第 28 回日本バーチャルリアリティ学会大会論文集（2023 年 9 月）

Media Pipe による口形素についての抽出とアバタ表現への応用

Extraction of visemes by Media Pipe and its application to avatar representation

戸田壮駿 [1)]，田中久弥 [2)]
Masatoshi TODA, and Hisaya TANAKA

1) 工学院大学大学院 工学研究科情報学専攻 （〒192-0015 東京都八王子市中野町 2665-1, em23030@ns.kogakuin.ac.jp）
2) 工学院大学 情報学部情報デザイン学科 （〒192-0015 東京都八王子市中野町 2665-1, hisaya@cc.kogakuin.ac.jp）

概要：コロナ渦を経て、メタバース空間でのコミュニケーションが盛んになった。そのような空間においてはアバタに表情を投影してコミュニケーションが行われる。そこで、本研究では Web カメラを入力として、MediaPipe を用いて口形素の抽出をリアルタイムで行いアバタへ反映するシステムを開発した。そのことで実際の映像と比較して非言語情報の伝達が行うことが出来るかを検証した。

キーワード：MediaPipe, 口形素, 非言語コミュニケーション

1. はじめに

2019 年に発生した新型コロナウィルスの世界的流行によって、対面でのコミュニケーションが困難になった。そのため、遠隔地でも同一の場所にいるようなコミュニケーションを取る手法の一つとして VR 技術が注目された。その中でも、メタバースというインターネット上でアバターを用いてコミュニケーションを行うことが盛り上がった。電通が 2022 年に実施した調査によると、前年度と比較して認知度が高まり、課金サービスへの課金額が 3 倍になったことが明らかになっている[1]。メタバース上に置いては、ユーザはアバタを用いてコミュニケーションを行う。船木らの研究によると、アバタをフェイストラッキングを用いて非言語情報を付与することで親近感が高まるという結果が得られた[2]。メタバースシステムにおいて、付与できる非言語情報は大きく分けて、開閉眼と口形素である。開閉眼の抽出には筆者の先行研究で Web カメラと MediaPipe を用いた手法で実装を行った[3]。口形素とは、池田らによると、「発話の際に生じる口の形の最小単位」されている[4]。福田らによると口形素の種類は 14 個存在する[5]。この口形素を表 1 にまとめる抽出には Meta 社が提供している OVRLipSync が存在するが、これはマイクからの音声を用いて動作する[6]。そのため周辺の音を拾って誤った判定を行ってしまう。本研究では、Web カメラからの映像から口形素を抽出を行うシステムを開発した。そしてアバタへ付与してその効果を検証する。

表 1 音素と口形素の対応表[4]

音素	口形素	音素	口形素	音素	口形素	音素	口形素
a	a	p	p	ty	sy	ts	s
a:		b		hy		z	
i	i	m		ry		s	
i:		w	w	py		k	vf
u	u	f		ch		g	
u:		j	sy	dy		h	
e	e	my		sh		q	無し
e:		ky		r	r	N	N
o	o	by		t	t		
o:		gy		d			
y	y	ny		n			

2. システム開発

2.1 システム概要

本研究で作成したシステムは unity 2022.3.0f1 上で構築した。MediaPipe は unity 上で動作するためのプラグイン[7]を使用して動かした。顔画像の入力としては、ビデオカメラ(SONY 社 HDR CX470)を Web カメラ化して unity 上の入力とした。その映像に対して MediaPipe を使用して顔の特徴点を抽出した。この時の特徴点を csv 形式で記録した。この特徴点を製作した口形素抽出システムの入力として口形素抽出し、そのデータを csv 形式に保存した。また、そのデータをリアルタイムでアバタに反映させ MP4 形式で映像として記録した。本研究で使用したアバタは unity Japan が提供している unity ちゃん[8]と vroid studio で作成した VRM アバタの二種類を使用して行った。

