리뷰 보고서

(First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations)

소개

이 논문에서는 세밀한 손 포즈 기능의 도움으로 일상 생활의 동적 동작을 수행하는 동안 손을 사용하는 방법을 이해하는데에 관심을 가지고 있습니다. 손 재활, 가상/증강 현실, 원격 조작, 로봇 모방 학습 등 있습니다. 1인칭 행동 인식에 대한 이전 작업은 일상적인 행동이 손을 보면 설명된다는 것을 발견했으며, 3인칭 시점에서 발견되는 유사한 관찰입니다. 이런 접근 방식에서 손에 대한 정보는 손 실루엣 or 낮은 수준의 이미지 특징을 사용하는 이산 파악 분류에서 추출됩니다.

전신 인간 동작 인식에서 신체 포즈와 같은 더 높은 수준 및 관점 불변 기능을 사용한다면 동작 인식에 도움이 될 수 있다고 알려져 있지만, 아직 손에 대해 자세히 연구되어 있지 않습니다.

전신 동작과 비교할 때 손 동작은 포즈를 단서로 사용하는 것이 분명하지 않은 고유한 차이점을 나타냅니다. 손가락의 높은 이동성으로 인해 대상 간의 스타일과 속도 변화가 더 뚜렷하고 동작이 매우 미묘할 수 있습니다. 동작 인식에 손 포즈를 사용하는데 걸림돌은 주로 실제 데이터 sequence에 손 포즈 주석이 없어서 전신과 달리 신뢰할 수 있는 포즈 추정기가 없습니다.

손가락 끝에 부착된 6개의 센서와 역운동학을 사용해 3D 손 포즈로 주석이 달린 100,00개 RGB-D frame이 있는 1인칭 손 동작 sequence의 새로운 데이터 셋입니다.

3개의 시나리오에서 26개의 서로 다른 개체를 조작하는 45개의 범주를 포함하여 1175개의 작업 샘플을 캡처하고, 여러 손 구성과 시간 역학을 다루기 위해 손 동작과 선택한 개체를 설계합니다.

데이터 셋에서 여러 기준선과 최신 RGB-D 및 포즈 기반 동작 인식을 평가하고 현재 최신 손 포즈 추정과 동작 인식에 미치는 영향을 테스트합니다.

스크린샷, 만화 영화, 콜라주이(가) 표시된 사진

자동 생성된 설명

액션 클래스 'pour juice'에 속하는 시퀀스에서 두 개의 프레임을 보여줍니다.

우리는 RGB-D 비디오와 3D 손 포즈 주석이 포함된 새로운 1인칭 동작 인식 데이터 세트를 제안합니다. 자기 센서와 역운동학을 사용하여 손 자세를 캡처합니다.

오른쪽에는 캡쳐된 깊이 이미지와 손 포즈가 표시됩니다.

또한 손 개체 동작의 하위 집합에 대한 6D 개체 포즈를 캡처했습니다.

논문 기여도

1. 데이터 셋: 우리는 자기 중심적인 동적 손 개체 동작 및 포즈 연구를 돕기 위해 완전히 주석이 달린 데이터 세트를 제안합니다.

2. 동작 인식: 제안된 데이터 세트를 사용하여 RGB-D 및 포즈 기반 동작 인식에서 18개의 기준선과 최신 접근 방식을 평가합니다.

3. 손 포즈: 실제 데이터 세트, 즉 손 개체 조작의 가려진 설정에서 최첨단 손 포즈 추정기를 평가하고 동작 인식에 대한 성능을 평가합니다.

관련된 연구

1. 자기 중심적 비전 및 조작 데이터 셋

물체를 조작하는 동안 손의 중요한 역할은 컴퓨터 비전 및 로봇 공학 커뮤니티입니다. 동작 인식 관점에서 RGB 큐만 사용하는 최근 연구는 일상적인 동작을 인식하고 조작된 물체와 손 모두 동작 인식 문제에 대한 중요한 단서라고 결정했습니다.

관련 작업 라인은 로봇 관점, 동작 인식, 힘 추정 및 인식 문제입니다.

최근에는 동작 인식을 위한 사용을 탐색하지 않고 더 쉽게 손을 감지할 수 있는 열화상 카메라를 사용하는 벤치마크를 제안했습니다. 이전 작업에서 손은 경험적 파악 분류법에 따른 하위 수준 기능 or 중간 표현을 사용하여 모델링되었으므로 3D 손 포즈 sequence에 비해 제한적입니다. 손 포즈 추정 관점에서는 정적 포즈의 합성 데이터 셋을 제안했고 더 큰 합성 데이터셋에 의해 완화된 데이터 알고리즘을 훈련할 수 없습니다. 개체의 하위 집합에 대해 6D 개체 포즈 및 3D 메시 모델도 제공한다는 점을 감안할 때 논문의 데이터 셋은 개체 포즈 및 공동 손-개체 추적 신흥 커뮤니티 모두에 관심이 있을 수 있습니다

