Hands

Overview

손의 모양과 움직임을 인식하는 능력은 다양한 기술 영역과 플랫폼에서 사용자 경험을 향상시키는 데 중요한 구성 요소가 될 수 있다. 예를 들어, 수화 이해와 손 제스처 제어의 기초를 형성할 수 있으며, 증강 현실에서 물리적 세계 위에 디지털 콘텐츠와 정보의 오버레이를 가능하게 할 수도 있다. 사람들에게 자연스럽게 다가오지만, 강력한 실시간 손 인식은 손이 종종 자신이나 서로를 가리고(예: 손가락/손바닥 폐색과 핸드 쉐이크) 높은 콘트라스트 패턴이 부족하기 때문에 확실히 도전적인 컴퓨터 비전 작업이다.

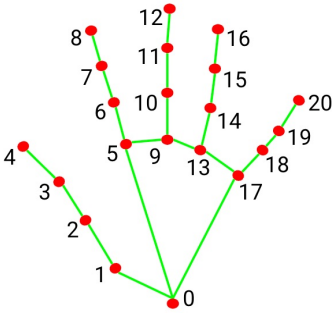
MediaPipe Hands는 충실도가 높은 손과 손가락 추적 솔루션입니다. 그것은 기계 학습(ML)을 사용하여 단 하나의 프레임에서 손의 21개의 3D 랜드마크를 추론합니다. 현재의 최첨단 접근 방식은 추론을 위해 주로 강력한 데스크톱 환경에 의존하는 반면, 우리의 방법은 휴대폰에서 실시간 성능을 달성하고 심지어 여러 손으로 확장합니다. 우리는 이 손 인식 기능을 더 넓은 연구 개발 커뮤니티에 제공하는 것이 창의적인 사용 사례의 출현을 초래하여 새로운 응용 프로그램과 새로운 연구 방법을 자극하기를 바랍니다.

ML Pipeline

MediaPipe Hands는 함께 작동하는 여러 모델로 구성된 ML 파이프라인을 사용합니다: 전체 이미지에서 작동하고 지향적인 손 경계 상자를 반환하는 손바닥 감지 모델. 손바닥 탐지기에 의해 정의된 잘린 이미지 영역에서 작동하고 고충실도 3D 손 키포인트를 반환하는 핸드 랜드마크 모델. 이 전략은 얼굴 랜드마크 모델과 함께 얼굴 탐지기를 사용하는 MediaPipe Face Mesh 솔루션에 사용된 것과 유사합니다.

손 랜드마크 모델에 정확하게 자른 손 이미지를 제공하면 데이터 증강(예: 회전, 번역 및 스케일)의 필요성이 크게 줄어들고 대신 네트워크가 대부분의 용량을 좌표 예측 정확도로 바칠 수 있습니다. 또한, 파이프라인에서 작물은 이전 프레임에서 식별된 손 랜드마크를 기반으로 생성될 수 있으며, 랜드마크 모델이 더 이상 손의 존재를 식별할 수 없는 경우에만 손의 국소화를 위해 호출된 손바닥 감지입니다.

참고: 그래프를 시각화하려면 그래프를 복사하여 MediaPipe Visualizer에 붙여넣으세요. 관련 하위 그래프를 시각화하는 방법에 대한 자세한 내용은 시각화 문서를 참조하십시오.



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Models

Palm Detection Model

초기 손 위치를 감지하기 위해, 우리는 MediaPipe Face Mesh의 얼굴 감지 모델과 유사한 방식으로 모바일 실시간 사용에 최적화된 단일 샷 탐지기 모델을 설계했습니다. 손을 감지하는 것은 확실히 복잡한 작업입니다: 우리의 라이트 모델과 전체 모델은 이미지 프레임에 비해 대규모 스팬(~20x)으로 다양한 손 크기로 작동해야 하며 가려지고 스스로 격리된 손을 감지할 수 있어야 합니다. 얼굴은 눈과 입 부위와 같이 콘트라스트 패턴이 높은 반면, 손에 그러한 특징이 없기 때문에 시각적 특징만으로도 안정적으로 감지하기가 비교적 어렵다. 대신, 팔, 신체 또는 사람의 특징과 같은 추가적인 맥락을 제공하는 것은 정확한 손 현지화를 돕는다.

우리의 방법은 다른 전략을 사용하여 위의 과제를 다룹니다. 첫째, 손바닥과 주먹과 같은 단단한 물체의 경계 상자를 추정하는 것이 관절 손가락으로 손을 감지하는 것보다 훨씬 간단하기 때문에 손 탐지기 대신 손바닥 탐지기를 훈련시킵니다. 또한, 손바닥은 더 작은 물체이기 때문에, 비최대 억제 알고리즘은 악수와 같은 양손 자기 포용 사례에서도 잘 작동한다. 게다가, 손바닥은 다른 종횡비를 무시하고 앵커 수를 3-5배 줄이는 정사각형 경계 상자(ML 용어의 앵커)를 사용하여 모델링할 수 있다. 둘째, 인코더 디코더 기능 추출기는 작은 물체에도 더 큰 장면 컨텍스트 인식에 사용됩니다(RetinaNet 접근 방식과 유사). 마지막으로, 우리는 대규모 분산으로 인한 많은 양의 앵커를 지원하기 위해 훈련 중 초점 손실을 최소화합니다.

