G-12

手指ジェスチャー認識に向けた Leap Motion と MediaPipe の比較検討

Comparison between Leap Motion and MediaPipe for Hand Gesture Recognition

生野 優輝†　　外村 佳伸‡
Masaki Shono　Yoshinobu Tonomura

1. はじめに

本研究は、聴覚障害者向けに指文字コミュニケーションを認識することに端を発し、将来的には障害の有無にかかわらず手指ジェスチャー認識を用いた汎用インタフェースにつなげることをめざしている。その一環として最初に Leap Motion[1]を用いて検討していたところ、昨年通常の単眼カメラを用いた画像処理により同様のことを可能にする MediaPipe[2]が登場した。そこで本報告では、手指ジェスチャー認識の観点から MediaPipe の適用性を検討するために両者を比較検討した初期の結果について述べる。

2. 手指ジェスチャー認識

聴覚障害者のコミュニケーション手段は、主に手話、指文字を使用する。しかし、健聴者相手とのコミュニケーションでは、一般的に聴覚障害者が音声言語を聞き取れず、健聴者は手話、指文字を理解することができないため、お互いに筆談、タブレット、スマートフォンで文字を打つなどテキストを介して意思疎通を行うことが多い。近年は、音声認識技術の進歩により、発話者の音声からの自動文字起こしが精度よくできるようになり、健聴者から聴覚障害者への意思疎通に対しては現実的な問題解決になる兆しが見えてきた。一方で聴覚障害者が健聴者に意志を伝える場合には、何らかの手段でテキストを打ってそれを相手に見せるか、音声合成することが現在の現実的な手段である。しかし操作の都合上、顔を見てのコミュニケーションになりにくく、また逐次的ゆえの遅延がスムーズなコミュニケーションを妨げる要因になる。

聴覚障害者の発言を手助けするための研究は以前より行われており、近年は機械学習を用いる手話認識も研究されている[3]。手話がその特性から意味的な概念をベースに意志を伝える手段であるため、それ自体は実用的で重要なものの、言葉自身をベースにしていないため細かな情報や専門用語が伝えにくく、言葉との対応を持つ指文字も併せて用いられる。もし指文字の認識が正確かつ高速にできるようになれば、手指による言葉ベースのコミュニケーションの有効な手段になる期待がある。

手話認識、指文字認識いずれにおいても必要なのが正確なハンドトラッキングである。そのため既存の研究ではハンドトラッキングに赤外線カメラ、深度カメラ、多視点カメラなど特殊なカメラが使われることが多かったが、手指認識に焦点を当てた Leap Motion が登場後、これを用いた様々な手指ジェスチャー認識が可能となり、指文字認識を試みたものもある。

例えば、舩坂らは、Leap Motion で取得できる指先、骨格の座標情報を使用し、手の平の向き、各指の曲げ伸びを Yes・No の二値で表す 19 種の条件分岐を作成し。19 種の条件分岐からなるランダムフォレストを構築した結果、動きの伴わない指文字 41 文字の認識において準最適解で、74.7%の精度を得たと報告している[4]。

一方で、近年機械学習技術の向上により、特殊なカメラを使うことなく、安価な単眼カメラを用いてハンドトラッキングすることが可能となってきた。小林ら[5]は OpenPose によって手指ランドマークを取得し、SVM による静的指文字 41 文字の認識及び、パタン認識による動的指文字 5 文字の認識を実施している。その結果、静的指文字で 82.5%、動的指文字で 96.0%の正解率が得られた。これらの研究により、手指ランドマーク情報を用いた指文字認識が現実的になってきたと言える。

Google 社が提供する MediaPipe は機械学習を用いた認識を比較的簡単に実現できる環境を研究者や開発者に提供するものである。MediaPipe Hands[6]は特に手指のトラッキングを可能とするもので、我々はこれを用いて Web カメラやスマートフォンなどに搭載されている単眼カメラでて指ジェスチャー認識を行なうことに関心を持っている。

そこで本報告では、これまで検討してきている LeapMotion を用いた場合と比較する中で MediaPipe Hands の特性を把握し、指文字認識を含む手指ジェスチャー認識への適用性について検討する。

