**리뷰 보고서**

*(First-Person View Hand Segmentation of Multi-Modal Hand Activity Video Dataset)*

**1. 소개**

손은 도구와 상호작용하며 증강, 가상 현실 or 인간과 컴퓨터 상호 작용과 같이 산업용 컴퓨터 비전에 매우 중요하며 많이 사용되고 있습니다. 비전 시스템에서 손을 인식하는 것은 사람과 장치 간의 상호 작용에 필요합니다. 손 위치, 자세, 제스처, 다중 시점 예측, 동작 분류 등을 비전 시스템으로 이해하는 것이 탐구되어있습니다. 이런 작업의 성능 정확도를 높이기 위해서는 배경에서 손을 분할합니다.

스크린샷, 국기, 종이접기이(가) 표시된 사진

자동 생성된 설명

배경에서 손을 분할하는 것이 어려운 이유

- 1인칭 시점에서 손가락 끝이 손등과 도구에 의해 가려지는 것

- 도구가 다양한 그립으로 고정된 것

- 도구 또는 개체의 모양이 무한한 것

RGB 손 분할 비디오 데이터셋을 생성하는 방법은 수동 픽셀 단위인 라벨링을 통하는 것이지만, PIP 분할의 시간으로 프레임 수가 증가하여 선형적으로 증가하여 확장성이 감소합니다. 결과적으로, 분할된 손 데이터셋을 생성하기 위해서는 효율적인 분할 방법을 개발해야 합니다.

라벨링 프로세스를 위해 손 온도를 활용할 것입니다. but LWIR 손의 열화상 픽셀은 주변 픽셀에서 잘못된 것을 포함할 수 있습니다. 이에 대한 해결방안은 cloud 작업을 활용하여 손의 세부 경계를 현지화하기 위해 AABB의 추적기를 제공하여 생성 작업을 완화시킵니다. 추적기는 손 모양 특징을 학습한 다음 손에 대한 OMBB를 추정합니다.

(OMBB -> AABB를 임의의 방향으로 회전한 박스)

AABB는 박스를 이루는 면의 normal vector들이 X, Y, Z축과 일치하는 모양이다. 모델을 이루는 다각형의 x, y, z 좌표의 최소 최대를 각 박스의 vertex로 생성한다. 따라서 회전함에 따라 크기가 계속 변화한다.

OBB는 AABB와 마찬가지로 박스를 이루는 세 면은 서로 수직이지만 해당 면의 normal vector가 X, Y, Z와 일치하지 않는 박스이다. X, Y, Z와 일치하지 않기 때문에 AABB보다 충돌 검출의 시간 복잡도가 높지만, 충돌 박스가 X, Y, Z 좌표를 기준으로 최대, 최소로 계속 변하는 AABB보다 항상 좀 더 fit한 bounding 박스를 만들 수 있다.

DNN을 최적화한다면 네트워크가 입력 소스에서 고유한 기능을 가져오지 못하여 오류가 발생할 수 있습니다. 극복하기 위해 여러 가지 방법들이 사용되고 있습니다.

LWIR의 장점에는 색상, 질감 및 조명 조건에 영향을 받지 않습니다. "hands using tools" sequence와 frames를 만듭니다. 새로운 데이터셋에는 세 가지의 다른 양식과 머리 장착형 카메라 IMU 정보를 포함합니다.

비디오 데이터셋을 사용해 DeepLabV3+와 3가지 양식 융합을 분석하고 5가지 최첨단 세분화 방법을 벤치마킹합니다. 3가지 양식은 LWIR, RGB, depth입니다.

이 세가지 양식의 기능은 손과 배경 간의 혼잡함을 방지합니다.

**논문에서의 주요 기여도**

1. LWIR, RGB, 깊이 및 IMU 정보를 포함하는 1인칭 시점의 대규모 액션 비디오를 수집합니다. 이 데이터 세트는 여러 양식을 사용하여 손 세분화 연구에 사용할 수 있습니다.

2. 우리는 사람이 도구를 들고 있을 때 픽셀 단위의 손 분할 기준 실측을 생성하기 위해 손 온도를 활용하여 분할 노력을 크게 줄일 수 있는 프레임워크를 개발합니다. 우리의 방법은 손 포즈 레이블이나 손 mesh 모델이 필요하지 않습니다.

3. 우리는 심층 신경망을 사용하여 손 세분화 작업을 위한 여러 양식의 효과를 분석하고 열(LWIR), RGB 및 깊이 양식을 융합하는 최적의 조합을 찾을 수 있습니다.

