리뷰보고서

(H2O: Two Hands Manipulating Objects for First Person Interaction Recognition)

**소개**

몇 년 동안 비디오 이해 및 동작 인식에 엄청난 발전이 있었습니다. 1인칭 시점의 동작 인식은 증강 현실, 로봇 공학 및 감시 분야에서 많은 응용 분야가 있지만 주로 크고 다양한 자기 중심적 데이터 셋의 부족으로 인해 3인칭 시점의 발전에 뒤떨어져 있습니다.

자기 중심적 관점에서 동작 인식은 대부분 손과 물체의 상호 작용을 이해하는 것입니다.

손과 조작 대상의 위치와 움직임에 대한 통일된 이해는 자기 중심적 상호 작용을 인식하는 데 중요합니다.

그러나 기존의 1인칭 상호 작용 데이터 세트는 대부분 손의 움직임과 조작된 개체에 대한 3D 추론 없이 2D 기능(예: 경계 상자, 손 분할)만 제공합니다.

동기화된 RGB-D 이미지, 카메라 포즈 오른쪽 및 왼쪽 손 포즈, 개체 포즈, 개체 메시, 장면 포인트 클라우드 및 동작 레이블을 포함하여 풍부한 주석이 달린 데이터 셋을 수집하여 3D 손-개체 상호 작용을 이해하기 위한 전례 없는 수준의 세부 정보를 제공합니다.

데이터 셋의 도움으로 컬러 이미지에서 두 손과 물체의 3D 포즈를 공동으로 추정하는 첫 번째 방법을 제시합니다.

3D 상호 작용 인식을 위한 적응형 그래프 convolution 네트워크를 사용하여 손과 물체 포즈 내외의 상호 의존성을 학습합니다.

동작 중인 손과 3D에서 조작된 개체를 공동으로 캡처하는 것은 상호 폐색으로 인해 어려운 문제입니다. 이 문제는 빠른 카메라 움직임, 큰 폐색, 배경 혼란 및 가장 중요한 데이터 세트 부족과 같은 자기 중심적 비전으로 인한 고유한 문제로 인해 1인칭 시점에서 더 어렵습니다.

스크린샷, 만화 영화이(가) 표시된 사진

자동 생성된 설명

1인칭 상호 작용 인식을 위해 두 손으로 개체를 조작합니다. 왼손 및 오른손의 3D 포즈, 6D 개체 포즈, 카메라 포즈, 개체 메시 및 장면 포인트 클라우드에 대한 풍부한 주석을 관련 상호 작용 레이블과 함께 제공하는 데이터 셋을 제안합니다. 데이터 셋을 활용하여 3D 상호 작용 인식을 위한 새로운 방법을 제안합니다.

주로 한 손 조작 시나리오에 중점을 둡니다. 한 손 조작은 일부 시나리오와 관련이 있지만 대부분의 경우 손-개체 상호 작용에는 개체를 조작하는 두 손이 포함됩니다.

2D 주석만 사용하여 손과 손 및 손과 물체의 상호 작용에 대한 데이터 셋을 제시했습니다. 그러나 손-물체 상호 작용의 복잡한 특성으로 인해 상호 폐색을 더 잘 해결하려면 2D가 아닌 3D 추론이 필요합니다.

손-객체 상호 작용의 맥락에서 초기 작업은 대부분 동작에 대한 추론 없이 3D 손과 객체 포즈의 공동 추정 문제를 다룹니다.

손과 물체에 대한 정확한 3D 위치 데이터는 로봇 공학 및 그래픽의 많은 응용 프로그램에 매우 중요하지만 포즈에 대한 지식만으로는 대상의 동작에 대한 의미론적 의미가 부족합니다. 손과 물체의 3D 주석을 포함하는 자기 중심적 동작 데이터 셋을 발표했습니다. 그러나 데이터는 방해가 되는 모션 캡처 시스템으로 캡쳐됩니다.

모션 캡처 데이터 셋은 정확한 3D 주석이 있는 많은 양의 교육 샘플을 제공할 수 있지만 제어된 설정에서만 캡처할 수 있으며 컬러 이미지에서 포즈 예측을 편향시키는 이미지에 가시적인 마커가 있습니다. 두 손의 3D 포즈와 조작된 물체의 6D 포즈의 마커 없는 주석을 포함하는 자기 중심적 3D 상호 작용 인식을 위한 통합 데이터 세트를 생성하기 위한 접근 방식을 제안합니다.