3. Leap Motion と MediaPipe

3.1. Leap Motion

図 1：Leap Motion （[1]より）

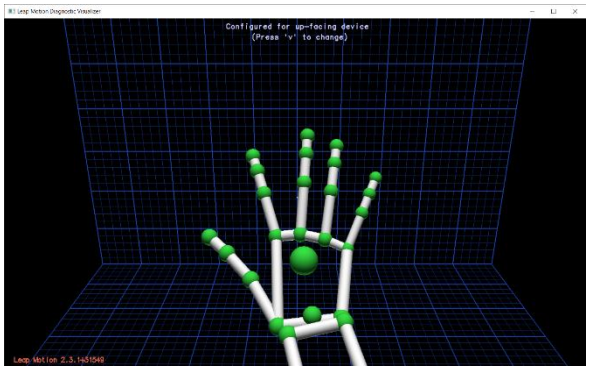

図 2：Leap Motion のビジュアライザ

図 1 の Leap Motion は 2012 年から販売されている、幅 80mm、縦 30mm、深さ 13mm と持ち運びしやすいサイズで、PC と USB 接続する手指のトラッキングに特化したデバイ

†龍谷大学大学院, Graduate school, Ryukoku University

‡ 龍谷大学, Ryukoku University

スである。構成は赤外線照射 LED と二基の赤外線カメラからなっており、赤外線 LED に照らされた手を二基の赤外線カメラで撮影し、画像解析することで図 2 のように 3D 手指ランドマークを取得することができる。

3.2. MediaPipe

MepipPipe とは Google 社が提供するオープンソースの機械学習（ML）ソリューションフレームワークである。MediaPipe で使用できるソリューションには複数あり、本稿では 2019 年に発表されたハンドトラッキングの MediaPipe Hands を使用する。

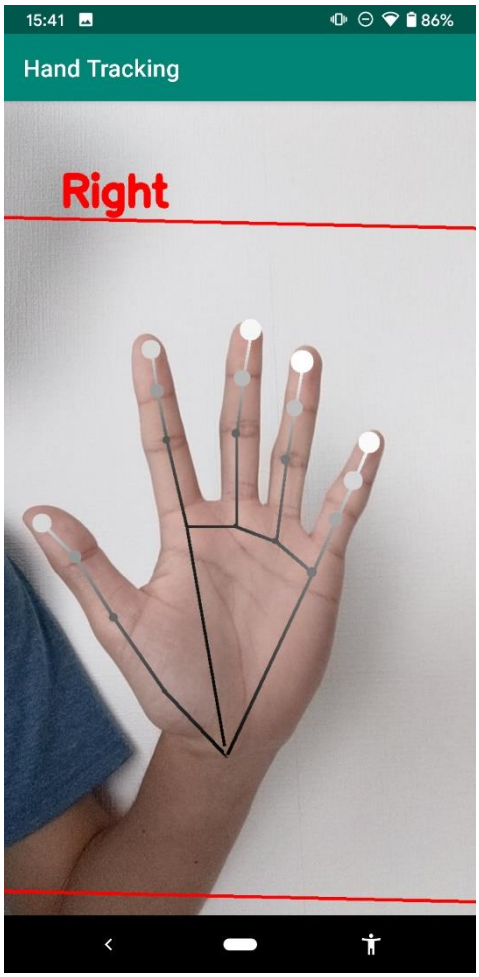

図 3：MediaPipe Hands in Mobile（Pixel 3a 上）

MediaPipe Hands では機械学習により、単眼カメラで撮影された 1 つのフレームから 21 個の 3D 手指ランドマークを推測し、高精度な手と指の追跡を提供する。Google AI Blog[7]によるとハンドトラッキングの平均精度は 95.7%である。強力なデスクトップ環境はもとより、モバイル環境で動くことも特徴のひとつとして挙げられている。

図 3 は MediaPipe で公開されている Android 版アプリ[8]を用いて実際にスマートフォン上で 3D 手指ランドマークを取得できることを示す。ランドマークのサイズによって奥行きを表しており、大きいほど手前で、小さいほど奥となる。なお、図 3 は右手がミラーリングされて表示された画像である。

4. Leap Motion と MediaPipe の比較検討

MediaPipe と Leap Motion は、3D 座標として取得できる指先と手の関節のランドマークのデータ種が近く、対応しやすい。本稿では Leap Motion をデスクトップ環境で動かしたもの、MediaPipe Hands をデスクトップ環境で動かしたもの、MediaPipe Hands を Pixel 3a で動かしたモバイル環境、以上の 3 つの環境で比較する。

4.1. 基本性能

指文字を含む繊細なジェスチャー認識をリアルタイムで実現するには、ハンドトラッキングの処理フレームレートが充分高速である必要がある。フレームレートの観点ではデバイス上で処理が行われる Leap Motion が優れており、MediaPipe Hands は使用するカメラからのデータフレームレートや処理アルゴリズムを実行する端末性能に大きく影響を受けることが考えられる。表 1 に手指ジェスチャー認識を行なう際に関わると思われる基本性能を 3 つの環境で実験比較したものを示す。

