장치는 적어도 하나의 프로세서가 명령어들(instructions)을 실행함으로써, 본 개시의 동작을 수행할 수 있도록 구성 및 연결된 컴퓨팅 장치이다. 컴퓨터 프로그램은 프로세서가 본 개시의 동작을 실행하도록 기술된 명령어들(instructions)을 포함하고, 비일시적-컴퓨터 판독가능 저장매체(non-transitory computer readable storage medium)에 저장될 수 있다. 컴퓨터 프로그램은 네트워크를 통해 다운로드되거나, 제품 형태로 판매될 수 있다.

[0037] 본 개시의 인공지능 모델(Artificial Intelligence model, AI model)은 적어도 하나의 태스크(task)를 학습하는 기계학습모델로서, 프로세서에 의해 실행되는 컴퓨터 프로그램으로 구현될 수 있다. 인공지능 모델이 학습하는 태스크란, 기계 학습을 통해 해결하고자 하는 과제 또는 기계 학습을 통해 수행하고자 하는 작업을 지칭할 수 있다. 인공지능 모델은 컴퓨팅 장치에서 실행되는 컴퓨터 프로그램으로 구현될 수 있고, 네트워크를 통해 다운로드되거나, 제품 형태로 판매될 수 있다. 또는 인공지능 모델은 네트워크를 통해 다양한 장치들과 연동할 수 있다.

[0039] 도 1은 실시예에 따른 인간 자세 추정 구조를 개략적으로 나타내고, 도 2는 종래의 방식에 따른 인간 자세 추정 구조를 나타내며, 도 1과 도 2를 참조하여 본 발명의 실시예와 종래의 기술을 비교한다.

[0040] 도 1에 따르면, 실시예에 따른 인간 자세 추정 구조는 두 인간이 포함된 잘려진 이미지(cropped input bounding box)를 입력 이미지로 사용하고, 입력 이미지로부터 지역 특징(local reasoning feature)과 전역 특징(global reasoning feature)을 추출한다. 이 구조는 추출한 특징들로부터 자세 특징(pose feature)을 추출하여 조화시키기 위하여 부착 가능한 특징 보정부(Attachable Feature Corrector)를 통해 특징 맵의 불필요한 부분을 필터링(Filtering)하고 정확한 지역화(localization)를 위해 부족한 부분을 보충(Supplementing)하여 자세 특징을 수정하고, 수정된 자세 특징을 출력한다. 이때, 출력되는 자세 특징은 각 관절에 대한 일련의 히트맵이다.

[0041] 여기서, 전역 특징은 트랜스포머(transformer)의 비로컬 계산 속성(non-local computation property)을 고려하

여 전신을 고려한 특징이다. 지역 특징은 컨볼루션 신경망(convolutional neural network)을 사용하여 개별 신체 부위들에 집중한 특징이다.

[0042] 도 2에 따르면, 종래의 일반적인 하향식 자세 추정(Top-down pose estimation) 방식은 인간 감지기를 사용해서 감지된, 인간이 포함된 잘려진 이미지(cropped input bounding box)를 입력으로 사용하고, 입력 이미지로부터 지역 특징을 추출하여 추출한 지역 특징으로부터 인간 자세를 추정하게 된다.

[0043] 이때, 일반적인 하향식 자세 추정 방식은 입력 이미지로부터 인간의 키포인트(keypoints)를 추정하는데, 입력 이미지 내에 인간이 한명만 있다고 가정한다. 루즈핏 바운딩 박스(loose-fitted bounding box)는 자세 추정 대상, 즉, 타겟 인간의 일부 신체 부위를 포함할 수 있다. 또한, 여러명 인간들 사이에 공간적 간섭으로 인해 잘려진 이미지에는 한 명 이상의 인간이 있을 수 있다. 그런데, 이러한 점을 일반적인 하향식 자세 추정 방식에서는 고려하지 않는다.

[0044] 따라서, 입력 이미지에 두명의 인간이 있는 경우, 도 2의 방식은 자세 추정 대상이 아닌 비타겟 인간의 신체 특징인 키포인트(keypoints)를 더 잘 식별하기 때문에, 비타겟 인간의 지역 특징을 추정한다.

[0045] 도 2의 방식은 지역 특징만을 고려하므로, 폐색(occlution)으로 인해 타겟 인간의 얼굴 키포인트를 제외한 키포인트들을 감지하는데 실패한다. 또한, 도 2의 방식은 타겟 인간과 오버랩되는 다른 인간의 키포인트를 타겟 인간의 키포인트로 추정하게 되므로, 두 명 이상의 키포인트를 한 명의 키포인트로 잘못 예측할 수 있다. 따라서, 도 2의 방식은 정답 자세(Ground Truth Pose)와 비교할 때, 잘못된 추정 결과를 초래하게 된다.

[0046] 반면, 도 1의 구조는 지역 특징 뿐만 아니라 전역 특징을 이용하여 인간의 자세 특징을 추출하고 추출한 특징을 필터링 및 보충을 통해 보정하여 최종적으로 인간 자세를 추정하므로, 타겟 인간의 키포인트에 집중할 수 있다.

[0047] 이처럼, 도 1의 구조는 입력 이미지의 지역 특징 뿐만 아니라 전역 특징을 고려하므로 타겟 인간의 모든 키포인트를 효과적으로 지역화할 수 있어 도 2에서 언급한 폐색 문제를 해결할 수 있다.

