Đại học Quốc gia Thành phố Hồ Chí Minh

Trường Đại học Khoa học Tự Nhiên

Khoa Công nghệ thông tin


# Real-time Hand tracking and recognition with Mediapipe

**Người hướng dẫn**

PGS.TS Lê Hoàng Thái

TS. Nguyễn Ngọc Thảo

Thầy Lê Thanh Phong

**Thông tin thành viên**

Lê Văn Đông - 19127363

Trần Đông Ba - 19127334

Đặng Nguyễn Minh Quân - 19127523

Ngày 14 tháng 4 năm 2022

# Mục lục

# 1 Bảng phân công công việc

Nhóm chia công việc đồng đều cho từng thành viên. Bên dưới là mô tả sơ bộ về công việc của từng thành viên.

| Tên | Công việc | Mức độ hoàn thành |
|---|---|---|
| Lê Văn Đông | Soạn báo cáo, tìm hiểu phần ứng dụng của Mediapipe, tìm hiểu source code demo | 100% |
| Trần Đông Ba | Tìm hiểu phần Hand Landmark model và dataset. Làm powerpoint và tìm hiểu source code demo. | 100% |
| Đặng Nguyễn Minh Quân | Tìm hiểu phần Blaze Palm và cách thiết lập Pipeline. Làm powerpoint. | 100% |

Bảng 1: Bảng phân công công việc

# 2 Giới thiệu

Theo dõi bàn tay (hand tracking) là một trong những thành phần quan trọng để hỗ trợ tương tác và giao tiếp tự nhiên trong môi trường thực tế ảo hay thực tế tăng cường (AR/VR). Vấn đề này đã được nghiên cứu rất nhiều trong nhiều năm trở lại đây. Tuy nhiên, trong thực tế, để theo dõi bàn tay theo thời gian thực (real-time hand tracking), một là cần những phần cứng chuyên dụng, hoặc là các phương thức không đủ nhẹ để chạy được trên các thiết bị di động và bị giới hạn bởi bộ vi xử lý. Và trong phần báo cáo này, nhóm sẽ đề cập tới 1 giải pháp có thể khắc phục đến những vấn đề kể trên. Đây chính là nghiên cứu của Google, được xuất bản vào ngày 18/06/2020.

# 3 Kiến trúc

Giải pháp theo dõi bàn tay này sử dụng pipeline của mô hình học máy gồm 2 mô hình sau kết hợp với nhau:

1. **Palm Detector**: dùng để định vị lòng bàn tay từ một hình ảnh đầu vào thông qua một khung giới hạn bàn tay (an oriented hand bounding box).
2. **Hand Landmark Model**: từ khung giới hạn bàn tay trả về hình ảnh 2.5D của lòng bàn tay.

Việc trả về ô định vị lòng bàn tay đưa vào Hand Landmark Model sẽ làm giảm thời gian xác định cấu trúc bàn tay cũng như tăng hiệu suất và độ chính xác của mô hình.

## 3.1 BlazePalm

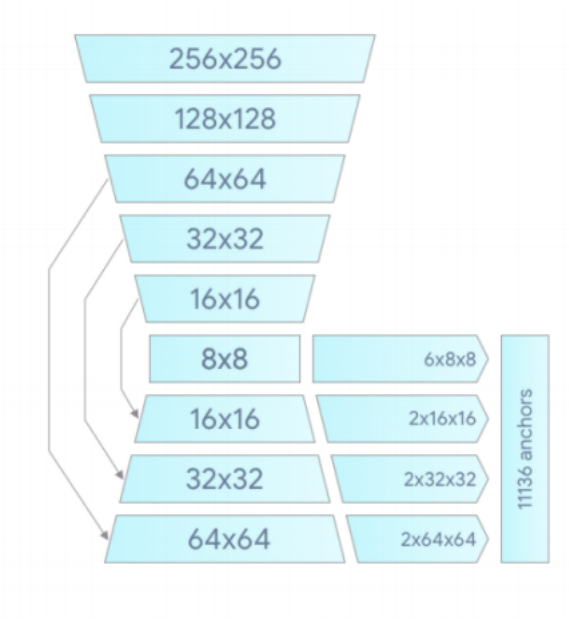
Để phát hiện vị trí ban đầu của bàn tay, nhóm nghiên cứu đã sử dụng mô hình máy dò tìm ảnh được tối ưu hóa trên các thiết bị đi dộng có sẵn trong MediaPipe. Và việc phát hiện bàn tay tưởng chừng như đơn giản, nhưng để thực hiện được thì mô hình phải hoạt động trên nhiều dữ liệu với các kích thước bàn tay khác nhau và khung bàn tay úp mở khác nhau. Khác với khuôn mặt có những đặc điểm tương phản khác nhau (xung quanh vùng mắt và miệng) thì bàn tay lại không có những đặc điểm kiểu đó, dẫn tới việc phát hiện đặc điểm bàn tay khó có độ tin cậy cao.