表 1 ：基本性能の比較

	Leap Motion	MediaPipe Hands on Desktop	MediaPipe Hands on Mobile
フレームレート (fps)	40~100	30	15
ハンドトラッキングの有効距離(mm)	30~500	150~2,000	150~2,000
暗所での使用	可	不可	不可
検出可能な手の数(個)	2	2※1	2※1
1 つの手から取得可能な指先、関節ランドマーク（個）	24	21	21

※1：Multi HandTracking のとき 2 個、Single HandTracking は 1 個のみ

ハンドトラッキングが可能な距離に関しては、Leap Motion がデバイスの上面から手の距離が 30~500mm でしか検出しなかったのに対し、MediaPipe Hands は Desktop、Mobile ともにカメラから手が 2,000mm 離れていても検出が可能であった。最低距離は、すべての指の指先から、手首までが撮影領域内に収まる必要があり、著者の手のサイズでは 150mm 以上離れる必要があった。

部屋に全く光源がない状況での使用は、Leap Motion が赤外線照射 LED と赤外線カメラを用いる特性上、精度に影響はあるものの問題なく扱えることが分かった。対する MediaPipe Hands は、撮影された画像内で手の輪郭を検出できないほど暗い場合は使用不可であるものの、暗くても手の領域がはっきりすれば手の座標を検出できることが分かった。

Leap Motion、MediaPipe Hands 両方とも検出できる手の数は最大 2 個（両手分）までである。さらに、MediaPipe Hands で Single HandTracking を使用する場合、1 個の手しか検出できなくなる。取得できる指先、関節ランドマークに関しては Leap Motion が手首付近に 4 つのランドマークがあ

るのに対しMediaPipe Handsでは手首に1つのランドマークのみの違いだけである。

4.2. ハンドトラッキング

Leap Motion、MediaPipe Handsの両方で50音の指文字をトラッキングさせ、取得できた手指ランドマークの3D形状が、指定されている指文字と一致しているかどうかを、指文字を読める人が視認する調査を行った。なお、後述の理由から、まず手を「パー」にして各環境で正確に手指をとらえたのちに指文字を表現している。

なお指文字とは図4に示すように、指の形を50音一つ一つに対応させた手話言語であり、発音の苦手な聴覚障害者のほとんどの人が手話と共に身に着けているものである。50音一つ一つに対応した決まった手の形を表現するため、手話と比べて個人差に影響されにくい。手話を使う聴覚障害者は手話主体で話すが、手話が分からない単語、例えば人名や地名などに指文字を使う。基本的には手話の方が早く表現ができるため手話を中心として手話で伝えにくい場面で指文字を使うことが多い。

指文字の多くは手、指の形、向きで決まるが、図4中矢印の記号があるものは動きを伴うものである。指文字の特徴には手の平が相手向きか、自分向きか、どの指がどのように曲がっているかなどの違いがあるため、こうしたことが認識できるかにも留意する必要がある。

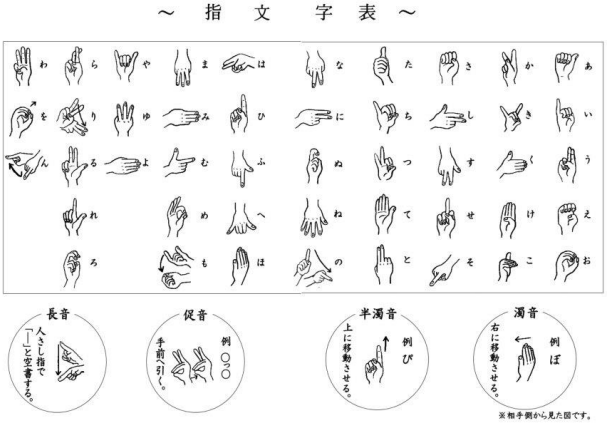

図4：指文字表 [9]

表2：トラッキング性能

	Leap Motion	MediaPipe Hands in Desktop	MediaPipe Hands in Mobile
一致率(%)	80.43	93.75	91.30

ハンドトラッキング性能（正確さ）はMediaPipe Handsのほうが高い結果になった。今回は全ての環境で、手を「パー」にしてから指文字を表現する手法をとったが、これはLeap Motionでは、手指の追跡に失敗すると、手を「パー」にしてかざすまで正確なトラッキング結果を得ることができないという課題があるためである。対するMediaPipe Handsではトラッキングに対する信頼度が悪くなれば、再度手検出をするためLeap Motionのように何度も「パー」をする必要が無かった。