[0048] 또한, 도 1의 구조는 입력 이미지의 지역 특징 뿐만 아니라 전역 특징을 고려하여 추출한 특징을 필터링 및 보충함으로써, 최종적으로 추정된 결과는 타겟 인간의 키포인트가 된다. 따라서, 추정된 인간 자세는 정답 자세(Ground Truth Pose)와 비교할 때, 추정 성능이 개선된다.

[0050] 도 3은 실시예에 따른 인간 자세 추정 장치의 개략적인 구성을 나타낸 블록도이고, 도 4는 실시예에 따른 인간 자세 추정 장치의 세부적인 구성을 나타낸 블록도이며, 도 5는 도 4에서 FFM(121)과 FSM(122)의 잔차 블록(Residual Bolck)의 구조를 나타낸다.

[0051] 먼저, 도 3은 도 1의 구조를 블록 구성 형태로 도시한 것으로서, 도 3에 따르면, 인간 자세 추정 장치(100)는 적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치이다.

[0052] 인간 자세 추정 장치(100)는 타겟 인간이 아닌 여러명의 인간이 존재하는 상황에서도 하향식 방식(top-down manner)으로, 타겟 인간의 전신을 이해하여 의미론적 키포인트를 추출하며, 이를 위해 키포인트의 지역화 작업을 히트맵 추정으로 변환한다.

[0053] 인간 자세 추정 장치(100)는 특징 추출부(110), 특징 결합부(120), 부착형 특징 보정부(130) 및 히트맵 추출부(130)를 포함할 수 있다.

[0054] 특징 추출부(110)는 지역 특징 추출부(111) 및 전역 특징 추출부(112)를 포함한다.

[0055] 지역 특징 추출부(111)는 사람이 포함된 잘려진 이미지(cropped input bounding box)(10)를 입력받고, 지역적 특성을 가지는 컨벌루션 연산(convolutional operations)을 통해 입력 이미지로부터 지역 특징을 추출한다. 컨벌루션 연산은 입력의 공간 영역에서 가중치를 공유하여 모델 복잡성을 제어하기 때문에 유리한 특성을 지니고 있다.

[0056] 지역 특징 추출부(111)는 이미 알려진 지역 특징 추출을 위한 컨벌루션 네트워크(convolutional neural networks, CNNs)를 사용할 수 있다.

[0057] 전역 특징 추출부(112)는 트랜스포머(transformer) 구조로 이루어지며, 입력 이미지의 전역 특징을 추출한다.

[0058] CNNs 기반의 키포인트 추정은 제한된 수용 필드로 인해 장거리 종속성 정보를 학습하기 어려운 한계가

있으므로, 전체 글로벌 이미지 컨텍스트를 고려하는데 적합하지 않기 때문에, 전역 특징 추출부(112)는 트랜스포머(transformer)를 사용하여 전역 특징을 추출한다.

[0059] 전역 특징 추출부(112)는 트랜스포머의 비로컬 계산을 통해 입력의 공간 도메인에서 픽셀 간의 모든 쌍별(pairwise) 상호 작용(interactions)을 고려하는 새로운 특징을 얻을 수 있다.

[0060] 이와 같이, 기존의 방법들처럼 CNNs(convolutional neural networks)를 사용해서 지역 특징을 추출하는 것과 더불어 트랜스포머를 사용해서 패치(patch) 단위로 특징맵을 분할하고, 모든 패치를 참고해서 새로운 특징맵(global-reaoning feature map)을 만들어내기 때문에 입력의 전체 영역을 참고한 특징을 규정할 수 있게 되므로 다른 관절과의 관계를 고려해 해당 관절의 위치를 추정할 수 있게 된다.

[0061] 부착형 특징 보정부(130)는 특징 추출부(110)의 추출 결과를 토대로, 불필요한 영역의 값을 필터링하고 필요한 영역의 값을 보충하는 역할을 한다.

[0062] 이때, 부착형 특징 보정부(130)는 종래에 키포인트 추정을 위한 지역 특징 추출 네트워크인 CNNs 기반의 특징 추출부(110)에 연결하여 구현될 수 있으므로, 부착형 특징 보정부(120)라 호칭한다.

[0063] 부착형 특징 보정부(130)는 특징 필터링 모듈(Feature Filtering Module, FFM)(131), 특징 보충 모듈(Feature Supplementing Module, FSM)(132) 및 연산부(133)를 포함할 수 있다.

[0064] 부착형 특징 보정부(130)는 특징 필터링 모듈(FFM)(131)과 특징 보충 모듈(FSM)(132)을 통하여, CNN으로 획득한 지역 특징과 트랜스포머를 통해 획득한 전역 특징을 통합하여 보정한다. 예컨대, 특징 필터링 모듈(FFM)(131)은 불필요하게 값이 높게 측정된 부분을 제거하거나 또는 줄이는 역할을 하고, 특징 보충 모듈(FSM)(132)은 원래는 높아야 할 곳이지만 높지 않은 곳에 대해서 값을 채워주는 역할을 하게 된다.

[0065] 특징 필터링 모듈(FFM)(131)은 전역 특징 추출부(112)에 의해 추출된 전역 특징을 입력받아, 이를 참조하여 입력 특징 맵(feature map)의 불필요한 부분을 필터링하여 자세 특징(pose feature)을 수정한다.

[0066] 특징 보충 모듈(FSM)(132)은 정확한 지역화를 위해 부족한 부분을 보충하여 자세 특징을 수정한다.

[0067] 연산부(133)는 특징 필터링 모듈(FFM)(131)과 특징 보충 모듈(FSM)(132)의 출력 결과에 대한 정해진 연산을 수행하며, 이에 대해서는 도 4를 참고하여 자세히 설명하기로 한다.

