This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.

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）

時系列と頭部姿勢を考慮した組み込み型光センサによる表情識別

Facial Expression Recognition by Photo-Reflective Sensors Considering Time Series and Head Posture

中林優樹 [1)]，中村文彦 [2)]，杉本麻樹 [1)]
Yuki Nakabayashi, Fumihiko Nakamura, and Maki Sugimoto

1) 慶應義塾大学大学院 理工学研究科（〒 223-0061 神奈川県横浜市港北区日吉 3-14-1）
2) 立命館大学 情報理工学部（〒 525-0058 滋賀県草津市野路東 1-1-1）

概要: HMD(Head-Mounted Display) 装着者の顔表情を識別する手法として，反射型光センサを HMD の顔接触部に配置し，その反射強度情報を機械学習する手法が提案されている．本研究では反射強度情報に加えて頭部姿勢情報を入力し，更にセンサ値の時間的な変化を考慮する学習モデルを使うことで表情の識別精度が上がるかを検討する．HMD に内蔵された IMU(Inertial Measurement Unit) から取得する頭部姿勢情報を識別器の入力に加え，LSTM(Long short-term memory) を使った時系列学習を行うことでユーザの表情識別精度が上がるかを検証した．

キーワード： 表情識別，時系列学習，マルチモーダル学習

1. はじめに

多くの表情認識システムはカメラによって撮像した顔画像を用いる．しかし，VR(Virtual Reality) 環境に没入する際に用いる HMD(Head-Mounted Display) はユーザの顔の大部分を遮蔽するので，従来のカメラによる表情認識の適用が難しい．そこで，HMD によって遮蔽されていない口周辺をステレオカメラで捉えて表情を認識する装置[1]や，組込小型カメラによる表情認識機能を搭載した HMD[2]が市場に流通している．しかし，画像処理では画像に含まれる高次元のデータを処理するために，高性能な計算資源が必要となり，経済的なコストの上昇を招く．

そこで，低次元のデータで HMD ユーザの表情を認識するために，HMD に反射型光センサを組み込むことで表情を認識する手法が提案されている [1]．反射型光センサは発光素子と受光素子で構成され，発光素子から照射された光が測定対象の物体によって反射された光の強度を受光素子で検出する．反射型光センサを HMD に組み込み，表情毎の顔面形状における反射強度情報を計測し機械学習することで，表情を識別できる．また，反射型光センサによる表情認識では，ユーザの様々な方向に視線や頭部を向けた時の表情を計測したデータを収集・学習することで表情認識の精度が向上することが示されている [2]．

しかし，これまでの研究では顔面形状からの反射強度情報のみを機械学習に用いていたが，頭部姿勢の情報を考慮した表情認識手法は検証されていない．また，これまでの研究において，表情の時系列変化も考慮されていなかった．そこで，本研究では HMD に組み込んだ反射型光センサを用いた表情識別において，反射強度情報に加え，HMD に内蔵された IMU から取得できる姿勢情報と，時系列を考慮した機械学習を行うことで，HMD 装着者の表情識別精度が向上するかを検証する．時系列学習には，モデルとして RNN(Recurrent Neural Network) と LSTM(Long short-term memory) を使用し，時系列を考慮しない DNN(Deep Neural Network) と結果を比較する．

[1]Vive Facial Tracker, https://www.vive.com/jp/accessory/facial-tracker/, 2023 年 7 月 5 日最終閲覧
[2]Meta Quest Pro, https://www.meta.com/jp/quest/quest-pro/, 2023 年 7 月 5 日最終閲覧

2. 関連研究

2.1 反射型光センサによる表情識別

HMD を装着したユーザの表情認識については様々な手法が提案されている．Suzuki らは HMD に組み込んだ 16 個の反射型光センサによって，表情毎の顔の表面とセンサの間の距離の変化を機械学習することで，ユーザを表情を認識した [1, 3]．また，反射型光センサを用いた表情認識では，様々な視線方向や頭部姿勢をとった時の表情の計測データを収集することで，表情認識の精度が向上できることが Nakamura らによって示された [2]．特に，視線や頭部を上下に動かした際のデータが識別機の性能の向上に貢献することが示されている．