また、指文字においてハンドトラッキングで得た形状を、人が見て一致していると判断できるかどうかを調査した結果、手の甲を相手に向ける指文字においてはLeap Motionで64.71%、MediaPipe Handsでは Desktop、Mobile両方で94.12%と一致しているように見えるということになった。また、「お」や「こ」などの様に正面から見て重なる指がある指文字では、Leap Motionが88.89%に対しMediaPipe Handsでは100%であった。手の平を向けた場合は両方とも「つ」の中指が曲げているにも関わらず、伸びているように見えるほかは両方ともよく認識できている。

Leap Motionでは手の甲を向けると、トラッキング追跡に失敗し正確なトラッキング結果を得ることができなくなった。対するMediaPipe Handsは手をどのように回転させても高精度のトラッキング追跡をしており、安定してトラッキング結果を得ることができた。

5. 考察

フレームレートはLeap Motionが最も高い結果となっている。対するMediaPipe Handsはカメラで撮影された画像を基に手指検出を行なっているため、カメラからのデータフレームレートが30fpsであるとき、MediaPipe Handsの処理フレームレートは30fps以上の速さを出すことができない。今回、デスクトップ環境において30fpsのWebカメラを用いており、処理フレームレートが30fpsとなっているため端末性能をフルに使えていない可能性がある。今後60fps以上のカメラを用いて、さらに処理フレームレートが上がるかどうか調査していく必要がある。モバイル環境はカメラからのデータフレームレートが30fpsに対し処理フレームレートが15fpsである。これに関しては端末のスペック不足が考えられる。

ハンドトラッキングの有効距離はLeap Motionが最大500mmなのに対し、MediaPipe Handsは2,000mmでもはっきりと捉えており、フレーム内に手の領域がはっきりと把握することができる限り、さらに離れても安定したハンドトラッキングを提供できることを確認できた。

Leap Motionは手の平をデバイスに向けている限り、高精度のトラッキング追跡が可能であるため、机の上に置いてマウス代わりに使うなどのほうが向いていると考えている。対するMediaPipe Handsは、動きに対するトラッキング追跡が優れているうえ有効距離が広く、ハンドジェスチャー認識に向いていると考えている。また、今後OpenPoseのような単眼カメラによる他のハンドトラッキング技術との比較検討をする必要がある。

6. 結論

本稿では聴覚障害者の発言を手助けするための指文字認識の実現をめざし、我々も含めて利用者の多いLeap Motionと、近年登場したMediaPipe Handsとを、ハンドトラッキングを用いた手指ジェスチャー認識の観点から基本性能比較を行った。

調査の結果Leap Motionでは、デバイスから50cm以内のところしか手を検出できないのに対し、MediaPipe Handsでは1m、2mと離れていても手を検出できることが分かった。さらに、手の平をデバイスに向ける時の認識精度に差は見られなかったが、デバイスに手の甲を向ける時の認識精度の差がLeap Motionが64.71%に対しMediaPipe Handsでは94.12%となり、Leap Motionが苦手な分野でもMediaPipe Handsは高精度のトラッキング結果を提供できることが分かった。まだ、指文字の詳しい認識性能としての比較には至っていない段階の初期検討ではあるが、利用環境の広さ

と認識精度の点で MediaPipe を今後の検討の中心に据えることを検討している。

参考文献

[1] Leap Motion. https://www.ultraleap.com/product/leap-motion-controller/ 2020 年

[2] MediaPipe https://google.github.io/mediapipe/ 2020 年

[3] 髙橋 佑汰, 木村 勉, 神田 和幸. 機械学習を用いた手話認識に関する研究. IEICE-WIT2018-60, Vol.IEICE-118, pp.IEICE-WIT-59-IEICE-WIT-64

[4] 舩阪真生子, 石川由羽, 高田雅美, 城和貴. Leap motion controller を用いた指文字認識. 情報処理学会研究報告. MPS, 数理モデル化と問題解決研究報告, Vol.2015,No.8, pp.1‐6, 2015

[5] 小林大起, 渡辺裕. 骨格推定と機械学習を用いたカナ指文字の分類. 早稲田大学 2018 年度修士論文

[6] MediaPipe Hands. https://google.github.io/mediapipe/solutions/hands.html 2020 年

[7] Google AI Blog “On-Device, Real-Time Hand Tracking with MediaPipe” https://ai.googleblog.com/2019/08/on-device-real-time-hand-tracking-with.html 2020 年

[8] MediaPipe Hands Example Apps for Android. https://drive.google.com/file/d/1uCjS0y0O0dTDItsMh8x2cf4-l3uHW1vE/view 2020 年

[9] 指文字表-京都府教育委員会 http://www.kyoto-be.ne.jp/rou-s/youjisyuwa/hyo/yubimoji.html 2020 年