[0068] 특징 필터링 모듈(FFM)(131)과 특징 보충 모듈(FSM)(132)을 통한 자세 보정을 통해 타겟 인간에게만 연관된 키포인트를 예측할 수 있게 된다.

[0069] 히트맵 추출부(140)는 1×1 컨벌루션 레이어를 사용하여, 부착형 특징 보정부(120)를 통해 통해 출력된 보정된 자세 특징으로부터 최종적인 자세 특징인타겟 인간의 각 관절에 대한 일련의 히트맵(20)을 추출하여 출력한다.

[0070] 전술한 인간 자세 추정 장치(100)의 구조에 대해 도 4를 참고하여 보다 자세히 설명하면, 다음과 같다.

[0071] 도 4를 참조하면, 지역 특징 추출부(111)는 백본 네트워크인 $\emptyset_B$, 그리고 이미 알려진 컨벌루션 네트워크(conventional convolution-based architecture)인 $\emptyset_{LR}$로 구성된다.

[0072] $\emptyset_B$는 딥러닝의 백본 네트워크로서, 입력 이미지(10)로부터 다양한 이미지 특징, F_B을 추출한다.

[0073] $\emptyset_{LR}$은 지역 특징을 추출하기 위한 네트워크로서, 다음 수학식 1에 정의된 바와 같이, 컨벌루션 연산을 통해 지역 특징, F_{LR}을 추출한다.

[0074] [수학식 1]

[0075] $$F_{LR} = \emptyset_{LR}(F_B)$$

[0076] 종래에는 $\emptyset_B$를 사용하여 특징을 추출하여 자세를 추정하고, $\emptyset_{LR}$을 사용하여 추출된 특징으로부터 히트맵을 생성한다.

[0077] 이러한 종래의 $\emptyset_B$와 $\emptyset_{LR}$을 본 발명의 실시예에서도 사용한다. 다만, 본 발명의 실시예에서는 종래와 달리,

$\emptyset_{LR}$을 사용하여 타겟 키포인트와 관련된 지역 특징, F_{LR}을 추출하는 점이 다르다.

[0078] 또한, 본 발명의 실시예는 종래와 달리, $\emptyset_{B}$를 사용하여 고해상도 특징(high resolution feature)인 F_{H}를 추출한다. F_{H}는 추가적인 세부 정보를 제공한다.

[0079] 전역 특징 추출부(112)는 트랜스포머 기반 네트워크(transformer-based architecture)인 $\emptyset_{GR}$을 사용하여 전체 타겟 인간에 포커싱한 전역 특징, F_{GR}을 추출한다. $\emptyset_{GR}$은 입력 특징 맵의 전역 특징들을 포함하는 F_{GR}을 생성한다.

[0080] $\emptyset_{GR}$은 트랜스포머에 기초한 인코더-디코더 구조(encoder-decoder architecture)로 이루어진다. 즉, $\emptyset_{GR}$은 인코더 파트인 $\emptyset_{GR}^{enc}$과 디코더 파트인 $\emptyset_{GR}^{dec}$을 포함한다.

[0081] 인코더는 시퀀스를 입력으로 예상한다. 따라서, $\emptyset_{GR}^{enc}$의 입력은 $F_{B} \in \mathbb{R}^{W_B \times H_B \times C_B}$의 공간 차원을 하나의 차원으로 축소한 $W_B \times H_B \times C_B$ 특징맵으로 구성된 시퀀스이다. 이때, W_B는 너비(width)를 나타내고, H_B는 높이(height)를 나타내며, C_B는 F_B의 채널을 의미한다.

[0082] 각각의 인코더 계층은 다른 트랜스포머 기반 네트워크와 마찬가지로 멀티헤드 셀프 어텐션 모듈(multihead self-attention module)과 피드 포워드 네트워크(feed forward network)로 구성된다. 이때, 지역성(locality) 누락을 방지하기 위해 위치 인코딩(positional encodings)이 각 어텐션 계층의 입력으로 추가된다.

[0083] 트랜스포머의 비-지역적 계산 속성(non-local computation property)을 사용하여, 인코더는 입력 특징, F_{B}를 전역적으로 이해하여 F_{GR}^{enc}를 추출하며, 이를 수식으로 표현하면, 수학식 2와 같다.

[0084] [수학식 2]

[0085] $$F_{GR}^{enc} = \emptyset_{GR}^{enc}(F_B, E_p)$$

[0086] 수학식 2에서, F_{GR}^{enc}는 인코더, 즉, $\emptyset_{GR}^{enc}$에서 추출한 특징이고, E_p는 위치 인코딩을 나타낸다.

[0087] 디코더는 인코딩된 특징, F_{GR}^{enc}를 사용하여 입력 임베딩(input embeddings)을 출력 임베딩(output embeddings)으로 변환한다. 이때, 디코더의 객체 쿼리(object queries)는 K 키포인트 임베딩으로 수정된다.

[0088] 디코더는 F_{GR}^{enc}를 사용하여 입력 임베딩을 각각의 키포인트를 위한 특징으로 디코딩한 후, 중선형보간법(Bilinear interpolation)을 통해 최종적으로 전역 특징, F_{GR}을 추출한다. 이를 수식화하면, 수학식 3과 같다.

[0089] [수학식 3]

[0090] $$F_{GR} = \text{Bilinear}(\emptyset_{GR}^{dec}(F_{GR}^{enc}, E_{key}))$$

[0091] 수학식 3에서, E_{key}는 키포인트 임베딩을 나타낸다.

[0092] 이상 설명한 바에 따르면, 특징 추출부(110)는 F_{GR}, F_{LR}, F_{H}를 출력하고, 특징 결합부(120)는 F_{GR}, F_{LR}, F_{H}를 결합한 F_{int}를 부착형 특징 보정부(130)로 출력한다.