2.2 口形素抽出システム

本研究において抽出を行った口形素は「あ」と「い」、「う」、「え」、「お」の 5 種類を使用した。今回は母音の 5 種類しか抽出を行わなかった理由としては、日本語の発音では必ず母音を使用するからである。動きには縦幅の変化と横幅の変化がある。縦幅の変化には人中の中心部分と顎の上部の 2 次元空間の距離を用いて算出した。横幅には Mouth Aspect Ratio(以下 MAR とする)と Mouth Wide Degree(以下 MWD とする)を用いて算出した。MAR は口のアスペクト比であり、計算方法は式(1)に示す通りに縦幅

$$MAR = \frac{Mouth\ hight\ distance}{Mouth\ wide\ distance} \quad (1)$$

で横幅を割ることで求められる。MWD は下唇の上部と唇の下部を結ぶ直線を余弦とし、左側の口角と下唇の中央を正弦とする三角形の角度であり、式(2)に示す式で計算させる。式(1)と式(2)で使用した距離の位置関係を図 1 に示す。

$$MWD = \tan^{-1}\left(\frac{Lower\ lip\ wide\ distance}{Lower\ lip\ hight\ distance}\right) \quad (2)$$

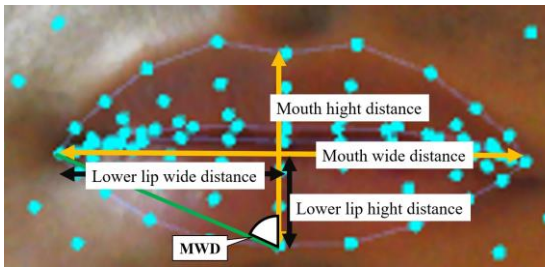

図 1 口形素を抽出する際に使用する距離関係

表 2 アバタに口形素を付与したときの様子とパラメータの変化

口形素	顔画像	MAR	MWD(°)	アバタ表現
あ		0.8353	27.19°	
い		0.6146	29.85°	
う		0.3190	50.22°	
え		0.6681	34.47°	
お		0.7014	33.55°	

「あ」は口の縦幅を用いて算出している。MAR が 0.5 以下のときかつ、MWD が 31 度以下のときに「い」であると推論し、MAR が 0.5 以上のときかつ、MWD が 31 度以上のときに「う」であると判定した。「え」は口の形の類似度の高かった「あ」と「い」を合成して作成し、「お」は口の形の類似度の高かった「あ」と「う」を合成して作成する。このように抽出をしてアバタに反映させた様子を表 2 に示す。

2.3 MeidaPipe における実装方法

MediaPipe とは、Google が公開しているオープンソースのライブラリである[9]。MediaPipe には複数の機能を実装したライブラリが存在するが、本研究では、顔の特徴点を 478 ヶ所取得できるモデルである Face Landmark Detection を使用した。Face Landmark detection を使用して顔の特徴点を取り出して顔に重畳した様子を図 2 に示す。MediaPipe

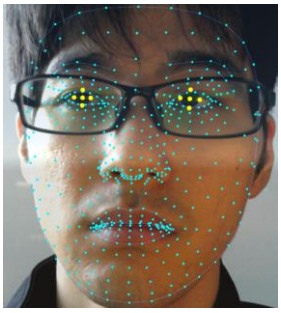

図 2 MediaPipe のランドマーク

では、アテンション・メッシュ・モデルというニューラルネットワークで作成したライブラリ用いて顔の特徴点を 3D で位置を推定する。Ivan らの研究によると、アテンション・メッシュ・モデルは Dlib で用いられている採用されているカスケードモデルと比較して 25%も高速であるということが発見されている[10]。

MediaPipe において縦幅の算出には人中の中心部分である特徴点 164 番と顎の上部である特徴点 18 番を用いた。MAR の算出には縦幅を上部は特徴点 15 番、下部を特徴点 13 番、左幅を特徴点 78 番、右幅を 308 番として算出を行った。MWD は下唇の下部を特徴点 18 番、下唇の上部である特徴点 14 番、左口角である特徴点 78 番を使用した。

2.4 アバタ表現法

抽出して得られた口形素をアバタへ反映するためには、Blend shapes という 3D モデルの移動を記録したパラメータに代入を行った。unity ちゃんの場合には「MTH_DEF」のオブジェクトの「blendShape1.MTH_A」と「blendShape1.MTH_I」、「blendShape1.MTH_U」に口形素を代入した。VRM アバタにおいては「Face」のオブジェクトの「Fcl_MTH_A」と「Fcl_MTH_I」、「Fcl_MTH_U」に抽出した口形素を代入した。