아래 표에 대한 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

HandNet 데이터셋 => Depth 포함

HOF 데이터셋, Hand - CNN => Both Hands 포함

EgoHands+, EYTH, EgoHands => Egocentric, Both Hands 포함

WorkingHands => Egocentric, Both Hands, Depth 포함

Ours => **Egocentric, Both Hands, Depth, LWIR, IMU** 포함

Ours에서 새로 만드신 데이터셋에는 다른 데이터셋에 포함되지 않은 것들을 모두 포함시켜 만든 데이터로 401,765 프레임을 구성합니다.

**2. 관련된 연구**

첫번째, 효율적인 주석 방법은 대규모 비디오 데이터셋에 대한 Label을 만드는 것입니다. 복잡한 경계를 추적할 수는 있지만 object mask로 데이터셋에 label을 지정하는 데 걸리는 시간 소비가 많습니다. 해결 방안은 정확성을 향상시키고 loop에 인간을 이용해 대화식으로 object의 polygon 주석을 생성하는 neural network를 가지게 합니다. 또한 pixel 단위의 모양 특징을 추출합니다. 이 논문에서의 주석 파이프라인은 색상과 텍스처에 대해 변하지 않고 레이블의 품질과 효율성을 높입니다.

두번째, 심층 신경망을 사용한 pixel 단위 hand 분할 데이터셋은 비디오에서 객체 분할하는 데 많이 사용되어 있습니다. 가장 많이 사용하는 구조는 고차원 이미지를 잠재 벡터에 투영하고 잠재 벡터를 클래스별 픽셀 공간으로 디코딩하는 encoder - decoder 구조입니다. 또한 다른 연구에서는 하드웨어 센서를 이용해 관절 위치를 예측하고 mesh model을 렌더링합니다. 하지만 하드웨어 센서는 RGB에서 사용할 수 있습니다. 연구는 손 포즈 추정 및 mesh model을 활용하여 프레임에서 손을 분할할 때 좋은 성능을 보여줍니다. 그러므로 이 연구에서는 심하게 가려지는 상황에서 손 동작 추정에 대해 높은 정확도를 생성하기 위해서는 hand pose network를 훈련할 필요가 없습니다.

세번째, Multi - model data는 color 및 texture 불변성 및 명시적 3D 정보 표현으로 인해 강력한 기능을 캡쳐하는데 사용됩니다. LWIR은 주변 온도가 다른 물체를 식별하는 능력이 뛰어납니다. LWIR 센서를 사용하여 손 온도를 임계 값으로 설정하고 손 분할 데이터셋 생성을 위한 공간 특징을 사용하여 유사한 열 복사 파장을 생략하는 배경을 제거합니다.

**3. 효율적인 multi - modal 손 분할**

**3-1 Mapping LWIR onto RGB-D**

LWIR 프레임으로 손 분할하기 위해 로 표시되는 LWIR 프레임과 으로 표시되는 열 Thermal Mask를 찾아 검색 공간을 줄인다. 이 식의 T는 LWIR 평면을 RGB 평면으로 변환하는 변환 함수이다.

폰트, 화이트, 서예, 타이포그래피이(가) 표시된 사진

자동 생성된 설명

의 공간 위치 (i, j)에 해당하는 대상 프레임의 픽셀에 대한 경계 값이다.

폰트, 텍스트, 화이트, 서예이(가) 표시된 사진

자동 생성된 설명

(손 온도 범위) => a - 최소값 b - 최대값 a,b 범위 밖은 0으로 정의

깊이의 map과 를 정렬하기 위해 와 깊이 카메라 간의 공간 관계를 찾아야 한다. 픽셀 공간에서 물체의 투영은 카메라 matrix와 물체점을 곱하면 파생이 된다.(LWIR 카메라의 경우 , 깊이 카메라의 경우 )

폰트, 텍스트, 타이포그래피, 화이트이(가) 표시된 사진

자동 생성된 설명

, 는 깊이 및 LWIR 카메라 픽셀 평면의 투영된 점이다.

와 는 깊이 및 LWIR 카메라 좌표의 객체 포인트이다.

는 scale vector이고, 는 카메라 공간의 깊이 값이다.

두 카메라 간의 공간적 관계는 R이 3D 회전 행렬이고, T가 변환 행렬인 아래 사진으로 정의한다.