[0093] 특징 결합부(120)는 컨벌루션 블록(convolutional block)을 포함한다. F_{GR}, F_{LR}, F_{H}는 채널 축(channel

axis)을 통해 연결되어(concatenated) 컨벌루션 블록을 통과함으로써, F_{int}가 되는데, 이를 수식화하면, 수학식 4와 같다.

[0094] [수학식 4]

[0095] $$F_{int} = \mathrm{Conv}\left(\mathrm{concat}(F_H, F_{GR}, F_{LR})\right)$$

[0096] F_{int}는 지역 특징 뿐만 아니라 전역 특징을 고려하여 자세 특징을 보정하는 부착형 특징 보정부(130)로 입력된다.

[0097] 이처럼, F_{GR}, F_{LR}, F_H를 결합한 F_{int}를 부착형 특징 보정부(130)의 입력으로 사용하게 되면, 해상도 특징(low-resolution feature)인 F_B로부터 추출된 F_{GR}, F_{LR}은 고해상도 특징인 F_H가 추가되어 이미지의 상세 정보를 보완하게 된다.

[0098] 부착형 특징 보정부(130)의 특징 필터링 모듈(FFM)(131)과 특징 보충 모듈(FSM)(132)은 F_{int}가 단지 타겟 인간에게만 포커싱하도록 하기 위한 구조와 목적을 가지지만, 다르게 동작한다.

[0099] 특징 필터링 모듈(FFM)(131)은 F_{int}를 시그모이드 함수(sigmoid function)를 통과시킨 후 그 출력을 곱하여 F_{int}의 불필요한 부분(unnecessary portions)을 제거한다. 특징 필터링 모듈(FFM)(131)의 출력은 필터링 특징(filtering feature)으로 호칭한다.

[0100] 특징 보충 모듈(FSM)(132)은 자체 출력을 추가함으로써, 필수적인 영역(necessary parts)을 보충한다. 특징 보충 모듈(FSM)(132)의 출력은 보충 특징(supplementing feature)으로 호칭한다.

[0101] 특징 필터링 모듈(FFM)(131) 및 특징 보충 모듈(FSM)(132)은 잔차 신경망(Residual neural network, ResNet)으로서, 복수개의 잔차 블록(residual blocks)으로 구성된다. 각각의 잔차 블록은 순차적으로 동작하는 얕은 컨벌루션 레이어(shallow convolution layers)와 어텐션 모듈(attention module)로 구성되며, 이는 도 5와 같다.

[0102] 도 5를 참조하면, 특징 필터링 모듈(FFM)(131) 및 특징 보충 모듈(FSM)(132)의 잔차 블록 구조는 3개의 잔차 블록들과 CBAM(Convolutional Block Attention Module)이 순차적으로 연결된 구조이다.

[0103] 3개의 잔차 블록들은 컨벌루션 레이어(CONV1×1, CONV3×1,), 그리고 어텐션 모듈인 NORM, RELU로 구성된다. CBAM은 잔차 합산 직전에 추가된다. 특징 필터링 모듈(FFM)(131) 및 특징 보충 모듈(FSM)(132)로 각각 입력된 이전 특징 맵은 각각의 잔차 블록들을 통과하여 처리되고, 최종적으로 CBAM의 출력 및 NORM의 출력은 요소별 합산(element-wise summation) 연산을 통해 다음 특징 맵으로 출력된다.

[0104] ρ_f^t는 t 단계에서 특징 필터링 모듈(FFM)(131)의 잔차 블록을 나타낸다. 특징 필터링 모듈(FFM)(131)은 F_{int}를 입력으로 사용하여 수학식 5와 같이 첫번째 필터링 특징(F_f^1)을 추출한다.

[0105] [수학식 5]

[0106] $$F_f^1 = \rho_f^1(F_{int})$$

[0107] T_f반복을 통해 시그모이드 함수를 통과한 $F_f^{T_f}$는 최종적인 필터링 특징, C_f이 된다.

[0108] ρ_s^t는 t 단계에서 특징 보충 모듈(FSM)(132)의 잔차 블록을 나타낸다. 특징 보충 모듈(FSM)(132)은 F_{int}를 입력으로 사용하여 수학식 6과 같이 첫번째 보충 특징(F_s^1)을 추출한다.

[0109] [수학식 6]

[0110] $$F_s^1 = \rho_s^1(\mathrm{F}_{int})$$

[0111] T_s 반복을 통해 시그모이드 함수를 통과한 $F_s^{T_s}$ 는 최종적인 보충 특징, C_s이 된다.

[0112] 필터링 특징, C_f 는 다시 시그모이드 함수를 통과한 후, F_{int}에 요소별 곱셈 연산(element-wise multiplication)되고, 그 연산 결과, 즉, 필터링된 F_{int}는 보충 특징, C_s에 요소별 합산 연산(element-wise summation)된다.

[0113] 이처럼, C_f 는 시그모이드 함수를 통과한 후, F_{int}에 요소별 곱셈 연산됨으로써, 각 히트맵에 하나 이상의 피크 포인트(peak point)가 존재하거나 또는 히트맵이 잘못 추정되었을 때 불필요한 특징을 축소(scale down)할 수 있게 된다.

[0114] C_s 는 필터링된 F_{int}에 요소 별로 추가되어 정확한 키포인트 추정을 위한 특징을 보충하게 된다.

[0115] 연산부(123)는 전술한 시그모이드 함수, 요소별 곱셈 연산, 요소별 합산 연산을 수행한다.