2. RGB-D 및 포즈 기반 동작 인식

깊이 센서를 사용하는 것은 대부분의 색상 접근 방식이 깊이 stream에 직접 적용될 수 없다는 점에서 기존의 색상 동작 인식과 다릅니다. 일반적으로 관점 변화에 민감한 로컬 기하학적 설명자와 view 불변 접근법을 사용하여 깊이 이미지에서 식별 특징을 추출하는 방법에 중점을 둡니다. 최근 추세는 깊이 채널을 활용하여 강력한 신체 포즈 추정을 얻고 포즈 or 골격 동작 인식으로 알려진 동작을 인식하는 기능으로 직접 사용하는 것입니다.

인기 있는 접근 방식 : 시간적 상태 공간 모델, key 포즈, 손으로 만든 포즈 기능 및 시간적 순환 모델

여러 데이터 stream을 갖는 것은 깊이 및 포즈, 색상 및 포즈 및 이들 모두와 같은 다양한 정보 소스를 결합하는 연구로 이어집니다. RGB-D 동작 인식의 대부분의 작업은 인간 - 컴퓨터 상호 작용을 위한 손 제스처와 같이 응용 지향적인 일부 예외를 제외하고는 전체 인체에 의해 수행되는 동작에 초점을 맞춥니다. 이러한 작업에서는 hand tracker의 낮은 품질로 인해 수행되는 동작 및 포즈 레이블이 매우 제한적인 반면, 현실적인 손 동작을 연구하기 위해 정확한 손 포즈 레이블을 제공합니다.

3. 3D 손 자세 추정

주로 RGB-D 센서의 최근 가용성으로 인해 이 분야는 물체 없는 3인칭 시점에서 상당한 진전을 이루었습니다. 그리고 1인칭 시점에서 좀 더 완만하게 발전했습니다.

중요한 제한 사항은 정확한 3D 손 포즈 주석을 얻기가 어려워 연구원이 합성, 수동 또는 반자동 주석이 달린 데이터 셋에 의존하여 결과적으로 비현실적인 이미지, 적은 수의 샘플, 종종 일관되지 않은 주석이 있습니다.

데이터셋 개요

데이터 셋에는 3개의 다른 시나리오에서 6명의 배우가 수행한 45개의 서로 다른 동작 범주에 속하는 1175개의 동작 비디오가 있습니다.

총 105,459개의 RGB-D frame에 정확한 손 포즈 및 동작 범주가 주석으로 추가됩니다.

작업 sequence는 스타일, 속도, 규모 및 관점의 개체 간 및 개체 내 변동성이 높습니다. 객체의 6차원 포즈, 3D 위치 및 각도, mesh 모델도 10가지 동작 카테고리가 포함된 4가지 객체에 대해 제공됩니다.

손 개체 작업

논문에서 26개의 다른 물체와 관련된 45개의 다른 손 동작 범주를 캡쳐했습니다. 또한 다른 손 구성에 걸쳐 있고 손 포즈와 행동 공간 모두에서 다양하도록 행동 범주를 설계했습니다. 각 개체에는 최소 하나의 관련 직업이 있습니다. 45개의 손 동작을 기록하고 주방, 사무실 및 소셜의 3가지 다른 시나리오로 그룹화했습니다.

센서 및 데이터 수집

시각적 데이터 : 피사체의 어깨에 카메라를 장착하고 색상 및 깊이 스트림에 대해 각각 30fps 및 해상도 및 sequence를 캡쳐했습니다.

포즈 주석 : 손과 물체 자세에 대한 품질 주석을 얻기 위해 사용자의 손에 부착된 6개의 자기 센서, 5개의 손가락 끝 및 1개의 손목을 사용하여 손 자세를 캡쳐합니다.

각 센서는 6자유도로 위치와 방향을 제공하고 정의된 21개 관절 손 모델에 대한 역운동학을 사용하여 전체 손 자세를 추론합니다.

각 센서의 너비는 2mm이며 사람의 손에 부착해도 깊이 이미지에 영향을 미치지 않습니다.

컬러 이미지는 센서와 이를 부착하는 테이프가 보이기 때문에 영향을 받지만 손은 완전히 보이고 컬러 이미지를 사용하여 동작을 구별할 수 있습니다.

물체의 자세와 관련하여 도달할 수 있는 질량 중심에 가장 가까운 지점에 센서를 하나 더 부착합니다.

기록 과정: 모두 오른손잡이인 6명에게 작업을 수행하도록 요청했습니다.

안전한 방식으로 작업을 수행하는 방법에 대한 지침이 제공되었지만 실제 데이터를 캡처하기 위해 스타일이나 속도에 대한 지침은 제공되지 않았습니다.

작업에 수동으로 레이블이 지정되었습니다.

