MediaPipe

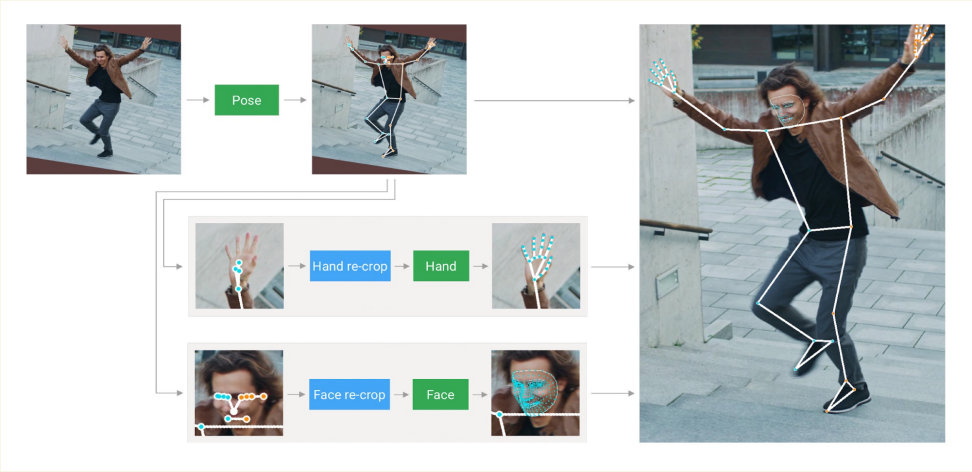
모바일 장치에서 실시간으로 동시 인간 포즈, 얼굴 랜드마크 및 손 추적에 대한 실시간 인식은 피트니스 및 스포츠 분석, 제스처 제어 및 수화 인식, 증강 현실 시도 및 효과와 같은 다양한 현대 생활 애플리케이션을 가능하게 할 수 있습니다. MediaPipe는 이미 이러한 작업에 대한 빠르고 정확하면서도 별도의 솔루션을 제공합니다. 그것들을 모두 실시간으로 의미론적으로 일관된 엔드투엔드 솔루션으로 결합하는 것은 여러 종속 신경망의 동시 추론을 요구하는 매우 어려운 문제이다.

< ML 파이프라인

MediaPipe Holistic 파이프라인은 포즈, 얼굴 및 손 구성 요소에 대한 별도의 모델을 통합하며, 각각은 특정 도메인에 최적화되어 있습니다. 그러나, 그들의 다른 전문 분야 때문에, 한 구성 요소에 대한 입력은 다른 구성 요소에 적합하지 않다. 예를 들어, 포즈 추정 모델은 더 낮은 고정 해상도 비디오 프레임(256x256)을 입력으로 사용합니다. 하지만 해당 이미지에서 손과 얼굴 영역을 잘라 각각의 모델로 전달한다면, 이미지 해상도는 정확한 조음에 비해 너무 낮을 것이다. 따라서, 우리는 지역 적절한 이미지 해상도를 사용하여 다른 지역을 처리하는 다단계 파이프라인으로 MediaPipe Holistic을 설계했습니다.

첫째, 우리는 BlazePose의 포즈 검출기와 후속 랜드마크 모델로 인간 포즈(그림 2의 상단)를 추정합니다. 그런 다음 추론된 포즈 랜드마크를 사용하여 각 손(2x)과 얼굴에 세 개의 관심 영역(ROI) 작물을 파생하고 ROI를 향상시키기 위해 재작물 모델을 사용합니다. 그런 다음 이러한 ROI에 전체 해상도 입력 프레임을 자르고 작업별 얼굴과 손 모델을 적용하여 해당 랜드마크를 추정합니다. 마지막으로, 우리는 모든 랜드마크를 포즈 모델의 랜드마크와 병합하여 전체 540개 이상의 랜드마크를 산출합니다.

얼굴과 손의 ROI 식별을 간소화하기 위해, 우리는 독립형 얼굴과 손 파이프라인에 사용하는 것과 유사한 추적 접근 방식을 활용합니다. 객체가 프레임 간에 크게 움직이지 않는다고 가정하고 이전 프레임의 추정치를 현재 프레임의 객체 영역에 대한 가이드로 사용합니다. 그러나, 빠른 움직임 중에 트래커는 대상을 잃을 수 있으며, 탐지기는 이미지에서 다시 현지화해야 합니다. MediaPipe Holistic은 빠른 움직임에 반응할 때 파이프라인의 응답 시간을 줄이기 전에 포즈 예측(모든 프레임)을 추가 ROI로 사용합니다. 이것은 또한 모델이 왼쪽과 오른손 또는 프레임에 있는 한 사람의 신체 부위 사이의 혼합을 방지함으로써 신체와 그 부분에 걸쳐 의미론적 일관성을 유지할 수 있게 해준다.



또한, 포즈 모델에 대한 입력 프레임의 해상도는 얼굴과 손의 결과 ROI가 여전히 너무 부정확하여 해당 영역의 재자리를 안내할 수 없으며, 가벼우우려면 정확한 입력 작물이 필요합니다. 이 정확도 격차를 좁히기 위해 우리는 공간 변압기의 역할을 하고 해당 모델의 추론 시간의 ~10%에 불과한 경량 얼굴과 손을 다시 자르기 모델을 사용합니다.