[0116] 히트맵 추출부(140)는 1×1 컨볼루션 레이어로 구성되고, 부착형 특징 보정부(130)의 출력으로부터 히트맵을 추출하여 출력한다. 즉, 히트맵 세트, $\{H_k\}_{k=1}^{K}$는 부착형 특징 보정부(130)의 출력인 F_{int}, C_s, C_f가 1×1 컨볼루션 레이어를 통과함으로써 추정된다. 여기서, C_s, C_f 는 F_{int}로부터 파생된 F_{int}와 별개의 특징맵이다.

[0117] 이를 수식화하면, 수학식 7과 같다.

[0118] [수학식 7]

[0119] $$\{H_k\}_{k=1}^{K} = \mathrm{Conv}_{1\times1}(F_{int} \times C_f \times C_s)$$

[0120] K는 총 키포인트의 개수를 의미한다.

[0122] 이상 설명한 내용을 기초로, 인간 자세 추정 장치(100)의 동작에 대해 설명하면, 다음과 같다.

[0123] 도 6은 한 실시예에 따른 인간 자세 추정 장치의 동작을 설명한 순서도로서, 도 1 ~ 도 5의 구성을 참고하여 설명한다.

[0124] 도 6을 참조하면, 인간 자세 추정 장치(100)의 지역 특징 추출부(111)는 적어도 한 명의 인간이 포함된 잘려진 이미지(도 1 ~ 도 5의 10)를 입력받는다(S101).

[0125] 지역 특징 추출부(111)는 컨벌루션 연산을 통해 입력 이미지(10)로부터 지역 특징 및 고해상도 특징을 추출하고, 전역 특징 추출부(112)는 트랜스포머를 통해 통해 입력 이미지로부터 전역 특징을 추출한다(S102).

[0126] 특징 결합부(120)는 지역 특징(F_{LR}), 전역 특징(F_{GR}) 및 고해상도 특징(F_{H})을 컨벌루션 블록을 통해 결합한 입력 특징(F_{int})을 생성한다(S103).

[0127] 특징 필터링 모듈(FFM)(131)은 S103의 입력 특징(F_{int})에서 불필요한 영역을 제거한 필터링 특징(C_f)을 출력하고, 특징 보충 모듈(FSM)(132)은 S102의 입력 특징에 필수적인 영역을 추가한 보충 특징(C_s)을 출력하며, 필터링 특징(C_f)과 보충 특징(C_s)은 시그모이드 함수, 요소별 곱셈 연산, 요소별 합산 연산된다(S104).

[0128] 히트맵 추출부(140)는 S104의 결과(F_{int}, C_s, C_f)를 히트맵 생성 네트워크, 예를 들어, 1×1 컨볼루션 레이어를 통과시켜 히트맵($\{H_k\}_{k=1}^{K}$)을 출력한다(S105). 최종적으로, 히트맵은 입력 이미지로부터 관절의 위치를 나타내는 특징맵이다. 히트맵을 통해 최종적으로 입력 이미지로부터 인간의 자세를 추정할 수 있다.

[0130] 한편, 도 7은 실시예에 따른 컴퓨팅 장치의 하드웨어 구성을 나타낸 블록도로서, 도 1 ~ 도 6에서 설명한 인간 자세 추정 장치(100)는 컴퓨팅 장치로 구현될 수 있다.

[0131] 도 7을 참조하면, 컴퓨팅 장치(200)는 하나 이상의 프로세서(210), 프로세서(210)에 의하여 수행되는 프로그램을 로드하는 메모리(220), 프로그램 및 각종 데이터를 저장하는 스토리지(230), 및 통신 인터페이스(240)를 포함하고, 이들은 버스(250)를 통해 연결된다. 다만, 상술한 구성 요소들은 본 개시에 따른 컴퓨팅 장치(200)를 구현하는데 있어서 필수적인 것은 아니어서, 컴퓨팅 장치(200)는 위에서 열거된 구성요소들 보다 많거나, 또는 적은 구성요소들을 가질 수 있다. 예컨대 컴퓨팅 장치(200)는 출력부 및/또는 입력부(미도시)를 더 포함하거나, 또는 스토리지(230)가 생략될 수도 있다.

[0132] 프로그램은 메모리(220)에 로드될 때 프로세서(210)로 하여금 본 개시의 다양한 실시예에 따른 방법/동작을 수행하게끔 하는 명령어들(instructions)을 포함할 수 있다. 즉, 프로세서(210)는 명령어들을 실행함으로써, 본 개시의 다양한 실시예에 따른 방법/동작들을 수행할 수 있다. 프로그램은 기능을 기준으로 묶인 일련의 컴퓨터 판독가능 명령어들로 구성되고, 프로세서에 의해 실행되는 것을 가리킨다.

[0133] 프로세서(210)는 컴퓨팅 장치(200)의 각 구성의 전반적인 동작을 제어한다. 프로세서(210)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit) 또는 본 개시의 기술 분야에 잘 알려진 임의의 형태의 프로세서 중 적어도 하나를 포함하여 구성될 수 있다. 또한, 프로세서(210)는 본 개시의 다양한 실시예들에 따른 방법/동작을 실행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.

[0134] 메모리(220)는 각종 데이터, 명령 및/또는 정보를 저장한다. 메모리(220)는 본 개시의 다양한 실시예들에 따른 방법/동작을 실행하기 위하여 스토리지(230)로부터 하나 이상의 프로그램을 로드할 수 있다. 메모리(220)는 RAM과 같은 휘발성 메모리로 구현될 수 있을 것이나, 본 개시의 기술적 범위는 이에 한정되지 않는다.