Giải pháp này sử dụng chiến lược khác để có thể phát hiện bàn tay tốt hơn.

Đầu tiên, huấn luyện mô hình phát hiện lòng bàn tay thay vì tay, vì ước lượng ô giới hạn lòng bàn tay hay nắm tay thì dễ hơn so với việc phát hiện bàn tay với các khớp tay. Ngoài ra, lòng bàn tay sẽ khiến thuật toán non-maximum suppression hoạt động tốt hơn. Hơn nữa, lòng bàn tay có thể được mô hình hóa chỉ bằng ô giới hạn, bỏ qua các tỷ lệ khung hình khác, từ đó làm giảm số lượng hình ảnh khác nhau của đối tượng đi 3 đến 5 lần.

Tiếp theo, sử dụng bộ trích xuất mã hóa - giải mã để có được bối cảnh lớn hơn của đối tượng (dù kích thước đối tượng nhỏ).

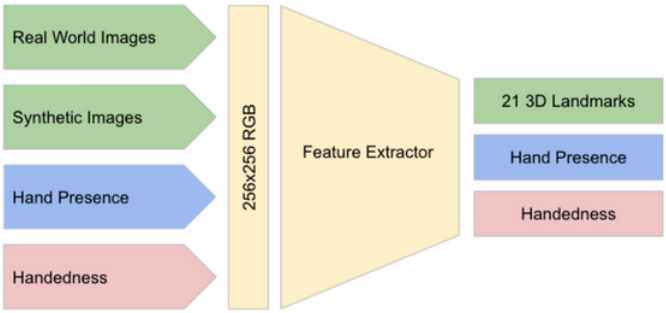
Cuối cùng, giảm thiểu focal loss trong khi huấn luyện để hỗ trợ cho các hình ảnh khác nhau của đối tượng với tỉ lệ phương sai lớn.




Hình 1: Kiến trúc của BlazePalm

## 3.2 Hand landmark model

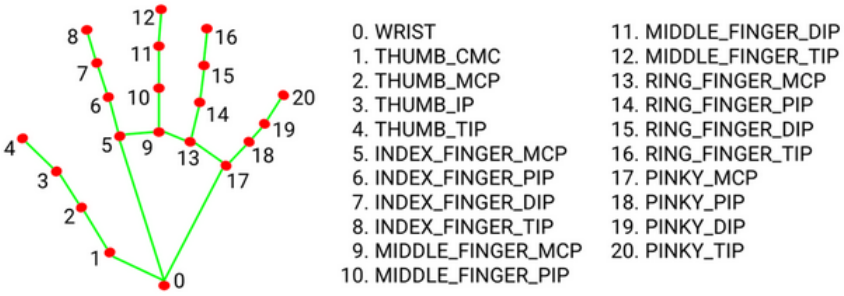
Sau khi chạy xong mô hình phát hiện lòng bàn tay trong toàn bộ hình ảnh đầu vào ta sẽ tiến hành tìm ra tọa độ của các điểm landmark, cụ thể trong trường hợp này là 21 điểm tọa độ trong không gian 2.5D. Mô hình có thể học được cách biểu diễn bàn tay thậm chí khi mà chỉ nhìn thấy 1 phần của bàn tay và khi bàn tay tự che đi một phần của chính nó. Mô hình có 3 đầu ra dùng chung bộ trích xuất đặc trưng, mỗi đầu ra sẽ được huấn luyện bởi tập dữ liệu tương ứng được đánh dấu bằng cùng màu:




Hình 2: Kiến trúc của mô hình hand landmark. Mô hình có 3 đầu ra dùng chung bộ trích xuất đặc trưng, mỗi đầu ra sẽ được huấn luyện bởi tập dữ liệu tương ứng được đánh dấu bằng cùng màu.

