ビデオ会議における参加者の表情分類とフィードバック手法

神場知成[†1]

概要：コロナ禍でオンラインのビデオ会議が急増したが，対面のコミュニケーションと比較してビデオ会議で伝わる情報量は少なく，参加者が居心地の悪さを感じる場合も多い．本論文では，参加者の表情から感情を分析して参加者にフィードバックするシステムについて述べる．感情の分析を深層学習のモデルを用いて行うには学習用のデータセットを必要とするが，表情に感情がタグ付けされたデータセットはそれほどなく，特に最近のような端末正面にセットされたカメラから映像を取得するビデオ会議に特化したものは見当たらない．プライバシーなどの面から多くのデータセットが公開されるとも考えづらく，むしろ会議をする人が場面に応じて容易に独自のデータを利用できることが望ましい．今回，一般のビデオ会議ツール（Zoom，Google Meet 等）を利用しながらその参加者の表情の分析結果を表示するとともに，データの収集，タグづけおよび結果修正を容易に行う機能を開発し，基礎的な有効性を確認した．さらに，表情分析結果の表示方法についても検討を行い，数値グラフ，アイコン化などの試みを行った．どのような利用局面において誰にどのようなフィードバックを行うかについてはさまざまなケースが考えられ，今後の検討課題である．

1. はじめに

コロナ禍で Zoom，Google Meet，Microsoft Teams などのツールを用いたインターネット上のビデオ会議が急増し，この流れは定着すると予想できる．しかし，対面の会議と比較してビデオ会議では特に非言語情報が伝わりづらい傾向があり，心理的な負担を感じる人も多い．筆者は従来から，ビデオ会議の映像から参加者の感情分析を行い，それを参加者にフィードバックするシステムについて検討を進めているが[1]，今回その実用性を高める機能を開発し，実験を行った．

2. 従来の研究と本論文の着眼点

対面で働く人々がオフィスにおいて活発なコミュニケーションをすることで，特にクリエイティブな生産性が上がることを示した研究として，Weber によるもの[2]や，矢野によるもの[3]がある．対面でのコミュニケーションの重要性は明らかであるが，コロナ禍で対面コミュニケーションが制限されるなか，ビデオ会議の利用は大きく広がった．ビデオ会議は対面と比較した制約も多いものの，地理的や時間的な制約が大幅に少ないことによるメリットも多く，今後もかなりの利用が定着するであろうことを考慮すると，技術的に従来のビデオ会議を補う手法を開発していくことは重要である．

ビデオ会議システムにおいて参加者の様子を見える化して会議を支援するものとして，Samrose 等によるもの[4]や Murali 等によるもの[5]がある．Samrose 等は，会議システムの画面上に音声認識にもとづく自動議事録に加え，各ミーティング参加者の発言量，声のトーンの状況，などを表示してそれを参加者が見ながら会議をすることができるようにし，会議において有効だったことを報告している[4]. Murali 等は，オンライン会議システム上で発表者がプレゼンテーションを行っているときに参加者の反応を見やすくするためのシステムとして，参加者の表情を自動認識し，15 秒ごとにもっとも特徴的な表情をした人の映像をスポットライトして画面上に表示する試みを行い，発表者から「より観客を意識し，その表情からの非言語的フィードバックにもとづきプレゼンテーション中にも対応しやすかった」等の反応があったことを述べている[5].

この 2 つの例は全員が相互に話をするミーティングと，誰かが一方的にプレゼンテーションをするという，ビデオ会議の異なる利用状況を想定しており，ビデオ会議における参加者の感情推定や表情のフィードバックという課題だけでも，さまざまなケースがあることを示しており興味深い．この 2 つの例以外にもさまざまな状況が想定される．たとえば「少人数の会議で全員がビデオをオンにしており，全員の表情がよく見えるケース」「そのなかで一部の人がビデオをオフにしているケース」「大人数であり，仮にビデオがオンでも，お互いに全員の状態が把握しづらいケース」「聴衆が多い講演などで自分が発表者であり，聴衆のビデオがオンでもオフでも聴衆の雰囲気が把握しづらいケース」などである．さらに，「参加者は少人数だが，画面共有される資料が会議画面の大部分を占めており，参加者の様子はほとんどわからないケース」などがある．たとえば「多人数の聴衆を対象とする講演で，参加者の様子がほとんどわからずプレゼンテーションがやりにくい」というケースなどは，多くの人が経験しているであろう．