원, 스크린샷, 텍스트, 도표이(가) 표시된 사진

자동 생성된 설명

데이터셋 통계

분류: 개체별로 다른 작업의 분포를 보여줍니다.

'숟가락'과 같은 일부 개체에는 여러 동작만 있습니다.

객체 자체는 아니지만 '악수', '하이파이브' 동작에 '손'을 객체로 포함시켰습니다.

작업 클래스당 비디오: 평균적으로 클래스 작업당 26.11개의 sequence와 개체당 45.19개의 sequence가 ​​있습니다.

동영상 길이: 45개의 액션 클래스에 대한 평균 동영상 길이를 보여줍니다.

'설탕 넣기' 및 '지갑 열기'와 같은 일부 작업 클래스는 평균 1초의 짧은 원자 이동을 포함하는 반면 '공개 서한'과 같은 다른 작업 클래스는 실행하는 데 더 많은 시간이 필요합니다.

파악: 우리는 가장 자주 연구된 것을 포함하여 34개의 다른 파악을 확인했습니다. 객체, 손 포즈 및 동작 간의 상관 관계에 대한 몇 가지 예를 보여줍니다.

시점: 손 시점 별 frame 분포를 보여줍니다.

시점을 카메라 방향과 손바닥 사이의 각도로 정의합니다.

데이터 세트는 3인칭 시점에서 일반적인 관점보다 자가 폐쇄되기 쉬운 관점을 제시합니다.

손 폐색: 동작 클래스당 보이는 손 관절의 평균 수를 보여줍니다.

대부분의 동작은 21개 관절 중 평균 10개 관절에서 높은 수준의 폐색을 나타냅니다.

개체 포즈: '우유 병', '소금', '주스 상자' 및 '액체 비누'의 10가지 동작이 포함된 개체에 대해 6D 개체 포즈 및 메시 모델이 제공됩니다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

동작 인식

데이터의 특성이 RGB-D이고 손 포즈가 있기 때문에 두 가지 RGB 동작 인식 방법도 평가하지만 RGB-D 및 포즈 기반 동작 인식 접근 방식에 중점을 둡니다. RGB-D 동작 인식의 대부분의 이전 작업은 손 대신 전신 포즈를 포함하고 일부는 손 동작에 맞게 조정되지 않았을 수 있습니다.

RGB 동작 인식의 현재 최첨단 기술이 데이터 세트에서 어떻게 수행되는지 평가하기 위해 하나의 기준선으로 시작합니다. 이를 위해 가장 성공적인 RGB 동작 인식 접근 방식인 ConvNets를 사용하여 색상 및 동작 흐름에서 descriptors를 학습한다는 점을 고려하여 데이터 셋에서 미세 조정된 최근 2개 stream 아키텍처를 평가합니다.

깊이 양식에 대해 먼저 두 가지 local 깊이 설명자 접근 방식인 HOG2 및 HON4D를 평가합니다. Global-scene depth descriptor로서 ConvNets를 사용하여 인체 포즈의 여러 합성된 깊이 뷰에서 view 불변 기능을 학습하는 접근 방식을 평가합니다. 포즈 기반 동작 인식 방법으로 평가를 따릅니다. 주요 기준으로 LSTM 모듈을 사용하여 순환 신경망을 구현했습니다.

시간 종속성을 학습하는 방법에 대해 HBRNN, Gram Matrix 및 TF를 평가합니다.

HBRNN은 신체 포즈에서 특징을 학습하도록 설계된 계층적 레이어가 있는 양방향 순환 신경망으로 구성됩니다.

Gram Matrix는 현재 바디 포즈에 가장 성능이 좋은 방법이며 Gram Matrix를 사용하여 동작의 역학을 학습합니다.

TF는 결정 포리스트를 사용하여 차별적인 정적 포즈와 포즈 간 전환을 모두 학습합니다.

결론적으로 모든 데이터 고려하여 기능을 학습하기 위해 반복 알고리즘을 사용하여 공동으로 학습하는 하이브리드 접근 방식을 평가합니다.

도표, 스크린샷, 그래프, 텍스트이(가) 표시된 사진

자동 생성된 설명

(a) t-SNE 데이터 셋에 포함된 손 포즈의 시각화.

각 색 점은 완전한 손 자세를 나타내고 각 궤적은 동작 순서를 나타냅니다.

(b) 물체, 잡기 및 동작 간의 상관 관계.

표시된 포즈는 특정 클래스의 모든 동작 시퀀스에 대한 평균 포즈입니다.

하나의 물체는 수행된 작업(예: '주스 상자' 및 '우유 병')에 따라 연결된 여러 개의 잡기를 가질 수 있으며 하나의 잡기는 연결된 여러 작업을 가질 수 있습니다(예: '뿌리기' 및 '깨끗한 유리잔'에 있는 측면 잡기).

(c) 손 동작 클래스당 동작 인스턴스 수.