Mô hình có 3 đầu ra như sau:

- 21 điểm landmark bao gồm x, y và độ sâu tương ứng.
- Xác suất cho thấy khả năng bàn tay có xuất hiện trong ảnh.
- Bộ phân loại nhị phân để xác định xem đó là tay trái hay tay phải.




Hình 3: 21 hand landmarks

Những điểm tọa độ 2D được học từ cả tập dữ liệu trong thực tế và tập dữ liệu nhân tạo như được thảo luận ở phần bên dưới. Với điểm cổ tay chỉ được học từ tập dữ liệu nhân tạo. Để khắc phục lỗi tracking, chúng ta sẽ phát triển thêm một đầu ra khác để tạo ra xác suất của trường hợp bàn tay được căn chỉnh hợp lý thực sự có mặt trong vùng đã được trích xuất ở bước trước đó. Nếu xác suất này thấp hơn một ngưỡng nào đó thì bộ dò sẽ được kích hoạt để tracking lại. Việc phân loại bàn tay trái hay bàn tay phải cũng là 1 thuộc tính quan trọng khi nó được sử dụng trong các ứng dụng AR và VR. Thuộc tính này đặc biệt hữu dụng khi được áp dụng và các ứng dụng mà mỗi tay có mỗi vai trò khác nhau, thực hiện các nhiệm vụ khác nhau. Đó là lí do vì sao ta đội ngũ của Google lại phát triển thêm bộ phân loại nhị phân để dự đoán liệu bàn tay trong ảnh là bàn tay trái hay bàn tay phải.

Mục tiêu của dự án là phát triển mô hình có thể chạy trên các thiết bị di động có GPU nhưng về sau cũng được thiết kế thành 2 phiên bản dùng GPU và CPU để có thể chạy trên các thiết bị di động, các máy tính cá nhân. Tuy khả năng tính toán bị hạn chế nhưng Google nói rằng độ chính xác đạt được vẫn cao.

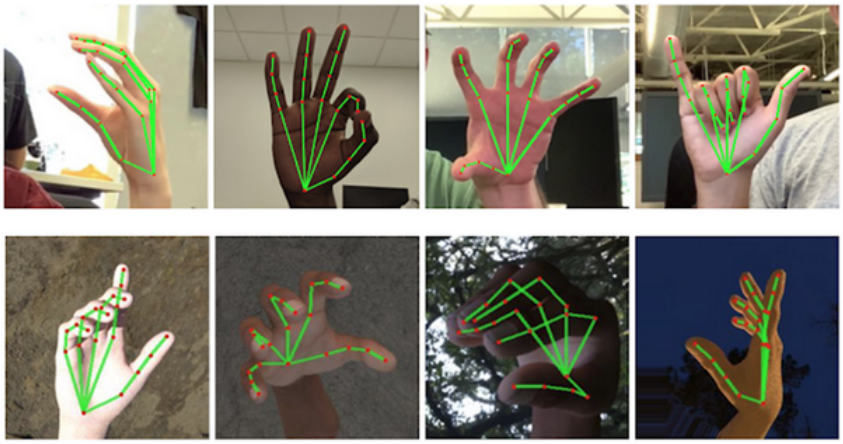
Hình 4: Kết quả đầu ra của toàn bộ quá trình. 2 bàn tay đã được phát hiện và định vị chính xác. Landmark cũng được phát hiện một cách chính xác và đầy đủ.