筆者は[1]において参加者の表情分析とそのフィードバック（表情分析値のグラフ表示）を行うシステムを示した

†1 東洋大学　情報連携学部

© 2023 Information Processing Society of Japan

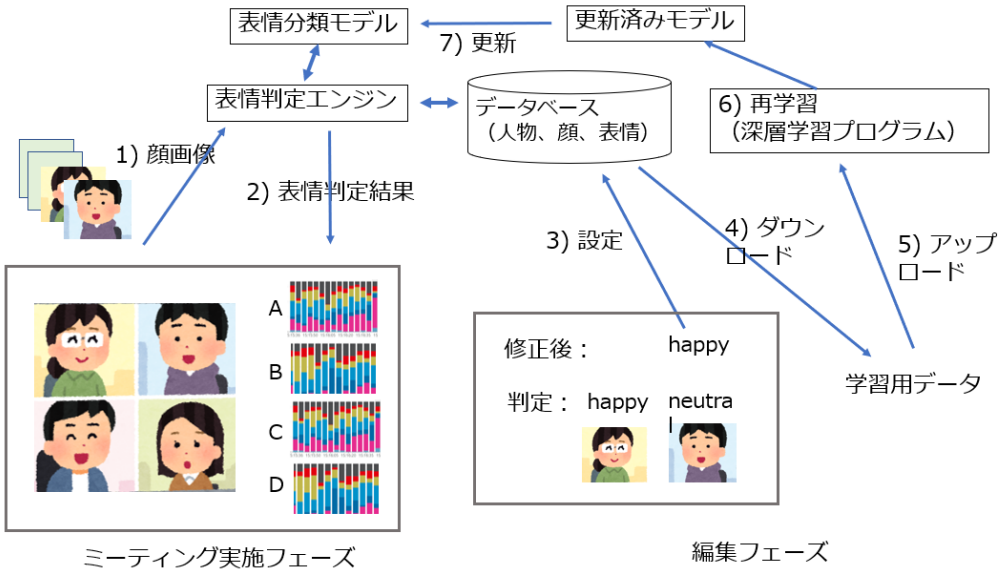


図 1. 本システムの構成

が，深層学習を用いた表情分析があまり正確ではないという課題や，分析結果を参加者にフィードバックする方法の検討が不十分であるという課題があった．表情分析が不正確な理由としては，深層学習のための学習データとして表情の映像に感情のタグ付けをしたものが少ないということが考えられた．顔画像に表情タグをつけたデータセットは存在し，たとえば上記システムでも用いていた，インド映画における俳優 100 人の画像 34512 枚と，それぞれの画像に対して人手によって 6 つの感情表現（Anger, Happiness, Sadness, Surprise, Fear, Disgust）が結びつけられた大規模データセット[6]などがあるが，日本人のものは見当たらないということや，特に会議中の顔画像，しかもビデオ会議中の顔画像はほとんどない．現在多く用いられるビデオ会議は，PC 等の正面カメラで撮影したものが多く，顔画像の向きもビデオ会議特有のものである．プライバシーの観点も考慮した場合，ビデオ会議中の顔画像が，利用可能なかたちで多数出てくるとも考えづらく，むしろ会議システムの利用者が，自分たちが利用する範囲内でみずから簡単に情報を収集して，それを分析ツールに容易に利用可能とする方が，有用性が高いであろう．

上記の考えにもとづき，一般的に用いられるビデオ会議システム（Zoom, Google Meet 等）を使いながら簡単に参加者の画像データを収集し，自ら設定した表情分類を深層学習のデータセットとして利用可能なシステムを実装した．

3. 実装システム

3.1 システム概要

図 1 に実装システムの構成を示す．ミーティング中，クライアント端末から一定時間ごとにビデオ会議画面の画像データがサーバに送信され (1)，サーバ上で顔検出，分類，表情分析を行った結果が本システムの画面に表示される (2)．ミーティング後に，システムが表示した表情判定結果が不適切な場合はそれを修正し(3)，修正後のデータもとに深層学習のモデルを更新できる（4〜7）．

