(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0143146
(43) 공개일자 2024년10월02일

(51) 국제특허분류(Int. Cl.)
G06T 7/50 (2017.01) *G06N 3/0455* (2023.01)
G06N 3/0464 (2023.01) *G06T 7/20* (2017.01)
G06T 7/80 (2017.01) *H04N 7/18* (2023.01)
(52) CPC특허분류
G06T 7/50 (2017.01)
G06N 3/0455 (2023.01)
(21) 출원번호 10-2023-0038053
(22) 출원일자 2023년03월23일
심사청구일자 2023년03월23일

(71) 출원인
국립한밭대학교 산학협력단
대전광역시 유성구 동서대로 125 (덕명동)
(72) 발명자
이승호
대전광역시 유성구 엑스포로123번길 26-30 스마트시티리버뷰오피스텔 102동 601호
라승탁
대전광역시 유성구 학하중앙로 59번길 7, 205호
오승진
대전광역시 대덕구 계족산로 136, 509동 1903호
(74) 대리인
이은철, 이우영

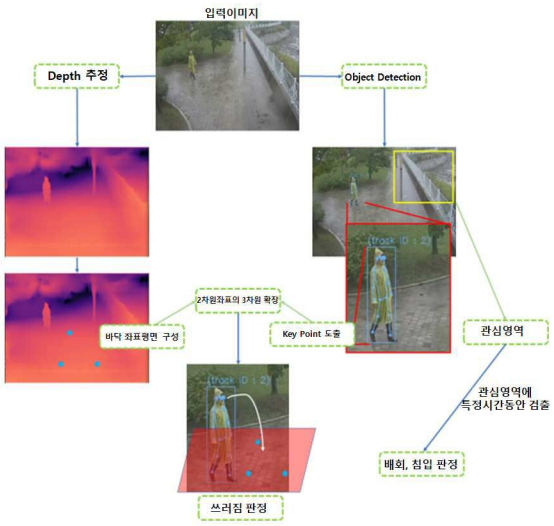
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템**

(57) 요 약

본 발명은 지능형 형태 식별 CCTV 솔루션 시스템에 관한 것으로써, 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템에 관한 것이다. 본 발명에 따르면, 입력된 RGB 이미지로부터 Depth 값을 추정하는 추정부, 입력된 RGB 이미지로부터 Object Detection을 수행하는 검출부 및 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 감지부를 포함하여 급증하는 고령자를 대상으로 부족한 요양보호사(간호사)를 보조하는 ICT 기술개발을 통한 고령자의 건강하고 안전한 삶에 기여할 수 있는 효과가 있다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/0464 (2023.01)
G06T 7/20 (2013.01)
G06T 7/80 (2017.01)
H04N 7/18 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711159933
과제번호 2022-0-00820-001
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 ICTR&D혁신바우처지원사업(예타)
연구과제명 단일 영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션
기 여 율 1/2
과제수행기관명 주식회사 공훈
연구기간 2022.04.01 ~ 2023.03.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711187165
과제번호 2022R1F1A1066371
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)
연구과제명 메타버스 컨텐츠에서 사용할 수 있는 음성에 따른 자연스러운 감정 표현이 가능한 3D 얼굴 자동 생성 솔루션 개발
기 여 율 1/2
과제수행기관명 한밭대학교
연구기간 2023.03.01 ~ 2024.02.29

명 세 서

청구범위

청구항 1

입력된 RGB 이미지로부터 Depth 값을 추정하는 추정부,

상기 입력된 RGB 이미지로부터 Object Detection을 수행하는 검출부 및

산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 감지부를 포함하는 것을 특징으로 하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템.

청구항 2

제1항에 있어서,

상기 추정부는 Window Shifted Transformer Layer와 Convolution Layer를 통한 Feature Map 추출모듈,

Deformable Self-Attention을 통한 Multi-level Hierarchical Feature Map 강화모듈 및

Decoder의 Hardtanh 함수 Loss 계산모듈을 포함하는 것을 특징으로 하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템.

청구항 3

제1항에 있어서,

상기 검출부는 Convolutional Layer를 통한 Feature Map 추출모듈,

Spatial Pyramid Pooling(SPP)과 Feature Pyramid Network(FPN)를 통한 데이터 크기를 줄여 매개변수 수를 줄이는 Multi-Scale의 Feature Pyramid 모듈 및

Decoupled Head를 통한 epoch 별 Converging 속도 개선모듈을 포함하는 것을 특징으로 하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템.

청구항 4

제1항에 있어서,

상기 감지부는 바닥 평면의 세 점을 구성하는 구성모듈 및

Object Detection에서 검출된 객체의 Key point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산, 계산된 상대거리 좌표를 이용한 쓰러짐을 감지하는 감지모듈을 포함하는 것을 특징으로 하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템.

청구항 5

단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템을 이용한 방법에 있어서,

(a)상기 시스템이 입력된 RGB 이미지로부터 Depth 값을 추정하는 단계,

(b)상기 시스템이 상기 입력된 RGB 이미지로부터 Object Detection을 수행하는 검출단계 및

(c)상기 시스템이 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 단계를 포함하는 것을 특징으로 하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 방법.

청구항 6

제5항에 있어서,

상기 (a)단계는 Window Shifted Transformer Layer와 Convolution Layer를 통한 Feature Map 추출단계,

Deformable Self-Attention을 통한 Multi-level Hierarchical Feature Map 강화단계 및

Decoder의 Hardtanh 함수 Loss 계산단계를 포함하는 것을 특징으로 하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 방법.

청구항 7

제5항에 있어서,

상기 (b)단계는 Convolutional Layer를 통한 Feature Map 추출단계,

Spatial Pyramid Pooling(SPP)과 Feature Pyramid Network(FPN)를 통한 데이터 크기를 줄여 매개변수 수를 줄이는 Multi-Scale의 Feature Pyramid를 구성하는 단계 및

Decoupled Head를 통한 epoch 별 Converging 속도를 개선하는 단계를 포함하는 것을 특징으로 하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 방법.

청구항 8

제5항에 있어서,

상기 (c)단계는 바닥 평면의 세 점을 구성하는 단계 및

Object Detection에서 검출된 객체의 Key point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산, 계산된 상대거리 좌표를 이용한 쓰러짐을 감지하는 단계를 포함하는 것을 특징으로 하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 지능형 형태 식별 CCTV 솔루션 시스템에 관한 것으로써, 보다 상세하게는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템에 관한 것이다.

배 경 기 술

[0002] 최근 인구 고령화가 빠르게 진행되면서 요양시설 내 돌봄에 대한 관심이 높아지고 있기 때문에 영상 기반 다중 객체(휴먼) 식별과 위치 측정 및 추적 기술은 요양 돌봄의 편리성을 높일 수 있는 매우 중요한 분야라고 할 수 있다.

[0003] 기존 영상 기반의 대표적인 위치 측정 및 추적 기술의 예로는 체스판 모양의 체커 보드를 이용하여 실세계의 좌표를 정의하고 카메라 내부 파라미터를 산출 및 렌즈 왜곡을 보정하여 영상 좌표로부터 3차원 공간좌표를 복원하는 수동 카메라 Calibration 방식이 존재한다.

[0004] 그러나 수동 카메라 Calibration 방식은 낙상 및 쓰러짐 감지를 위하여 요양원 등에 적용하기에는 체커 보드와 같은 기준점이 카메라를 설치할 때 항상 필요하다는 점이고, 카메라와 체커 보드의 파라미터를 실측을 통해 구해야한다는 단점이 존재한다. 따라서 범용적인 요양원 등과 같은 시설에 적용하기에는 많은 제약이 있었다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 등록특허 1743689(2017.05.30. 등록)

발명의 내용

해결하려는 과제

[0006] 본 발명은 상술한 문제를 해결하고자, 기존의 수동 CCTV Caibration이 필요한 기법을 딥러닝 네트워크로부터 산출된 Depth 추정을 이용하여 대체하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템을 제공함에 목적이 있다.

[0007] 또한 항상 실제 해당 지역에 가서 수동으로 Calibration을 진행해야하던 문제를 Window Shifted Transformer를 통한 Depth 추정을 이용하여 자동화하므로 시간적, 인적인 자원을 확보하는데 있다.

과제의 해결 수단

[0008] 상기 목적을 달성하기 위하여 본 발명은 입력된 RGB 이미지로부터 Depth 값을 추정하는 추정부, 상기 입력된 RGB 이미지로부터 Object Detection을 수행하는 검출부 및 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 감지부를 포함한다.

발명의 효과

[0009] 본 발명에 따르면, 급증하는 고령자를 대상으로 부족한 요양보호사(간호사)를 보조하는 ICT 기술개발을 통한 고령자의 건강하고 안전한 삶에 기여할 수 있는 효과가 있다. 개발 ICT 기술을 이용한 안전관제를 통해 고령환자, 제조 근로자의 안전사고로 인한 재정지출(의료보험)의 감소에 기여할 수 있는 효과가 있다.

도면의 간단한 설명

[0010] 도 1은 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템의 전체 개요도이다.

도 2는 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템의 학습 과정에 대한 흐름도이다.

도 3은 Window Shifted Transformer Layer와 ViT의 Patch 크기에 따른 연산 차이를 나타낸 것이다.

도 4는 Window Shifted Transformer Layer를 사용한 Feature Map의 추출과정 및 구조를 나타낸다.

도 5는 Deformable Self-Attention에 입력된 Feature Map이 연산되는 과정을 나타낸 것이다.

도 6은 Hardtanh함수를 나타낸 것이다.

도 7은 Object Detection 학습과정에 대한 흐름도이다.

도 8은 Shortcut Connection이 적용된 Object Detection 모델의 네트워크의 구조이다.

도 9는 Spatial Pyramid Pooling의 구조를 나타낸 도면이다.

도 10은 Feature Pyramid Network의 구조를 나타낸 것이다.

도 11은 MS COCO Dataset을 기본적인 YOLOv5 구조와 Decoupled Head를 적용한 구조들의 Epoch별 AP 달성 비교 그래프이다.

도 12는 2차원 Pixel 좌표로부터 Depth 값을 이용한 3차원 좌표 변환 과정이다.

도 13은 각 Key Point에 대한 3차원 좌표로 변환을 나타낸 도면이다.

도 14는 배회, 침입 감지를 나타낸 도면이다.

도 15는 제안된 솔루션의 실행 과정의 흐름도이다.

도 16은 NYU Depth Dataset 예시를 나타낸 도면이다.

도 17은 Pascal VOC 2007 및 2012의 예시를 나타낸 도면이다.

도 18은 Depth 추정 모델의 Test 결과를 나타낸다.

도 19는 Object Detection 모델의 Test 결과를 나타낸다.

도 20은 이미지 객체 Depth 추정과 다중객체검출 및 추적에 대한 실험 결과를 나타낸 것이다.

도 21은 각각 이미지 객체 Depth 추정과 다중객체검출 및 추적에 대한 공인시험기관의 시험성적서이다.

도 22는 Patch Size에 따른 Feature Map 추출의 예시를 나타낸다.

도 23은 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템의 구성도이다.

도 24는 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 방법의 전체흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0011] 이하, 첨부된 도면들에 기재된 내용들을 참조하여 본 발명을 상세히 설명한다. 다만, 본 발명이 예시적 실시 예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일 참조부호는 실질적으로 동일한 기능을 수행하는 부재를 나타낸다.

[0012] 본 발명의 목적 및 효과는 하기의 설명에 의해서 자연스럽게 이해되거나 보다 분명해질 수 있으며, 하기의 기재만으로 본 발명의 목적 및 효과가 제한되는 것은 아니다. 또한, 본 발명을 설명함에 있어서 본 발명과 관련된 공지 기술에 대한 구체적인 설명이, 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다.