파이프라인은 전체적인 랜드마크 모듈의 전체적인 랜드마크 서브그래프를 사용하고 전용 전체론적 렌더러 서브그래프를 사용하여 렌더링하는 MediaPipe 그래프로 구현됩니다. 전체적인 랜드마크 서브그래프는 내부적으로 포즈 랜드마크 모듈, 핸드 랜드마크 모듈 및 얼굴 랜드마크 모듈을 사용합니다. 구현 세부 사항은 확인해 주세요.

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

모델

랜드마크 모델

MediaPipe Holistic은 각각 MediaPipe Pose, MediaPipe Face Mesh 및 MediaPipe Hands의 포즈, 얼굴 및 핸드 랜드마크 모델을 활용하여 총 543개의 랜드마크(33개의 포즈 랜드마크, 468개의 얼굴 랜드마크 및 손당 21개의 핸드 랜드마크)를 생성합니다.

손 재작업 모델

포즈 모델의 정확도가 충분히 낮아서 결과 손의 ROI가 여전히 너무 부정확한 경우 공간 변압기의 역할을 하고 핸드 모델 추론 시간의 ~10%에 불과한 추가 경량 핸드 리크롭 모델을 실행합니다.

솔루션 API

크로스 플랫폼 구성 옵션

명명 스타일과 가용성은 플랫폼/언어마다 약간 다를 수 있습니다.

STATIC_IMAGE_MODE

False로 설정하면, 솔루션은 입력 이미지를 비디오 스트림으로 취급합니다. 그것은 첫 번째 이미지에서 가장 눈에 띄는 사람을 감지하려고 노력할 것이며, 성공적인 탐지가 되면 포즈와 다른 랜드마크를 더욱 현지화합니다. 후속 이미지에서는 계산과 대기 시간을 줄이기 위해 트랙을 잃을 때까지 다른 탐지를 호출하지 않고 이러한 랜드마크를 추적합니다. True로 설정하면 사람 감지는 모든 입력 이미지를 실행하며, 정적, 관련이 없는 이미지 배치를 처리하는 데 이상적입니다. 기본값

MODEL_COMPLEXITY

포즈 랜드마크 모델의 복잡성: 0, 1 또는 2. 랜드마크 정확도와 추론 대기 시간은 일반적으로 모델 복잡성에 따라 올라간다. 기본값은 1입니다.

SMOOTH_LANDMARKS.

True로 설정하면, 솔루션 필터는 지터를 줄이기 위해 다른 입력 이미지에 랜드마크를 포즈를 취하지만 static_image_mode도 true로 설정하면 무시됩니다. 기본값은 true입니다.

ENABLE_SEGMENTATION

True로 설정하면, 포즈, 얼굴 및 손 랜드마크 외에도 솔루션은 세분화 마스크를 생성합니다. 기본값은 false입니다.

REFINE_FACE_LANDMARKS

눈과 입술 주위의 랜드마크 좌표를 더욱 다듬고 홍채 주변에 추가 랜드마크를 출력할지 여부. 기본값은 false입니다.

최소_검출_신뢰 MIN_DETECTION_CONFIDENCE

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

최소 추적_신뢰 MIN_TRACKING_CONFIDENCE

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

Output

명명 스타일은 플랫폼/언어마다 약간 다를 수 있습니다.

포즈_랜드마크 POSE_LANDMARKS

포즈 랜드마크 목록. 각 랜드마크는 다음과 같이 구성되어 있다:

X와 y: 랜드마크 좌표는 각각 이미지 너비와 높이에 의해 [0.0, 1.0]으로 정규화되었다.

Z: 현재 모델이 깊이를 예측하도록 완전히 훈련되지 않았기 때문에 폐기해야 하지만, 이것은 로드맵에 있습니다.

가시성: 이미지에서 랜드마크가 보일 가능성을 나타내는 [0.0, 1.0]의 값.

포즈_월드_랜드마크 POSE_WORLD_LANDMARKS

세계 좌표의 또 다른 포즈 랜드마크 목록. 각 랜드마크는 다음과 같이 구성되어 있다:

X, y 및 z: 엉덩이의 중심에 원점이 있는 미터의 실제 3D 좌표.

가시성: 해당 pose_landmarks에 정의된 것과 동일합니다.

FACE_LANDMARKS

468개의 얼굴 랜드마크 목록. 각 랜드마크는 x, y 및 z로 구성됩니다. x와 y는 각각 이미지 너비와 높이에 의해 [0.0, 1.0]으로 정규화됩니다. z는 머리 중앙의 깊이가 원점인 랜드마크 깊이를 나타내며, 값이 작을수록 랜드마크가 카메라에 더 가까워집니다. Z의 크기는 x와 거의 같은 스케일을 사용한다.

왼쪽_손_랜드마크

왼쪽에 있는 21개의 랜드마크 목록. 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.

오른쪽_손_랜드마크

Left_hand_landmarks와 같은 표현으로 오른쪽에 있는 21개의 손 랜드마크 목록.

세분화_마스크

Enable_segmentation이 true로 설정된 경우에만 예측되는 출력 세분화 마스크. 마스크는 입력 이미지와 너비와 높이가 동일하며, 1.0과 0.0이 각각 "인간"과 "배경" 픽셀의 높은 확실성을 나타내는 [0.0, 1.0]의 값을 포함합니다. 사용 세부 사항은 아래의 플랫폼별 사용 예제를 참조하십시오.