Layer (type)	Output Shape	Param #
conv2d_4	(None, 64, 64, 32)	320
activation_6	(None, 64, 64, 32)	0
conv2d_5	(None, 62, 62, 32)	9,248
activation_7	(None, 62, 62, 32)	0
max_pooling2d_2	(None, 31, 31, 32)	0
dropout_3	(None, 31, 31, 32)	0
conv2d_6	(None, 31, 31, 64)	18,496
activation_8	(None, 31, 31, 64)	0
conv2d_7	(None, 29, 29, 64)	36,928
activation_9	(None, 29, 29, 64)	0
max_pooling2d_3	(None, 14, 14, 64)	0
dropout_4	(None, 14, 14, 64)	0
flatten_1	(None, 12544)	0
dense_2	(None, 512)	6,423,040
activation_10	(None, 512)	0
dropout_5	(None, 512)	0
dense_3	(None, 4)	2,052
activation_11	(None, 4)	0

図 2. 表情認識のモデル

システムはコンテナ仮想化のプラットフォーム Docker（https://www.docker.com/）上で，下記の構成から成る．

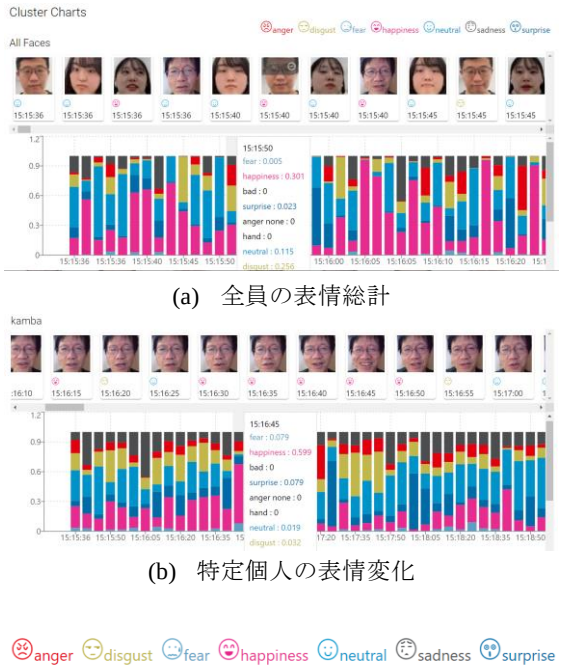
- ビデオ会議ツール画面をキャプチャし，一定時間ごとにその画像をサーバ送信する Chrome 拡張プログラム [7]．現在，画面は 10 秒ごとに取得している．
- 受信した画像から，顔の切り出し，分類，モデルに基づく表情認識を行って，表情のパラメータ値を JSON 形式でデータベースに登録する API サーバ. Python スクリプトで構成され，実装のフレームワークとして FastAPI を利用（https://fastapi.tiangolo.com/ja/）. 顔からの特徴抽出には Google が開発した顔認識モデル FaceNet [8]を用い，利用者が会議の参加者人数をあら

© 2023 Information Processing Society of Japan

かじめ指定した場合の顔分類は K-means 法で行い，その他の場合は X-means 法で行う．表情認識は，図 2 に示すような CNN(Convolutional Neural Network)の簡易モデルで事前に学習させたもの（hdf5 形式）を用いる．学習済みの h5 フォーマットファイルを入れ替えるだけで異なる表情認識モデルを用いることができるので，たとえば会議参加者ごとに異なる表情認識モデルを簡単に切り替えることができる．

- UI を制御するフロントエンド（フレームワーク React を利用）．サーバでの表情分析の結果として各画像に対して表示された表情分類が違っていると感じた場合に，ユーザが正しいと感じる分類を設定する機能を持つ．設定しなおした画像と表情とのセットは，深層学習用のネットワークに学習させる形式のデータセットとしてダウンロード可能．
- ダウンロードデータを利用してバッチ処理で表情の学習を行い分類モデル（h5 フォーマット）を作成するためのスクリプト（Google Colaboratory 等で動作可能）．学習モデルは前述の図 2 の通りである．

3.2 ユーザインタフェース



(a) 全員の表情総計

(b) 特定個人の表情変化

(c) 表情の一覧

図 3. 参加者表情の変化グラフ

図 3 に利用者画面の例を示す．図中に表示されていないが，前画面ではこのツールを用いて行ったすべての会議の一覧が表示されており，そのうちの一つを選択することで，この画面表示になる．(a)，(b)いずれも画面上部には，会議画面から自動検出した参加者の顔が時系列で並んでおり，各顔画像には時刻と，表情検出結果がアイコンで示されている．グラフ中の各表情に該当する色は，表情アイコンの色と一致している．

