7ZB-05

特定フレーズに対する顔のモーションデータを用いたユーザの継続的識別

寺田 和仙 †　　豊浦 正広 ‡
山梨大学 †　　山梨大学 ‡

1　はじめに

仮想空間においてユーザのなりすましを防止することは難しい．たとえば VR であれば，ユーザは相手の顔を直接確認することができず，ヘッドマウントディスプレイ (HMD) が別のユーザへ渡ったとしても，VR 空間に投影されたアバタは変わらず存在し続けることから，その変化に気付くことができない．アバタを介して操作することは，顔を相手に見せることを強要しないという点で利用価値があると考えられるが，その反面，それぞれの操作がユーザ本人によるものであるかについては，操作ごとに確認を求めなければ分からない．

Kuang ら [1] は，口の静的な生理的特徴と動的な運動特徴を組み合わせてユーザの認証を行うことを提案した．予め登録しておいた特徴との差分を求め，差分が閾値以下の場合に認証を通過させる．これにより，50 人のボランティアを 0.9924 の正解率で識別することに成功した．この方法は，高精度にユーザを識別することが可能だが，唇を閉じた状態で笑うという制約があり，ユーザは唇を開かないよう意識しなければならない．

本研究では，日常会話における顔のモーションデータからユーザを識別することを目標とする．日常の顔の動きからユーザを識別することができれば，図 1 へ示すように，VR サービス提供会社はユーザのなりすましを検知できるようになり，ユーザは意識的に認証することなく，自身による行動であると証明しながらアバタを操作できるようになる．

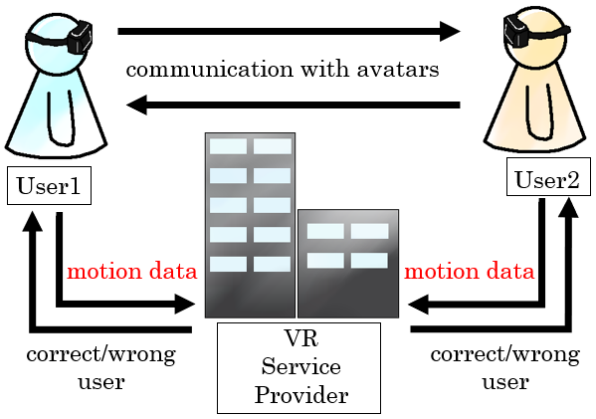

図 1　ユーザの継続的識別

2　提案手法

“Hello”，“Nice to meet you”，“Excuse me” の 3 フレーズからユーザの識別を行う．提案手法として，顔のモーションデータを，時系列データの解析に有効なニューラルネットワークへ学習させることを提案する．

2.1　識別ネットワーク

ユーザの識別には，LSTM と Transformer を含む 2 つの深層学習モデルを用いる．

LSTM は, RNN の長期学習によって起こる勾配消失の問題を解決したモデルである．LSTM を多層化することによって，時系列データの複雑な依存関係を学習できる．

Transformer では，Encoder 部のみを用いてユーザの推定を行う．Attention 機構により，意味が強く結びついている時刻のデータをそ

Continuous identification using facial motion data for specified phrases
† Kazuhito Terada, University of Yamanashi
‡ Masahiro Toyoura, University of Yamanashi

れぞれ抽出する．抽出されたデータを残差接続によって元の特徴ベクトルと足し合わせることで，時系列データの前後関係を捉える．

2.2 複数フレーズによる学習

前述のユーザの識別ネットワークの精度を上げるためには，該当フレーズの発音回数を増やし，データ数を増加させる方法が挙げられる．しかしながら，モデルを訓練させるために必要なフレーズの発音回数が多くなると，ユーザにとって負担が大きくなることから，必要最小限の回数で済ませることが望ましい．そこで，実験では複数のフレーズをそれぞれ一定回数分を収録し，フレーズごとにモデルを構築するのではなく，複数のフレーズに対してひとつのモデルを構築することを試みた．

3 実験

実験では，17 名の実験協力者を対象に実施した．ノートパソコンに搭載されている Web カメラによって実験協力者の顔をキャプチャした状態で，指定する 3 つのフレーズをそれぞれ 30 回発音してもらったときのモーションデータを抽出した．ラベル付けは，図 2 に示すソフトウェアを実装して行った．それぞれのフレーズの開始と終了の時刻を与えることにより，その区間において，誰が言ったときの動きであるかのラベルが自動で付けられる．

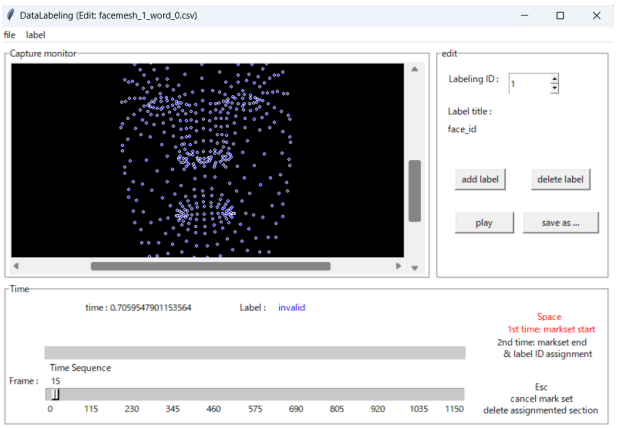

図 2 ソフトウェアによるラベル付け

得られたモーションデータを分割して，15 回分を学習用，6 回分を検証用，9 回分をテスト用とした．テスト用データに対するモデルの正解率により識別精度を評価した．単一フレーズと複数フレーズをそれぞれモデルへ学習させたときの識別精度は表 1 に示す通りとなった．

表 1 単一フレーズと複数フレーズでの学習による識別精度

フレーズ	識別手法	正解率
“Hello”	LSTM	0.727
	Transformer	0.888
“Nice to meet you”	LSTM	0.775
	Transformer	0.936
“Excuse me”	LSTM	0.719
	Transformer	0.927
all phrases	LSTM	0.828
	Transformer	0.978

ユーザ 17 人を “Hello”，“Nice to meet you”，“Excuse me” の 3 フレーズから，LSTM モデルでは 0.828 の正解率で，Transformer モデルでは 0.978 の正解率で識別することができた．複数のフレーズに対して，Transformer を含むモデルを学習させることによって，ユーザの識別精度を向上できることを確認した．

4 まとめ

実験では，ユーザを顔のモーションデータから高精度に識別できることを確かめた．提案手法は，日常会話からのユーザ識別を可能とし，仮想空間におけるユーザのなりすまし防止及び検出に役立てることができる．

参考文献

[1] L. Kuang et al., “ LipAuth: Securing Smartphone User Authentication with Lip Motion Patterns,” IEEE IoT, vol.11, no.1, pp.1096–1109, 2023.