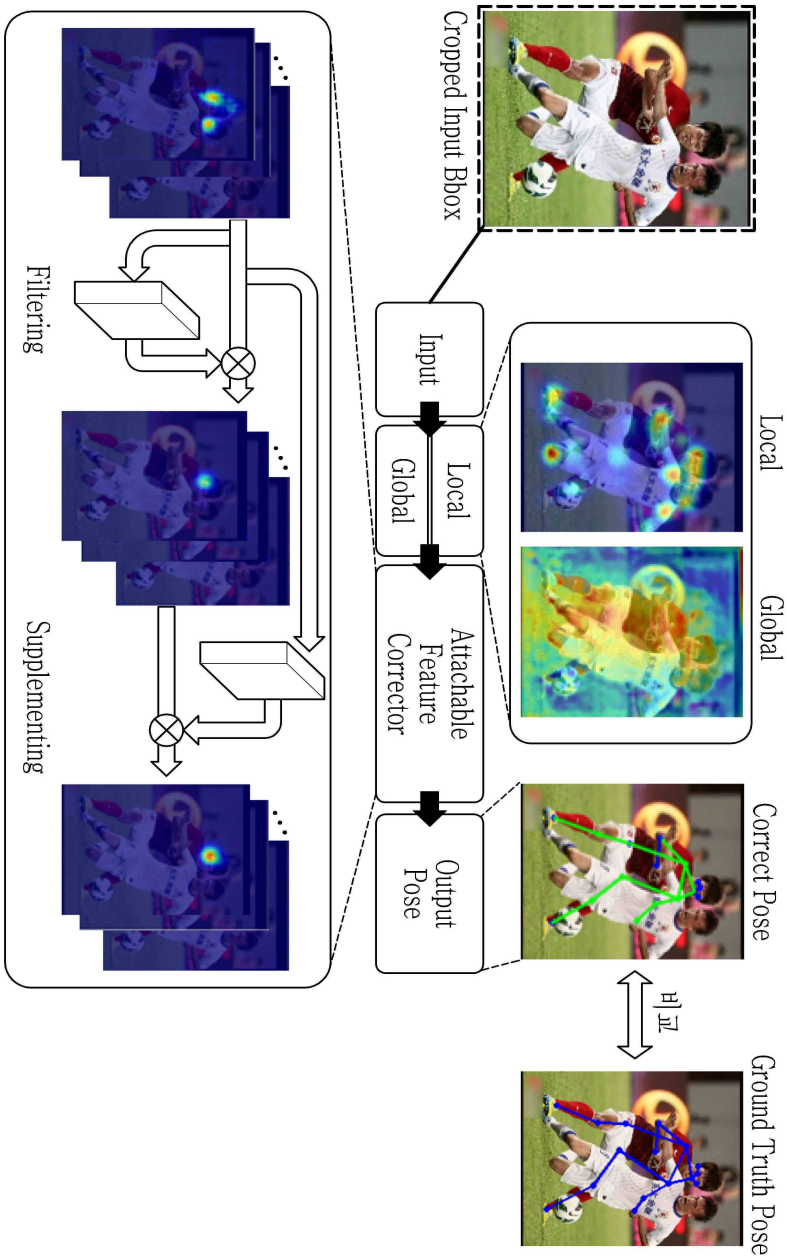
[0135] 스토리지(230)는 프로그램을 비임시적으로 저장할 수 있다. 스토리지(230)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크, 착탈형 디스크, 또는 본 개시가 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체를 포함하여 구성될 수 있다. 통신 인터페이스(240)는 유/무선 통신 모듈일 수 있다.

[0137] 이상에서 설명한 본 발명의 실시예는 장치 및 방법을 통해서만 구현이 되는 것은 아니며, 본 발명의 실시예의 구성에 대응하는 기능을 실현하는 프로그램 또는 그 프로그램이 기록된 기록 매체를 통해 구현될 수도 있다.

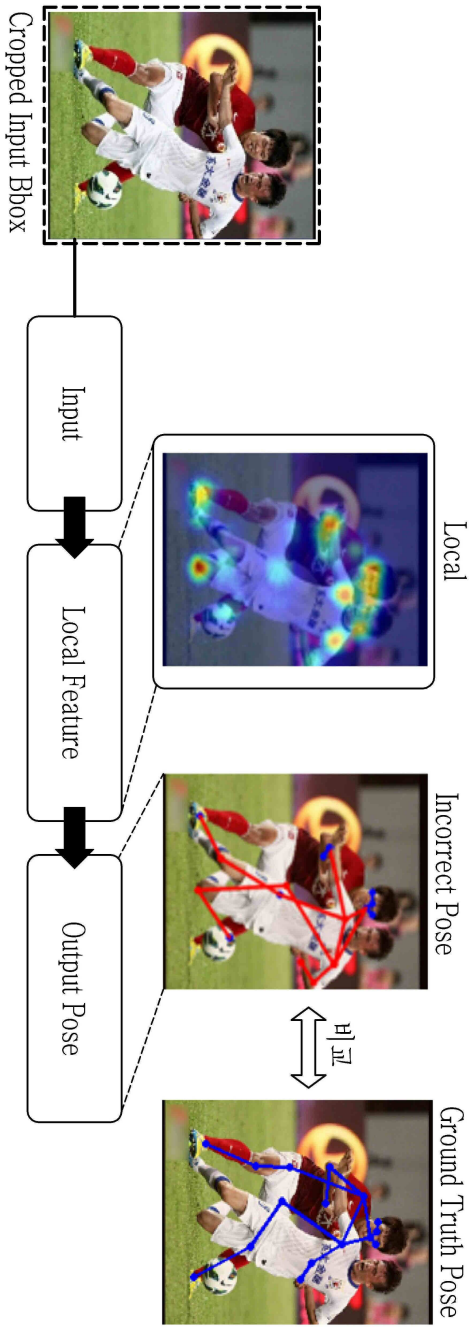
[0138] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