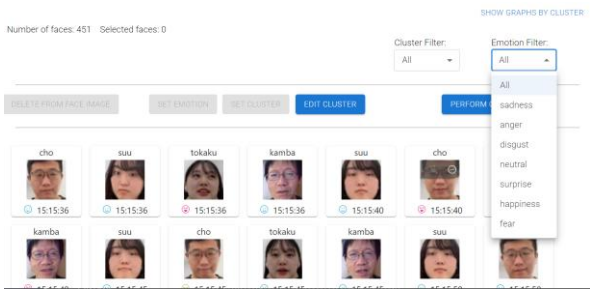


図 4. 顔検出，表情分類の編集画面

図 3(a)のグラフは，参加者全員の 7 つそれぞれの感情の表情値を合計したものを，横軸を時間とする積み上げ棒グラフで示している．つまり，ある時刻において，全員の happiness（楽しさ），neutral（中立）など感情の数値を合計したときの，感情ごとの比率を表示している．色分けをしているので，見た時に何の色が多いかで，そのときの会議の雰囲気がわかる．グラフ上で特定の時刻にカーソルを置くと，その時刻の各表情のパラメータ値が表示される．特定の時刻の表情のパラメータを見て，該当する顔画像を確認したいときには，時刻が一致する画像を見れば良い．

表情を示すアイコン一覧を図 3(c)に示す．表情は，Ekman 等が社会の文化等とかかわりなく普遍的に人間の表情にあるとした 6 分類「喜び，悲しみ，驚き，怒り，嫌悪，恐れ」[9] に「中立」を加えた 7 分類とし，それぞれアイコンを割り当てた．アイコンは「いらすとや」の表情マークを用いている(https://www.irasutoya.com/2013/03/blog-post_25.html)．表情をアイコンとして示しているのは，後に認識結果としての表情変化だけをアニメーション等で示すことを想定したためである．「中立」を加えた理由は，会議などもっとも多いのは，Ekman が示した 6 つの感情のどれも強く示していない状態と考えられるからである．

図 4 は，顔分類結果や表情認識結果の編集画面である．画面右上に「Cluster Filter」「Emotion Filter」というボタンがあり，それぞれ「特定の参加者の表示」「特定の表情の表示」の選択機能を持つ．分類結果がおかしいと感じた画像があればそれを選択すると，図 4 ではグレー表示されている「Set Cluster」または「Set Emotion」のボタンの色が変わってアクティブになるため，それを押せば，正しい分類に選択し直すことができる．顔画像表示領域には，まれに OpenCV により顔画像以外のものが顔と誤判定されて表示されることがあり，その場合はその画像を選択し，「Delete from Faces」のボタンを押せばよい．

上記において，画像の表情分類結果を修正したときは，その時点では他の画像の表情分類結果に影響を与えないが，

© 2023 Information Processing Society of Japan

メニューからこの会議のデータセットをダウンロードすると，深層学習用のデータセットとして，取得した画像および修正済みの表情のタグづけ結果がすべてダウンロードされ，そのまま学習に利用可能である．そのデータを用いて学習しなおした深層学習のモデル（h5 フォーマット）をサーバ側に設定すれば，次回以降の会議における参加者表情は新しいモデルにもとづき分類される．

4. 評価と考察

4.1 顔表情の学習と分類表示

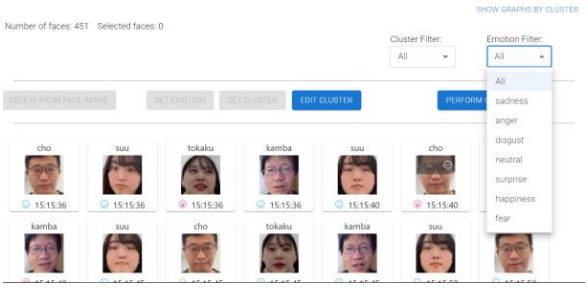
顔分類をするにあたり，初回のミーティングの利用ではすべてを人手で分類する手間を軽減するため，前述したインド映画の俳優表情とその分類結果のデータセット[6]を用いて分類用のモデル（.h5 フォーマット）を作成し，表情分類を行った．その表示結果において筆者自身の主観で誤りと思われるものをすべて修正した上で，深層学習の学習用データセットとして用いて学習を行ってモデルを作成し，そのモデルを利用して再度ミーティングでの利用を行った．いずれもメンバーは筆者を含む共通の 4 名だが，ミーティングの途中でマスク着用と非着用を切り替え，その 2 種類はデータを分けて学習に用いた．