(c) 손 동작 클래스당 각 비디오의 평균 프레임 수.

데이터 셋에는 원자적이고 시간적으로 더 복잡한 작업 클래스가 모두 포함되어 있습니다.

(d) 방향 사이의 각도로 정의되는 손 관점의 분포

*T분포의 의미*

*t-분포는 표준 정규분포와 유사하게 0을 중심으로 좌우 대칭 형태를 이루며 표준 정규분포보다 평평하고 긴 꼬리 형태를 가집니다.*

*즉, 양쪽 꼬리 형태가 두터운 형태를 가집니다. 이와 같은 형태를 가지는 이유는 표준 정규분포보다 분산이 크기 때문입니다.*

*정규분포는 평균과 분산을 통하여 그 형태가 달라지게 됩니다. 반면 t-분포는 자유도에 따라 다른 모양을 나타냅니다. 이 점은 카이 제곱 분포와 유사하다고 말할 수 있습니다. 자유도 = 표본의 수 -1로 정의되며 자세한 의미는 위에서 언급한 상세 내용 링크에서 확인할 수 있습니다. 이 자유도가 점점 더 커질수록 표준 정규분포에 가까워지고 대개 자유도가 30이 넘으면 표준 정규분포와 가까워지기 때문에 표본이 30이 넘어가게 되면 정규 분포를 사용하고 표본이 30 보다 작으면 t-분포를 사용하는 것이 일반적인 사용법입니다.*

*SNE의 의미와 기본적 활용 방법*

*t-distributed stochastic neighbor embedding 소위 t-SNE이라고 불리는 방법은 높은 차원의 복잡한 데이터를 2차원에 차원 축소하는 방법입니다. 낮은 차원 공간의 시각화에 주로 사용하며 차원 축소할 때는 비슷한 구조끼리 데이터를 정리한 상태이므로 데이터 구조를 이해하는 데 도움을 줍니다.*

*t-SNE는 manifold 학습의 하나로 복잡한 데이터의 시각화가 목적입니다. 높은 차원의 데이터를 2차원 또는 3차원으로 축소시켜 시각화 합니다.*

*t-SNE를 사용하면 높은 차원 공간에서 비슷한 데이터 구조는 낮은 차원 공간에서 가깝게 대응하며, 비슷하지 않은 데이터 구조는 멀리 떨어져 대응됩니다.*

*위 그림을 보면 왼쪽의 스위스 롤과 같은 3차원 데이터를 오른쪽과 같이 2차원 데이터로 차원을 축소해 볼 수 있습니다. 이 때 사용되는 방법이 t-SNE입니다.*

*t-SNE에 대한 수식적 의미*

*PCA는 Feature Extraction의 방법이고 linear한 방법을 사용합니다.*

*PCA는 원본 데이터를 저차원으로 linear mapping 합니다. 이 방법으로 저차원에 표현되는 데이터의 variance가 최대화 됩니다.*

*기본적인 방법은 공분산 행렬에서 고유벡터를 계산하는 것입니다.*

*가장 큰 고유 값을 가지는 고유벡터를 principal component로 생각하고 새로운 feature를 생성하는 데 사용합니다.*

*위 방법을 이용하여 PCA는 입력 받은 데이터 들의 feature를 결합합니다.*

*feature들을 결합할 때, 가장 중요하지 않은 feature들은 제거해가고 가장 중요한 feature들은 남깁니다.*

*새로 생성된 feature들은 기존의 feature들과 독립적입니다. 즉 기존 feature들의 단순 선형 결합으로 만들어진 것은 아닙니다.*

*t-SNE는 비선형적인 방법의 차원 축소 방법이고 특히 고차원의 데이터 셋을 시각화하는 것에 성능이 좋습니다.*

*t-SNE는 다양한 분야에서 시각화 하는 데 사용되고 있습니다.*

*t-SNE 알고리즘은 고차원 공간에서의 점들의 유사성과 그에 해당하는 저차원 공간에서의 점들의 유사성을 계산합니다.*

*점들의 유사도는 A를 중심으로 한 정규 분포에서 확률 밀도에 비례하여 이웃을 선택하면 포인트 A가 포인트 B를 이웃으로 선택한다는 조건부 확률로 계산됩니다.*

*그리고 저 차원 공간에서 데이터 요소를 완벽하게 표현하기 위해 고차원 및 저 차원 공간에서 이러한 조건부 확률 간의 차이를 최소화하려고 시도합니다.*

*조건부 확률의 차이의 합을 최소화하기 위해 t-SNE는 gradient descent 방식을 사용하여 전체 데이터 포인트의 KL-divergence 합계를 최소화합니다.*

*Kullback-Leibler divergence 또는 KL divergence는 한 확률 분포가 두 번째 예상 확률 분포와 어떻게 다른지 측정하는 척도입니다.*

*정리하면 t-SNE는 두가지 분포의 KL divergence를 최소화 합니다.*

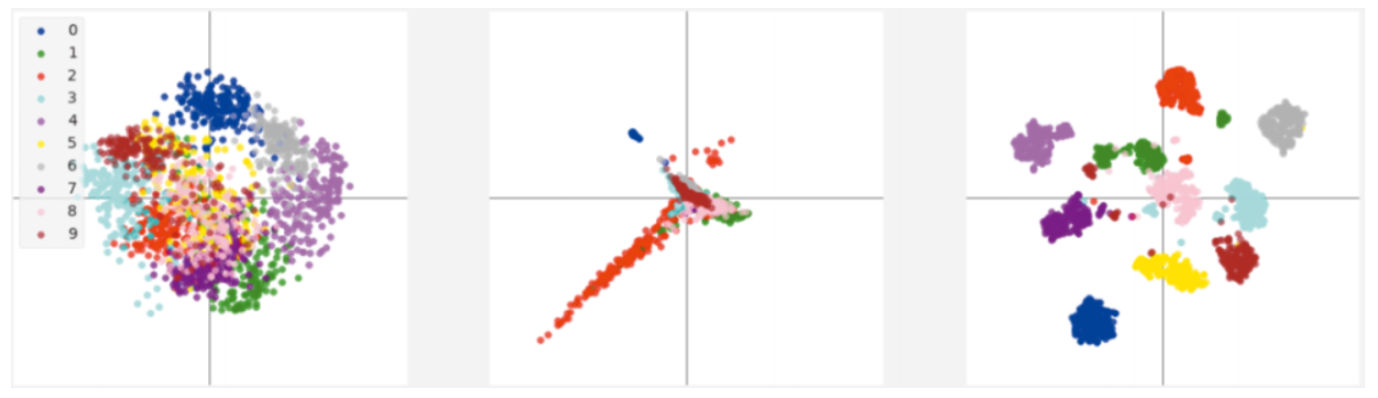
*입력 객체(고차원)들의 쌍으로 이루어진 유사성을 측정하는 분포*

*저 차원 점들의 쌍으로 유사성을 측정하는 분포*

*t-SNE는 다차원 데이터를 보다 낮은 차원 공간으로 매핑하고, 다수의 특징을 갖는 데이터 포인트의 유사성을 기반으로 점들의 클러스터를 식별함으로써 데이터에서 패턴을 발견합니다.*

*하지만 t-SNE 과정이 끝나면 input feature를 확인하기가 어렵습니다.*

*따라서 t-SNE는 주로 시각화 툴로 사용 됩니다.*



손 포즈 추정

최첨단 손 포즈 추정을 평가하기 위해 ConvNet을 사용합니다.

해석하기 쉽고 교차 벤치마크 평가에서 좋은 성능을 제공하는 것으로 나타났기 때문에 이 접근 방식을 선택합니다.

선택한 방법은 추적 실패 시 초기화 및 수동 복구가 필요하지 않은 프레임 단위로 작동하는 차별적인 접근 방식입니다.

벤치마크 평가 결과

- 동작 인식

손 자세가 주어진다고 가정합니다.

즉, 자기 센서와 역운동학을 사용하여 얻은 손 자세 주석을 사용합니다.

동작 인식을 위한 센서의 도움 없이 추정된 손 포즈의 사용을 평가합니다.

전신 포즈 동작 인식의 일반적인 관행에 따라 관절 쌍 사이의 거리가 동일하도록 포즈를 정규화하고 손목을 좌표 중심으로 정의하여 의인화 및 시점 차이를 보상합니다.

LSTM

우리는 단순하지만 강력한 기준선인 장기 단기 메모리 모듈(LSTM)이 있는 반복 신경망으로 실험 평가를 시작합니다. 양방향 대신 일반적인 단방향 네트워크를 사용합니다.

뉴런 수를 100으로 설정하고 dropout 확률을 0.2로 설정했습니다.

TensorFlow와 Adam optimizer를 사용합니다.

교육 및 테스트 프로토콜: 두 가지 프로토콜을 실험

첫 번째 프로토콜은 훈련을 위해 데이터의 다른 파티션을 사용하고 테스트를 위해 나머지 데이터를 사용하는 것으로 구성되며 시퀀스 수준에서 1:3, 1:1 및 3:1의 세 가지 다른 훈련:테스트 비율을 시도

두 번째 프로토콜은 6겹 'leave-one-person-out' 교차 검증입니다.

각 폴드에서 3:1 설정과 유사한 훈련/테스트 비율을 갖는다는 점을 고려할 때 교차인 프로토콜을 따르는 것이 최악의 결과를 낳는다는 것을 알 수 있습니다.

이는 피험자 간의 손 동작 스타일의 차이로 설명할 수 있습니다.

이 논문의 나머지 부분에서는 훈련을 위한 600개의 작업 sequence와 테스트를 위한 575개의 작업 sequence로 1:1 설정을 사용하여 실험을 수행합니다.

결과 토론: 하위 집합 동작에 대한 범주별 인식 정확도를 보여주고 동작 혼동 행렬을 표시합니다.

'숟가락 뿌리기', '티백 넣기', '주스 붓기'와 같은 일부 동작은 쉽게 식별할 수 있는 반면, '지갑 열기' 및 '계산기 사용'과 같은 동작은 일반적으로 손 포즈가 유사하지 않고 더 미묘하기 때문에 혼동되기 쉽습니다.

동작 인식 성능에 대한 각 손가락 동작의 기여도를 보여주며 검지가 가장 유익한 손가락임을 발견했습니다.

엄지와 검지 포즈를 결합하면 정확도가 높아집니다. 대부분의 잡기가 이 두 손가락으로 설명되기 때문일 수 있습니다. 손가락 끝만으로도 관절이 가장 높고 손 자세를 '설명'할 수 있기 때문에 높은 정보 소스이기도 합니다.

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

다양한 교육/테스트 프로토콜에 대한 동작 인식 결과

최첨단 평가

다양한 데이터 형식의 최신 접근 방식에 대한 결과는 다음 표에서 보여줍니다.

Two-stream 방법은 공간 및 시간 큐를 결합할 때 잘 수행됨을 관찰합니다.

깊이 방법은 나머지 방법보다 성능이 약간 떨어지는 경향이 있으며, 이는 개체 신호나 손 포즈를 완전히 캡처할 수 없음을 나타냅니다.

Novel View의 경우 신체의 여러 합성 보기에 대해 훈련된 네트워크에서 심층 기능을 추출했는데, 이는 손 포즈에 잘 일반화되지 않을 수 있으며 데이터 셋의 미세 조정이 도움되지 않습니다.

모든 접근 방식에서 논문은 손 포즈를 사용하는 것이 최고의 성능을 달성하는 것을 관찰했습니다. Gram Matrix와 Lie 그룹이 특히 잘 수행되어 신체 포즈 동작 인식에서 보고된 결과와 일치합니다.

아래 그림에서 우리는 가장 대표적인 몇 가지 방법을 선택하고 그 성능을 자세히 분석합니다.

논문은 특히 top k 동작 가설을 검색할 때 동작 인식에 손 포즈를 사용하는 이점을 보여주는 포즈 방법 Gram Matrix가 대부분의 측정에서 나머지를 능가한다는 것을 관찰했습니다.

객체가 크고 동작이 많은 동작을 포함하지 않는 일부 범주(예: '계산기 사용' 및 '종이 읽기')에서 Two-stream이 나머지 방법보다 성능이 우수함을 관찰할 수 있습니다. 이 좋은 성능은 큰 이미지 인식 데이터 세트에 대한 공간 네트워크의 사전 훈련 때문일 수 있습니다.

논문에서는 예측에 의해 주어진 top k 가설을 분석하고 예측된 동작에 조작 중인 개체가 포함되어 있는지 살펴봄으로써 네트워크가 개체를 올바르게 인식하지만 시간 역학을 캡처하지 못한다는 것을 제안합니다.

*Two-stream Approach*

*spatial info와 temporal info을 별개의 stream에서 각각 학습한 뒤 마지막에 fusion하는 모델이다.*

*텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명*

*여기서 optical flow는 다음 프레임으로 넘어갈 때 각 픽셀이 어디로 이동했는지를 벡터로 나타낸 것이라고 간단히 생각하면 된다. 이를 모델 입력에 사용되는 프레임마다 구하여 temporal stream convNet에 태운다. 논문에서는 수평 방향(u)과 수직 방향(v)로 나눠 계산한다.*

*optical flow를 계산할 때 기본 가정이 있는데, 다음 프레임으로 갈 때*

*Brightness Consistency: 각 object의 같은 지점은 밝기가 거의 같게 유지된다.*

*Temporal Persistence: 각 object는 먼 거리를 이동하지 않는다.*

*Spatical Coherence: 인접한 점들은 거의 같은 방향으로 이동한다.*

*도표, 라인, 평면도, 폰트이(가) 표시된 사진

자동 생성된 설명*

*모델 설명 : Spatial stream은 이미지 한 장을 사용하므로 2D convNet을 사용한다.*

*Temporal stream은 image sequence를 입력으로 받는다.*

*Optical flow stacking : L frame에 대해 각 방향 u, v가 있으므로 2L개의 channel이 있다.*

*Trajectory stacking : optical flow vector를 그냥 쌓는 대신 flow를 따라가면서 sampling한다. 채널 수는 같게 유지된다.*

*Two-Stream Inflated 3D convNet*

*Two-stream 방법에서, spatial stream을 3차원으로 확장한다.*

*Temporal stream은 미리 계산된 optical flow를 입력으로 하고 early fusion 대신 3D conv 방식을 사용한다.*

*텍스트, 폰트, 도표, 라인이(가) 표시된 사진

자동 생성된 설명*

*Architecture는 GoogLeNet에 기초하지만 바뀐 것은 2D가 아니라 3D로 확장한 것이다.*

Hand pose 대 depth 대 color : 각 데이터 양식의 기여도를 세분화하여 JOULE 접근 방식을 사용하여 한 가지 추가 실험을 수행했습니다.