폰트, 타이포그래피, 화이트, 상징이(가) 표시된 사진

자동 생성된 설명

위의 모든 식들을 조합하여 최종적으로 추출된 식은 다음과 같다.

폰트, 친필, 타이포그래피, 서예이(가) 표시된 사진

자동 생성된 설명

이 방정식을 풀어 깊이 평면을 RGV 평면으로 변환합니다. RGB의 깊이 frames는 API 문서를 정렬합니다.

두 카메라 사이의 서로 다른 해상도와 시야는 고유 및 외부 카메라 매개변수를 사용하여 조정됩니다.

정렬 후 그림 b에 묘사된 대로 인체 온도의 최소 및 최대를 설정하여 손 온도 임계 값을 지정합니다.

폰트, 화이트, 텍스트, 타이포그래피이(가) 표시된 사진

자동 생성된 설명 (손 온도 범위) => a - 최소값 b - 최대값

이 범위는 수동으로 캡쳐되며 주변 배경 및 손에 들고 있는 물체에서 손을 분할하는 순으로 사용합니다. 그 후, 정확한 범위를 생성하기 위해 깊이 maps에서 I를 중첩시켰습니다.

최종적으로, 열 mask를 필터링하여 분할된 손을 추출합니다.

스크린샷, 다채로움, 예술이(가) 표시된 사진

자동 생성된 설명

**3-2 레이블이 잘못하여 픽셀 제거 및 방향 식별**

손과 비슷한 온도를 가진 배경에서 잘못된 레이블이 지정된 픽셀을 관찰할 수 있습니다.

tracker가 손 모양 특징을 학습하는 순서

1. 잘못된 레이블이 지정된 픽셀을 제거하기 위해 'SiamMask'라는 추적 알고리즘을 사용하여 그림 3에 표시된 OMBB로 손을 표시합니다.

2. 및 AABB로 tracker를 훈련시킵니다.

3. AMT를 사용하여 수동 AABB생성을 crowdsource합니다.

4. AABB가 주어지면 추적기는 OMBB를 예측하고 frames을 통해 순차적으로 OMBB를 왼손 또는 오른손으로 분류합니다.

스크린샷, 라인, 도표, 종이접기이(가) 표시된 사진

자동 생성된 설명

여기서 OMBB는 및 OMBB를 교차하여 레이블이 잘못된 픽셀을 제거하는 데 사용됩니다.

SiamMask는 바이너리 세그먼트 작업으로 Loss를 Augmentation시켜 객체 추적을 위해 Full-convolutional Siamese 접근법을 적용하여 학습 과정을 개선시킨다.

학습된 SiamMask는 단일 경계 상자 초기값에만 의존을 받으며, 초당 35 프레임의 회전이 가능한 경계 상자를 생성한다. 단순하며, 추적이 가능하고, 세그먼트 까지 가능함에도 불구하고 빠른 속도를 보장한다.

**4. 실험**

손 분할 문제를 두 class 분할 작업으로 간주하고 예측 mask 생성을 위해 픽셀당 두 class, 배경 및 손 사이의 최대 확률을 플로팅합니다. 평가 지표로는 손(hIoU)과 배경(bIoU)의 IoU(Intersection over Union)를 사용합니다. 평균 IoU(mIoU)를 이 두 클래스 IoU의 평균으로 정의합니다. 테스트 경우 훈련 세트나 검증 세트에서 사용되지 않는 sequences로 구성된 수동 주석 레이블을 사용합니다.

IoU는 물체 검출 분야에서 얼마나 겹쳐져 있는지를 표시하는 지표이다. IoU는 주로 Object Detection에서 예측된 bounding box의 정확도를 평가하는 지표 중 하나로 사용됩니다. 일반적으로 임계 값을 지정하여 해당 임계 값을 초과하는 경우에는 해당 위치에 대한 예측을 유효한 것으로 간주합니다.

임계 값을 높이면 더 정밀하게 일치한 예측만을 맞는 것으로 간주하는 것이므로 전체적인 성능평가 수치는 하락할 것입니다.

텍스트, 도표, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

**4-1 데이터셋 개요**

데이터 세트는 401,765개의 프레임과 790개의 시퀀스로 구성되어 있습니다. 테스트 데이터셋을 만들기 위해서 수동 주석을 달고, 비디오 데이터셋에는 5개의 주제, 15개의 작업 및 23개의 도구를 포함합니다.

마지막으로 손의 픽셀 레이블에 주석을 달기 위해 를 사용하고 손의 방향과 손이 왼손인지 오른손인지 식별은 추적기를 사용합니다.