表 1.初回ミーティング（マスクなし）表情の修正数

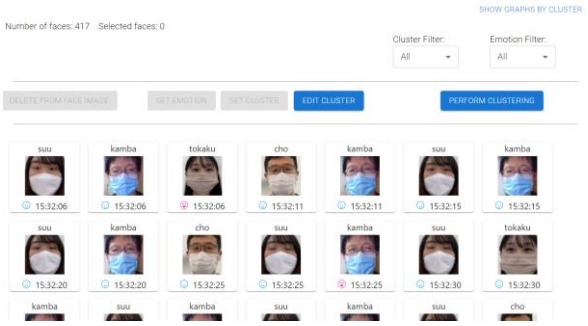
	全数	人 A	人 B	人 C	人 D
検出顔画像数	451	95	108	103	145
うち，表情修正数	164	29	26	78	31

初回は 4 人のメンバーによりマスク着用状態，非着用状態でのミーティングをそれぞれ約 15 分間行った．検出された顔画像数は，マスクなしのミーティングが 451，マスクありのミーティングが 417 である．この違いはミーティングの実施時間中の違い，顔以外のものを顔として誤検出したものを除いた個数（いずれも少数だが発生）の違いによる．いずれの場合も，4 人の顔を自動分類した結果，人物分類の誤りは 1 件もなかった．次に，それぞれの顔に対して判定された表情を，筆者自身の主観的評価で修正した．作業中の画面例を図 5 に示す．マスクなしの場合の表情判定結果を修正した件数は表 1 の通りであり，全体で 40%の修正を行っている．修正作業を主観的に言っても，「かなり当たっているものも多いが，誤りも多い」というレベルである．また，マスク着用状態に対する判定結果については，筆者自身が見てもかなり判定しづらく，修正作業も非常に大雑把なものとなったのでその内容については後述する．

このデータをもとに前述の深層学習を用いて作成したモデル（h5 フォーマット）を用いて，第 2 回のミーティングを行った．学習時の学習曲線を図 6 に示す．このモデルを用いて行った表情判定を第 1 回と同様に修正した結果を

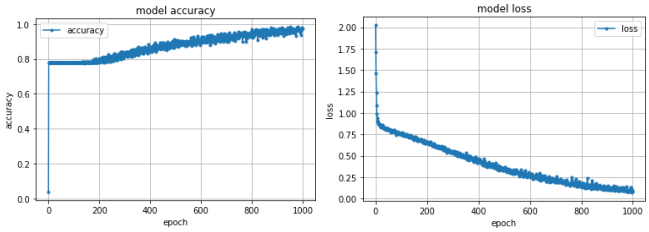


(a) マスクなし



(b) マスクあり

図 5. 表情検出結果と編集作業画面



(a) accuracy (b) loss

図 6. 表情学習時の学習曲線

表 2 に示す．修正を要したものは約 3%と大きく減っており，第 1 回のデータを用いることで各人の表情学習はかなり正確に行われたと考えられる．

表 2. 第 2 回ミーティング（マスクなし）表情の修正数

	全数	人 A	人 B	人 C	人 D
検出顔画像数	451	110	112	114	115
うち，表情修正数	15	3	9	0	3

表情判定処理のなかで，学習モデルにデータが反映されていない初回ミーティングの場合は，人によって，neutral（中立）に判定すべきであろう通常の表情が happiness（喜び）に判定されがちな人，sadness（悲しみ）に判定されがちな人，anger（怒り）に判定されがちな人などがいて，それをいったん修正して再学習を行うだけで，2 回目以降は大幅に精度が向上した．つまり，1 回目で通常の表情が sadness（悲しみ）に分類されてしまったとしても，それを修正した学習データセットを与えて一度学習するだけで，

© 2023 Information Processing Society of Japan

2 回目以降はほぼ正確に表情が分類されることになる．前述のように，本システムは利用のたびに表情分類用のモデルを簡単に入れ替えることができるので，参加者によってモデルを入れ替えることで，かなり正確な表情分類が期待できる．もちろん，第 1 回目と第 2 回目の修正をいずれも筆者自身の主観で行っており，たとえばこれを別の評価者が行った場合，あるいは第 1 回目と第 2 回目の評価者が異なる場合にどのようになるかは現時点で不明だが，少なくとも学習がかなり有効にしたということはできるであろう．