**기여한 방법**

두 손의 마커 없는 3D 주석과 조작된 개체의 6D 포즈를 사용하여 자기 중심적 상호 작용 인식을 위한 최초의 통합 데이터 셋을 제시합니다. 2개의 손과 물체를 나타내는 H2O라고 하는 데이터 세트는 포괄적인 자기 중심적 장면을 생성할 수 있는 3D 손 물체 포즈 및 모양, 동작 레이블, 카메라 포즈, 장면 포인트 클라우드 및 개체 mesh에 대한 풍부한 실측 주석을 제공합니다.

다양한 배경을 가진 실용적인 다중 카메라 시스템을 사용하여 접촉하는 물체 뿐만 아니라 상호 작용하는 두 손의 포즈와 동작 레이블이 있는 손 물체 상호 작용 데이터 세트를 선별하는 반자동 파이프라인을 제안합니다. 자세한 검증을 통해 주석의 충실도와 정확성을 입증합니다.

처음으로 상호 작용하는 두 손의 3D 포즈와 조작된 개체의 6D 포즈, 동작 및 개체 클래스를 동시에 예측하는 RGB 이미지에서 손 개체 상호 작용을 인식하는 통합 접근 방식을 소개합니다.

데이터 셋을 활용하여 토폴로지 인식 그래프 컨볼루션 네트워크를 통해 두 손과 객체 간의 상호 의존성을 학습하는 3D 상호 작용 인식을 위한 새로운 방법이 있습니다. 손과 물체 포즈를 개별 그래프로 매개변수화하고 단일 다중 그래프 아키텍처에서 결합합니다. 그런 다음 적응형 아키텍처를 사용하여 서로 다른 그래프 엔터티 간의 상호 의존성과 연결을 학습하고 3D 손-개체 상호 작용을 인식하기 위한 다중 그래프 구조의 토폴로지를 계산합니다.

논문의 데이터 셋에 의해 촉진된 포즈 예측을 사용하여 최신 기술을 능가하는 상호 작용을 인식하는 데 있어 전반적으로 더 나은 성능을 달성한다는 것을 입증합니다.

우리는 이 데이터 셋에 대한 추가 벤치마킹을 가능하게 하기 위해 손 및 물체 포즈 추정 및 상호 작용 인식을 위한 기준선을 추가로 제공합니다.

**관련된 연구**

자기 중심적 행동 인식 및 손 물체 포즈 추정을 위한 데이터 셋입니다. 3인칭 행동 인식을 위한 많은 데이터 세트가 수년에 걸쳐 제안되었지만, 최근에는 자기 중심적 시나리오를 대상으로 하는 데이터에 대한 관심이 급증하고 있습니다. 대부분 2D 기능을 포함합니다.

이러한 데이터 셋은 제한된 다중 보기 데이터만 제공하고 장면을 포괄적으로 이해하는 데 유용한 단서로 표시되는 손 및 개체 포즈를 제공하지 않습니다. 몇몇 데이터 셋은 자동화 또는 반자동 방식으로 획득한 손 자세 지상 실측 데이터를 수집합니다. 그러나 개체와의 상호 작용은 고려하지 않습니다.

최근에 GRAB은 mocap 시스템과 객체를 사용하여 해당 이미지를 제공하지 않고 장면과 상호 작용하면서 몸과 손 포즈를 추적합니다.

HOnnotate는 멀티 뷰 RGB-D 데이터에서 정확한 손과 물체 포즈를 추정하기 위해 최적화 프로세스에 의존합니다.

ObMan은 물체를 들고 있는 손의 순수 합성 이미지를 수집합니다. 그러나 이 모든 작업은 한 손 시나리오만 고려하고 동작 인식에 중점을 두지 않습니다. 논문과 마찬가지로 FPHA는 동작, 손 및 개체 포즈 주석이 있는 자기 중심적인 RGB-D 프레임을 수집합니다. 그러나 데이터 셋은 RGB 이미지를 오염시키는 자기 센서에 의존하며 다시 점 데이터와 양손 포즈를 포함하지 않습니다.

표에서 볼 수 있듯이 우리의 데이터 셋은 자기 중심적 3D 상호 작용 인식을 위한 두 손의 3D 포즈, 개체 포즈 및 동작 레이블에 대한 실제 다시 점 RGB-D 데이터와 정확한 주석을 포함하는 최초의 데이터 셋입니다.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

손 상호 작용을 위한 3D 주석이 있는 기존 관련 이미지 기반 데이터 세트의 비교. H2O는 5개의 다른 보기에서 캡쳐된 총 571k 프레임을 제공합니다.