上記のように，反射型光センサによる表情や顔のジェスチャの識別が行われており，特に，顔のジェスチャの認識では，時系列データを活用することで識別精度の向上や転移学習などへの応用も可能であることが示されている．その一方で，表情認識を行う研究では，時系列データを活用し

© 2023 日本バーチャルリアリティ学会

た識別手法についてはほとんど検討されていない．

2.2 姿勢情報や時系列を考慮した学習

IMU から得られる角速度や加速度のセンサ値を他のセンサデータと同時に入力し，マルチモーダルな学習を行う研究が多く実施されている．Georgi ら [4] は，手のジェスチャーの認識に EMG(ElectroMyoGraphy) や IMU を用いて，それぞれ個別に用いた場合と，そのどちらも使ったマルチモーダルなジェスチャーの識別性能を評価した．その結果，EMG と IMU を個別に使うより，どちらも使ったマルチモーダルな識別方法が最も高い精度となった．この研究によって，EMG や IMU を併用したジェスチャー認識の実現可能性が示された．また，前後のデータとの時間的関係が重要なセンサデータに対しては，時系列を考慮する学習モデルを使った研究が多く存在する．Zhang ら [5] は，EMG と IMU に加え，圧力センサを用いたウェアラブルデバイスを使い，時系列的なジェスチャーの識別を LSTM により行った．この研究により，時系列を考慮する，IMU を用いたマルチモーダルな学習を行うシステムが，ジェスチャー識別の高い識別精度を実現することが示唆された．

3. 理論

3.1 反射型光センサによる HMD 装着者の表情識別

本研究では，Suzuki らの研究 [3] で用いられた，反射型光センサが組み込まれた HMD と同一のデバイスを使用する．そのデバイスを図 1 に示す．取得した 16 個のセンサデータを 16 次元データとし，1 時刻分のデータの一部として使う．HMD 装着者が表情を変えた際に顔表面と反射型光センサの距離が変化するため，そのセンサ値の組み合わせを機械学習することで表情を識別する．識別する表情は，Neutral(真顔)，Smile(笑顔)，Surprised(驚き)，Sad(悲しみ)，Angry(怒り) の 5 つである．

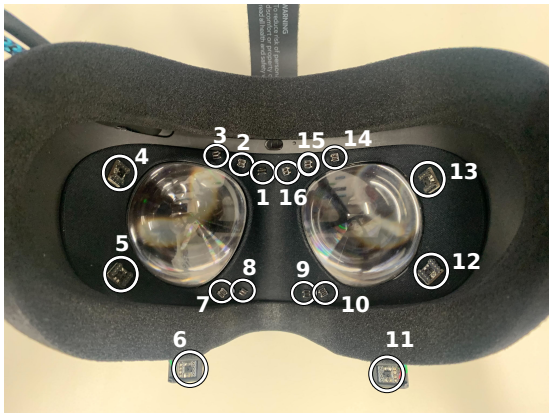

図 1: HMD に配置された 16 個の反射型光センサ

3.2 頭部姿勢がセンサ値に与える影響

頭部姿勢を変化させることでセンサ値が影響を受けることは，Nakamura らの研究 [2] で明らかになっている．HMD 装着者が頭部姿勢を変更すると，HMD が重力の影響でずれることで HMD と顔表面の距離が変わるため，反射型光センサのセンサ値も変わる．本研究でも Nakamura らの研究 [2] と同じ手法でデータ収集を行うため，頭部姿勢の影響を受ける．頭部姿勢の変化によるセンサ値の変化の様子を図 2 の左側に示す．この図は顔が正面を向いている状態から下方向に向いた際のセンサ値の変化であるが，下方向を向くと HMD が重力の影響で顔表面から遠くなるため，センサ値が全体的に小さくなっている．反対に，上方向を向くと HMD が重力の影響で顔表面との距離が近くなるため，センサ値は全体的に大きくなる．

3.3 表情遷移時のセンサ値の連続的な変化

Neutral から Smile に表情遷移する際のセンサ値の変化を図 2 の右側に示す．表情を変えると各センサ値が連続的に変化し，次の表情の定常状態に落ち着く．時系列学習の際は，表情遷移中の各センサ値の変化量を見ることで，次に来る表情を予測する．