아래 표에서 우리는 손 포즈 특징이 RGB 및 깊이 단서와 결합하여 성능을 높일 수 있지만 가장 차별적인 특징임을 보여줍니다. 이 결과는 손 포즈가 RGB 및 깊이 기능에 대한 보완 정보를 캡처함을 시사합니다.

Object pose : 논문에서는 개체 포즈에 주석이 달린 동작의 하위 집합을 사용하여 동작 인식을 위한 기능으로 개체 포즈를 사용하여 추가 실험을 수행했습니다.

논문은 LSTM 기준선을 시퀀스의 절반과 세 가지 다른 입력(hand pose, object pose 및 둘 다 결합)을 사용하여 훈련했습니다.

도표, 그래프, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

(a) 클래스의 하위 집합에 대한 일부 대표적인 방법의 클래스 정확도.

(b) Top-k 동작 정확도: true 동작 레이블은 top-k 동작 예측 가설에 있습니다.

(c) Top-k 개체 정확도: 조작된 개체는 top-k 동작 예측 가설에 있습니다.

(d) 동작 인식에 대한 다섯 손가락 각각, 이들의 조합 및 손가락 끝의 영향.

Hand pose estimation

Training with objects 대 no objects : 논문은 실험을 설계하는 동안 제기된 한 가지 질문은 손의 동적 동작을 실험하기 위해 실측 정확도에 가까운 손 포즈에 실제로 주석을 달아야 하는지 여부였습니다. 우리는 액션 분할에서와 같이 데이터를 분할하는 두 가지 방법으로 손 동작 데이터 세트의 손 포즈를 추정하여 이 질문에 답하려고 합니다.

결과는 트레이닝 세트에 hand - object 이미지를 갖는 것이 최첨단 hand pose 추정기를 훈련하는 데 중요하다는 것을 시사합니다. 모양은 추정자가 미리 볼 필요가 있습니다.

이를 확인하기 위해 두 가지 추가 실험을 수행했습니다. 교차 주제(사용자의 절반은 교육에, 절반은 테스트에, 두 분할에서 모든 개체가 표시됨) 및 교차 개체(교육에 있는 개체의 절반과 테스트에 절반, 모든 개체가 두 분할 모두에서 볼 수 있음).

논문은 네트워크가 보이지 않는 주제로 일반화할 수 있지만 보이지 않는 물체에 대해서는 그렇게 하기 위해 고군분투하고 있음을 관찰합니다. .

이것은 개체와 상호 작용하는 주석이 달린 손 포즈의 필요성과 손 포즈 커뮤니티에서 우리의 데이터 세트가 관심을 가질 수 있는 이유를 보여줍니다.

논문은 제안된 데이터 세트에서 손 포즈 추정의 일부 정성적 결과를 보여주고 완벽하지는 않지만 동작 인식에 충분하다는 것을 관찰합니다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

*HOG(Histogram of Oriented Gradient)*

*HOG는 보행자 검출을 위해 만들어진 descriptor이다. HOG는 이미 경계의 기울기 벡터 크기와 방향을 히스토그램으로 나타내 계산한다. 이미지 안의 픽셀들의 방향의 변화를 나타낸다.*

*도표, 라인, 폰트, 원이(가) 표시된 사진

자동 생성된 설명*

*Histogram of Gradient (Keypoint Descriptor)는 픽셀 값의 변화량을 나타내는 척도로 주로 이미지의 texture를 표현한다.*