(\*): 프레임의 하위 집합에 대해서만 제공되는 개체 포즈입니다.

(\*\*) 손과 물체에 마커가 없는 방법입니다.

(\*\*\*) 질감이 없는 인쇄된 개체입니다.

**손 및 물체 포즈 추정**

상당한 양의 연구가 손 또는 물체의 자세를 단독으로 예측하는 데 초점을 맞추었지만, 관절 손과 물체의 상호 작용에 대한 이해는 훨씬 덜 주목을 받았습니다. 손과 물체를 함께 고려하면 추론이 필요한 여러 가지 문제가 추가됩니다.

Pioneer는 느리고 조정하기 어려울 수 있는 최적화 프레임워크에 의존하는 손-손 및 손-개체 상호 작용을 조사합니다. RGB 이미지에서 직접 손 및 물체 포즈를 효율적으로 추정합니다. 그러나 그들은 한 손 시나리오만을 고려합니다.

**상호 작용 인식**

동작 인식은 컴퓨터 비전 커뮤니티에서 많은 관심을 받았습니다. 딥 러닝의 출현과 대규모 데이터 세트의 가용성으로 인해 3인칭 행동 인식에서 상당한 진전이 이루어졌습니다.

그러나 1인칭 시점에서 상호 작용을 인식하는 것은 큰 폐색, 빠른 카메라 움직임 및 배경 혼란과 같은 많은 특정 문제를 제기합니다. 처음에는 많은 양의 데이터 부족으로 인해 효과적인 DNN 기반 방법의 개발이 다소 방해를 받았지만 최근 몇 년 동안 이 문제에 대한 새로운 관심이 생겼습니다. 일부 접근법은 머리 움직임 및 시선 응시와 같은 다중 모달 입력을 활용합니다.

또한 CNN을 사용하여 기능을 추출하고 동작, 손 위치, 개체 위치 또는 개체 클래스와 관련된 추가 2D 단서를 활용하는 것이 일반적입니다. 이러한 모든 방법이 2D 기능에 초점을 맞추는 동안 최근 연구는 3D 단서(손 및 물체 포즈와 같은)가 자기 중심적 행동 인식의 맥락에서 효과적일 수 있다고 제안합니다.

그러나 기존 방법은 한 손 추적에 중점을 두었고 실제 시나리오에서 발생하는 상호 작용을 더 잘 나타내는 시나리오인 객체와 상호 작용하는 두 손의 포즈를 추정하는 데 지금까지 관심을 기울이지 않았습니다.

스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명

그림 2

(a) IR 구체 마커와 PnP를 사용하여 카메라를 보정합니다.

(b) RGB-D 캡처에서 BADSLAM을 사용하여 개체 메시를 만듭니다.

(c) Mask R-CNN의 RGB-D 이미지 및 마스크 이미지에서 DenseFusion을 사용하여 개체 포즈를 추정합니다. 그런 다음 5대의 카메라 중에서 가장 신뢰도가 높은 포즈를 선택합니다.

(d) 결과적으로 우리는 OpenPose로 손 관절을 감지하고 이를 사용하여 손 모양을 최적화합니다.

(e) 마침내 시간적으로 부정확한 포즈를 감지하고 부드럽게 합니다.

**주석 방법**

그림은 주석 파이프라인의 개요를 보여줍니다.

5대의 Azure Kinect 카메라를 사용하여 여러 보기에서 동기화된 RGB-D 프레임을 캡처합니다.

카메라 중 하나는 자기 중심적 프레임을 캡처하기 위해 다른 피사체가 착용하는 헬멧에 장착됩니다. 우리는 반자동 방식으로 실제 손과 물체 포즈를 획득합니다.

먼저 Kinect로 각 개체를 스캔하여 완전한 3D 모델을 얻습니다.

이 모델은 DenseFusion을 통해 각 프레임에서 개체 6D 포즈를 추적하는 데 사용됩니다.

손을 추적하기 위해 MANO 파라메트릭 손 모델[64]을 각 프레임의 멀티뷰 깊이 데이터에 맞춥니다. 이 자동화된 추적 프로세스는 (자체) 폐색, 흐림 및 어수선한 배경과 같은 문제로 인해 일부 프레임에서 실패할 수 있습니다. 따라서 실패 사례를 수동으로 감지하고 해당 포즈를 제거합니다. 이러한 포즈는 임시 스무딩을 통해 대체할 수 있습니다.