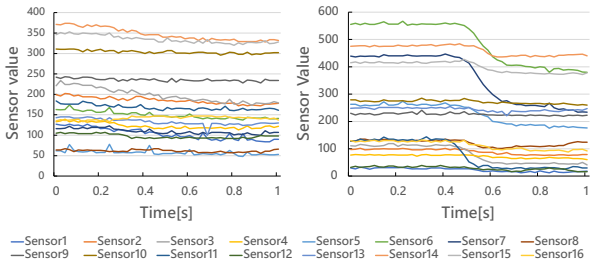

図 2: 頭部姿勢 (左) と表情遷移 (右) の変化によるセンサ値の変化

3.4 LSTM による表情識別

本研究では時系列を考慮する学習手法として，LSTM を使った学習モデルによる識別を行う．モデルの入力には，0.02 秒ごとに取得したデータを時系列順に 20 個用いて作った 1 まとまりの時系列データを用いた．本研究における LSTM を用いた識別器のネットワークを図 3 の左側に示す．

このモデルは Pytorch 1.13.0 [6] によって実装した．識別器は t_i から t_{i+19} までの 18 次元の表情データからなる時系列データを受け取り、t_{i+20} に対する 5 次元の表情予測ラベルを出力する。正規化は時間列方向の変化量が重要であるため、時間列方向に行う。t_{i+20} での予測ラベルだけが必要なので、LSTM 層からの最後の出力だけが次の層に渡される。損失関数は交差エントロピー誤差を用いた。また、最適化手法には Adam [7] を用いた。

3.5 DNN による表情識別

時系列を考慮しない表情識別の手法としては DNN を用いる．本研究における DNN のネットワークを図 3 の右側に示す．このモデルは t_i の表情データを受け取り、t_i の 5 次元表情予測ラベルを出力する。実装方法や損失関数，最適化手法は 3.4 と同じである．

4. 実験

4.1 データ収集

データは Unity にて作成した環境で収集した．データ収集は実験参加者 10 名（20 代男性 8 人，20 代女性 2 名）に対して行った．実験参加者はデータ収集の前に，データ収集

© 2023 日本バーチャルリアリティ学会

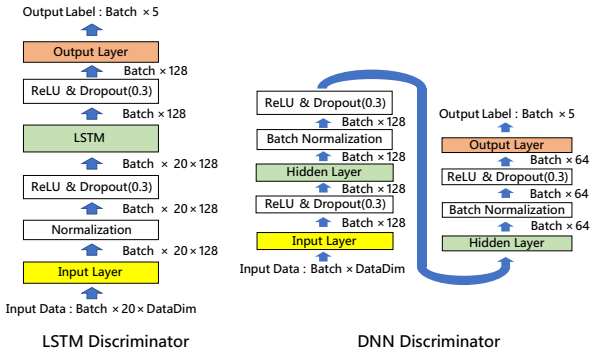

図 3: 識別器として用いた LSTM(左) と DNN(右) の構造

に関する説明を受け，その後，計測の練習を 1 回行った．計測 1 回で 5 種類の表情の遷移を全通り試すために，表情の推移は 1 回で 20 通り計測する．遷移の順番は固定である．表情の状態遷移図を図 4 に示す．また，頭部姿勢が識別結果に与える影響を調べるため，図 5 で示す 5 種類の頭部姿勢でデータを収集した．この 5 種類それぞれの頭部姿勢で 20 種類の表情遷移を取得する流れを 3 セット行った．計測時間は，センサ取得の仕様上，最初の表情のみ 5 秒間表情を作り，その後は 1 表情あたり 3 秒間の計測を 20 個の表情に対して行うため，計測 1 回に要する時間は 65 秒である．尚，最初の表情もデータを保存するのは後半 3 秒のみである．データ収集は各回の始まりのみ実験者が合図をし，その後は 1 回の計測が終わるまで全て自動でデータ収集を行う．

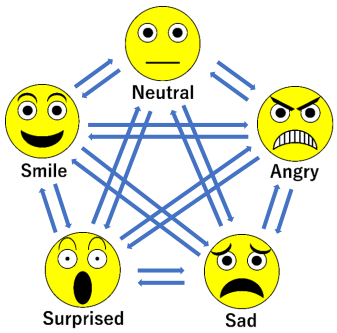

図 4: 表情の状態遷移図

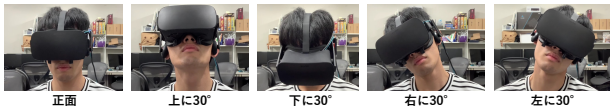

図 5: 計測時の頭部姿勢

4.2 データセット

データセットは 1 次元の正解ラベル、16 次元の反射型光センサの値、2 次元の頭部姿勢 (Pitch 角, Roll 角) から構成され、1 回あたり 19 次元である。Yaw 角は重力による HMD の変位とは無関係であるため、本研究では使用しない。表情遷移の瞬間の前後のラベルは、参加者の反応速度を考慮すると不正確であるため、表情遷移の前後の 2 つのクラスに対して k-means 法を用いてラベルの付け替えを行った。時系列データは、長さ 20 固定のスライディングウィンドウ法を用いて作成した。すべての学習モデルにおいて、参加者 1 人当たり 42,500 時刻分のデータを学習、評価、テストに使用した。

4.3 実験

図 3 の LSTM，DNN と，比較用に RNN を用いた学習モデルを用いて実験を行った．個人に対する識別精度を確認するため，実験参加者 10 人分のデータセットに対し，10 人それぞれのデータのみで学習，評価を行った．1 人当たりの学習に使った総データ数は 42,500 時刻分である．学習時はデータセット 1 人分を訓練データ，評価データ，テストデータに分け，交差検証を行った．各データの比率は，訓練データ：評価データ：テストデータ = 7 : 1 : 2 であり，交差検証は $k = 10$ の k 分割交差検証を用いた．結果は正解率 (Accuracy)，適合率 (Precision)，再現率 (Recall)，F 値 (F-measure) を用いた評価した．各指標は 5 つの表情別に導出され，最終的な結果にはそれらのマクロ平均を使用した．全ての学習モデルにて，バッチサイズは 64，学習率は 1.0×10^{-4} であり，LSTM と RNN の時系列サイズは 20 とした．

5. 結果

1 人ずつのデータセットで学習，テストを行った結果の平均値を表 1，図 6 に示す．Accuracy はテストデータの Accuracy のみを記載している．

表 1: 実験参加者 10 人の各評価指標の平均値

	Mean(%)			
Learning Condition	Accuracy	Precision	Recall	F-measure
LSTM with Head Data	99.2	99.2	99.2	99.2
LSTM without Head Data	99.2	99.2	99.2	99.2
RNN with Head Data	98.5	98.5	98.5	98.5
RNN without Head Data	98.4	98.4	98.4	98.4
DNN with Head Data	97.6	97.5	97.6	97.5
DNN without Head Data	97.2	97.2	97.2	97.2

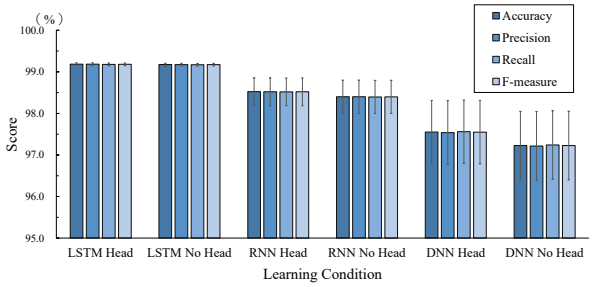

図 6: 実験参加者 10 人の各評価指標の平均値

1 人ずつのデータで学習，テストを行った結果，4 つの評価指標全てにおいて，LSTM，RNN，DNN の順番でスコアが高かった．また，標準誤差は LSTM，RNN，DNN の順番で低かった．最も精度の高かった学習条件は頭部姿勢ありの LSTM であり，最も低かった頭部姿勢なしの DNN を全指標で 2.0%上回った．時系列のみの影響を考えるため，頭部姿勢なしの 3 つの学習モデルの結果を比較すると，LSTM

© 2023 日本バーチャルリアリティ学会

は DNN を全指標で 2.0%上回り，RNN は DNN を全指標で 1.2%上回った．頭部姿勢に関しては，同じモデルで比べた時，LSTM は頭部姿勢の有無で精度の差はなく，RNN は全指標で 0.1%上回り，DNN は 4 つの平均で 0.4%上回った．実際のデモの様子を図 7 に示す．DNN は表情遷移の際に識別を間違えているが，表情の時間的変化を考慮する LSTM は遷移を正しく識別していることがわかる．

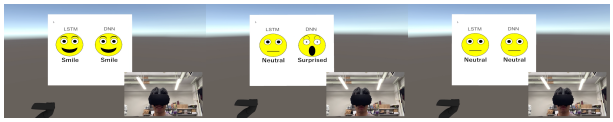

図 7: Smile から Neutral への遷移と各モデルの識別結果

6. 考察

6.1 時系列考慮の有無による識別精度の違い

実験結果から，時系列を考慮すると表情識別精度が上がることがわかった．これは表情遷移中の中間的な表情も，過去のデータを使うことで表情が予測できるためであると考えられる．時系列を考慮しない場合はその時刻のセンサ値のみで表情を識別しなければならないが，時系列を考慮する場合は，その時刻までのセンサ値の変化がわかるため，その時刻のセンサ値が中間的でも表情をより正確に識別できていると考えられる．

また，何らかの要因でセンサ値の一部が飛び値をとった場合，時系列を考慮しない識別はその飛び値が識別結果へ大きな影響を与える可能性があるが，時系列的な識別の場合は人の表情が非連続的に変化しないことを考慮できるため，飛び値に対してロバストである可能性がある．実際に，図 6 のエラーバーを見ると，LSTM は全参加者において約 99%の精度であるのに対し、DNN は精度に幅があることがわかる．この結果は、時系列学習が高精度かつ安定した表情の識別を可能にすることを示唆している。

6.2 頭部姿勢データの有無による識別精度の違い

同じ学習モデルで比較すると，LSTM, RNN, DNN の全てで，頭部姿勢を考慮する場合の方が精度が高かった．これは 3.2 で示したように，頭部姿勢によって反射型光センサの値は全体的に影響を受けるが，頭部姿勢がデータとして識別器に入力されていることで，識別器がセンサ値を補正して識別できることが要因と考えられる．頭部姿勢が無い場合は, センサ値の大小が生じる原因に関するデータを識別器が取得できないが，頭部姿勢をデータとして入力することで，全体的なセンサ値の大小に関するデータを識別器が参照できるため，精度に差が出た可能性がある．

7. 結論

本研究では，反射型光センサによる表情識別の際に表情データの時系列性を考慮し，頭部姿勢情報を新たに加えたマルチモーダルな学習を行った．反射型光センサの値を時系列データとして学習した結果，非時系列学習より識別精度を上げることができ，参加者ごとの精度の差が小さくなった．また，頭部姿勢情報を入力に加えると，時系列学習と非時系列学習のどちらでも僅かに精度が上がり，非時系列学習に対する精度向上効果は時系列学習よりも高くなった．

参考文献

[1] Katsuhiro Suzuki, Fumihiko Nakamura, Jiu Otsuka, Katsutoshi Masai, Yuta Itoh, Yuta Sugiura, and Maki Sugimoto. Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display. In *2017 IEEE Virtual Reality (VR)*, pages 177–185. IEEE, 2017.

[2] Fumihiko Nakamura, Masaaki Murakami, Katsuhiro Suzuki, Masaaki Fukuoka, Katsutoshi Masai, and Maki Sugimoto. Analyzing the effect of diverse gaze and head direction on facial expression recognition with photo-reflective sensors embedded in a head-mounted display. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[3] Masaaki Murakami, Kosuke Kikui, Katsuhiro Suzuki, Fumihiko Nakamura, Masaaki Fukuoka, Katsutoshi Masai, Yuta Sugiura, and Maki Sugimoto. Affectivehmd: Facial expression recognition in head mounted display using embedded photo reflective sensors. In *ACM SIGGRAPH 2019 Emerging Technologies*, SIGGRAPH '19, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Marcus Georgi, Christoph Amma, and Tanja Schultz. Recognizing hand and finger gestures with imu based motion and emg based muscle activity sensing. In *Biosignals*, pages 99–108, 2015.

[5] Xiaoliang Zhang, Ziqi Yang, Taiyu Chen, Diliang Chen, and Ming-Chun Huang. Cooperative sensing and wearable computing for sequential hand gesture recognition. *IEEE Sensors Journal*, 19(14):5775–5783, 2019.

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

© 2023 日本バーチャルリアリティ学会