**4-3 데이터셋 생성의 효율성**

수동으로 레이블이 지정된 프레임으로 의 정확도를 평가합니다. hIoU(손)에서 0.849의 상당히 합리적인 정확도를 보여줍니다. 손이 아닌 영역이 손과 동일한 온도를 갖는 sequence에서 이 거짓 양성을 감소시키는 것을 알아낼 수 있습니다. 이러한 잘못된 레이블이 지정된 픽셀은 hIoU(손) 저하의 주요 원인입니다.

초반 AABB가 제공한 tracker를 사용하여 레이블이 잘못 지정한 픽셀을 제거시키고, 을 이용해 레이블을 지정하는 프레임의 효율성을 높일 수 있습니다.

픽셀 단위 세분화 라벨링을 위한 방법에는 4가지가 있습니다.

1. 주석자는 와 함께 PolyRNN++를 사용

2. 주석 작성자는 가 지정된 태블릿 펜으로 태블릿에 손에 레이블을 지정

3. 주석 작성자는 마스킹된 = ⊙ 이 지정된 태블릿 펜으로 태블릿에 손을 표시

4. 주석자는 의 상단 rgb에 AABB를 그립니다.

**속도 비교 결과**

our의 방법이 에서 AABB와 교차하기 때문에 PolyRNN++보다 24배 더 빠르다는 것을 알 수 있습니다.

세 번째 방법은 두 번째 방법보다 2배 빠릅니다.

PolyRNN++를 사용하는 것보다 6배 빠릅니다.

으로 손을 마스킹하면 주석의 영역이 상당히 좁아져 라벨링 시간이 단축됩니다. 주석 방법의 품질을 검증하기 위해 무작위로 136개 시퀀스를 샘플링하고, 13,792개 프레임에 수동으로 주석을 작성합니다. 이러한 수동 주석 프레임을 사용하여 AABB 레이블이 있거나 없는 의 IoU를 평가합니다.

**4-3 tracker의 성능**

이 데이터셋은 훈련을 위한 7,882개의 프레임이 있는 441개의 sequences와 검증을 위한 3,836개의 프레임이 있는 77개의 sequences로 나뉩니다.

tracker는 프레임과 프레임의 두 가지 방식으로 훈련됩니다.

아래 표를 보았을 때, 프레임이 다른 것보다 성능이 우수하다는 것을 알 수 있습니다.

은 보다 팔뚝을 더 많이 감지하는 경우가 있습니다. 이것은 가 있는 tracker가 가 있는 tracker보다 손목의 볼록한 모양을 찾는 데 더 민감하여 손의 방향이 더 좋고 OMBB가 더 타이트하다는 것을 말합니다.

스크린샷, 다채로움, 예술이(가) 표시된 사진

자동 생성된 설명

만약 tracker가 다음 프레임을 예측하지 못한다면 AABB로 초기화합니다.

또한, 두 손이 많이 겹친다면 추적기는 손을 추적할 수 없습니다. 이 경우에는 OMBB를 그려야 합니다.

결론적으로 our의 방법은 픽셀 단위로 레이블을 지정할 수 있으며, 부드러운 곡선을 나타낼 수 있습니다. 그러므로 PolyRNN++의 mIoU는 0.895[1]이며 이는 our 방법인 0.923만큼 정확하지 않습니다.

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

**4-4 Multi - modal sequence 분석**

{, , }의 모든 가능한 조합을 입력 양식으로 사용하여 7개의 절제 연구를 수행하여 손 분할에 대한 Multi - modal sequence {,,}의 효과를 알아냅니다.

이 어떤 기여를 하는지 7가지의 연구를 실행합니다.

모든 실험에서 random으로 샘플링된 50k 프레임을 사용하고, 40K 프레임은 훈련 데이터셋으로, 10K 프레임은 테스트 데이터셋으로 분할합니다.

DeepLabV3+를 사용하는 이유는 RGB만 사용하여 손 분할 벤치마크 실험에서 다른 방법보다 성능이 우수하기 때문입니다. 또한 매개 변수가 가장 적습니다.

모든 실험은 입력 양식의 수와 동일한 수의 인코더를 사용합니다. 실험에서 우리는 가 사용될 때 손실 감소와 성능 향상을 모두 관찰하여 네트워크가 더 나은 최소값을 찾도록 안내합니다.

를 포함하면 {, }}에 비해 hIOU 점수에서 손 세분화 성능이 5% 더 높습니다.