# 4 Dataset

Để có được ground truth data, nhóm đã tạo ra các tập dữ liệu sau để giải quyết nhiều khía cạnh của vấn đề:

- **In-the-wild dataset**: tập dữ liệu này chứa khoảng 6.000 hình ảnh đa dạng khác nhau: đa dạng về vị trí, điều kiện ánh sáng khác nhau và cả sự xuất hiện của bàn tay. Hạn chế của tập dữ liệu này là nó không chứa các khớp nối phức tạp của bàn tay.
- **In-house collected gesture dataset**: tập dữ liệu chứa khoảng 10.000 ảnh thể hiện đa dạng mọi góc độ khác nhau của mọi tư thế có thể của bàn tay. Hạn chế của tập dữ liệu này là nó chỉ được thu thập từ 30 người nên không thật sự đa dạng về nhiều loại bàn tay. Tập dữ liệu in-the-wild và in-house bổ sung qua lại cho nhau để cải thiện độ chất lượng của tập dữ liệu.
- **Synthetic dataset**: để bao phủ tốt hơn các tư thế tay có thể có và có khả năng giám sát sâu hơn, nhóm tạo ra mô hình bàn tay nhân tạo chất lượng cao trên các loại bàn tay khác nhau và ánh xạ nó thành các tọa độ 3D tương ứng. nhóm sử dụng một mô hình bàn tay 3D thương mại được gắn với 24 xương và bao gồm 36 hình dạng pha trộn, giúp kiểm soát các ngón tay và độ dày của lòng bàn tay. Mô hình cũng tạo ra 5 kết cấu với các tông màu da khác nhau. nhóm đã tạo chuỗi video về sự chuyển đổi giữa các tư thế tay và lấy mẫu 100K hình ảnh từ video. nhóm kết xuất mỗi tư thế

với một môi trường ánh sáng dải động cao ngẫu nhiên và ba camera khác nhau. Xem Hình 5 để biết các ví dụ.



Hình 5: Trên - Hình ảnh các bàn tay đã được canh chỉnh được truyền qua mạng cùng với ground truth annotation. Dưới - hình ảnh bàn tay được tổng hợp với ground truth annotation.

Đối với mô hình phát hiện lòng bàn tay chúng ta chỉ sử dụng in-the-wild dataset bởi vì tập dữ liệu này đủ để định vị được vị trí của các bàn tay và có được sự đa dạng cao nhất về sự xuất hiện của bàn tay. Tuy nhiên tất cả các tập dữ liệu này đều được sử dụng cho việc huấn luyện mô hình để landmark bàn tay. nhóm cũng đã chú thích tập dữ liệu hình ảnh thực tế bằng 21 landmark và sử dụng các khớp 3D được cho các hình ảnh tổng hợp. Đối với sự xuất hiện của bàn tay chúng ta sử dụng tập con của tập dữ liệu thực tế làm các mẫu dương và các mẫu trên khu vực không bao gồm chú thích bàn tay là các mẫu âm. Đối với phân loại tay phải hay tay trái, nhóm chú thích một tập hợp con các hình ảnh trong thế giới thực có độ thuận tay để cung cấp dữ liệu cần thiết.

# 5 Kết quả

Trong hand landmark model, các thí nghiệm cho thấy việc kết hợp hai bộ dữ liệu real-world và synthetic cho kết quả tốt nhất. Quan sát này là tiền đề để cho ta có thể mở rộng real-world dataset để đạt được hiệu quả tốt hơn. Xem Bảng 2.

Mục đích của chúng ta là làm sao đạt được real-time performance trên thiết bị di động. Khi làm nhiều thí nghiệm khác nhau với nhiều model thì "Full" model khá tốt để chúng

| Dataset | MSE normalized by palm size |
|---|---|
| Only real-world | 16.1% |
| Only synthetic | 25.7% |
| Combined | 13.4% |

Bảng 2: Kết quả việc train trên nhiều dataset khác nhau

ta có thể đánh đổi giữa chất lượng và tốc độ. Việc tăng hiệu suất model, chúng ta chỉ tăng một ít về chất lượng nhưng giảm khá nhiều về tốc độ.

| Model | Param | MSE | Time(ms) Pixel 3 | Time(ms) Samsung S20 | Time(ms) Iphone11 |
|---|---|---|---|---|---|
| Light | 1 | 11.83 | 6.6 | 5.6 | 1.1 |
| Full | 1.98 | 10.05 | 16.1 | 11.1 | 5.3 |
| Heavy | 4.02 | 9.817 | 36.9 | 25.8 | 7.5 |