마지막으로 시퀀스에 작업 레이블을 수동으로 추가합니다. 다음 섹션에서는 파이프라인의 각 단계를 자세히 설명합니다.

**카메라 보정**

논문에서의 설정은 4개의 정적 카메라와 1개의 머리 장착형 RGB-D 카메라로 구성됩니다. Azure Kinect DK를 통해 액세스할 수 있는 공장에서 보정된 고유 매개 변수를 사용합니다. 외부 매개변수는 IR 반사 구에 의존하는 보정 방법으로 얻습니다. 논문은 설정을 이식 가능하고 배포하기 쉽게 만들기 위해 이 방법을 선택합니다.

장면의 임의 위치에 9개의 IR 반사 구를 배치하여 모든 카메라에서 각 구를 볼 수 있도록 합니다.

깊이를 재구성하기 위해 카메라로 캡처한 IR 이미지에서 이러한 구체는 자동화된 방식으로 쉽게 감지할 수 있는 밝은 원으로 표시됩니다. 각 구의 중심을 계산한 다음 깊이 이미지의 해당 픽셀을 고려하여 3D 위치를 얻습니다. 각 프레임에서 9개의 구의 3D 위치가 주어지면 PnP를 통해 카메라 포즈를 해결합니다. 여러 프레임에서 구체를 일관되게 식별하기 위해 첫 번째 프레임에서 초기 매핑을 정의한 다음 시간 경과에 따라 추적합니다. 헤드 마운트 카메라에 대해 계산된 포즈는 jitter를 나타낼 수 있습니다. 머리가 일정한 속도로 움직인다는 가정 하에 Kalman 필터링을 통해 부드럽게 합니다. 전체 프레임워크를 사용하면 주석 중에 여러 카메라를 사용할 수 있으므로 결국 주석의 충실도와 정확도가 높아집니다.

**개체 포즈 주석**

카메라 포즈 정보와 함께 점 이미지를 사용하여 정확한 perframe 개체 6D 포즈를 얻습니다.

먼저 각 객체에 대한 3D mesh 모델을 재구성합니다. 이를 위해 주변을 움직이는 휴대용 Kinect 카메라로 RGB-D 프레임을 캡처하여 개체를 스캔합니다. 이러한 프레임을 최첨단 RGB-D SLAM 방법인 BADSLAM [68]에 공급하여 3D mesh를 재구성합니다. 각 객체에 대한 텍스처를 얻습니다.

BADSLAM에서 반환한 카메라 포즈를 사용하여 스캐닝 시간에 얻은 RGB 이미지를 mesh 표면에 투영합니다. 이러한 모델을 활용하여 객체 포즈 추적기를 훈련합니다.

먼저 Mask R-CNN을 기반으로 객체 마스크 예측기를 훈련합니다. 교육 데이터로 BADSLAM 기반 재구성에 사용된 이미지에 3D 모델을 투영하여 얻은 마스크를 사용합니다. 그런 다음 해당 RGB-D 이미지와 함께 마스크 예측을 DenseFusion에 공급하여 개체 포즈를 추정합니다. 각 카메라 뷰에 대한 포즈 예측을 얻고 가장 신뢰도가 높은 것을 선택합니다.

즉, 5개의 깊이 이미지 각각에서 포인트 클라우드를 계산하고 카메라 포즈 정보를 사용하여 단일 포인트 클라우드로 병합합니다. 그런 다음 DenseFusion의 예측을 초기화로 사용하여 개체 모델을 이 포인트 클라우드에 맞춥니다.

**손 포즈 주석**

손 포즈 추정을 위해 널리 사용되는 MANO 손 모델에 의존합니다.

MANO는 사람의 손 모양을 식별 매개변수 세트 β ∈ R10과 포즈 매개변수 세트 θ ∈ R51로 분해하여 15개의 골격 관절에 대한 각도와 전역 회전 및 변환을 저장합니다.

공식적으로 우리는 MANO를 꼭짓점을 가진 삼각형 메시를 반환하는 (θ, β) 함수로 정의할 수 있습니다. 또한 MANO 스켈레톤을 = 21개의 관절 위치를 반환하는 함수 (θ, β)로 정의합니다.

위와 같이 추정된 물체 포즈를 취하여 손 포즈를 추적할 때 활용합니다. 손실 함수의 각 frame에서 최소화하여 손을 추적합니다.

는 카메라 수입니다. 는 실루엣 기반 오류 향이고, 및 는 각각 2D 및 3D에서 관절 오류를 측정하고, 및 는 포즈에 대한 정규화이며, 는 손과 물체 사이의 물리적으로 타당하지 않은 상호 침투에 페널티를, 은 3D에서 거리에 페널티를 줍니다. λ는 각 오류 용어의 기여도에 가중치를 부여합니다. 주제별 매개 변수를 얻기 위해 식을 최소화합니다.

텍스트, 폰트, 화이트, 친필이(가) 표시된 사진

자동 생성된 설명

2D joint error

MANO 관절과 OpenPose 추정 사이의 2D 거리에 페널티를 부여합니다. 여기서 J2D는 OpenPose로 사전 계산된 2D 관절 위치를 나타내고 HJ(θ)[i]는 MANO 골격의 i번째 3D 관절 위치를 반환합니다.

폰트, 텍스트, 화이트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

3D joint error

2D 관절 오류와 마찬가지로 OpenPose 추정치를 삼각 측량하여 3D에서 페널티를 계산합니다.

우리는 이 오차항을 사용하는 것이 더 빠른 수렴을 달성하고 안정성을 높이는 데 도움이 된다는 것을 발견했습니다.

3D mesh surface error

서로 다른 뷰에서 각 깊이 이미지에서 얻은 포인트 클라우드를 병합하고 위와 같이 계산된 손 마스크에 투영되지 않는 포인트를 분할하여 손 데이터에 대한 포인트 클라우드를 얻습니다.

우리의 3D 표면 오류 항은 이 포인트 클라우드와 MANO 표면 사이의 거리에 불이익을 줍니다. 여기서 는 포인트 클라우드의 j번째 포인트이고 HV ⊥ (θ)는 핸드 메시 정점 i의 법선을 나타냅니다.

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명

식에 나타낸 바와 같이. 1, 우리의 최적화 기능에는 관절 각도 제한 및 물리적 제약에 대한 실루엣 오류 용어 및 정규화가 추가로 포함됩니다. 자동화된 파이프라인을 실행한 후 모든 프레임을 검사하여 부정확한 포즈를 식별하고 제거합니다.

마지막 단계로 Kalman 필터링을 통해 포즈를 매끄럽게 연결합니다.

**임시 작업 주석**

동작 레이블을 동사-명사 쌍으로 제공합니다.

논문은 11개의 동사 클래스를 고려합니다 : 잡기, 놓기, 열기, 닫기, 붓기, 꺼내기, 넣기, 적용하기, 읽기, 뿌리기 및 짜기.

명사에 관해서는 책, 에스프레소, 로션, 스프레이, 우유, 코코아, 칩, 카푸치노의 8가지 클래스를 고려합니다.

동사와 명사를 결합하여 데이터 세트에 표시되지 않은 쌍을 제외하고 총 36개의 동작 클래스를 얻습니다.

모든 프레임에 대해 하나의 동사와 하나의 명사만 선택하므로 겹치는 작업 레이블이 없습니다.

VIA 주석 도구를 사용하여 전체 데이터 세트에 대한 작업 레이블을 수동으로 선택합니다.

아래 그림은 몇 가지 주석 예를 보여줍니다.

스크린샷, 만화 영화이(가) 표시된 사진

자동 생성된 설명

**H2O 데이터 셋**

논문은 주제가 양손을 사용하여 8개의 다른 물체와 상호 작용하는 실내 설정에서 H2O 데이터 세트의 이미지를 획득했습니다. 데이터 셋에는 571,645개의 RGBD 프레임이 포함되어 있으며 3가지 다른 환경에서 36개의 개별 작업 클래스를 수행하는 4명의 참가자가 있습니다. 왼쪽 및 오른쪽 손 포즈, 6D 개체 포즈, 카메라 포즈 및 동작 레이블에 대한 정확한 실측 데이터에 주석을 답니다.

데이터 세트에서 우리는 왼손과 오른손 모두에 대한 MANO 핸드 핏과 고품질 개체 mesh를 추가로 제공합니다.

또한 카메라 포즈와 동기화된 RGBD 데이터를 사용하여 장면 포인트 클라우드를 계산합니다.

전체적으로 선별된 데이터 세트를 통해 자기 중심적인 장면을 포괄적으로 이해할 수 있습니다.

**데이터 셋 통계**

데이터 을 교육 및 테스트 셋으로 나눕니다.

교육 및 테스트 데이터를 주제 기반 분할로 분할하여 하나의 주제는 테스트에, 나머지는 교육에 사용합니다. 모델 선택을 위한 유효성 검사 데이터 셋으로 한 주제의 교육 데이터 일부를 추가로 사용합니다.

다중 뷰의 데이터는 훈련용 344,645개 프레임, 검증용 73,380개 프레임, 테스트용 153,620개 프레임으로 구성됩니다.

액션당 인스턴스 수와 각 액션 클래스의 평균 프레임 수를 그림에 표시합니다.

가장 빈도가 낮은 작업이 21회 나타나는 작업 인스턴스는 데이터 세트 전체에 잘 분산되어 있습니다. 데이터 셋에서 양손을 사용하는 비율은 57.8%, 왼손만 사용하는 비율은 12.4%, 오른손만 사용하는 비율은 29.8%이다. 동작 클립의 길이는 광범위하며 느린 동작과 빠른 동작을 모두 포함하는 데이터 셋의 다양성을 보여줍니다.

스크린샷, 라인, 텍스트, 도표이(가) 표시된 사진

자동 생성된 설명

**3D 손-개체 상호 작용 인식**

H2O의 풍부한 주석을 감안할 때 우리의 목표는 이미지 시퀀스에서 자기 중심적 장면에 대한 포괄적인 해석을 구성하여 인간 상호 작용을 이해하는 것입니다.

두 손과 조작 대상의 포즈를 공동으로 추정하고 자기 중심적 상호 작용을 인식하는 통합 프레임워크를 제안합니다. 이 프레임워크를 사용하여 1인칭 상호 작용 인식과 손 및 물체 포즈 추정에 대한 기준선을 설정합니다.

한 손 시나리오만 다루지만, 우리의 경우 양손의 포즈를 예측하는 것을 목표로 합니다. 이를 위해 시퀀스의 각 프레임은 YOLOv2의 백본이 있는 완전 convolution 네트워크를 통해 전달됩니다. 2D 그리드를 생성하는 대신 완전 convolution 네트워크의 출력으로 3D 그리드를 생성합니다. 손과 물체의 자세를 동시에 예측할 수 있도록 각 출력 그리드 셀을 왼손, 오른손 및 조작된 물체에 대한 3개의 벡터와 연결합니다. 이러한 벡터에서 왼손(), 오른손() 및 개체 포즈()에 대한 목표 값과 개별 포즈 예측에 대한 전체 신뢰도 값()이 포함됩니다.

신뢰도 값은 예측된 포즈와 실측 포즈의 거리 함수로 훈련 중에 즉석에서 정의됩니다. 단일 샷 네트워크의 마지막 계층은 각 셀 i에 대해 왼손(), 오른손() 및 객체()에 대한 예측을 관련 전체 신뢰 값 와 함께 생성합니다. , 와 각 프레임에 대해 네트워크를 훈련시키는 손실 함수는 다음과 같이 정의됩니다.

**상호작용 인식**

RNN은 동작을 인식하기 위해 이전에 성공적으로 사용되었습니다. 그러나 그들은 손과 물체의 상호 작용을 위해 골격 데이터의 특수한 그래프 구조를 완전히 활용하지 않습니다. 왼쪽 관절, 오른쪽 관절 및 개체 경계 상자를 개별 그래프로 매개 변수화하고 다중 그래프 구조로 결합합니다. 그런 다음 상호 작용에 관여하는 손과 개체 위치에 걸친 링크를 학습하여 그래프 convolution 네트워크(GCN)를 사용하여 다중 그래프 구조의 토폴로지를 계산합니다. 단일 그래프 내에서 내부 종속성을 모델링하는 동안 이 프레임워크는 왼손-오른손, 왼손-객체 및 오른손-객체 간의 상호 종속성을 학습할 수도 있습니다.

특히 ST-GCN 및 2s-AGCN에서와 같이 공간 및 시간 정보를 모두 인코딩하기 위해 시공간 그래프를 사용합니다.

폰트, 텍스트, 화이트, 그래픽이(가) 표시된 사진

자동 생성된 설명

인간 행동 인식 모델을 위한 표준 ST-GCN ∈ 은 입력 특징 맵, ∈ 은 골격 연결을 나타내는 인접 행렬, 를 사용하여 신체 골격 관절 간의 구조화된 정보 ∈ 은 1 x 1 convolution의 가중치 벡터이고 ∈ 은 attention map입니다.

여기서 j는 convolution 커널에 의해 정의된 정점 이웃을 나타내고, C는 채널 수, T는 시간 길이, N은 정점 수입니다. ST-GCN은 단일 그래프 entity에서 작동합니다. 인간 골격, 고정된 인접 행렬과 골격 내 연결을 모델링합니다.

우리의 경우 그래프 내 종속성 외에도 손과 개체 간의 그래프 간 종속성을 모델링하는 것을 목표로 합니다.

서로 다른 손과 물체 부분이 상호작용에 관여할 때마다 상호 종속성을 모델링하기 위한 고정 인접 행렬은 최적의 결과를 산출하지 못합니다.

따라서 왼손, 오른손 및 개체에 대해 개별적으로 종속성을 모델링할 수 있도록 다음을 사용합니다.

에서 와 같은 역할을 한다. 왼손, 오른손 및 물체에 대한 모델은 대칭 손 부분과 물체 중심 사이의 정적 연결을 통해 손과 물체 사이의 상호 의존성을 모델링합니다. 여기서 이 두 행렬은 모두 ST-GCN에서와 같이 고정된 인접 행렬입니다. 왼손과 물체, 오른손과 물체 사이의 상호 연결을 나타냅니다.

와 달리 는 고정되지 않고 매개변수화 됩니다.

그 값은 제약이 없으며 다른 네트워크 매개변수와 공동으로 최적화됩니다.

즉, 그래프 토폴로지와 에지 가중치가 교육 데이터에서 완전히 학습됩니다.

외에도 상호 작용하는 동안 단일 그래프 entity(예: 왼손, 오른손 또는 개체) 내에서 내부 관련 종속성을 적응적으로 학습하는 추가 매개 변수화된 인접 행렬 도 사용합니다.

이 데이터 기반 모델을 통해 손-개체 상호 작용 작업을 완전히 대상으로 하는 그래프를 학습할 수 있습니다.



**평가**

이 섹션에서는 먼저 ground truth 주석의 정확성을 확인합니다.

그런 다음 데이터 세트에 대한 손 및 물체 포즈 추정 및 자기 중심적 행동 인식에 대한 기준 결과를 제시합니다.

후자의 경우, 우리는 또한 기본 접근 방식을 동작 인식의 최신 기술과 비교하고 기존 방법과 관련하여 손 개체 포즈를 기반으로 하는 접근 방식의 명확한 이점을 보여줍니다.

**데이터 셋 분석**

데이터 세트의 무작위 분할에서 손 개체 포즈 주석의 정확성을 확인합니다.

이를 위해 손끝과 조작된 객체의 미리 정의된 키포인트를 사용하여 5개의 다른 카메라 보기에서 500개의 이미지에 주석을 답니다. 2D 점을 삼각 측량하여 손과 물체에 대한 수동 3D 주석을 얻습니다. 포즈의 정확성을 측정하기 위해 주석과 수동으로 생성된 주석의 거리를 계산합니다.

검증 결과는 표에 나와 있습니다. 양손과 물체에 대해 오차는 약 1cm 범위 내이며 이는 데이터 세트의 높은 정밀도를 보여줍니다.

H2O는 정확한 손 및 개체 포즈 주석 및 메시를 통해 손-개체 접촉 모델링을 더욱 용이하게 합니다. 이를 위해 손 메시의 각 정점에 대해 특정 임계값(예: 2cm) 내에서 객체에서 가장 가까운 정점을 찾습니다. 그런 다음 MANO 메시의 각 정점에 대한 이웃 수를 계산하는 히스토그램을 계산하고 이를 정규화하여 가까이 있는 접촉 hotspot을 모델링합니다. 개체 mesh에 대해서도 동일한 절차를 반복하여 개체 표면에 접촉 map을 만듭니다. 그림에서 데이터 세트의 예제 연락처 map을 시각화합니다.

**실험 결과**

두 손과 조작 대상의 3D 포즈를 함께 예측합니다. 설명된 훈련, 검증 및 테스트 분할을 사용하여 방법을 훈련하고 평가하고 그림에서 손과 물체에 대한 기본 포즈 추정 정확도를 보고합니다. 올바르게 추정된 포즈의 백분율을 사용하여 손 및 물체 포즈 추정 정확도를 평가합니다. 손 포즈 추정을 위해 3D PCK matric을 사용하고 객체 포즈 추정을 위해 2D 재투영 및 ADD matric을 사용합니다. 논문은 우리의 방법이 낮은 오류 마진으로 두 손의 자세와 조작된 물체를 안정적으로 예측할 수 있고 물체와 상호 작용하는 두 손의 관절 자세 추정을 위한 강력한 기준선을 구성한다는 것을 입증합니다. 또한 우리의 접근 방식은 단일 RGB 이미지에서 개체와 상호 작용하는 두 손의 포즈를 추정하기 위한 첫 번째 방법 및 기준선을 구성합니다. 비교 목적으로 단일 손 개체 포즈 추정 방법에 대한 접근 방식을 표에서 평가하고 그림 5에서 포즈 예측의 정성적 예를 추가로 제공합니다.

손 포즈와 객체 키포인트는 단일 패스 네트워크를 통해 예측됩니다.

논문은 오른쪽 및 왼쪽 손 포즈의 조합과 손 및 물체 포즈의 조합이 전체 동작 인식 점수를 크게 향상시키는 것을 보여줍니다. 이는 각 입력 양식의 개별 기여 및 보완적인 특성을 보여줍니다.

표에서 손과 물체 간의 상호 의존성 모델링의 중요성을 더 평가하고 왼손과 오른손, 손과 물체 간의 상호 의존성 모델링이 상호 작용 인식의 정확도를 높인다는 것을 입증합니다.

표에서 식의 다른 항의 영향을 평가하고 모든 그래프를 함께 추가하면 모델이 기준선과 비교하여 최상의 결과를 얻는다는 것을 보여줍니다.

또한 PySlowFast 라이브러리를 사용하는 C2D, I3D 및 SlowFast의 최신 이미지 기반 학습 방법과 우리의 동작 인식 정확도를 비교하고 포즈 기반 학습 방법은 다음과 같습니다. H+O 및 ST-GCN 및 표에 결과를 보여줍니다.

우리는 ResNet-50 백본을 사용하고 학습률이 0.1인 SGD를 사용하여 네트워크를 교육합니다.

상호 작용 인식에 대한 우리의 접근 방식은 H2O 데이터 세트에서 가장 높은 검증 및 테스트 정확도를 달성하여 우리 방법의 효율성과 H2O에 의해 촉진되는 3D 포즈 예측의 중요성을 보여줍니다.

라인, 그래프, 도표, 폰트이(가) 표시된 사진

자동 생성된 설명

(a) 3D PCK matric이 있는 손 및 (b) 2D 재투영 및 (c) matric을 추가합니다.

텍스트, 스크린샷, 폰트, 영수증이(가) 표시된 사진

자동 생성된 설명

서로 다른 (a) 입력 양식, (b) 상호 연결 및 (c) 상호 작용 인식 정확도에 대한 그래프 용어의 영향.

스케치, 종이접기이(가) 표시된 사진

자동 생성된 설명

다양한 계층에 대한 학습된 그래프 연결.

각 열의 레이어 1, 5 및 9에서 각각 상위 20개의 학습된 내부(상단) 및 내부(하단) 연결을 보여줍니다.

연결의 두께는 학습된 연결 값의 가중치에 해당합니다.

손-개체 연결은 개체와 상호 작용하는 동안 손-손 연결보다 더 많은 가중치가 부여됩니다.

우리 모델은 조작에 더 일반적으로 관여하는 손가락 끝과 DIP 관절에 더 많은 중요성을 부여합니다.

**정확히 모르는 개념 정리**

Object Recognition

Object recognition은 대체로 Object detection과 같은 의미로 쓰입니다. 그러나 detection은 object의 존재 유무만 의미하고 recognition은 이 object의 종류를 아는 것이라 해석하여 Object detection이 Object Recognition보다 작은 개념이 될 수도 있다고 합니다.

Object Segmentation

Object Segmentation이란 위 이미지와 같이 검출된 object의 형상에 따라 영역을 표시하는 것입니다. 보통 이미지의 pixel을 classification 하여 위와 같은 결과를 도출합니다. 단순히 foreground와 background를 구분하는 용도로 쓰이기도 합니다.

Instance Segmentation

Instance segmentation이란 semantic segmentation 에서 한발 더 나아가서, 같은 class이어도 서로 다른 instance 들을 구분해주는 것입니다.

BBOX

Bbox는 이미지 내에서 물체의 위치를 사각형으로 감싼 형태의 도형으로 정의하고 꼭짓점의 좌표로 표현하는 방식입니다.

포유류, 개, 개 품종, 애완동물이(가) 표시된 사진

자동 생성된 설명

이 이미지로 보면 좌표가 left top과 right bottom 부분의 좌표로 표현되어 있습니다. 해당 방법 이외에 Bbox의 width와 height로 정의하는 방식이 있습니다. 이 경우 left top의 점에 대한 상대적인 위치로 물체의 위치를 정의할 수 있습니다.