[0013] 최근에는 딥러닝 네트워크를 이용한 단일영상 이미지로부터 Depth 값을 추정하고 해당 Depth 값을 이용하여 상대거리 좌표를 구하는 연구가 꾸준히 진행되고 있다. 공간좌표를 도출하기 위해 사용되던 기존의 카메라 Calibration 방식에 비해 단일 이미지만을 사용하여 Depth 값과 상대거리 좌표를 도출할 수 있어 인적 자원과 초기 설치 비용이 저렴하고 확장성 및 범용성 확보에 유리하며 무엇보다 단일 이미지로 상대거리 좌표를 획득하기 쉽다는 큰 장점이 있다.

[0014] 본 발명에서는 기존에 체스판 모양의 체커 보드를 이용하여 실세계의 좌표를 정의하고 카메라 내부 파라미터를 산출 및 렌즈 왜곡을 보정하는 작업을 필요로 했던 기존의 카메라 Calibration 방식을 대체하기 위해, 딥러닝 기반의 Depth Estimation 방법 및 단일 이미지만을 사용한 고령자 등의 낙상과 쓰러짐을 판단하는 지능형 형태 식별 CCTV 솔루션을 제안하였다. 딥러닝 기반의 Depth Estimation을 통해 수동으로 구역마다 적용해야하는 카메라 Calibration을 생략하고 상대거리 좌표를 구할 때 딥러닝 네트워크로부터 산출된 Depth 값을 이용하여 해결하고자 하였다.

[0015] 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템은 기존의 수동 CCTV Caibration이 필요한 기법을 딥러닝 네트워크로부터 산출된 Depth 추정을 이용하여 대체하는데 있다. 이로써, 항상 실제 해당 지역에 가서 수동으로 Calibration을 진행해야하던 문제를 Window Shifted Transformer를 통한 Depth 추정을 이용하여 자동화하므로 시간적, 인적 자원을 확보할 수 있다.

[0016] 또한 2차원 평면 좌표를 추정한 Depth 값을 이용하여 3차원 좌표를 구축함으로써, 딥러닝 네트워크로부터 산출된 Depth 값과 SPP, FPN 및 Decoupled Head가 적용된 Object Detection 모델로 검출한 객체의 2차원 평면 좌표로부터 3차원 가상의 상대거리 좌표를 구축하여 쓰러짐 판단의 정확성을 증대시킬 수 있다.

[0017] 본 실시예에 따르면, 기존에 센서나 CCTV Calibration 등을 필요로 했던 기존의 쓰러짐 판단 방법에서 딥러닝 기반의 Depth Estimation 방법 및 단일 이미지만을 사용하여 고령자 등의 쓰러짐을 판단하는 지능형 형태 식별 CCTV 솔루션을 제공함에 있다.

[0018] 도 1은 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템의 전체 개요도이다. 도 1에 도시된 바와 같이, 본 발명에서 제안하는 기법은 단일 RGB 이미지만을 사용하여 Depth 값을 추정하고 3차원 가상의 상대거리 좌표를 구축한다. 제안된 기법은 크게 3가지 과정으로 구성되며 각 과정은 도 23에 도시된 바와 같이, 입력된 RGB 이미지로부터 Depth 값을 추정하는 추정부(110), 입력된 RGB 이미지로부터 Object Detection을 수행하는 검출부(120), 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point(관절 부위)를 이용하여 배회, 침입, 쓰러짐을 감지하는 감지부(130)로 이루어진다.

[0019] 먼저 첫 번째로, 입력된 RGB 이미지로부터 Depth 값을 추정하는 추정부(110)는 딥러닝 네트워크를 이용한 학습을 통해 산출할 수 있는데, Window Shifted Transformer Layer와 Convolution Layer를 통한 Feature Map 추출 모듈(111), Deformable Self-Attention을 통한 Multi-level Hierarchical Feature Map 강화모듈(112),

Decoder의 Hardtanh 함수 Loss 계산모듈(113)로 구분된다.

[0020] Window Shifted Transformer Layer와 Convolution Layer를 통한 Feature Map 추출 과정에서는 입력 RGB 이미지에 대한 특징점인 Feature Map을 추출한다. Feature Map의 내용은 입력된 RGB 이미지가 어떤 모양의 사물을 촬영한 것인지, 어떤 방향으로 촬영되었는지, 어떤 색상인지 등으로 인해 달라진다.

[0021] Deformable Self-Attention을 통한 Multi-level Hierarchical Feature Map 강화 과정에서는 Window Shifted Transformer Layer와 Convolution Layer에서 추출된 Feature Map를 정합하여 강화한다.

[0022] Window Shifted Transformer Layer와 Convolution Layer로부터 각각 전달받은 Feature Map을 계층적 구조로 강화하는 것의 이점은 각 RGB 이미지가 Window Shifted Transformer Layer와 Convolution Layer를 통과하여 생성된 Feature Map이 각각 다른 정보들에 가중치를 두어 Feature Map의 생성을 더욱 유의미하게 만들어준다. 결과적으로 더 많은 깊이 정보의 값이 가중치를 가지고 학습을 진행하여 범용성과 확장성을 가지고 딥러닝 네트워크의 정확성을 높여준다.

[0023] Decoder의 Hardtanh 함수 Loss 계산과정에서는 입력 RGB 이미지로부터 Window Shifted Transformer Layer와 Convolution Layer 및 Deformable Self-Attention을 통과하여 생성된 Predicted Depth Map를 Upconvolutional Layer를 통해서 작은 차원의 Predicted Depth Map을 확장시키고, 학습 과정 중에 사용되는 검증 Depth Map 데이터인 Ground Truth Depth Map을 비교하여 가중치 값을 수정한다. Hardtanh 함수는 가중치 값의 범위를 min_val과 max_val로 조정할 수 있으며, 가중치 값의 범위가 제한적인 모델에서 사용하기 좋은 함수이다. 생성되는 Depth Map의 범위가 한정되어 있기에 Hardtanh 함수를 사용하여 Loss를 업데이트하였다.

[0024] 두 번째로, 입력된 RGB 이미지로부터 Object Detection을 수행하는 검출부(120)는 딥러닝 네트워크 학습을 통하여 수행할 수 있는데, 53층의 Convolutional Layer를 통한 Feature Map 추출모듈(121), Spatial Pyramid Pooling(SPP)과 Feature Pyramid Network(FPN)를 통한 데이터 크기를 줄여 매개변수 수를 줄이는 Multi-Scale의 Feature Pyramid 모듈(122), Decoupled Head를 통한 epoch 별 Converging 속도 개선모듈(123)로 구분된다.

[0025] 53층의 Convolutional Layer를 이용하여 입력 RGB 이미지로부터 Feature Map을 추출한다. 또한, Residual Block을 적용하여 가중치가 높은 Feature Map을 골라 학습을 함으로 학습이 쉬워지고 Gradient를 더욱 잘 수렴할 수 있게 해준다.

[0026] SPP는 53층의 Convolutional Layer를 통과한 각기 다른 사이즈의 Feature Map을 고정된 크기로 변환시켜주며 이는 Head(Decoder)에서 이미지를 따로 변환시켜줄 필요가 없어 학습 시간 단축에 효과적이다. 또한 FPN은 피라미드 구조의 각기 다른 해상도의 Feature Map을 쌓아올린 형태이며 여러 가지 조건에서의 가중치가 높은 Feature Map을 사용하여 Object Detection에 기여할 수 있도록 한다.

[0027] Decoupled Head는 기존에 사용되던 하나의 Head 구조와 달리 여러 개의 Head를 사용하여 학습을 수행하면서 Epoch(학습 수행 횟수) 별 Gradient 수렴의 속도가 더욱 빠르게 진행될 수 있도록 한다.

[0028] 마지막으로, 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 감지부(130)는 바닥 평면의 세 점을 구성하는 구성모듈(131), Object Detection에서 검출된 객체의 Key point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산, 계산된 상대거리 좌표를 이용한 쓰러짐을 감지하는 감지모듈(132)로 구분된다.

[0029] 바닥 평면의 세 점을 구성하는 과정은 Depth Map을 산출하는 학습 모델에 사용된 공개된 표준 데이터베이스인 NYU Depth V2 데이터셋을 구성하는데 사용된 카메라의 내부 파라미터를 적용시켜 2차원인 세 점을 3차원으로 차원확장을 수행한다. 그 후, 3차원으로 확장된 세 점을 이용하여 평면의 방정식을 구성, 가상의 평면 좌표를 생성한다.

[0030] 또한 Object Detection에서 검출된 객체의 2차원 Key Point 좌표를 3차원으로 차원확장을 수행한 후, 이전에 구성된 평면의 방정식에 대입하여 객체의 각 Key Point 별 바닥으로부터의 상대거리를 계산한다.

[0031] 배회 및 침입은 추적중인 객체가 지정된 관심영역에 특정 시간동안 검출될 시 감지된다. 쓰러짐은 계산된 상대거리와 구성된 바닥 평면과의 거리를 실시간으로 계산하여 감지된다.

[0032] 본 발명에서 제안된 기법의 객관적인 신뢰성을 평가하기 위해 공개된 표준 데이터베이스인 NYU Depth V2 데이터셋과 Pascal VOC 2007 데이터셋을 이용하여 다른 발명에서 발표한 기법들과 비교 평가하였다.

[0033] 실험결과, 본 발명에서 제안하는 기법의 Depth 추정의 평가지표이며 낮을수록 에러율이 적은 RMSE값이 0.40으로 나타내어 다른 발명에서 발표한 기법들보다 낮은 수치를 나타내었다. 또한 Object Detection의 평가지표이며 높을수록 예측 값의 신뢰도가 높은 AP의 수치가 0.8로 다른 발명에서 발표한 기법들보다 높은 수치를 나타내었다. RMSE와 AP 모두 다른 발명에서 발표한 기법들보다 향상된 수치를 나타내어 그 효율성이 입증되었다.

[0034] 마지막으로, 본 발명에서 제안하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션을 KISA(한국인터넷진흥원)의 지능형 CCTV 인증시험의 배회, 침입, 쓰러짐 부문 배포 데이터셋으로 실험한 결과 70개의 영상 중 65개의 형태 식별을 탐지하여 그 효율성이 입증되었다.

[0035] 도 2는 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템의 학습 과정에 대한 흐름도이다. 본 실시예의 Depth 추정 기법의 학습 과정은 Window Shifted Transformer Layer와 Convolutional Layer를 사용하여 Feature Map을 추출하고 Deformable Self-Attention으로 Multi-level Hierarchical Feature Map을 강화, 그 후 Decoder의 Hardtnah 함수로 Loss를 계산한다.

[0036] 먼저, 입력된 RGB이미지로부터 Depth값을 추정하는 과정을 설명하기로 한다.

[0037] 도 2의 Window Shifted Transformer Layer와 Convolution Layer를 통한 Feature Map 추출 과정(S1)에서, 이미지는 Text에 비해 Patch를 어떻게 분할하느냐에 따라 다양한 variant가 존재하고, 이미지의 Resolution이 크다는 차이점이 있다. 따라서 이 차이점을 다루기 위해 기존의 Vision Transformer(ViT)가 아닌 Window Shifted Transformer Layer는 Patch를 Window로 나누어 해당 Window 안에서만 Self Attention을 수행하고 그 윈도우를 Shift한 뒤 다시 self attention을 수행한다. 도 3은 Window Shifted Transformer Layer와 ViT의 Patch 크기에 따른 연산 차이를 나타낸다.

[0038] 도 3은 Window Shifted Transformer Layer와 ViT의 Patch 크기에 따른 연산 차이를 나타낸 것이다. 본 실시예의 Window Shifted Transformer Layer와 Convolutional Layer를 이용한 Feature Map 추출 과정에서, 4x4의 작은 Patch에서 시작해 점점 Patch들을 Merge 해나가는 방식을 취하여 피라미드 구조의 연산 방식을 채택한다.

[0039] 입력 이미지의 사이즈를 224x224로 가정하였을 때, ViT는 Patch Size를 16pixel x 16pixel로 만들어 총 (244/16)2 = 196개의 Patch를 가진 상태를 유지하고 각 Patch와 나머지 전체 Patch에 대한 Self-Attention을 수행한다. 반면, Window Shifted Transformer Layer는 마치 Feature Pyramid Network(FPN)처럼 작은 Patch 4x4에서 시작해 점점 Patch들을 Merge 해나가는 방식을 취한다. 이런 특징으로 Window Shifted Transformer Layer는 Computation Cost를 줄이고 입력 데이터의 중요한 특징을 추출하는 것에 높은 성능을 발휘하며 많은 논문에서 사용되고 있다.

[0040] 본 발명에서는 4개의 스테이지로 구성된 Window Shifted Transformer Layer를 사용하여 640×480 크기의 입력 RGB 이미지로부터 15×20×24 크기의 Feature Map을 추출한다.

[0041] 도 4는 Window Shifted Transformer Layer를 사용한 Feature Map의 추출과정 및 구조를 나타낸다. 본 실시예의 Window Shifted Transformer Layer와 Convolutional Layer를 이용한 Feature Map 추출 과정에서, Patch를 이용한 피라미드 구조의 연산 방식 때문에 Computation Cost가 적고 입력 데이터의 중요한 특징을 추출하는 것에 높은 성능을 발휘한다.

[0042] Deformable Self-Attention에 Window Shifted Transformer Layer의 Feature Map 및 Convolution Layer를 이용하여 별도의 Feature Map을 전달하기 위하여 Convolution Layer의 Feature Map을 별도로 추출한다. 본 발명에서 사용되는 Conv Layer의 사이즈는 1x1 이며, 1x1로 설정한 이유는 Window Shifted Transformer Layer Branch에서 추출된 Feature Map이 Conv Layer에 비해 상대적으로 특징점이 많이 소실되므로, Conv Layer에서 원본 이미지의 특징점을 최대한 살리는 것에 목적이 있다.

[0043] 다음으로 도 2의 Deformable Self-Attention을 통한 Multi-level Hierarchical Feature Map 강화 과정(S2)을 설명하면 다음과 같다.

[0044] Window Shifted Transformer Layer와 Conv Layer에서 추출된 Feature Map을 Deformable Self-Attention으로 전달하여 Deformable self Attention을 수행한다. Deformable Self-Attention은 앞서 Window Shifted Transformer Layer와 Conv Layer에서 추출된 Feature Map을 좀 더 유의미한 영역으로 Sampling 한다. 또한 입력 Patch의 수보다 적은 Reference Point에 대해 Offset을 계산하는 것으로 Self-Attention의 연산량을 줄여 학습 속도에 기여하게 한다. Deformable Self-Attention에 입력된 Feature Map이 연산되는 과정은 도 5와 같다.

[0045] 도 5는 Deformable Self-Attention에 입력된 Feature Map이 연산되는 과정을 나타낸 것이다. 본 실시예의 Deformable Self-Attention을 이용한 Multi-level Hierarchical Feature Map 강화에서, Deformable Self-Attention은 앞선 Window Shifted Transformer Layer와 Convolutional Layer에서 전달 받은 Feature Map을 좀 더 유의미한 영역으로 Smapling한다.

[0046] 다음으로 도 2의 Decoder의 Hardtanh 함수 Loss 계산과정(S3)을 설명하면 다음과 같다.

[0047] Window Shifted Transformer Layer와 Conv Layer에서 추출된 Feature Map이 Deformable Self-Attention을 통과하여 특징점이 강화되면 Upsampling을 수행하는 Decoder에 입력되게 된다.

[0048] Decoder는 Upconvolutional Layer가 4층으로 구성되어있으며 Window Shifted Transformer Layer에서 생성된 Feature Map은 1x1 Convolution 연산을 거쳐 2배의 크기로 Upsampling 된다. 이후 Deformable Self-Attention을 통과한 Convolutional Layer에서 생성된 Feature Map과 결합 후 3x3 convolution 연산하여 640x480 이미지를 생성하게 된다. Loss 함수로는 Hardtanh 함수를 사용하며 도 6과 같다.

[0049] 도 6은 Hardtanh함수를 나타낸 것이다. 본 실시예의 Decoder의 Hardtanh 함수를 통한 Loss 계산에서, Hardtanh 함수는 Depth 값의 범위 같은 제한적인 영역에서의 결과가 좋다.

[0050] Hardtanh 함수를 사용하여 계산된 Loss로 가중치 값을 업데이트하고 생성된 Depth 이미지와 Ground Truth 이미지의 RMSE를 계산한다. RMSE 계산식은 식 1과 같고 계산된 RMSE가 목표치에 도달하면 학습을 종료한다.

[0051] (식 1)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

[0052]

[0053] 여기서, RMSE는 Root Mean Square Error(RMSE)로서, 추정 값 또는 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 사용되는 척도이며, 0에 가까울수록 높은 성능을 갖는다. y_i와 $\hat{y}_i$는 각각 추정된 Depth 값과 실제 Depth 값인 Ground Truth(GT) 값이다.

[0054] 성능평가를 위한 값으로는 Depth 추정은 RMSE를, Object Detection은 AP를 사용한다. 여기서, AP는 Average Precision(AP)로서, 물체의 검출과 모델의 비교에 적용되는 방식이며, 1에 가까울수록 높은 성능을 갖는다.

[0055] 도 7은 Object Detection 학습과정에 대한 흐름도이다. 도 7의 Object Detection 학습과정에서, 53 층의 Convolutional Layer를 통하여 Feature Map을 추출한 후 SPP와 FPN으로 Multi-Scale의 Feature Pyramid를 구성하여, Decoupled Head를 적용하여 Epoch별 Converging 속도를 개선한다.

[0056] 입력된 RGB 이미지로부터 Object Detection을 수행하는 검출부(120)를 도 7을 참조하여 설명하면 다음과 같다.

[0057] 도 7에 따르면, 53층의 Convolutional Layer를 통한 Feature Map 추출 과정(S71)에서, 사용된 네트워크는 53층의 Convolution Layer로 구성되어 있으며, Resnet에서 제안된 Shortcut Connection을 사용한다. 일반적인 학습 네트워크는 추출되는 Feature가 뒤 쪽까지 영향이 직접적으로 전달되는 것이 아니라, 중간을 거쳐 전달되기 때문에 학습의 과정에서 매우 큰 변화가 일어난다.

[0058] 하지만 Shortcut Connection을 적용하면 이전으로부터 얼마만큼 변하는지 잔차(Residual)만 계산하는 문제로 바뀌게 된다. 따라서, 현재 Layer의 출력 값과 이전 Scale의 Layer 출력 값을 더해 입력을 받기 때문에 그 차이를 볼 수 있고 더 빠르게 학습을 할 수 있다는 장점이 있다. Shortcut Connection이 적용된 Object Detection 모델의 네트워크의 구조는 도 8과 같다.

[0059] 도 8은 Shortcut Connection이 적용된 Object Detection 모델의 네트워크의 구조이다. Convolutional Layer 53개를 통한 Feature Map 추출에서, 본 발명에서 사용된 Object Detection 네트워크의 Backbone은 도 8과 같다.

[0060] 도 7의 Spatial Pyramid Pooling(SPP)과 Feature Pyramid Network(FPN)를 통한 데이터 크기를 줄여 매개변수 수를 줄이는 Multi-Scale의 Feature Pyramid 구성(S72)을 설명하면 다음과 같다.

[0061] 53층의 Convolutional Layer로 구성된 Encoder를 통과한 Feature Map은 Fully Connected Layer(FCL)을 통과하여 Vectorize되기 전에 Pooling Layer를 통과하게 된다. Pooling은 데이터의 크기를 줄이며 이 과정으로 인해 모델의 전체 매개변수 수를 줄이는 과정이다. Spatial Pyramid Pooling은 Feature Map이 FCL을 통과하기 전에 동일한 크기로 조절해주는 역할을 한다. Darknet-53을 통과한 Feature Map을 미리 정해져 있는 4x4, 2x2, 1x1의 피라미드 영역으로 나누어준다. 해당 피라미드 한 칸을 bin이라고 하며 64x64x256 크기의 Feature Map이 들어오면 4x4 피라미드의 bin 크기는 16x16이 된다. 각 bin에서 가장 큰 값만을 추출하는 Max Pooling을 수행한 후, 그 결과 값들은 이어붙여 Vectorize 시켜준다. SPP의 구조는 도 9와 같다.

[0062] 도 9는 Spatial Pyramid Pooling의 구조를 나타낸 도면이다. 도 9의 SPP와 FPN을 이용한 Multi-Scale의 Feature Pyramid 구성에서, Spatial Pyramid Pooling(SPP)는 데이터의 크기를 줄이며 모델의 전체 매개변수를 줄일 수 있도록 Feature Map이 Fully Connected Layer(FCL)를 통과하기 전에 동일한 크기로 조절해주는 역할을 한다.

[0063] FPN은 Computation Cost를 낮추면서 다양한 크기의 객체를 인식할 수 있도록 한다. Darknet-53을 통과하는 Feature Map들이 각 Size에 따라 FPN을 통과하는데, 해당 Size에 따라 Feature Map을 쌓아 올린 형태를 피라미드 구조라고 한다. SPP를 통과한 Feature Map은 동일한 해상도를 가지게 되는데, 이 Feature Map을 FPN에 입력시켜 2배씩 작아지는 Feature Map을 추출한다. FPN의 구조는 도 10과 같으며, FPN으로 인해서 매우 강한 Feature들만 남아 Object Detection을 수행할 때 오검출을 줄여주는 장점을 갖는다.

[0064] 도 10은 Feature Pyramid Network의 구조를 나타낸 것이다. 도 10의 SPP와 FPN을 이용한 Multi-Scale의 Feature Pyramid 구성에서, Feature Pyramid Network(FPN)는 Computation Cost를 낮추면서 다양한 크기의 객체를 인식할 수 있도록 하며 강한 Feature들을 걸러내어 Object Detection 시에 오검출을 줄여주는 역할을 한다.

[0065] 도 7의 Decoupled Head를 통한 epoch 별 Converging 속도 개선 과정(S73)을 설명하면 다음과 같다.

[0066] 기존 Decoder들은 하나의 Head를 가지고 있었지만 본 발명에서는 여러 개의 Head를 가지는 Decoupled Head를 Decoder에 적용시켜 Object Detection의 성능지표인 AP를 더욱 빠른 시간 안에 도달할 수 있도록 하였다. 기존 Decoder는 Classification, Regression, IoU에 대한 Prediction을 전부 합쳐 하나의 Vector로 만들고 하나의 Head로 학습을 진행하였다. Decoupled Head의 장점은 Epoch 별 학습이 진행됨에 따라 단일 Head를 가진 Decoder보다 가중치의 수렴 속도가 월등히 빠르다는 점이다.

[0067] 도 11은 MS COCO Dataset을 기본적인 YOLOv5 구조와 Decoupled Head를 적용한 구조들의 Epoch별 AP 달성 비교 그래프이다. 그래프 상에 보여지다시피 단일 Head를 사용한 일반적인 YOLOv5의 구조보다 Decoupled Head를 적용한 구조가 Converging 속도 면에서 월등한 것을 볼 수 있다.

[0068] Decoupled Head를 적용하여 Epoch 별 Converging 속도 개선에 있어서, 본 발명에서는 여러 개의 Head를 Decoder에 적용시켜 Object Detection의 성능지표인 목표 AP를 더욱 빠른 시간 안에 도달할 수 있도록 한다.

[0069] 본 실시예의 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 과정은 Depth 이미지의 바닥 평면 세 점을 구성하고, Object Detection에서 검출된 객체의 Key Point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산하며, 계산된 상대거리 좌표를 이용한 쓰러짐을 감지한다.

[0070] 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point(관절 부위)를 이용하여 배회, 침입, 쓰러짐을 감지하는 감지부(130)에 대한 설명을 하면 다음과 같다.

[0071] (1) 바닥 평면의 세 점을 구성

[0072] 단일 입력 RGB 이미지로부터 Depth 추정이 완료되어 Depth Map이 생성되면, 생성된 Depth Map에 세 점을 찍어 기존 2차원 좌표에서 3차원 좌표로 변환한다. 이 때, 카메라 내부 파라미터가 필요하므로 NYU Depth V2 Dataset을 구성할 때 사용했던 표 1의 카메라 내부 파라미터를 적용하여 식 2와 같이 3차원 좌표를 도출해낸다. 도 12는 2차원 Pixel 좌표로부터 Depth 값을 이용한 3차원 좌표 변환 과정이다. 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 과정에서, Depth 이미지의 바닥 평면 세 점을 구성하고, 도출된 평면의 방정식을 기반으로 법선벡터를 계산한다.

[0073] 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 과정에서, Depth 이미지의 바닥 평면 세 점을 구성하며, 본 발명에서 2차원 좌표를 3차원 좌표로 재구축을 하기 위해서는 카메라 내부 파라미터가 필요하므로 NYU Depth V2 Dataset을 구성할 때 썼던 파라미터

를 적용하여 아래 식과 같이 3차원 좌표를 도출한다.

[0074] 표 1은 NYU Depth V2 데이터셋의 카메라 내부 파라미터를 나타낸다.

[0075] (표 1)

초점 거리		광학 중심	
FX	FY	CX	CY
5.8262448167737955e+02	8.8269103270988637e+02	3.1304475870804731e+02	2.3844389626620386e+02

[0076]

[0077] (식 2)

$$Z = img_{depth}[py, px]$$
$$X = ((px - CX) \times Z)/FX$$
$$Y = ((py - CY) \times Z)/FY$$

[0078]

[0079] 여기서, Z는 픽셀좌표 px, py 위치의 추정된 Depth 값임과 동시에 3차원으로 확장될 좌표의 Z축 값이고, X는 3차원으로 확장될 좌표의 X축 값, Y는 3차원으로 확장될 좌표의 Y축 값이다. px와 py는 각각 픽셀좌표의 x, y값이고, CX, CY는 광학 중심의 X, Y 값이며, FX, FY는 각각 초점 거리의 X, Y 값이다.

[0080] 또한 바닥 평면 3개의 3차원 좌표로부터 평면의 방정식을 도출해낸다. 예를 들어 점 P, Q, R이 (x, y, z)과 같은 형식이고 P(-3.99, 0.619, 210), Q(-10.17, 44.88, 227), R(76.27, 47.23, 243)의 값을 가진다고 가정하면 식 3과 같이 평면의 방정식을 도출할 수 있다.

[0081] (식 3)

$$A = \begin{pmatrix} 1 & Py & Pz \\ 1 & Qy & Qz \\ 1 & Ry & Rz \end{pmatrix}, B = \begin{pmatrix} Px & 1 & Pz \\ Qx & 1 & Qz \\ Rx & 1 & Rz \end{pmatrix}, C = \begin{pmatrix} Px & Py & 1 \\ Qx & Qy & 1 \\ Rx & Ry & 1 \end{pmatrix}, d = -\begin{pmatrix} Px & Py & Pz \\ Qx & Qy & Qz \\ Rx & Ry & Rz \end{pmatrix}$$
$$A = Py(Qz - Rz) + Qy(Rz - Pz) + Ry(Pz - Qz)$$
$$B = Pz(Qx - Rx) + Qz(Rx - Px) + Rz(Px - Qx)$$
$$C = Px(Qy - Ry) + Qx(Ry - Py) + Rx(Py - Qy)$$
$$-d = Px(QyRz - RyQz) + Qx(RyPz - PyRz) + Rx(PyQz - QyPz)$$

[0082]

[0083] 식 3에서, A, B, C는 평면 그 자체에 대한 법선 벡터이고, d는 법선벡터의 크기이고, Px, Py, Pz는 P점의 3차원 좌표이며, Qx, Qy, Qz는 Q점의 3차원 좌표이고, Rx, Ry, Rz는 R점의 3차원 좌표이다.

[0084] (2) Object Detection에서 검출된 객체와 해당 객체의 Key point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산을 설명하면 다음과 같다.

[0085] Object Detection 학습 네트워크를 사용하여 사람이 검출되면 사람의 2차원 Key Point 좌표를 Depth Map에 투영시키고 식 2를 이용하여 도 13과 같이 각 Key Point에 대한 3차원 좌표로 변환시킨다.

[0086] 도 13은 각 Key Point에 대한 3차원 좌표로 변환을 나타낸 도면이다.

[0087] 본 실시예의 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 과정에서, Object Detection에서 검출된 객체의 Key Point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산하고, 검출된 객체의 2차원 Key Point 좌표를 Depth Map에 투영시키고 3차원 좌표로 변환한다.

[0088] 또한 검출된 객체와 바닥 평면 사이의 상대거리를 구하기 위해 도 12의 3차원 세 점을 이용하여 계산된 평면의 방정식을 바탕으로 법선벡터를 구한다. 점 P(-3.99, 0.619, 210), Q(-10.17, 44.88, 227), R(76.27, 47.23, 243)의 법선벡터 V는 식 4와 같이 구할 수 있다.

[0089] (식 4)

$$\overrightarrow{PQ} = Q - P = (Qx - Px,\ Qy - Px,\ Qz - Pz)$$
$$\overrightarrow{PR} = \overrightarrow{R} - P = (Rx - Px, Ry - Py, Rz - Pz)$$
$$PQ \times \overrightarrow{PR} = \overrightarrow{V} = [\overrightarrow{Vx},\ \overrightarrow{Vy},\ Vz]$$
$$Vx = \overrightarrow{PQ}y\,\overrightarrow{PR}z - \overrightarrow{PQ}z\,\overrightarrow{PR}y$$
$$Vy = \overrightarrow{PQ}z\,\overrightarrow{PR}x - \overrightarrow{PQ}x\,\overrightarrow{PR}z$$
$$Vz = PQx\,PRy - PQy\,PRx$$

[0090]

[0091] 여기서, $\overrightarrow{PQ}$ 는 점 P와 점 Q를 잇는 벡터이고, $\overrightarrow{PR}$ 는 점 P와 점 Q를 잇는 벡터이며, V는 법선벡터이고, Vx, Vy, Vz는 법선벡터V의 좌표이다.

[0092] 식 2의 법선벡터 V와 식 3의 d값을 이용하여 평면을 식 5와 같이 최종적인 평면의 방정식을 정의할 수 있다.

[0093] (식 5)

$$(Vx)x - (Vy)y + (Vz)z = d$$

[0094]

[0095] 식 5에서, (Vx)x는 법선벡터 V의 x값(Vx)과 정의된 평면과 상대거리 좌표를 측정할 점의 3차원 좌표 중 x값이고, (Vy)y는 법선벡터 V의 y값(Vy)과 정의된 평면과 상대거리 좌표를 측정할 점의 3차원 좌표 중 y값이며, (Vz)z는 법선벡터 V의 z값(Vz)과 정의된 평면과 상대거리 좌표를 특정할 점의 3차원 좌표 중 z값이고, d는 식 3에서 산출된 법선 벡터의 크기이다.

[0096] 마지막으로 정의된 평면의 식 5와 객체의 Key Point 점 T(x, y, z) 사이의 상대거리는 식 6과 같이 구할 수 있다.

[0097] (식 6)

$$Distance = \frac{Vx\,Tx + Vy\,Ty + Vz\,Tz - d}{\sqrt{Vx^2 + Vy^2 + Vz^2}}$$

[0098]

[0099] 본 실시예의 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 과정에서, Object Detection에서 검출된 객체의 Key Point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산하고, 최종적으로 정의된 평면의 방정식과 객체의 Key Point 점 T 사이의 상대거리를 계산한다.

[0100] (3) 계산된 상대거리 좌표를 이용한 쓰러짐 감지

[0101] 도 13에서 계산된 객체의 각 Key Point 별 3차원 좌표와 식 5를 이용하여 가상의 바닥 평면과 객체의 신체 부위 Key Point 간의 상대거리를 구할 수 있다. 각 Key Point들과 바닥 평면과의 상대거리를 구한 후, 도출된 상대거리들 중 최대값과 최소값을 선별하고, 최대값에서 최소값을 뺀 값을 검출된 사람의 고유값으로 설정한다. 식 7과 같이 해당 고유값의 크기 및 비율에 따라 일정 비율의 Threshold를 설정하고 고유값에 Threshold 값을 곱하여 도출된 상대거리들의 평균값 보다 작을 때 쓰러짐 정의 변수의 값을 True로 적용시킨다. 고유값과 Threshold 연산은 영상의 최근 10 프레임 동안 진행하며 10 프레임 중 7 프레임 이상의 쓰러짐 정의 변수의 값이 True일 때 쓰러짐이 발생했다고 판단한다.

[0102] (식 7)

$$V_{eigen} = D_{\max} - D_{\min}$$
$$Th = 0.9,\ V_{eigen} < 20$$
$$Th = 0.8,\ V_{eigen} > 20,\ V_{eigen} < 30$$
$$Th = 0.7,\ V_{eigen} > 30,\ V_{eigen} < 50$$
$$Falldown = Th^{*}\ V_{eigen} < D_{mean}$$

[0103]

[0104] 식 7에서, V_{eigen}은 검출된 사람의 고유값이고, $D_{\max}$는 상대거리 최대값, $D_{\min}$는 상대거리 최소값, Th는 고유값의 크기 및 비율에 따라 설정된 Threshold 값이며, $Falldown$은 쓰러짐 발생이고, $Th^{*}V_{eigen}$은 Threshold 값과 검출된 사람의 고유값을 곱한 것이며, D_{mean}은 상대거리 평균값이다.

[0105] 본 실시예의 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 과정에서, 계산된 상대거리 좌표를 이용한 쓰러짐 감지는 객체의 Key Point로부터 도출된 상대거리들 중 최대값과 최소값을 선별, 최대값에서 최소값을 뺀 값을 검출된 객체의 고유값을 설정하고, 고유값의 크기 및 비율에 따라 일정 비율의 Threshold를 설정하며, 고유값에 Threshold 값을 곱하여 도출된 상대거리들의 평균 값보다 작을 때 쓰러짐 정의 변수의 값을 True로 적용한다. 고유값과 Threshold 연산은 영상의 최근 10프레임 동안 진행하며 10프레임 중 7프레임 이상의 쓰러짐 정의 변수의 값이 True일 때 쓰러짐 발생으로 판단한다.

[0106] (4) 관심영역을 통한 배회, 침입 감지

[0107] 배회 및 침입은 산출된 Depth 값 및 Object Detection에서 검출된 객체와 Key Point(관절 부위)를 이용하여 도 14와 같이 추적중인 객체가 지정된 관심영역에 특정 시간동안 검출될 시 감지된다.

[0108] 도 14는 배회, 침입 감지를 나타낸 도면이다. 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 과정에서, 배회 및 침입은 산출된 Depth 값 및 Object Detection에서 검출된 객체와 Key Point(관절 부위)를 이용하여 추적 중인 객체가 지정된 관심영역에 특정 시간 동안 검출될 시 감지된다.

[0109] 다음으로 실행과정에 대해 설명하기로 한다.

[0110] 실행과정은 학습과정에서 학습된 Depth 추정 모델과 Object Detection 모델을 이용하여 수행된다.

[0111] 도 15는 제안된 솔루션의 실행 과정의 흐름도이다. 실행 과정에서는 학습 과정에서 학습된 Depth 추정 모델과 Object Detection 모델을 이용하여 단일 RGB 이미지로부터 Depth를 추정 및 객체를 검출하고 검출된 객체의 2차원 Key Point 좌표를 3차원 상대거리 좌표로 변환하여 배회, 침입, 쓰러짐을 판단한다.

[0112] 도 15는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태식별 CCTV 솔루션의 실행 과정을 나타낸 것으로, 학습 과정에서 학습된 Depth 추정 모델과 Object Detection 모델을 이용한다. 단일 RGB 이미지로부터 Depth를 추정 및 객체를 검출한다. 검출된 객체의 2차원 Key Point 좌표를 3차원 상대거리 좌표로 변환하여 쓰러짐을 판단한다. 배회 및 침입은 추적 중인 객체가 지정된 관심영역에 특정 시간 동안 검출될 시 감지된다.

[0113] 본 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템의 실험환경은 <하드웨어>Intel(R) Core(TM) i7-8700k 3.70GHz CPU, RAM 16GB, NVIDIA GeForce RTX 3080Ti(VRAM 12GB)이고, <운영체제>Ubuntu 20.04, <개발도구>Visual Studio Code, Python 3.8, <라이브러리>Pytorch 1.11.0, CUDA 11.3, cuDNN 8.4.1이다.

[0114] 데이터 셋과 관련하여, 본 발명에서는 제안하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태식별 CCTV 솔루션의 정확도를 평가하기 위해 NYU Depth V2 Dataset을 이용하여 Depth 추정 모델을 학습하고,

Pascal VOC 2007, 2012를 이용하여 Object Detection 모델 학습했다.

[0115] NYU Depth V2 Dataset은 New York University의 연구원들이 협업하여 19개의 카테고리의 실내 이미지들로 구축하여 공개된 표준 데이터셋으로서, 기존의 논문들에서 비교 벤치마크로 사용되었다. NYU Depth V2 Dataset은 19개의 카테고리를 가진 각각 24,231개의 RGB 이미지와 Ground Truth 이미지로 이루어져 있으며, 이미지 객체 Depth 추정 실험에 사용하였다.

[0116] Pascal VOC 2007 및 2012는 카테고리의 수는 20개로 동일하고 각각 9,963개, 11,530개의 이미지로 이루어져 있으며 다중객체검출 및 추적 실험에 사용하였다. 각 이미지에 할당된 Annotation 파일이 존재한다. 본 발명에서는 비교 논문들에서 수행한 결과를 바탕으로 유사한 환경을 구성하기 위해 [표 2]와 같은 카테고리를 사용하였다. Depth 추정과 Object Detection의 Test Dataset은 각각 654개, 4,952개로 이루어져 있으며 도 16과 도 17은 각각 본 발명에서 실험에 사용한 NYU Depth Dataset과 Pascal VOC 2007 및 2012 Dataset의 예시 이미지를 나타내고 있다.

[0117] 표 2는 학습에 사용된 데이터셋의 카테고리를 나타낸 것이다.

[0118] (표 2)

분류	내용
Depth 추정	Basement, Bathroom, Bedroom, Bookstore, Cafe, Living Room, Dining Room, Furniture Store, Home Office, Kitchen, Library, Studies, Misc, Office, Office Kitchen, Playroom, Reception Room, Study Room
Object Detection	Background, Aeroplane, Bicycle, Bird, Boat, Bottle, Bus, Car, Cat, Chair, Cow, Dining table, Dog, Horse, Motorbike, Person, Potted Plant, Sheep, Sofa, Train, TV Monitor

[0119]

[0120] 도 16은 NYU Depth Dataset 예시를 나타낸 도면이다.

[0121] NYU Depth V2 Dataset은 19개 카테고리의 실내 이미지들로 구축하여 각각 24,231개의 RGB 이미지와 Ground Truth 이미지로 구성된 표준 데이터셋으로서 이미지 객체 Depth 추정 실험에 사용하였다.

[0122] 도 17은 Pascal VOC 2007 및 2012의 예시를 나타낸 도면이다.

[0123] Pascal VOC 2007, 2012 Dataset은 20개의 동일한 카테고리 수를 가지며 각각 9,963개, 11,530개의 이미지들과 각 이미지에 할당된 Annotation 파일로 구성된 표준 데이터셋으로서 다중객체검출 및 추적 실험에 사용하였다.

[0124] 본 실시예에 따른 실험결과를 설명하면 다음과 같다.

[0125] 성능평가를 위한 값으로는 Depth 추정은 RMSE를, Object Detection은 AP를 사용한였다. Root Mean Square Error(RMSE)는 추정 값 또는 모델이 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 사용되는 척도로서, 0에 가까울 수록 높은 성능을 나타낸다. Average Precision(AP)은 물체의 검출과 모델의 비교에 적용되는 방식으로서, 1에 가까울 수록 높은 성능을 나타낸다.

[0126] 도 18은 Depth 추정 모델의 Test 결과를 나타낸다. 도 18에서, Depth 추정 모델 학습에 사용된 Train Dataset을 제외하고 654개의 Test Data로 실험을 수행한 결과를 나타낸 것이다.

[0127] 도 19는 Object Detection 모델의 Test 결과를 나타낸다. 도 19는 Object Detection 모델 학습에 사용된 Train Dataset을 제외하고 4,952개의 Test Data로 실험을 수행한 결과를 나타낸 것이다.

[0128] 본 발명에서 제안하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템은 공개된 표준 데이터셋인 NYU Depth V2 Dataset 및 Pascal VOC 2007, 2012를 이용하여 실험하였다. 도 20과 같이 실험 결과, 이미지 객체 Depth 추정의 RMSE는 0.40을 나타내어 세계최고 수준인 0.824 이하보다 우수한 결과를 산출하였다. 또한, 다중객체검출 및 추적의 AP는 80을 나타내어 세계 최고 수준인 70이상 보다 우수한 결과를 산출하였다.

[0129] 도 20은 이미지 객체 Depth 추정과 다중객체검출 및 추적에 대한 실험 결과를 나타낸 것이다.

[0130] 한편, 본 발명에서 제안하는 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션

시스템의 객관적인 성능을 평가하기 위하여, 공인시험기관의 시험성적서를 발급받았다. 도 21은 각각 이미지 객체 Depth 추정과 다중객체검출 및 추적에 대한 공인시험기관의 시험성적서이다.

[0131] 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템은 Window Shifted Transformer를 Encoder로써 사용하여 Depth 추정을 수행한 것은 일반적인 ViT를 수행한 것 보다 FPN 또는 Segmentation의 U-Net처럼 계층적인 정보를 학습에 활용할 수 있다는 장점이 있다. 각 Window 내에 여러 Patch들로 구성되어 해당 Window에 대해서만 Self-Attention을 계산하기 때문에 Window Shifted Transformer Layer에 Bias를 가하여 학습 모델에 범용성을 가질 수 있다.

[0132] 또한 Stage를 진행할수록 각 Patch들은 병합되어 더 큰 Patch로 이루어진 Window를 학습에 사용하므로 더 큰 관점에서의 Feature Map을 추출할 수 있고, 반대로 초기 Stage 및 계층에서는 작은 Patch로 이루어진 Window로 Feature Map을 추출하기 때문에 더욱 세부적인 Feature Map을 추출할 수 있다는 장점이 있다.

[0133] 도 22는 Patch Size에 따른 Feature Map 추출의 예시를 나타낸다. 도 22의 예시를 부연 설명하면, 일반적인 Vision Transformer(ViT)를 사용하지 않고 Window Shifted Tranformer를 Encoder로써 딥러닝 네트워크를 구성한다. 이때, Feature Pyramid Network(FPN) 또는 Segmentation의 U-Net처럼 계층적인 정보를 학습에 활용할 수 있다는 장점이 있다. 그리고 Stage를 진행할수록 Window Shifted Transformer의 Patch들은 병합되어 더 큰 Patch로 이루어진 Window를 학습에 이용하므로 다양한 관점에서의 Feature Map을 추출할 수 있다는 장점이 있다.

[0134] 도 23은 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템의 구성도이다. 도 23에 도시된 바와 같이, 본 실시예에 따른 구성은 추정부(110), 검출부(120), 감지부(130)를 포함한다.

[0135] 추정부(110)는 입력된 RGB 이미지로부터 Depth 값을 추정한다.

[0136] 이러한 추정부(110)는 Window Shifted Transformer Layer와 Convolution Layer를 통한 Feature Map 추출모듈(111), Deformable Self-Attention을 통한 Multi-level Hierarchical Feature Map 강화모듈(112) 및 Decoder의 Hardtanh 함수 Loss 계산모듈(113)을 포함한다.

[0137] 검출부(120)는 입력된 RGB 이미지로부터 Object Detection을 수행한다.

[0138] 이러한 검출부(120)는 Convolutional Layer를 통한 Feature Map 추출모듈(121), Spatial Pyramid Pooling(SPP)과 Feature Pyramid Network(FPN)를 통한 데이터 크기를 줄여 매개변수 수를 줄이는 Multi-Scale의 Feature Pyramid 모듈(122) 및 Decoupled Head를 통한 epoch 별 Converging 속도 개선모듈(123)을 포함한다.

[0139] 감지부(130)는 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지한다.

[0140] 이러한 감지부(130)는 바닥 평면의 세 점을 구성하는 구성모듈(131) 및 Object Detection에서 검출된 객체의 Key point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산, 계산된 상대거리 좌표를 이용한 쓰러짐을 감지하는 감지모듈(132)을 포함한다.

[0141] 도 24는 본 발명의 일 실시예에 따른 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 방법의 전체흐름도이다. 도 24에 도시된 바와 같이, 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템을 이용한 방법에 있어서, (a)상기 시스템이 입력된 RGB 이미지로부터 Depth 값을 추정하는 단계, (b)상기 시스템이 상기 입력된 RGB 이미지로부터 Object Detection을 수행하는 검출단계 및 (c)상기 시스템이 산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 단계를 포함한다.

[0142] 여기서, (a)단계에서, 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템은 Window Shifted Transformer Layer와 Convolution Layer를 통한 Feature Map 추출단계, Deformable Self-Attention을 통한 Multi-level Hierarchical Feature Map 강화단계 및 Decoder의 Hardtanh 함수 Loss 계산단계를 포함한다.

[0143] 또한 (b)단계에서, 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템은 Convolutional Layer를 통한 Feature Map 추출단계, Spatial Pyramid Pooling(SPP)과 Feature Pyramid Network(FPN)를 통한 데이터 크기를 줄여 매개변수 수를 줄이는 Multi-Scale의 Feature Pyramid를 구성하는 단

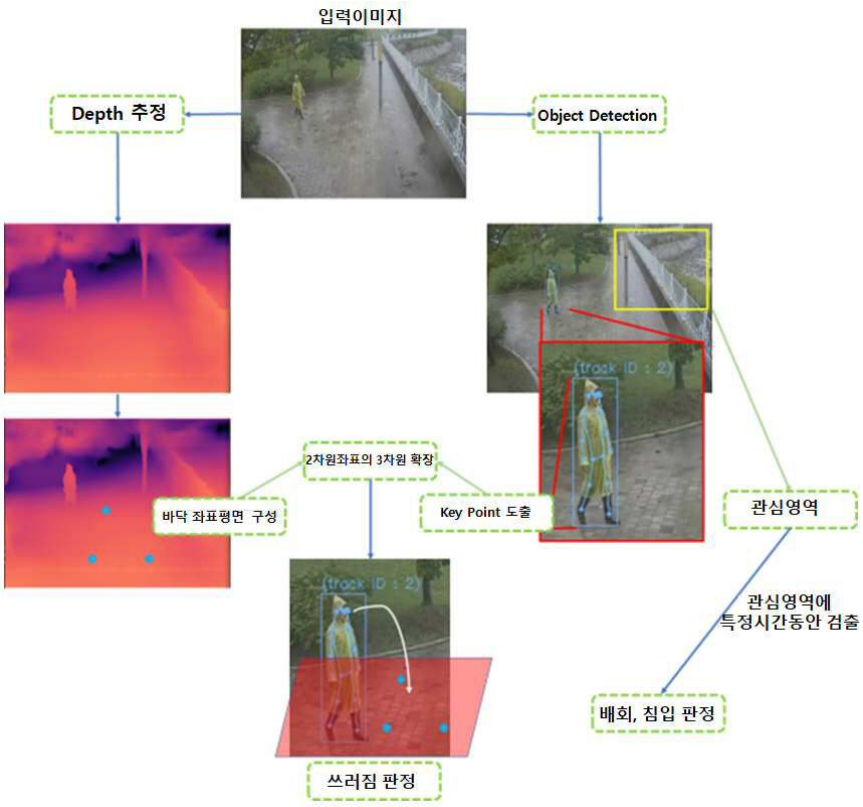
계 및 Decoupled Head를 통한 epoch 별 Converging 속도를 개선하는 단계를 포함한다.