*위 그림은 가우시안 필터(σ = 0.1)를 사용해서 weighted magnitude를 표현한 그림을 나타낸다*

*가우시안 필터는 커널의 weights가 중심점에서부터 가우시안 분포를 따르는 것이다.*

*다채로움, 스크린샷이(가) 표시된 사진

자동 생성된 설명*

*σ=2 인 30x30 Gaussian kernel(왼쪽 그림)*

*σ=5 인 30x30 Gaussian kernel(오른쪽 그림)*

*텍스트, 폰트, 스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명*

*1번 픽셀은 중심점 (c, r)으로부터 한 칸 위쪽인 (c, r-1)인 곳에 위치한다*

*2번 픽셀은 중심점 (c, r)으로부터 한 칸 왼쪽인 (c-1, r)인 곳에 위치한다*

*3번 픽셀은 중심점 (c, r)으로부터 한 칸 아래쪽인 (c, r+1)인 곳에 위치한다*

*4번 픽셀은 중심점 (c, r)으로부터 한 칸 오른쪽인(c+1, r)인 곳에 위치한다*

*최종적으로 Gradient orientation & Gradient magnitude를 이용해 HOG를 구하는 방법*

*도표, 디자인이(가) 표시된 사진

자동 생성된 설명*

*첫번째 이미지를 살펴보면 object(사람)의 경계면을 따라 gradient orientation이 다르다는 것을 확인 할 수 있다.*

*3x3 필터가 슬라이딩 하면서 각 픽셀마다 gradient orientation을 구한다.*

*두번째 그림의 격자는 수는 히스토그램의 수를 의미이다.*

*위 예제는 3x3 local image를 하나의 히스토그램으로 생성한다.*

*일반적으로 8x8 local image에 대해서 생성한다.*

*각 히스토그램을 Concatenation하여 하나의 벡터로 만든다.*

*그 벡터는 Keypoint Descriptor로 이미지 분석의 설명변수로 사용된다.*

Hand pose estimation and action recognition : 논문은 '현재 손 포즈 추정이 손 동작 인식에 얼마나 좋은가?'라는 핵심 질문에 답하려고 합니다.

hand pose label을 테스트 세트의 추정 label로 교체하여 손 동작 인식 결과를 보여줍니다.

hand pose error를 두 배로 줄이면 동작 인식이 두 배 이상 향상된다는 것을 관찰했습니다. 테스트에서 hand pose label을 사용한 것과 추정된 레이블을 사용한 손 동작 인식의 차이는 6.67%입니다.

또한 이전 섹션에서 가장 성능이 좋은 두 가지 방법인 Lie 그룹과 Gram Matrix를 테스트했습니다.

Lie 그룹의 경우 69.22%의 정확도를 얻었지만 Gram Matrix의 경우 잡음 분포에 대한 강한 가정으로 인해 32.22%의 좋지 않은 결과를 얻었습니다.

반면에 논문의 LSTM 기준선은 잡음이 많은 손 포즈 추정이 있을 때 더 강력한 동작을 보여줍니다. 그림 논문에서 hand occlusion pose 추정 품질과 클래스 인식 정확도에 미치는 영향에 어떻게 영향을 미치는지 보여줍니다.

일부 클래스는 포즈 오류와 행동 정확도 저하 사이에 명확한 상관관계를 제시하지만 LSTM은 여전히 ​​시간 패턴에서 행동을 추론할 수 있기 때문에 허용 가능한 인식률을 얻을 수 있습니다.

더 많은 통찰력을 얻기 위해 손가락당 자세 오류를 분석했습니다: T: 12.45, I: 15.48, M: 18.08, R: 16.69, P: 18.95, 모두 mm 단위입니다. 엄지 관절과 검지 관절은 일반적으로 자기 중심적 설정에서 덜 가려지기 때문에 가장 낮은 추정 오류를 나타냅니다.

노이즈가 많은 손 포즈 추정을 가지고 있으면서도 여전히 좋은 동작 인식 성능을 얻을 수 있는 이유에 대한 설명이 될 수 있습니다.

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

상단: 추정된 손 포즈를 사용하는 LSTM 기준선에 대한 집단 행동 인식 정확도(정확도 포즈의 정확도는 검은색 삼각형으로 표시됨).

하단: 데이터 세트에서 손 동작에 대한 보이는 관절의 평균 수와 손 포즈 추정에 미치는 영향.

(b) LSTM 기준선에 대한 손 동작 혼동 행렬.

(c) 다른 손 포즈 추정 오류 임계 값에 대한 프레임의 백분율.

(d) 손 자세 추정에 대한 정성적 결과.

텍스트, 스크린샷, 폰트, 영수증이(가) 표시된 사진

자동 생성된 설명

평균 손 포즈 추정 오류, 다른 프로토콜 및 동작 인식에 미치는 영향에 대한 자기 포즈와 추정 사이의 모든 21개 관절에 대한 3D 거리.

결론

1인칭 설정에서 RGB-D 및 포즈 기반 손 동작 인식에 대한 새로운 벤치마크를 제안하고 실험적 평가를 제시했습니다. 이 벤치마크는 임시 작업 레이블과 전체 3D hand pose label을 모두 제공하고 데이터 셋의 일부에 추가로 6D 개체 포즈 레이블을 제공합니다. RGB-D 동작 인식과 3D hand pose 추정은 모두 비교적 새로운 분야이며, 두 가지 모두를 전체 인체와 유사하게 연관시키려는 첫 번째 시도입니다.

데이터 셋에서 여러 기준선을 평가했으며 hand pose 기능이 조작 동작을 인식하기 위한 풍부한 정보 소스라는 결론을 내렸습니다.

데이터 셋과 실험이 동작 인식, hand pose 추정, object pose 추정 및 공동 hand - object pose 추정과 같은 새로운 분야를 포함한 여러 분야에서 향후 작업을 장려할 수 있다고 믿습니다.