위의 기술로, 우리는 손바닥 감지에서 평균 95.7%의 정밀도를 달성합니다. 정기적인 교차 엔트로피 손실과 디코더를 사용하면 86.22%의 기준선에 불과합니다.

Hand Landmark Model

전체 이미지에서 손바닥을 감지한 후 우리의 후속 핸드 랜드마크 모델은 회귀를 통해 감지된 손 영역 내부의 21개의 3D 핸드 너클 좌표의 정확한 키포인트 현지화를 수행합니다. 즉 직접 좌표 예측입니다. 이 모델은 일관된 내부 손 포즈 표현을 배우고 부분적으로 보이는 손과 자기 포함에도 견고하다.

지상 진실 데이터를 얻기 위해, 우리는 아래와 같이 21개의 3D 좌표로 ~30K개의 실제 이미지에 수동으로 주석을 달았습니다 (상응하는 좌표에 따라 존재하는 경우 이미지 깊이 맵에서 Z-값을 가져옵니다). 가능한 손 포즈를 더 잘 커버하고 손 기하학의 특성에 대한 추가 감독을 제공하기 위해, 우리는 또한 다양한 배경에서 고품질 합성 손 모델을 렌더링하고 해당 3D 좌표에 매핑합니다.

Solution APIs

Configuration Options

명명 스타일과 가용성은 플랫폼/언어마다 약간 다를 수 있습니다.

STATIC_IMAGE_MODE

False로 설정하면, 솔루션은 입력 이미지를 비디오 스트림으로 취급합니다. 그것은 첫 번째 입력 이미지에서 손을 감지하려고 시도할 것이며, 성공적인 탐지가 되면 손 랜드마크를 더욱 현지화합니다. 후속 이미지에서, 모든 max_num_hands 손이 감지되고 해당 손 랜드마크가 현지화되면, 손의 추적을 잃을 때까지 다른 탐지를 호출하지 않고 랜드마크를 추적하기만 하면 됩니다. 이것은 대기 시간을 줄이고 비디오 프레임 처리에 이상적입니다. True로 설정하면, 손 감지는 모든 입력 이미지에서 실행되며, 관련이 없을 수 있는 정적 이미지 배치를 처리하는 데 이상적입니다. 기본값은 false입니다.

MAX_NUM_HANDS

감지할 수 있는 최대 손 수. 기본값은 2입니다.

MODEL_COMPLEXITY

손 랜드마크 모델의 복잡성: 0 또는 1. 랜드마크 정확도와 추론 대기 시간은 일반적으로 모델 복잡성에 따라 올라간다. 기본값은 1입니다.

MIN_DETECTION_CONFIDENCE

탐지가 성공한 것으로 간주될 손 감지 모델의 최소 신뢰 값([0.0, 1.0]). 기본값은 0.5입니다.

MIN_TRACKING_CONFIDENCE:

손 랜드마크가 성공적으로 추적되는 것으로 간주될 랜드마크 추적 모델의 최소 신뢰 값([0.0, 1.0]) 또는 다음 입력 이미지에서 손 감지가 자동으로 호출됩니다. 더 높은 값으로 설정하면 더 높은 대기 시간을 희생시키면서 솔루션의 견고성을 높일 수 있습니다. Static_image_mode가 참이면 무시되며, 손 감지는 단순히 모든 이미지에서 실행됩니다. 기본값은 0.5입니다.

Output

명명 스타일은 플랫폼/언어마다 약간 다를 수 있습니다.

MULTI_HAND_LANDMARKS

감지/추적된 손의 컬렉션, 각 손은 21개의 손 랜드마크 목록으로 표시되며 각 랜드마크는 x, y 및 z로 구성됩니다. x와 y는 각각 이미지 너비와 높이에 의해 [0.0, 1.0]으로 정규화됩니다. z는 손목의 깊이가 원점인 랜드마크 깊이를 나타내며, 값이 작을수록 랜드마크가 카메라에 가까워집니다. Z의 크기는 x와 거의 같은 스케일을 사용한다.

MULTI_HAND_WORLD_LANDMARKS

감지/추적된 손 컬렉션, 각 손은 세계 좌표에서 21개의 손 랜드마크 목록으로 표시됩니다. 각 랜드마크는 x, y 및 z로 구성됩니다: 손의 대략적인 기하학적 중심에서 원점이 있는 미터의 실제 3D 좌표.

MULTI_HANDEDNESS

감지/추적된 손의 손 수집 (즉, 왼쪽 또는 오른손). 각 손은 라벨과 점수로 구성됩니다. 라벨은 "왼쪽" 또는 "오른쪽" 값의 문자열입니다. 점수는 예측된 핸드레드의 예상 확률이며 항상 0.5보다 크거나 같습니다 (그리고 반대 핸디네도에는 예상 확률은 1 - 점수입니다).

입력 이미지가 미러링된 경우, 즉 이미지가 수평으로 뒤집힌 전면/셀카 카메라로 촬영했다고 가정하면 수작성이 결정됩니다. 그렇지 않다면, 응용 프로그램에서 수작업 출력을 바꾸세요.