[0144] 그리고 (c)단계에서, 단일영상 이미지 기반의 Depth 추정 기법을 사용한 지능형 형태 식별 CCTV 솔루션 시스템은 바닥 평면의 세 점을 구성하는 단계 및 Object Detection에서 검출된 객체의 Key point 2차원 좌표를 기반으로 Depth Map에 전달하여 3차원 가상 상대거리 좌표 계산, 계산된 상대거리 좌표를 이용한 쓰러짐을 감지하는 단계를 포함한다.

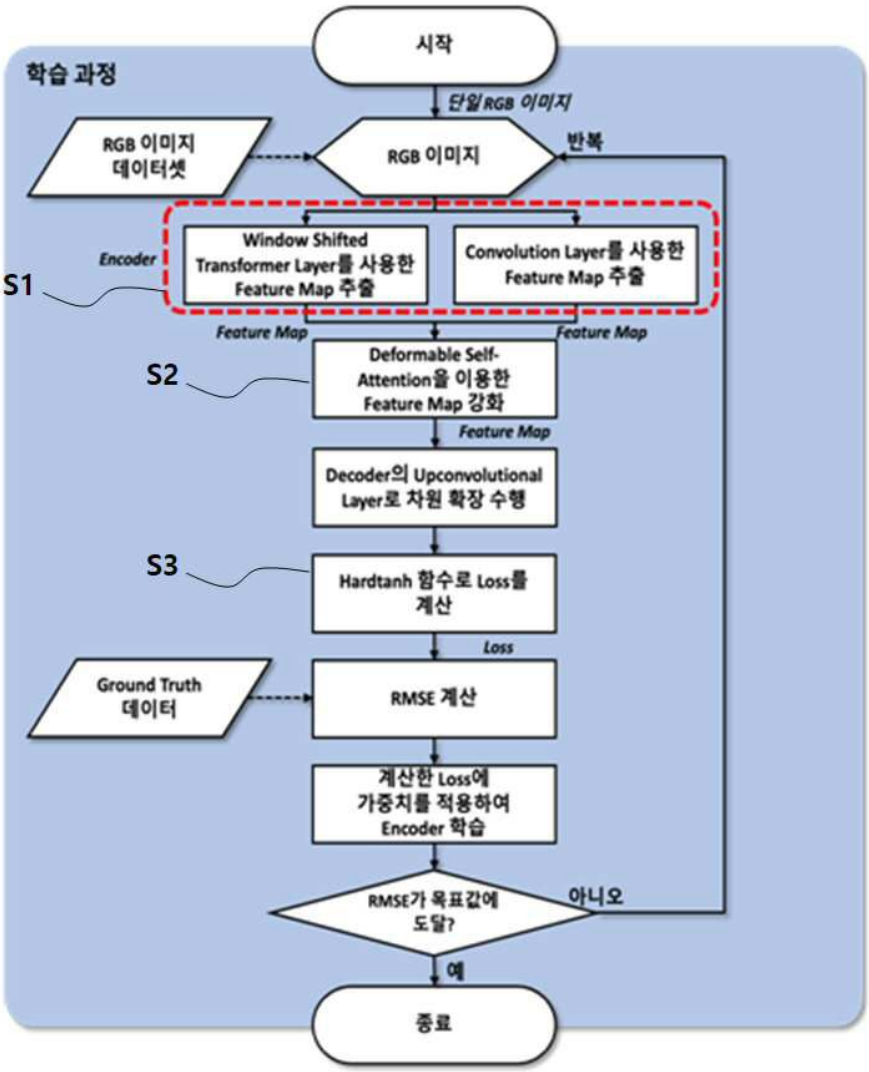
[0145] 본 발명에 따르면, 급증하는 고령자를 대상으로 부족한 요양보호사(간호사)를 보조하는 ICT 기술개발을 통한 고령자의 건강하고 안전한 삶에 기여할 수 있는 효과가 있다. 개발 ICT 기술을 이용한 안전관제를 통해 고령환자, 제조 근로자의 안전사고로 인한 재정지출(의료보험)의 감소에 기여할 수 있는 효과가 있다.

도면

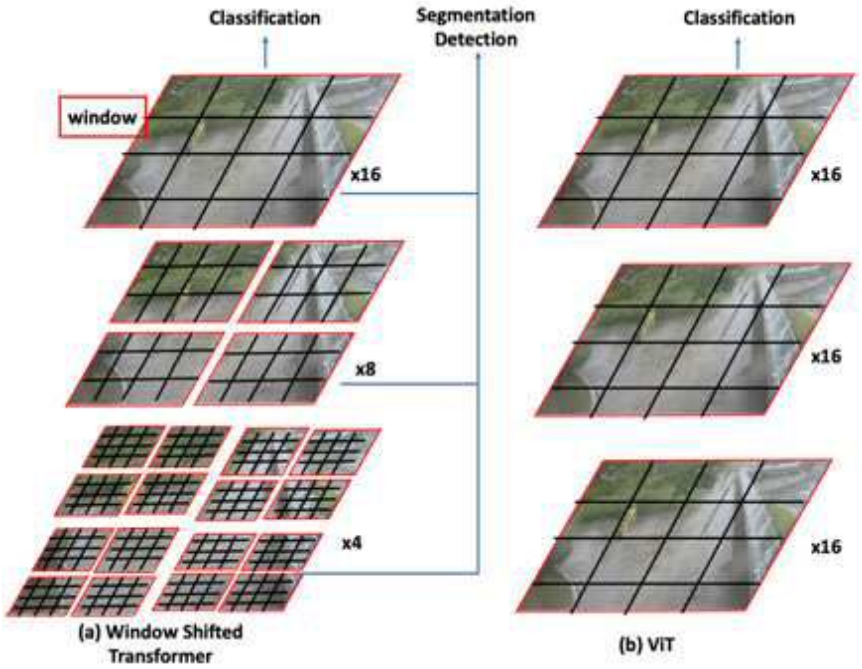
도면1



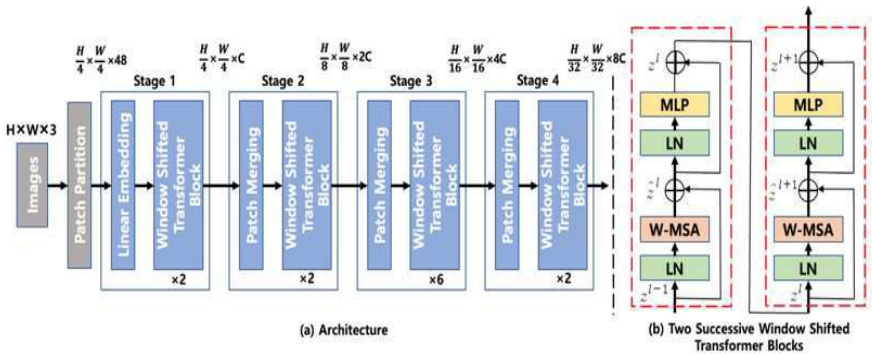
도면2


시작
학습 과정
단일 RGB 이미지
RGB 이미지 데이터셋
RGB 이미지
반복
Encoder
S1
Window Shifted Transformer Layer를 사용한 Feature Map 추출
Convolution Layer를 사용한 Feature Map 추출
Feature Map
Feature Map
S2
Deformable Self-Attention을 이용한 Feature Map 강화
Feature Map
Decoder의 Upconvolutional Layer로 차원 확장 수행
S3
Hardtanh 함수로 Loss를 계산
Loss
Ground Truth 데이터
RMSE 계산
계산한 Loss에 가중치를 적용하여 Encoder 학습
RMSE가 목표값에 도달?
아니오
예
종료

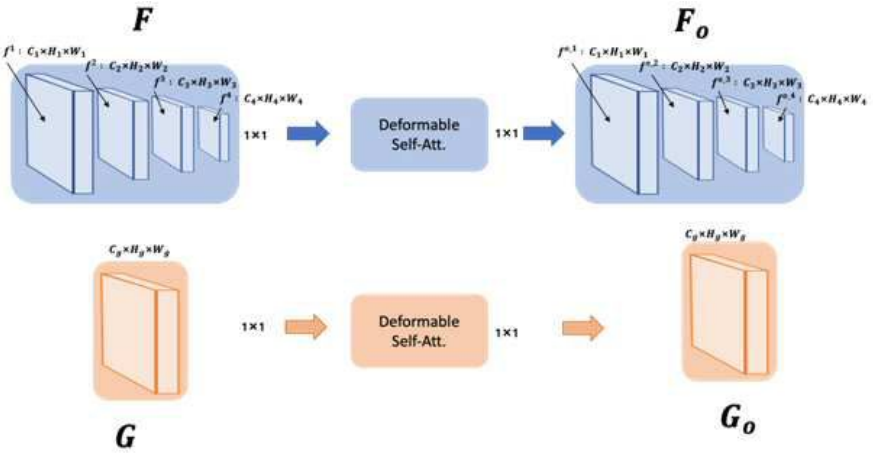
도면3


Classification
Segmentation
Detection
Classification
window
x16
x8
x4
x16
x16
x16
(a) Window Shifted Transformer
(b) ViT

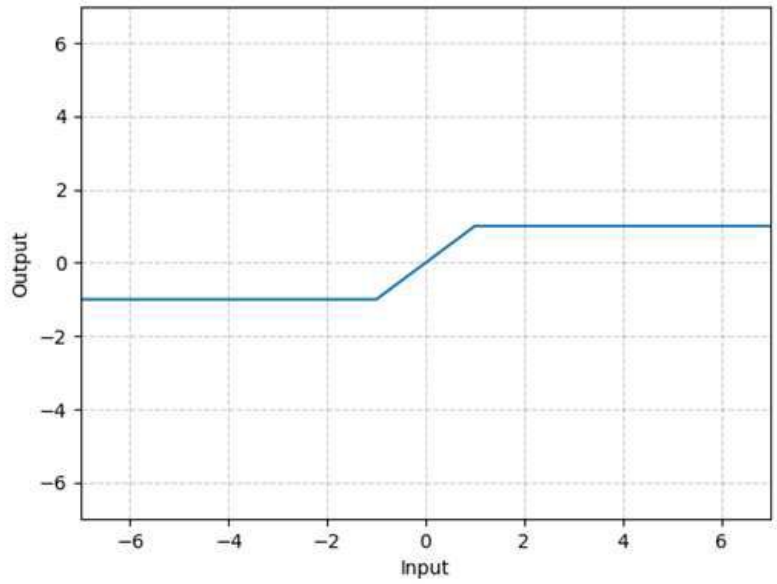
도면4


H×W×3
Images
Patch Partition
Stage 1
Stage 2
Stage 3
Stage 4
Linear Embedding
Window Shifted Transformer Block
Patch Merging
×2
×2
×6
×2
MLP
LN
W-MSA
LN
(a) Architecture
(b) Two Successive Window Shifted Transformer Blocks

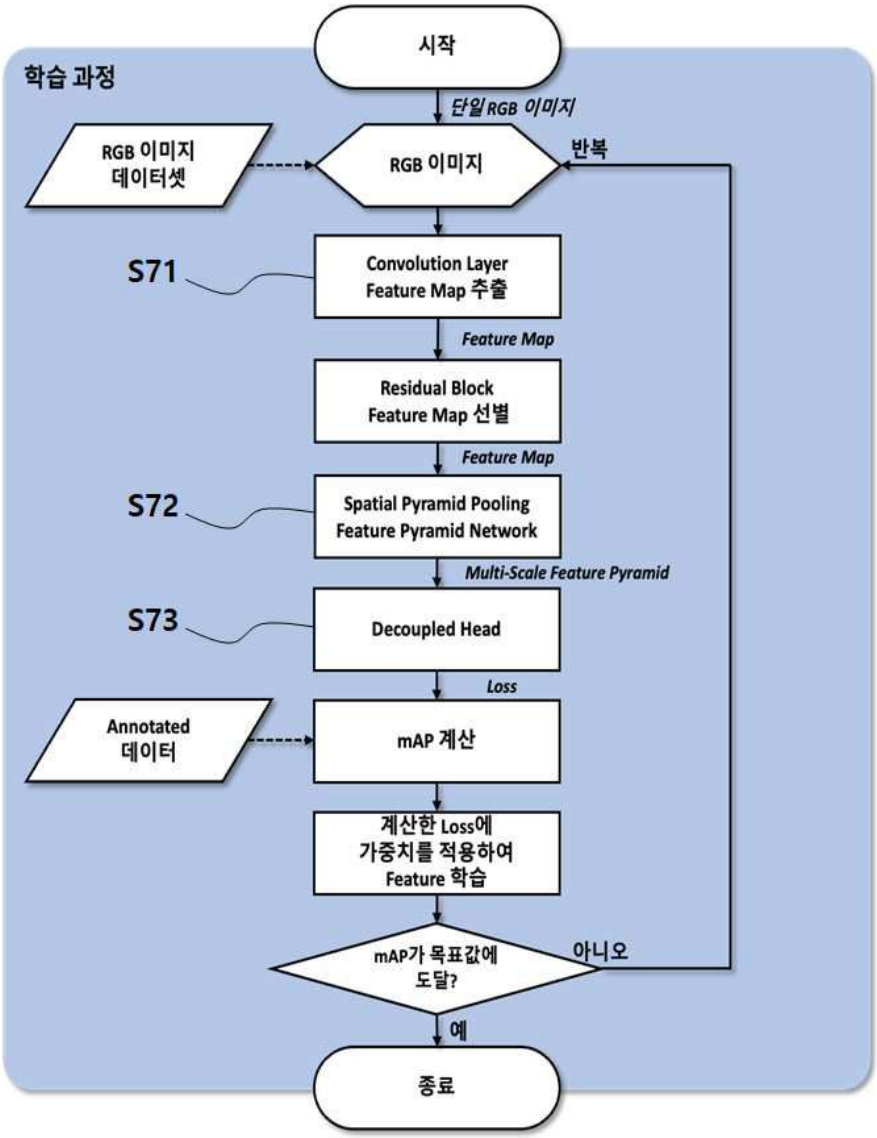
도면5


F
F_o
1×1
Deformable Self-Att.
1×1
1×1
Deformable Self-Att.
1×1
G
G_o

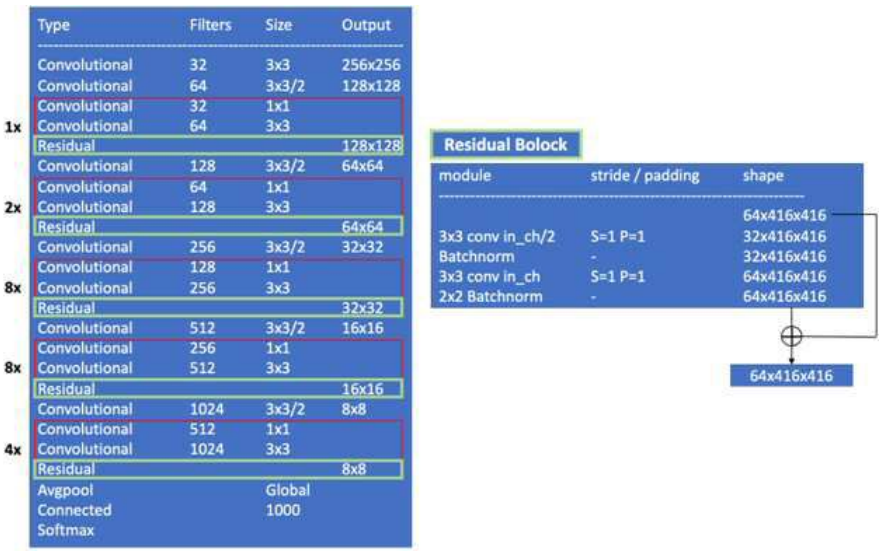
도면6



도면7

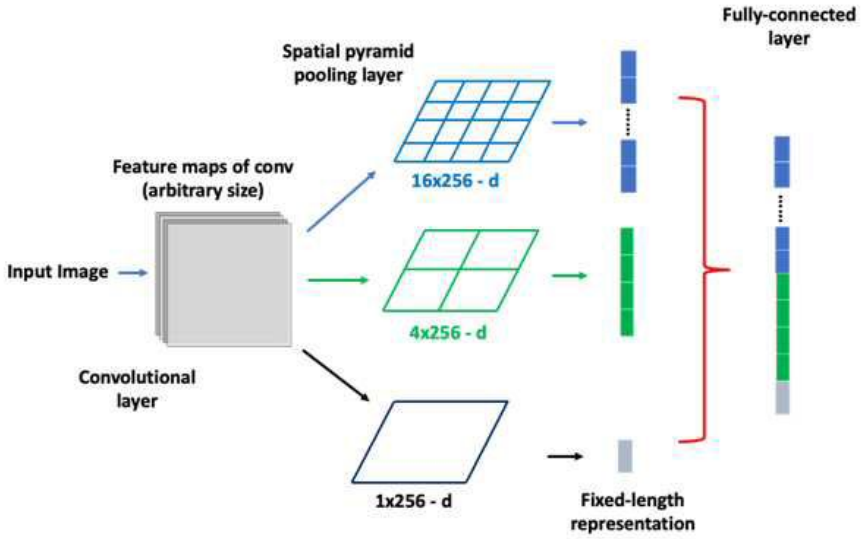


도면8

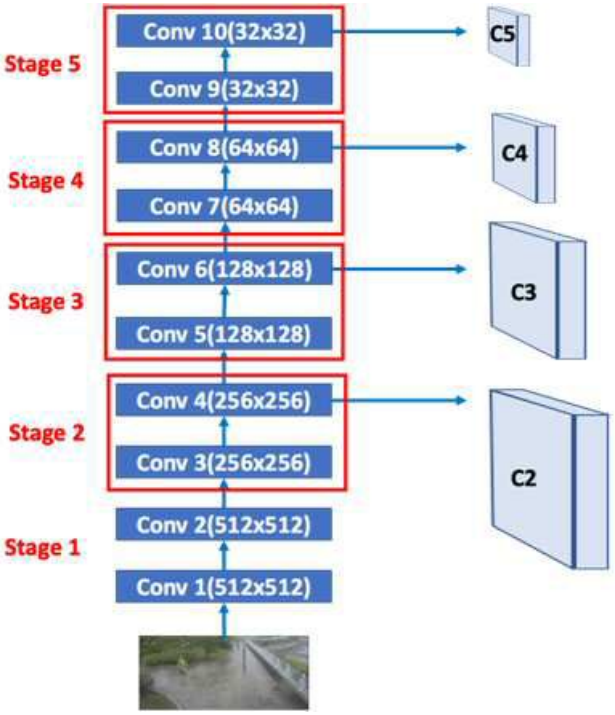


	Type	Filters	Size	Output
	Convolutional	32	3x3	256x256
	Convolutional	64	3x3/2	128x128
1x	Convolutional	32	1x1	
	Convolutional	64	3x3	
	Residual			128x128
	Convolutional	128	3x3/2	64x64
2x	Convolutional	64	1x1	
	Convolutional	128	3x3	
	Residual			64x64
	Convolutional	256	3x3/2	32x32
8x	Convolutional	128	1x1	
	Convolutional	256	3x3	
	Residual			32x32
	Convolutional	512	3x3/2	16x16
8x	Convolutional	256	1x1	
	Convolutional	512	3x3	
	Residual			16x16
	Convolutional	1024	3x3/2	8x8
4x	Convolutional	512	1x1	
	Convolutional	1024	3x3	
	Residual			8x8
	Avgpool		Global	
	Connected		1000	
	Softmax			

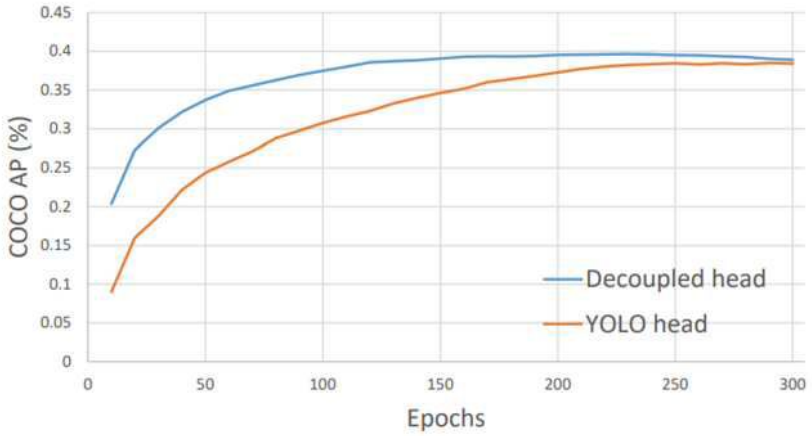
도면9


Fully-connected layer
Spatial pyramid pooling layer
Feature maps of conv (arbitrary size)
16x256 - d
Input Image
4x256 - d
Convolutional layer
1x256 - d
Fixed-length representation

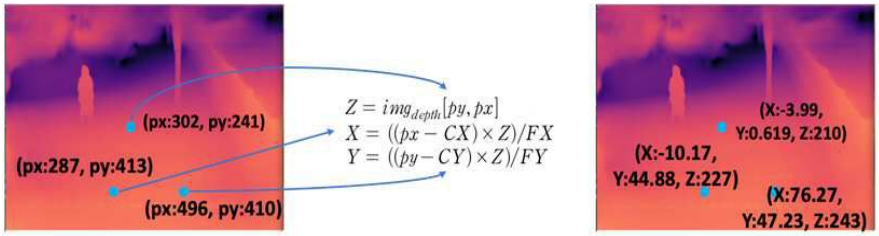
도면10


Stage 5
Conv 10(32x32)
Conv 9(32x32)
C5
Stage 4
Conv 8(64x64)
Conv 7(64x64)
C4
Stage 3
Conv 6(128x128)
Conv 5(128x128)
C3
Stage 2
Conv 4(256x256)
Conv 3(256x256)
C2
Stage 1
Conv 2(512x512)
Conv 1(512x512)