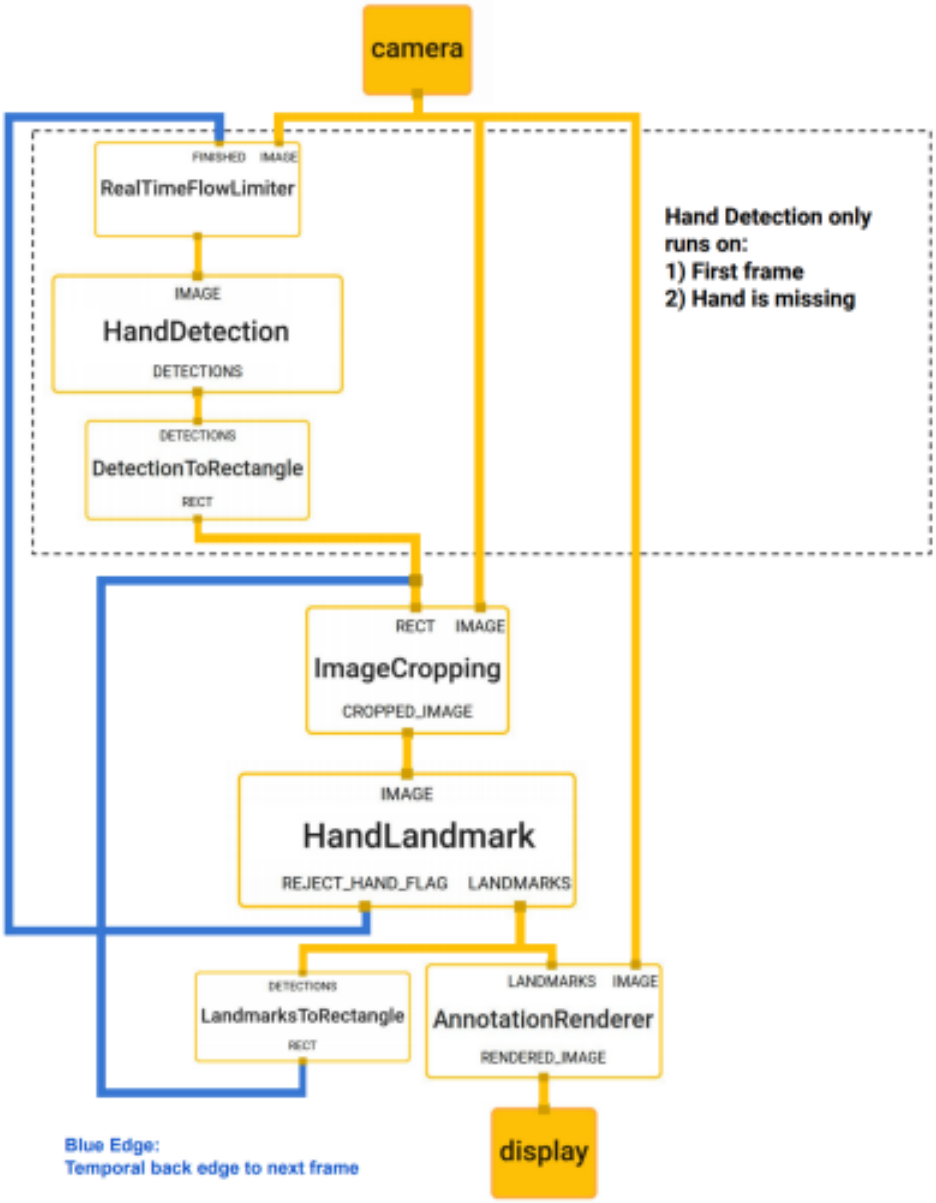
Bảng 3: Hand landmark model performance characteristics

# 6 Thiết lập Pipeline

Với MediaPipe, pipeline theo dõi bàn tay có thể được xây dựng nhưng là một đồ thị có hướng với từng thành phần mô-đun, gọi là các Calculator. Mediapipe kết hợp với các bộ Calculator mở rộng để giải quyết các vấn đề như thực thi mô hình, xử lý dữ liệu đa phương tiện, và chuyển đổi dữ liệu trên rất nhiều thiết bị và nền tảng khác nhau. Từng Calculator riêng biệt như cắt, xuất ảnh và tính toán mạng nơ-ron được tối ưu hơn để tận dụng khả năng tăng tốc của GPU. Ví dụ, nhóm nghiên cứu cài đặt trình thực thi GPU - TFLite trên hầu hết các điện thoại hiện đại.