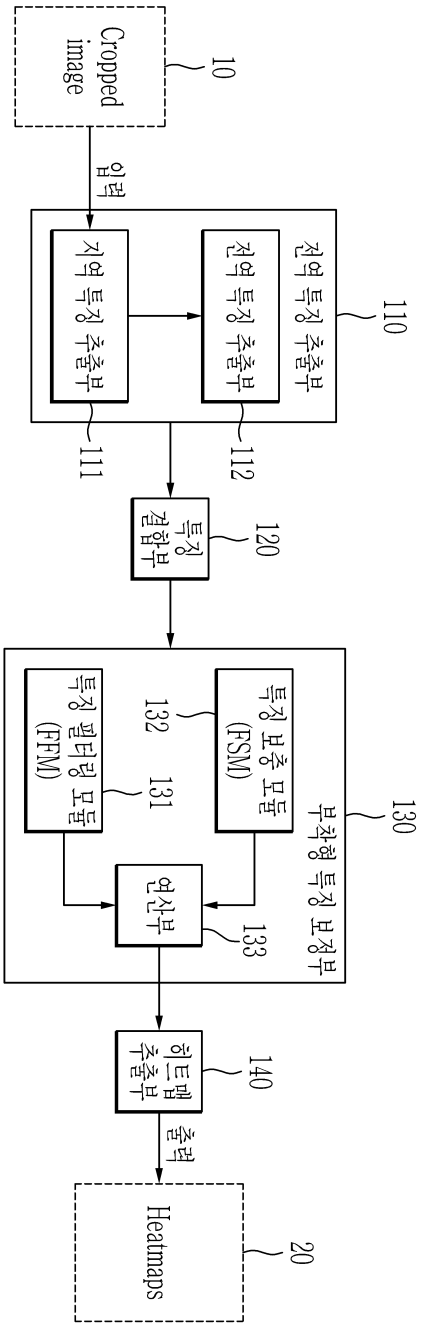
도면1


Cropped Input Bbox
Input
Local
Global
Attachable Feature Corrector
Output Pose
Local
Global
Correct Pose
비교
Ground Truth Pose
Filtering
Supplementing

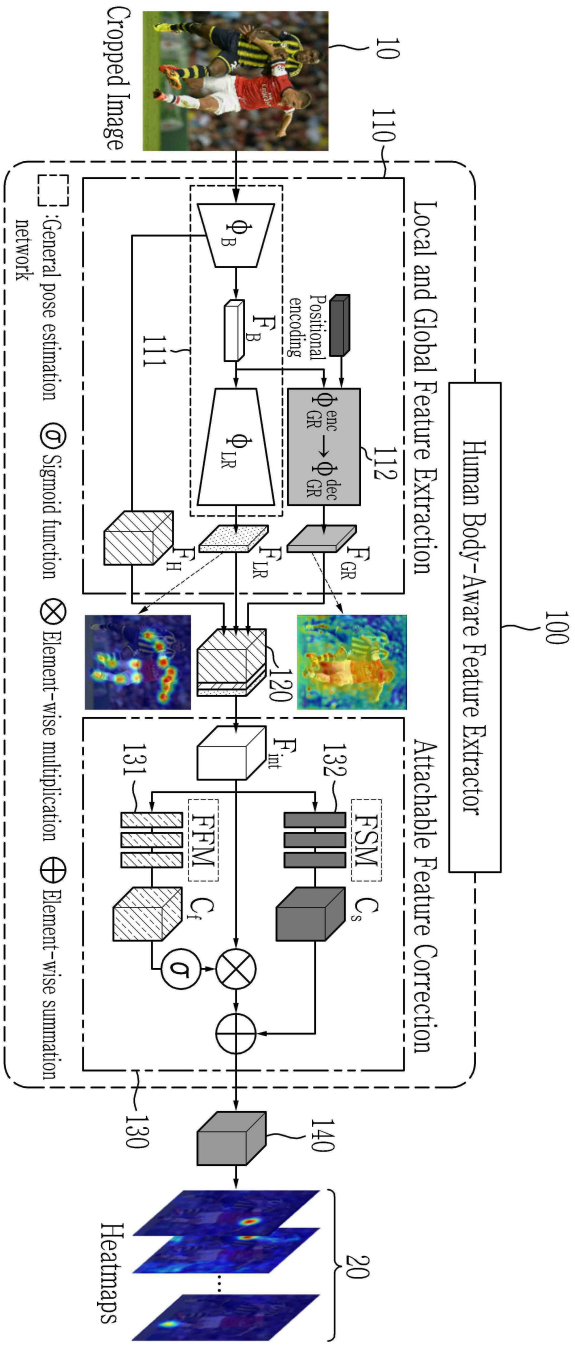
도면2


Cropped Input Bbox
Input
Local Feature
Output Pose
Local
Incorrect Pose
비교
Ground Truth Pose

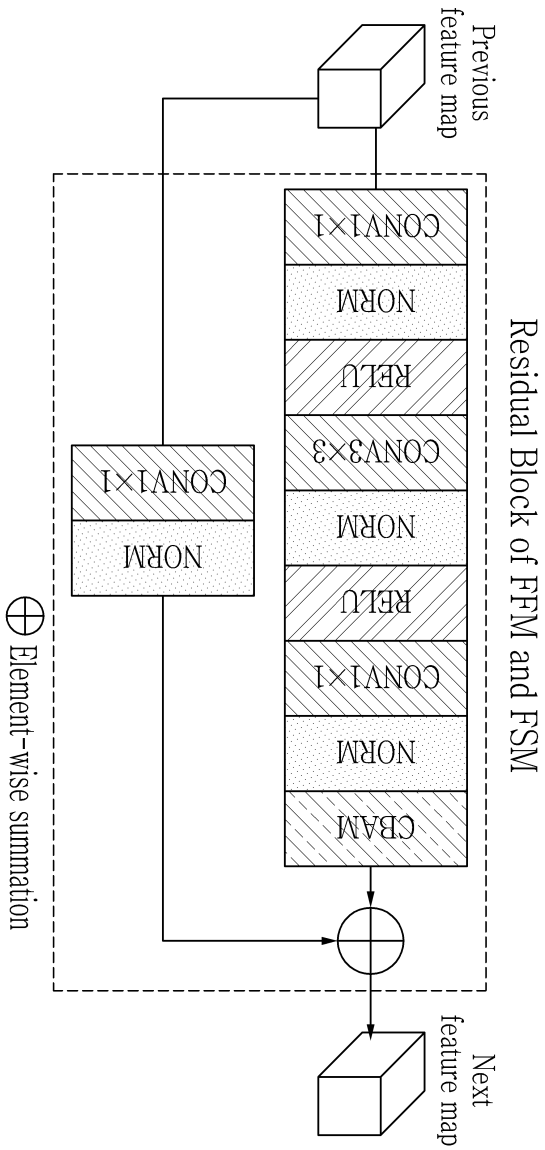
도면3


Cropped image
10
입력
전역 특징 추출부
110
지역 특징 추출부
111
전역 특징 추출부
112
특징 결합부
120
부착형 특징 보정부
130
특징 필터링 모듈 (FFM)
131
특징 보충 모듈 (FSM)
132
연산부
133
히트맵 추출부
140
출력
Heatmaps
20

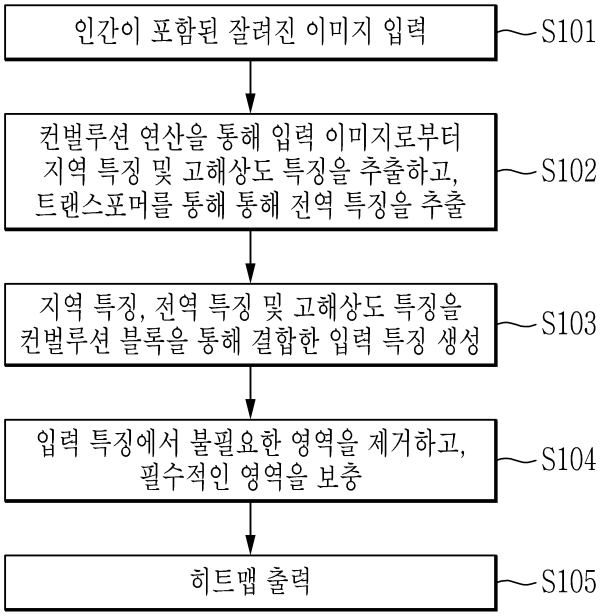
도면4


Cropped Image
10
110
Local and Global Feature Extraction
Human Body-Aware Feature Extractor
100
:General pose estimation
Sigmoid function
Element-wise multiplication
Element-wise summation
network
Positional encoding
111
112
120
Attachable Feature Correction
131
132
FFM
FSM
130
140
Heatmaps
20

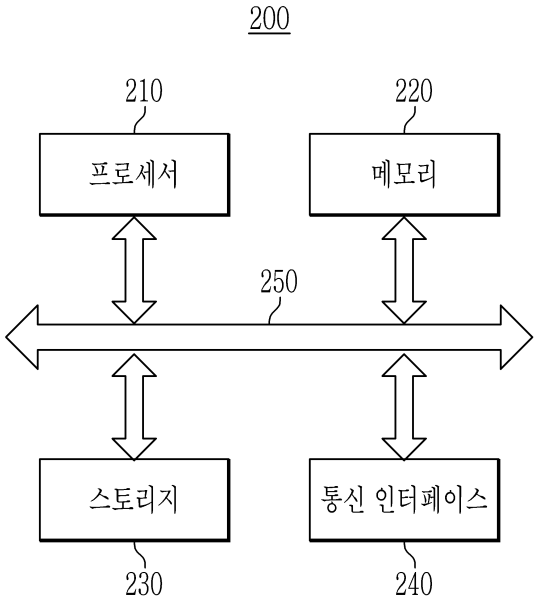
도면5


Previous feature map
CONV1×1
NORM
RELU
CONV3×3
NORM
RELU
CONV1×1
NORM
CBAM
CONV1×1
NORM
Residual Block of FFM and FSM
Element-wise summation
Next feature map

도면6


인간이 포함된 잘려진 이미지 입력
S101
컨벌루션 연산을 통해 입력 이미지로부터 지역 특징 및 고해상도 특징을 추출하고, 트랜스포머를 통해 통해 전역 특징을 추출
S102
지역 특징, 전역 특징 및 고해상도 특징을 컨벌루션 블록을 통해 결합한 입력 특징 생성
S103
입력 특징에서 불필요한 영역을 제거하고, 필수적인 영역을 보충
S104
히트맵 출력
S105

도면7


200
210
프로세서
220
메모리
250
스토리지
230
통신 인터페이스
240