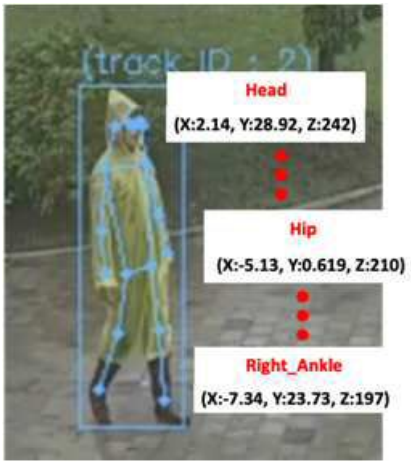
도면11



도면12



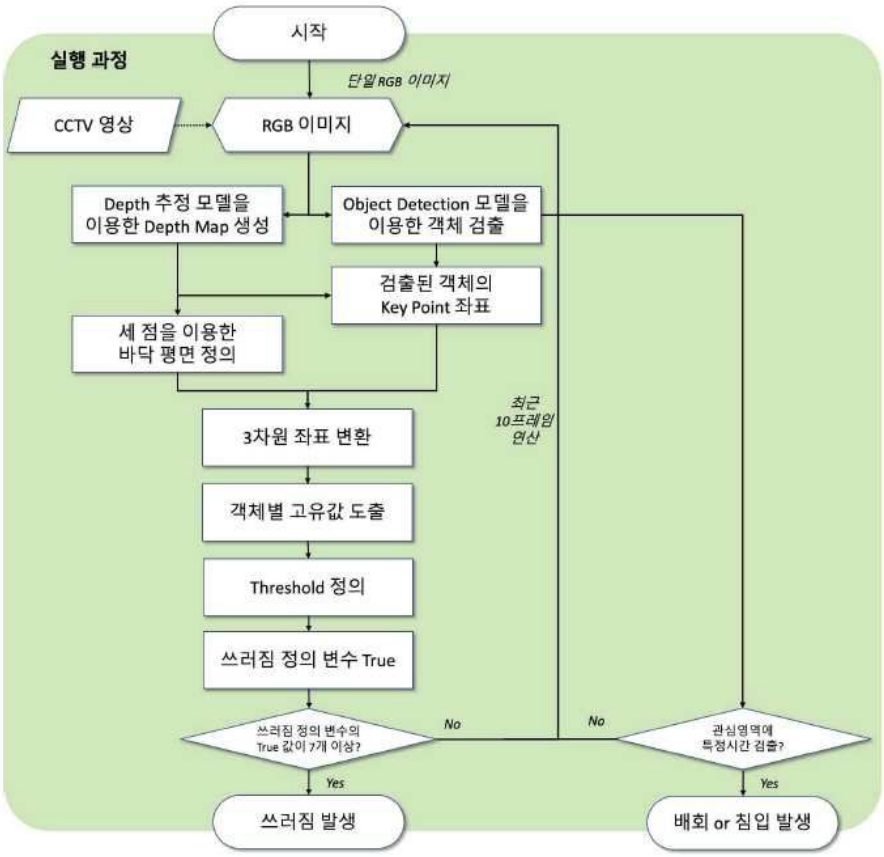
도면13



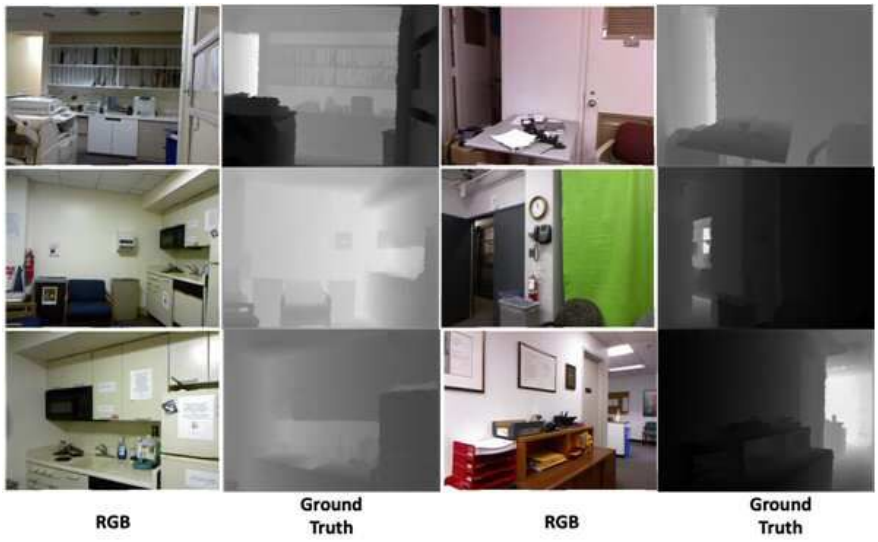
도면14



도면15


시작
실행 과정
단일 RGB 이미지
CCTV 영상
RGB 이미지
Depth 추정 모델을 이용한 Depth Map 생성
Object Detection 모델을 이용한 객체 검출
검출된 객체의 Key Point 좌표
세 점을 이용한 바닥 평면 정의
최근 10프레임 연산
3차원 좌표 변환
객체별 고유값 도출
Threshold 정의
쓰러짐 정의 변수 True
쓰러짐 정의 변수의 True 값이 7개 이상?
No
No
관심영역에 특정시간 검출?
Yes
Yes
쓰러짐 발생
배회 or 침입 발생

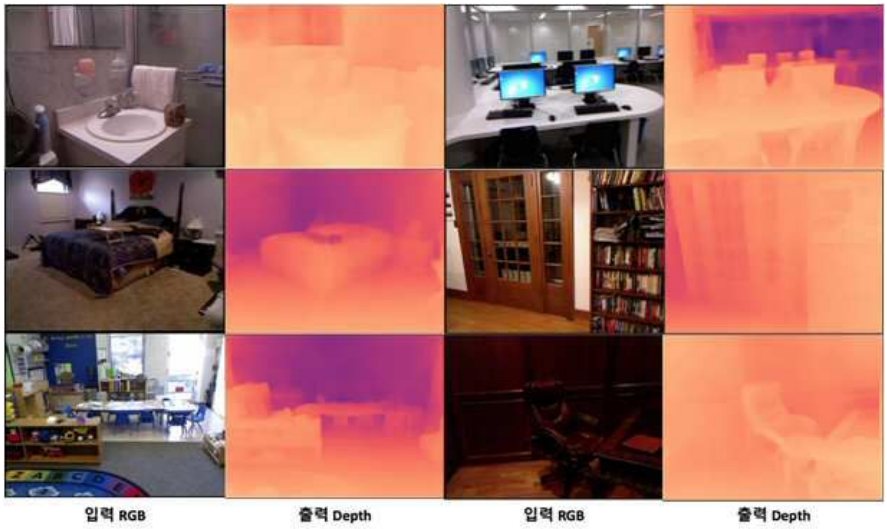
도면16



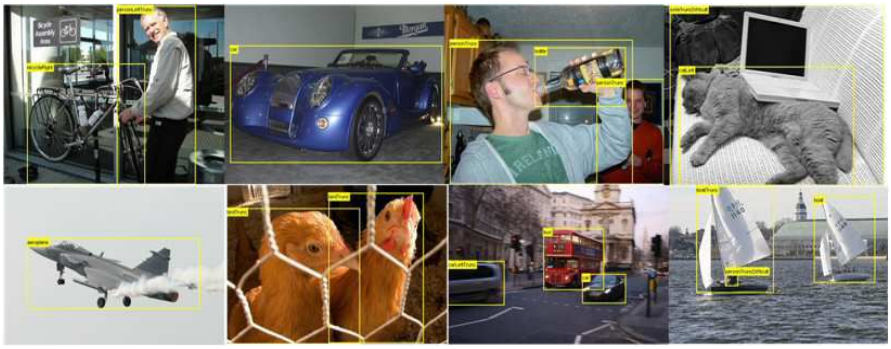
도면17



도면18



도면19



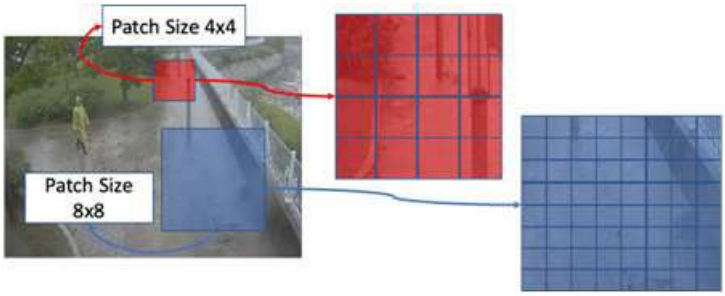
도면20

평가 항목	단위	세계 최고 수준	본 발명 결과
이미지 객체 Depth 추정	RMSE	0.824 이하	0.40
다중객체 검출 및 추적	AP	70 이상	80

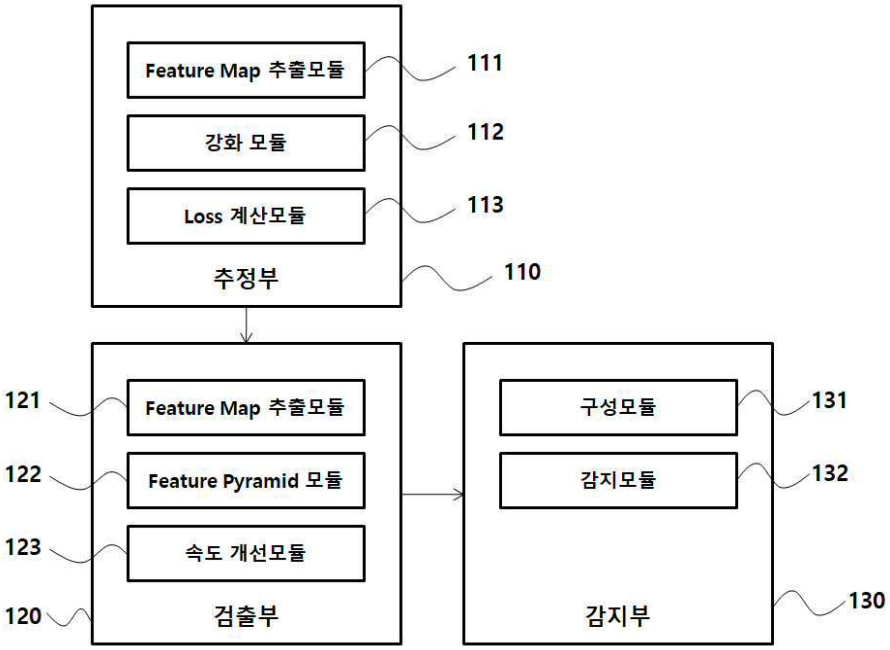
도면21



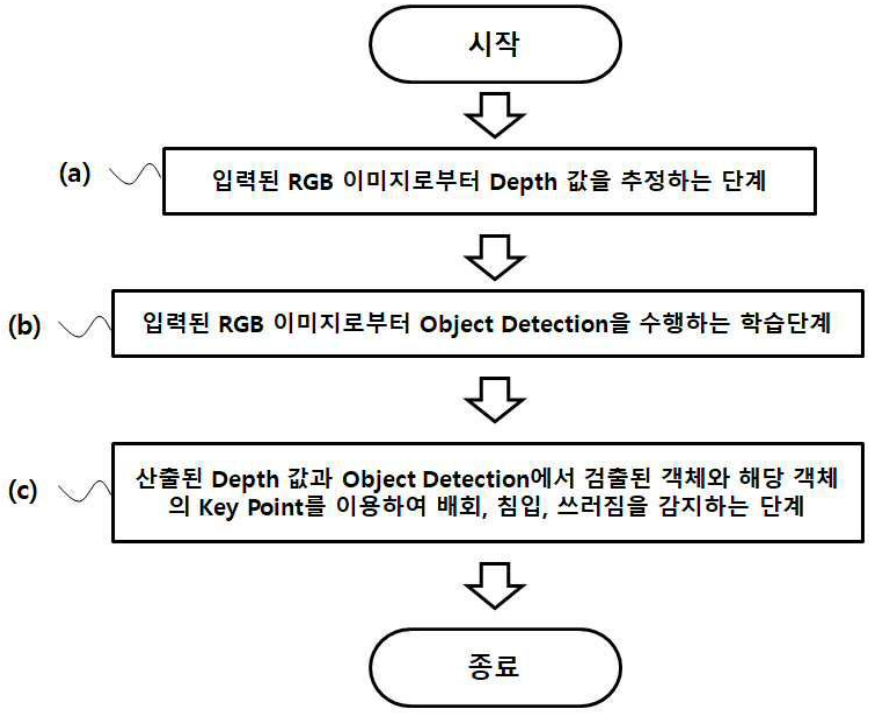
도면22


Patch Size 4x4
Patch Size
8x8

도면23


Feature Map 추출모듈
111
강화 모듈
112
Loss 계산모듈
113
추정부
110
121
Feature Map 추출모듈
122
Feature Pyramid 모듈
123
속도 개선모듈
120
검출부
구성모듈
131
감지모듈
132
감지부
130

도면24


시작
(a)
입력된 RGB 이미지로부터 Depth 값을 추정하는 단계
(b)
입력된 RGB 이미지로부터 Object Detection을 수행하는 학습단계
(c)
산출된 Depth 값과 Object Detection에서 검출된 객체와 해당 객체의 Key Point를 이용하여 배회, 침입, 쓰러짐을 감지하는 단계
종료