Sơ đồ MediaPipe cho theo dõi bàn tay gồm 2 sơ đồ nhỏ: 1 là cho nhận diện bàn tay, còn lại là để tính toán cấu trúc của bàn tay. Khả năng tối ưu hóa mà MediaPipe mang lại là nhận diện bàn tay chỉ hoạt động khi cần thiết, từ đó tiết kiệm đáng kể những tính toán. Đạt được điều này bằng cách xác định vị trí của bàn tay tại khung video hiện tại từ việc tính toán đặc điểm tay của khung video trước đó. Sự chắc chắn được thể hiện rõ hơn khi mô hình xuất ra một khoảnh khắc vô hướng độ tin cậy rằng bàn tay xuất hiện và nằm căn chỉnh hợp lý trong miền đầu vào. Và chỉ khi nào độ tin cậy giảm dưới ngưỡng chắc chắn

thì mô hình phát hiện bàn tay mới được áp dụng tiếp ở khung hình tiếp theo.

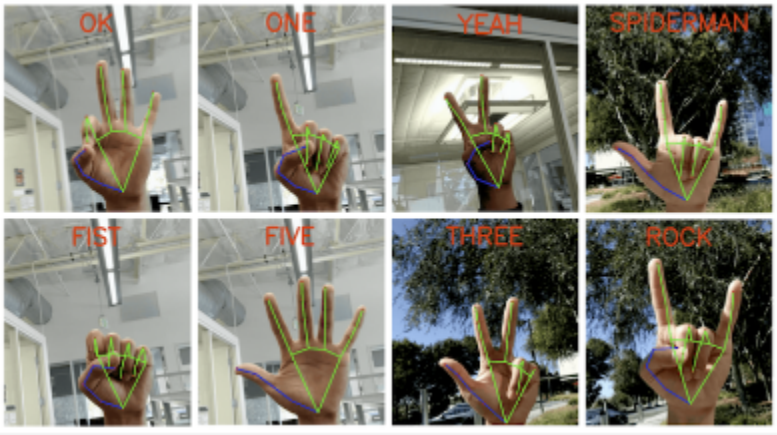



Hình 6: Pipeline của Mediapipe

# 7 Ứng dụng

Các cử chỉ tay có rất nhiều ứng dụng, đặc biệt là AR. Dựa vào khung bàn tay, người ta có thể sử dụng các thuật toán để dự đoán các cử chỉ, xem Hình 7. Đầu tiên, họ sẽ xác định trạng thái của mỗi ngón tay xem nó đang cong hay thẳng. Sau đó, họ ánh xạ các trạng

thái ấy vào tập hợp các cử chỉ đã xác định trước. Kỹ thuật đơn giản này, giúp cho họ có thể dự đoán được các cử chỉ với độ tin cậy hợp lý.




Hình 7: Ảnh chụp nhận dạng cử chỉ real-time

# 8 Kết luận

Mediapipe Hands là một giải pháp để theo dõi và nhận diện cử chỉ của tay đạt được mức độ thời gian thực. Nó có thể dự đoán được 2.5D mà không cần phần cứng chuyên dụng. Vì vậy, nó có thể dễ dàng triển khai cho các thiết bị nhỏ gọn như điện thoại hoặc máy tính yếu không có GPU. Ngoài ra, MediaPipe cũng open-source để khuyến khích các nhà phát triển có thể dễ dàng sử dụng cho các dự án của họ.

# 9 Tham khảo

1. MediaPipe Hands
2. Hand landmarks detection on an image using Mediapipe
3. MediaPipe Hands: On-device Real-time Hand Tracking
4. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs