JSiSE Research Report
vol.34,no.1(2019-5)

PC カメラ映像を利用した学習状態推定手法に関する研究

長谷川 忍[*1]，平子 温[*2]，卯木 輝彦[*3,*4]
*1 北陸先端科学技術大学院大学 情報社会基盤研究センター
*2 北陸先端科学技術大学院大学 先端科学技術研究科
*3 IMAGICA GROUP，*4 フォトロン

A Study on Learning State Estimation using PC built-in Camera

Shinobu Hasegawa [*1], Atsushi Hirako[*2], Teruhiko Unoki [*3,*4]
*1 Research Center for Advanced Computing Infrastructure, JAIST
*2 Graduate School of Advanced Science and Technology, JAIST
*3 IMAGICA GROUP, *4 Photron

The purpose of this research is to estimate learners' states, such as difficulty, interest, fatigue, and concentration, to their learning tasks through the built-in camera on the laptop/tablet PCs. We did an experiment which joined 19 learners, took the videos of facial expression in learning, and got parameters like emotion, eye gaze, head pose, face rectangle, mouse status, etc. using a vision library, Face++. We finally developed a neural network from the gathered parameters and compared the average accuracy of prediction by some traditional machine learning methods. The results show the potential of the proposed method and improvement plan from the accuracy point of view.

キーワード: PC カメラ，学習状態推定，ニューラルネットワーク

1. はじめに

ICT 活用教育の特徴の一つとして，非同期の e-learning や講義中の実習を通じたアクティブラーニング等といった，学習者それぞれの主体的な学習活動が重視される点が挙げられる．こうした self-directed(主体的)/self-regulated(自己調整)な学習環境では，学習者がそれぞれのペースで学習を行うことで学習プロセスが個別化されるため，個々の学習者の学習活動に対する意欲や取り組み方，理解状態などを把握しながら適切な教育的支援を行うことは容易ではない．一方，こうした ICT 活用教育における学習用端末として一般に利用されるノート／タブレット PC の大部分にはカメラが標準で内蔵されているが，遠隔会議等の特定の用途のみでしか活用されていないことが現状である．

本研究の目的は，こうした主体的学習活動において利用されるノート／タブレット PC に内蔵されているカメラの映像を活用して，学習プロセスにおける学習者の状態を分析するためのプロトタイプを開発することである．本研究における学習状態とは，学習時のタスクに対して学習者が感じる難しさや興味，疲労度，集中度などを指し，学習者と学習環境の間で時系列に従って変化する心的状態であると捉える．こうした学習状態を推定することは，e-learning 等の学習活動における質の高い学びを支援する上で重要であり，本研究では，PC 内蔵カメラ映像という外的状態からいかに学習状態を推定できるかがポイントとなる(1).

2. 関連研究

学習時の学習者の状態／行動全般を推定する手法としては，以下のように様々な研究が行われてきた．

1. 姿勢情報の利用：Yokoyama らの USB カメラ映像から学習者の机上姿勢を判定し，その時系列データに基づいて意識状態を推定する手法(2)や，手塚らの圧力センサや距離センサを用いた姿勢計測による e-learning 受講者の行動推定手法(3)，山本らの Kinect を用いた頭と手の 3 次元座標から学習者の行動を推定する手法(4)など．
2. 視線情報の利用：坪倉らのアイカメラによる視線

行動に基づく学習者の習熟度把握手法[5]や，講義中の受講者の撮影映像から受講者が前を向いている比率を特徴量として講義への関心度を推定する手法[6]など．

3. 映像に基づくエンゲージメントの推定 : Grafsgaard らの Web カメラと Kinect カメラによる PC 前の学習者の顔動作符号の推定に基づく感情やパフォーマンスとの関係づけの調査[7]，Whitehill らの Web カメラによる iPad 上の認知スキルトレーニングにおける SVM による推定[8]，など．
4. その他のセンサ情報の利用：竹花らの脳波計や脈波計などから収集した生体信号による心的状態の推定[9]や，李らの腕に装着するモーションセンサを利用した学習活動の推定手法[10]など．

本研究では市販のノート／タブレット PC を活用した主体的な学習プロセスにおいて，追加コストなしで利用可能な PC 内蔵カメラの映像ストリームを用いるアプローチを採る．また，学習者のエンゲージメントは，学習内容や作業負荷等によって時々刻々変化するため, ある瞬間の学習者の状態を推定するだけでなく，状態の変化をきっかけとして近い将来の学習意欲の増減を予測したり，同様の傾向を持つ学習者群をグルーピングしたりするといったことが期待される．本研究では，こうした学習中の映像ストリームに対する時系列データの分析・推定を実現するために準備したデータセットについても報告する．

3. データ収集

学習中の顔映像から学習者のエンゲージメントを推定する研究はいくつか存在しているが，ベンチマークとなるデータセットは存在していない．そこで本研究ではまず推定を行うためのトレーニングデータおよびテストデータとして利用するデータセットの収集を行った．

3.1 CAB テスト

本研究では学習課題として，日本エス・エイチ・エル社製の就職試験問題 CAB の「法則性」模擬問題[11]を使用した．この問題はある法則性に基づいて並んだ 5 つの図形の内，1 つだけ欠けている物があり，これを残り 4 つの図形から法則性を予想して合致する図形を 5 つの選択肢から選ぶという課題である．この例題を図 1 に示す．

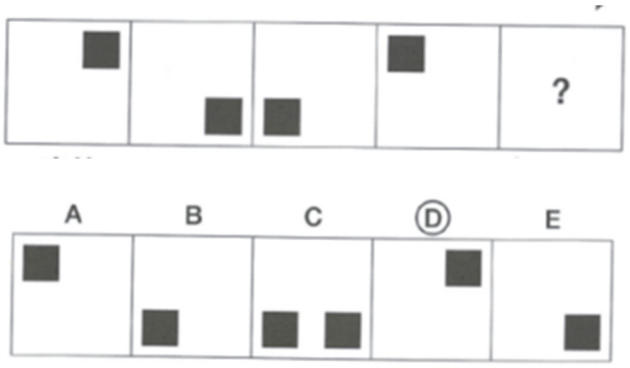

図 1 CAB テストの例[11]

この課題の選定理由は以下の通りである．まず表情から心的状態に関連する特徴量を時系列的に抽出するという性質上，事後アンケートで収集する学習状態について時系列の単位が長すぎると実験協力者が正常に思い出せない可能性があり，逆に時系列の単位が短すぎれば実験協力者の負担が重くなりすぎることが考えられる． CAB の法則性問題は，他の「暗号」や「命令表」,「暗号解読」等の CAB 問題と比較して問題毎の回答時間のばらつきが少なく，SPI 等の模擬試験と比較して回答時間が長すぎないことが挙げられる．

さらに，他の問題は 1 つの出題に対して問が 3 つ有るなど，アンケートにおける問題の指定において混乱が生ずる可能性があった．例えば，CAB の暗号解読問題では，ある図形を変形する系統図に対してその系統がいかなる処理を行っているかを推測し問いに回答するが，1 つの系統図に対して 3 つの問いが存在するため，問題全体に対する難易度と，それぞれの問いに対する難易度が混同されやすいという課題があった．

3.2 データ収集の手続き

本研究におけるデータ収集の手続きは以下の通りである．

1. 実験協力者に実験の流れの説明
2. カメラ付きノート PC の前に着席
3. CAB 問題の低難易度の例題の回答
4. PC のインカメラを用いて実験協力者の顔を写した動画（以下，顔動画）と回答中の PC の画面を写した動画（以下，PC 動画）の撮影の開始
5. CAB 問題の回答（問題は全部で 30 問，制限時間を

12 分とし，それ以降は残りの問題が有っても回答を終了）

6. 回答終了後，実験協力者にアンケートの説明
7. 顔動画および PC 動画の撮影の終了
8. 上記の手順 4 で回答した問題について，各問題と実験協力者の回答および正誤，正解と解説を見せ，各問題の「難しさ」「面白さ」「疲労度」「集中度」について 1〜5 の 5 段階で評価

なお，本データ収集プロセスで用いた機材は Lenovo 製 E560 ノートパソコンであり，そのスペックは以下の通りである．

- CPU : Intel Celeron 3855U 1.60GHz
- RAM : 8 GB
- HDD : 1 TB
- モニタ ：15.6 型フル HD
- Web カメラ ：2D カメラ搭載

3.3 動画データからのパラメータ抽出

3.2 節で取得した顔動画から学習者の心的状態を推定するにあたって，動画像を直接利用するにはトレーニングデータが不十分である．そこで，機械学習のための素性をいかに抽出するかは重要な課題である．本研究では，既存の顔画像認識ライブラリを利用してパラメータを抽出することとし，感情推定，目線，顔の向き，位置（画面に対するパーセンテージ），眼鏡の有無，性別等のデータを抽出可能であった Face++[12]を利用することとした．顔動画からのパラメータ抽出手順を図 2 に示す．

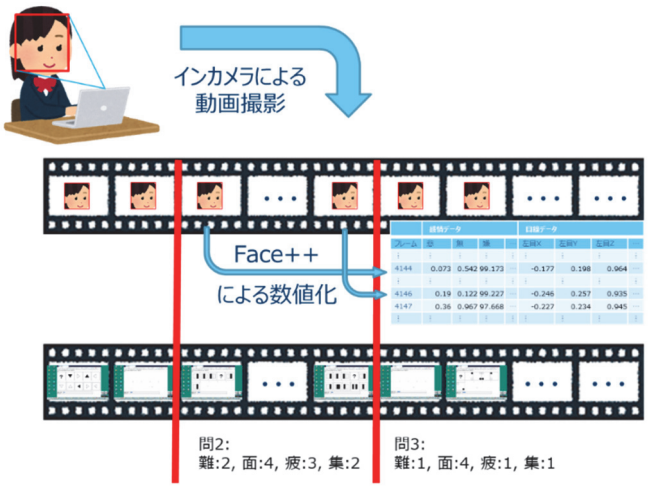

図 2 顔動画からのパラメータ抽出手法の概要

1. 入手した顔動画フレームの静止画化
2. 手順 1 で得られた静止画に対して Face++によるパラメータの取得
3. PC 動画から実験協力者が各問題を解いているフレームを計測
4. 手順 3 で得られたフレームを元に顔動画における各問題回答時のフレームのデータセットに分解
5. 手順 4 で得られたデータセットをアンケートと紐づけ

3.4 データ収集

実際のデータ収集は 2018 年 6 月〜2018 年 12 月の間に実施した．実験協力者として 19 名（男性 13 名、女性 9 名、年齢 20〜46 歳）が参加した．動画フレームに顔が写っていない等の理由で Face++による分析が行えなかった表情データや，アンケートが未回答のデータ等を削除した結果，表情データとアンケートの有効なデータセットとして「難しさ : 453 問」「面白さ : 453 問」「疲労度 : 452 問」「集中度 : 285 問」を得た．

4. 推定モデルのプロトタイプ開発

4.1 開発環境

本研究では，分析モデルのプロトタイプ開発を行うライブラリとして TensorFlow[13]と Keras[14]を使用した．TensorFlow はニューラルネットワークの設計において高い自由度を持つことが特徴である．Keras はそれ単体が機械学習を行うライブラリではなく，TensorFlow や Theano といった別のライブラリ上で動作するもので，下位のライブラリに渡すコードが関数としてまとめられている．このため，ニューラルネットワークを構成する上で簡便な記述で済み，画像分析を意識した関数を多数備えている事から汎用性が高いという理由で採用した．

4.2 ニューラルネットワークによる推定

ニューラルネットワークとは，人間の脳神経系の働きを数理モデルにしたもので，同様の動作を行わせることで人間と同じような問題解決能力をもたせようとするものである[14]．具体的には，複数の入力値に対して重み係数を掛けた合計値を計算し，この合計値が閾値を超えたときに出力するパーセプトロンと呼ばれる

関数をネットワーク状に組み合わせたものをニューラルネットワークと呼ぶ．近年では，ニューラルネットワークの技術を拡張した DNN（ディープニューラルネットワーク・深層学習）や CNN（畳み込みニューラルネットワーク）などの技術により，従来分類が困難だったデータも分類ができるようになったことで注目を集めている．

第 3 章で得られたデータをニューラルネットワークに学習させるに当たり，入出力の数を一致させる必要がある．実験協力者のそれぞれの問題にかかる回答時間は異なるため，得られたデータは実験協力者と問題毎に異なっている．そこでここでは，1 問の回答に含まれる全てのフレーム内データの各系列（表情の怒りの成分の数値や目線の X 方向の数値など）に対して「平均値・分散値・最低値・最高値」の 4 つの統計値を取ることで 1 個の入力データに圧縮した．最大値や最小値を利用したのは，驚きや嫌悪などの一瞬しか現れない表情だと平均値に反映されにくいためである．また，分散については，例えば目線が左右に繰り返し動いている状態と目線が中央から全く動いていない状態では，双方とも目線の平均値が中央になってしまい区別がつかないためである．

次に，圧縮後のデータをそのままニューラルネットワークに用いた場合，ニューラルネットワークで使用する活性化関数の関係上，極端な値の振れを起こしてネットワーク内の各ノードの振る舞いが単純パーセプトロンのような動きになり学習を阻害する恐れがある．そのため，圧縮後のデータにおける各系列に対して，図 3 に示すような正規化を行った．

図 3 データ正規化の例

本研究では以下のような条件で NN（ニューラルネットワーク）の学習を行った。

- 入力層ノードは 83 個（入力データの次元と同数
- 隠れ層はシグモイド関数を用いノードは 83 個（入力データの次元数と同じ）
- 出力層はソフトマックス関数を用い，ノードは 5 個（5 段階評価に合わせて）
- 学習率は 0.2
- 隠れ層の層数を 0～5 層に変更
- 30 個をランダムにテストデータとして抽出し，残りをトレーニングデータに割当
- 抽出テストデータは 50 回変更し，テストデータごとの正解率の平均値を算出

NN における Confusion Matrix を図 4 に示す．ここでは正解を青色，不正解のうち推定値が 1 違いのものを黄色，それ以外の不正解を赤色に着色している．

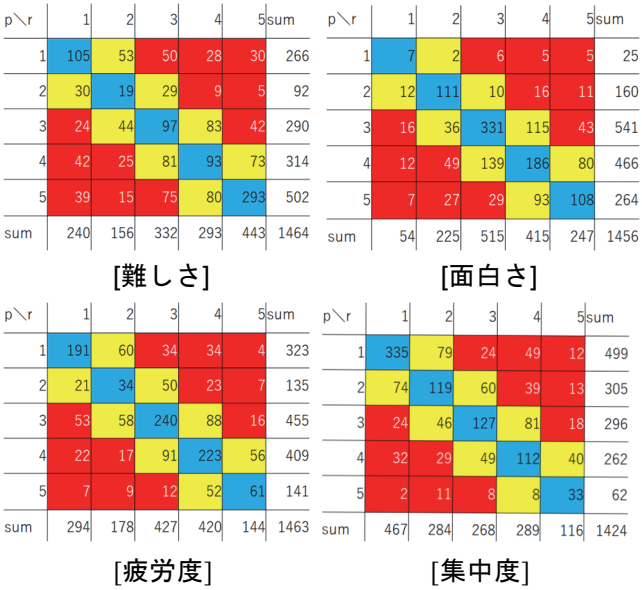

図 4 NN 条件における Confusion Matrix

加えて，SVM（サポートベクターマシン）と KNN（k 近傍法）を用いて同一のデータセットからアンケートデータを推定した際の正解率を比較したものを表 1 に示す．この 2 法についても，テストデータの抽出をランダムに 50 回行い，各テストデータの正解率の平均値を比較した．また，KNN 法では K 値（近傍個数）を 1～199 まで変化させ，最も高い平均正解率を比較の対象とした．

表 1 各手法の平均正解率の比較

手法	難しさ	面白さ	疲労	集中度
NN	0.42 (隠れ 2 層)	0.55 (隠れ 0 層)	0.52 (隠れ 2 層)	0.51 (隠れ 0 層)
SVM	0.32	0.50	0.42	0.52
KNN	0.38 (k=4)	0.53 (k=18)	0.51 (k=1)	0.66 (k=1)

表 1 より,「集中度」以外は NN が最も高い平均正解率となっている．また，NN においてはすべて隠れ層 2 層以下という特徴が見られた.

4.3 統計的手法による分析

4.2 節にて述べた正規化済み表情データの各系列に対して，アンケートデータとの関連性を分析するために，MIC(Maximal information coefficient) [15]を用いて非線形相関係数を算出した結果の概要を表 2 に示す.

表 2 MIC による非線形相関係数の概要

難しさ	MIC > 0.3	—
	MIC > 0.2	怒り，恐れ，右目線 X 軸，右目線 Z 軸，左目線 X 軸，左目線 Z 軸，顔の幅，顔の高さ
面白さ	MIC > 0.3	右目線 Z 軸，左目線 X 軸，左目線 Z 軸
	MIC > 0.2	不快さ，驚き，恐れ，右目線 X 軸, 右目線 Y 軸, 左目線 Y 軸, 顔の幅，顔の高さ，頭部ピッチ, 頭部ロール, 頭部ヨー, 閉口
疲労度	MIC > 0.3	—
	MIC > 0.2	怒り，幸せ，右目線 X 軸，右目線 Y 軸，右目線 Z 軸，左目線 X 軸, 左目線 Y 軸, 顔の幅, 顔の高さ, 頭部ピッチ, 頭部ロール, 頭部ヨー, 閉口, 他の口
集中度	MIC > 0.3	右目線 Y 軸，右目線 Z 軸，顔の幅，顔の高さ
	MIC > 0.2	悲しさ, 無表情, 不快さ, 驚き, 幸せ，右目線 X 軸，左目線 X 軸, 左目線 Y 軸, 左目線 Z 軸, 頭部ピッチ, 頭部ロール, 頭部ヨー，開口，他の口

表 2 より，全ての心的状態に対して，目線の項目に関する弱い相関係数が確認されている. このことから，より精度の高い予測を行うためには，目線のデータを重視すべきであると考えられる.

4.4 考察

機械学習の分野，特にニューラルネットワーク関連の研究においては使用するデータは何万個単位である場合が多い. 本研究において入手したデータは約 30 万個であるが，問題数の観点からは 453 個に過ぎない. この課題の解決策としては，調査の内容を専用の Web アプリなどを開発して自動化し，インターネットを通じて協力者を募るなどして幅広くデータを集めることなどが考えられる．ただし後述のデータの質の問題と引き換えになることに留意しなければならない.

今回の実験を通じて，アンケート項目の定義や基準に対する質問があったから改善が必要である．特に課題となるケースとして，アンケートにおける難しさの項目は調査協力者の主観による難しさを表現するものを想定していたが，ある協力者は回答開始前に見せる低難易度の例題を基準として難易度の判断を行っていた．このような課題を防ぐためには，アンケートの項目について，用語の定義を更に厳密化することや，実際にアンケートに回答する際の例示を準備するなどして回答基準の個人差をできるだけ小さくすることが必要である.

また，アンケートの回答時の心理状態と実際の回答中の心理状態に差がある可能性も考えられる．本研究では全ての CAB 問題を回答した後にそれを思い出しながらアンケートを書くこととなっていたが, この際, 思い出す為の補助として見せていた回答や解説が無意識下で記憶に影響を与え，記憶が変質した可能性がある．例として，調査協力者が回答中にある問題を難しいと感じたが，アンケート回答中に解説を見たことでアハ体験を起こし，この結果難易度の低い問題と誤認してしまい，これが無意識下で行われるため後の聞き取りでも発覚しないというものである．調査の方法をアンケートではなく聞き取りにし，調査協力者が中立的な聞き取り調査の手法について技能を持つことでこの問題を解決できるとしているが，これは調査にかかる労力との兼ね合いを考慮する必要がある.

正答と予測を組み合わせた Confusion Matrix により，表情からアンケート結果を正しく推定できなかったデータであってもその誤差範囲が 1 である「惜しい」データが多く，完全に予測ができていないわけではないことが明らかになった．また，本研究ではニューラル

ネットを用いたアンケートデータの予想において，ニューラルネットワークの出力層にソフトマックス関数という離散的なクラス分類に用いる関数を使用しているが，これを連続的な出力にした場合予測と正答の誤差が縮まるのではないかと考えられる. しかしながら，いかなる出力層・活性化関数が適当なのかという問題と，出力が連続値の場合，単純に正解・不正解を区別することが出来ないため，いかなる方法によってモデルの精度を評価したり，他の機械学習手法と比較したりするかという問題が有り，定性的・定量的な評価基準を新たに開発しなければならない.

本研究では，NN の出力層にソフトマックス関数を用い 5 クラス分類問題として解いている．前述した通り，誤差範囲が 1 である「惜しい」データが多い事から，クラス分けを減らすことで相対的に正解率を上げることが可能である．本来の目的である学習者の心的状態の変化のタイミングを検出することが目的であれば, その変化の程度を厳密に測定する必要はなく, 3 クラス程度でも効果が期待できると考えられる.

5. おわりに

本研究では，学習中の表情のデータから学習者の心的状態を推定するという目的を達成するため，学習状態と表情には一定の関係性があるという仮定の元，調査協力者 19 名に対し，CAB 模擬問題を PC で解いてもらい，その作業中の表情と PC の画面を動画で撮影した．CAB 模擬問題の回答終了後に各問題に対しての「難しさ」「面白さ」「疲労度」「集中度」といった学習状態に関係・影響すると思われる 4 つの項目についてアンケートを取り回答を得た.

このようにして得た表情のデータとアンケートのデータに対し，ニューラルネットワークを用いて表情のデータからアンケートのデータを推定することができるかを実験した．その結果，それぞれのアンケート項目が推定出来ているかどうかの指標となる正解率はそれぞれの項目の最大値で「難しさ:0.41」「面白さ:0.55」「疲労：0.522」「集中度：0.505」であった.

今後の課題としては，推定のためのデータセットを増やすとともに，時系列データを活用した推定手法を適用することが考えられる．RNN(Recurrent Neural Network・再帰ニューラルネットワーク)は，隠れ層において過去に入力されたデータを記憶し，不正解時の勾配（修正量）の計算を遡って行うことで時系列的なデータの予測・分類が可能なニューラルネットワークの一種である．RNN の特徴として「時系列データの分類ができる」事と「入出力の数が一致していなくても良い」という特徴がある．この特徴は自然言語処理において文章を入力して分類や連続する文書の生成をさせたり，音声解析において各周波帯の音の大きさを入力として音声認識や続く音声の予想をさせたりするなどの形で利用されている．RNN に対して 4.2 節で加工する前のデータが動画の連続したフレームという時系列なデータであるため，これをそのまま RNN に入力することでアンケート結果を予想することができると期待される．そこで，本研究では RNN の一手法であり，過去の入力値を隠れ層に記憶することで時系列的予測を行うとともに, LSTM の入力・出力・忘却を 1 つに統合して簡易化した手法である GRU（Gated Recurrent Unit）を利用して推定を行うことを検討している.

参 考 文 献

(1) 長谷川 忍，卯木 輝彦：PC 内蔵カメラを利用した学習者のエンゲージメント分析に関する検討，2018 年度人工知能学会全国大会（第 32 回）講演論文集, 4H1-OS-9a-04，(2018).

(2) N. Yokoyama, T. Yamaguchi, and S. Hashimoto: "Care Giving System Based on Consiousness Recognition", Human Interface, Part I, HCII 2011, LNCS 6771, pp.659-668, (2011).

(3) 手塚太郎，清野悠希，古谷遼平，佐藤哲司: "姿勢計測による e-learning 受講者の行動推定"，知能と情報 28(6), pp.952-962, (2016).

(4) 山本千尋，天野直紀: "Kinect を用いた学習行動計測システムの研究"，情報処理学会第 75 回全国大会講演論文集, pp.585-586, (2013).

(5) 坪倉篤志，松原伸人，林敏浩，西野和典 : "視線行動を用いた対話型学習環境における学習者習熟度 : 対話型環境の構築と対話分解能"，電子情報通信学会技術研究報告 114(441), pp.33-38, (2015).

(6) 村井文哉，角所考，小島隆次，村上正行 : "授業映像に基づく雰囲気認識のための基本特性と観測特徴量，教育システム情報学会誌 32(1), pp.48-58, (2015).

(7) 竹花和真，田和辻可昌，村松慶一，松居辰則 : "学習時における学習者の生体情報と心的状態の関係の形式化の試み"，人工知能学会研究会資料 SIG-ALST-B501-07, pp.34-39, (2015).
(8) 李凱，熊崎忠，三枝正彦: "モーションセンサを用いた学習活動の状態推定手法の開発", 教育システム情報学会誌 33(2), pp.110-113, (2016).
(9) Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester: Automatically Recognizing Facial Expression: Predicting Engagement and Frustration, Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp.159-165, (2013).
(10) Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan: The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions, IEEE Transactions on Affective Computing, Vol.5, No.1, pp.86-98, (2014).
(11) SPI ノートの会, '20　必勝・就職試験！CAB・GAB 完全突破法！, 洋泉社, (2018).
(12) Megvii Technology Inc, “Face++ Cognitive Services”, https://www.faceplusplus.com/. (2019 年 4 月 8 日確認)
(13) Google, “TensorFlow の概要,” https://www.tensorflow.org/. (2019 年 4 月 8 日確認).
(14) Keras, “Keras Documentation,” https://keras.io/ja/. (2019 年 4 月 8 日確認).
(15) David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti: Detecting Novel Associations in Large Data Sets, Science Vol. 334, Issue 6062, pp.1518-1524, (2011).