3. 実験

3.1 実験方法

実験の被験者は 22 歳の読唇術の能力を有していない日本語を母語とする男性 3 名に対して行った。本実験は工学院大学のヒトを対象とする研究倫理審査「新しいインタフェース開発のための心理生体計測 2021-A-29」に基づいて実施した。被験者には書面での同意の下実験を行った。

3.2 実験手段

実験は二つに分けて行った。1 つ目には顔の撮影を行った。被験者の顔の正面から 65cm の位置にカメラを設置した。そして、顔全体がビデオカメラに写っていることを確認した後に、5 文字のア行だけの無意味語の一文を発話してもらい記録した。2 つ目には撮影した映像から文字起こしを行ってもらった。この時の映像は他の被験者の撮影した映像と抽出した口形素を付与したアバタをランダムに再生した。評定映像の視聴は一度だけとした。

4. 結果

ビデオカメラによる実際の映像から文字起こしをさせた正答率は 86.67%であった。unity ちゃんに口形素を付与した映像から文字起こしをさせた正答率は 16.67%であった。VRM アバタに口形素を付与した映像から文字起こしをさせた正答率は 10.00%であった。

5. 考察

今回の結果から実際の映像だけが正答率が 86.67%と他の条件と比較して大幅に良い結果が出たが、これは、撮影の際に口の動きを撮影すると明示的に説明したために、意識的に口の動きをはっきりと動かすことをしたと考えられる。そのため、評定の正答率が向上した可能性が示唆される。評定した映像を再度確認したところはっきりと口を動かしていたことが確認された。口形素を付与したアバタによって正答率が変化しているが、これは、アバタの口の動きが unity ちゃんの方が自然な開閉度合いだったからではないかと考えられる。

6. 結論

本研究では現状のメタバースシステムに使われている音声入力型の口形素分析システムから置き換えるために Web カメラからの入力で行えるシステムを開発した。その結果、口形素を抽出し、アバタに付与するシステムの実装ができた。しかし、付与するアバタによって正答率が異なることが明らかになった。そのため、システムの改良を行い、正答率を向上させ、伝達の効果を高めることを目指す。加えて、今回は母音 5 種類だけの口形素抽出を行ったが、池田らによると口形素は全部で 14 種類ある[5]。そのため今回抽出を行わなかった他の口形素を抽出するアルゴリズムの構築を行い、さらにアバタに付与できる情報を増やしたいと考える。最終的な展望としてはこのシステムをメタバースシステムに組み込み、その有用性を検証したいと考えている。

謝辞

実験に協力して頂いた、被験者の皆さんと本研究について積極的な議論を行ってくれた工学院大学生体情報処理研究室の皆さんに感謝の意を示します。

参考文献

[1] 電通株式会社: メタバースに関する意識調査 2022; https://www.dentsu.co.jp/news/item-cms/2022054-1222.pdf, (2023 年 7 月 20 日参照)

[2] 舟木 烈, 物部 寛太郎, :フェイストラッキングを用いたアバターの親近感を高める手法に関する研究, 第 27 回日本バーチャルリアリティ学会大会論文集 (2022 年 9 月)

[3] Masatoshi TODA, Hisaya TANAKA: Extraction of Eye Open/Close State by MediaPipe and its Application to Avatar Expression, The 9th International Society of Affective Science and Engineering, 2023

[4] 池田 大輔, 桂田 浩一, 入部 百合絵, 新田 恒雄: 認識に使用する顔領域の違いによる読唇性能の比較, HAI シンポジウム 2011, 2011

[5] Yumiko Fukuda, Shizuo Hiki: Characteristics of the mouth shape in the production of Japanese-Stroboscopic observation, Journal of the Acoustical Society of Japan, 1982

[6] Meta: ネイティブ開発のための Oculus リップシンクの設定 https://developer.beta.oculus.com/documentation/native/audio-ovrlipsync-native/, (2023 年 7 月 20 日参照)

[7] GitHub. https://github.com/homuler/, (2023 年 7 月 20 日参照)

[8] Unity Japan: UNITY CHAN!, https://unity-chan.com/, (2023 年 7 月 21 日参照)

[9] Google: https://google.github.io/ (2023 年 7 月 20 日参照)

[10] Ivan Grishchenko Artsiom Ablavatski Yury Kartynnik Karthik Raveendran Matthias Grundmann: Attention Mesh: High fidelity Face Mesh Prediction in Real time, Google Research, 2020