는 세 가지 양식에는 강력한 기능을 생성하고 서로의 약점을 보완하며 장점을 활용하는 보완적인 속성을 가지고 있습니다.

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

7가지의 실험 IOU 비교 결과

높은 값들은 IOU에서 더 좋고 낮은 값들은 parameter에서 더 좋습니다.

**4-5 손 세분화 벤치마크**

여러 양식을 사용하는 성능을 검증하기 위해 및 분할 네트워크[49]를 사용하고 및 를 입력 양식으로 공동 사용하는 5가지 최첨단 분할 방법[13, 13, 50, 51, 61]과 방법을 비교합니다. 섹션 4.4와 동일한 데이터 세트와 원본 논문에 나열된 hyper -parameters를 사용합니다.

우리는 RTFNet이 모든 방법 중에서 두 번째로 우수하다는 것을 알았습니다. 이는 가 프레임에서 손을 분할하는 데 가장 의미 있는 사전 지식을 제공한다는 것을 나타냅니다. 3가지 방식의 DeepLabV3+\*는 두 번째로 좋은 방법인 RTFnet보다 hIoU에서 4% 더 우수하고 매개변수는 30% 더 적습니다.

Image Segmentation

이미지의 유사한 영역 or 부분을 해당 class label로 그룹화하는 것이다.

전체 프로세스가 디지털이기 때문에, 아날로그 이미지를 pixel 형태로 표현할 수 있으므로, 부분(segment)를 구성하는 작업이 픽셀을 그룹화하는 작업과 동일하다.

Recognition외에도 localization을 작업할 수 있다.

Encoder : Input의 latent space representation을 encoding하고

Decoder : encoder로부터 encoding된 정보로부터 decoding하고 segment map을 형성한다.

Decoder로 나온 segment map은 이미지에서 각 개체의 위치를 나타내는 지도라고 할 수 있다.

Image Segmentation tasks는 전달하는 정보의 양과 유형에 따라 세 그룹으로 분류가 가능하다.

1. Semantic segmentation

- image의 pixel을 semantic으로 분류하는 것이다.

- 특정 클래스에 속하는 pixel은 단순히 해당 클래스로 분류한다.

- 위와 같은 특징 때문에 이미지에 동일한 클래스의 여러 instance가 밀접하여 단일로 그룹화된다는 단점이 있다.

결론적으로, 이미지에 대한 심층적인 세부 정보가 제공되지 않는다.

2. Instance segmentation

- 클래스가 아닌 Instance를 기반으로 픽셀을 범주로 분류한다.

- 분류된 영역이 속한 클래스에 대한 개념이 없지만 boundary를 기반으로 겹치거나,

매우 유사한 object region을 분리할 수 있다.

- 만약 군중의 이미지가 있다면 각각의 사람을 모두 분리할 수 있지만, 각 영역/객체가 어떤 instance인지는 예측할 수 없다.

3. Panoptic segmentation

- Object의 각 Instance를 분리하고 Object의 identity을 예측하는 semantic segmentation과 instance segmentation의 조합이다.

- 주변 환경에 대한 방대한 양의 정보를 캡처해야 하는 환경에서 적용 가능성이 있다.

**5. 결론**

세 가지 다른 이미지 양식과 픽셀 현명한 손 및 동작 레이블이 있는 1인칭 시점 이미지의 IMU 정보로 풍부한 sequence를 기록합니다.

손 분할을 위한 기존의 최첨단 방법과 비교할 때 4% 더 나은 hIoU를 달성한다는 것과 LWIR 및 RGB-D 프레임을 융합한 Multi - modal 데이터셋이 RGB-D 프레임만 사용하는 것보다 5% 더 나은 hand IOU 성능을 달성한다는 것을 알 수 있습니다.

만 사용하면 RGB 및 깊이와 같은 다른 양식을 사용하는 것보다 결과가 좋지 않습니다.

그 이유는 손의 열 신호가 다른 신체 부위와 공유되기 때문일 수 있습니다.

our의 방법은 수동으로 레이블이 지정된 프레임의 품질이 비슷한 PolyRNN++보다 24배 빠릅니다.

제한 사항에는 두 손이 많이 겹치면 추적기가 제대로 작동하지 않는다는 것입니다. 향후 개발은 손 물체 포즈 추정, 사람이 물체를 잡고 있을 때 물체 재구성, 손 동작 인식과 같은 보다 다양한 손 관련 작업을 위한 데이터셋을 개선해야합니다.