またこのことは，本システムのように表情分類の結果を利用者が簡単に修正し，それを学習用データセットとして反映できる機能の有用性を示す．一般に会議は種類によって参加者がほぼ決まっていることが多く，たとえば 10 名の会議において，参加したことのないメンバーが新規に参加するのは稀で, 多くても 1, 2 名というような場合が多い．メンバーによって分類用のモデルを入れ替えてシステムを利用し，分類結果がおかしいと感じた場合のみ修正をして学習用データセットを用意することは，実際に会議システムで用いる場合でも現実的と考えている．

次に，マスク着用時の結果であるが，前述のように，初回のミーティングで行われた表情判定結果を筆者自身が編集する作業がきわめて困難かつ大雑把なものとなった．つまり, 各画像を見てもその表情がたとえば neutral なものであるが happiness を示すものであるかの判定が極めて難しかった．筆者以外の 3 名もすべて筆者自身がよく知っているメンバーではあったが，画像だけからの表情判定は非常に難しく, 行うためには別途, ミーティング中の会話内容, 声のトーン，場合によっては本人に対するインタビューなどを行うことが必要であろう．そのような作業を行った上で，初回ミーティングの結果をもとに学習モデルを作成して第 2 回ミーティングを行うことにより，マスク着用時の表情判定を，人間の目視よりも深層学習により正確に行うかどうかの確認は可能だが，それについては将来の課題とする．

4.2 利用者へのフィードバック方法

現在の実装では，参加者表情をグラフまたはアイコンで表示したが，次のようにさまざまな検討要素がある．

1) 表情数値の積み上げ棒グラフ
積み上げ棒グラフの表示は直観的にもわかりやすいが, 表示の適切な時間間隔は要検討である．グラフの時間間隔をズームイン・アウトする機能を設けたとしてもデフォールトで表示する時間間隔として適切なものは検討余地がある．たとえば「参加者の表情が徐々に否定的なもの（悲しみや怒り）になってきている」というようなことをリアルタイムで把握するためには，1, 2 分程度の間隔で表示をするのが良いであろうし，会議の全体的な雰囲気をつかみたいだけであれば 5 分程度の方が見やすい場合もあるだろう．

2) 表情のアイコン表示
ここでは利用者人数と同じだけの顔アイコンを並べて表示し，それぞれのアイコン表示を表情にあわせて 10 秒ごとに切り替えた．これは主としてリアルタイムに全員の表情を一覧することを目的としたもので，人数分のアイコンを見ているだけで「今は全員，笑顔である」というようなことを一目で把握できる．

3) 表情分析の結果を誰に表示するか
次のようにいくつかの観点がある．

a) 他の人に見せるか本人に見せるか？
b) 他の人にも見せるとした場合, 本人がビデオをオフにしていても見せるか？

まず a)については，目的によって異なる．たとえば，何人かで会議をしていて全員の顔がお互いに見えているときに，それぞれのその人の表情分類の結果をビデオ画像の横にアイコン表示するようなことも考えられる．これはたとえばテレビ番組等で演出に用いられるような手法に近く，その人の表情を強調する効果があるだろう．大勢の聴衆を対象として一人がプレゼンテーションをしているときに，参加者全員の表情をアイコンだけで並べて表示すると，話しては全体の雰囲気を把握するのに役立つだろう．逆に，プレゼンテーションをしている人に対し，自分自身の顔が他の人にどのような印象を与えているかをフィードバックするために表示することも考えられる．

b)について「本人がビデオをオフにしているのにその表情分類結果を他の人に見せる」というのは一見違和感があるかも知れないが，実際にはこれに類似したものはすでに存在する．たとえば VTuber 等は「顔は出さないが，顔の表情を取得してアバター表情に反映させて配信する」ということを行っている[10]．会議等においても,「ビデオをオフにするが，表情だけは相手に伝わっても良い」または「伝わってほしい」という人はいるだろう．

4.3 高臨場感通信やメタバース内の会議との比較

本研究では，ビデオ会議で感情が伝わりにくいことを補うために表情分類結果を表示する試みを行っているが，これはいわば現実の対面会議ではない機能を加えることにより，現実の対面会議を超えようとする，あるいは対面会議とは異なる価値をオンライン会議に加えようとするものである．

ビデオをできる限り高精細にしたり立体映像や立体音響にすることにより対面の会議を忠実に再現することを目標とする研究もあるが[11]，それとは異なるアプローチである．

© 2023 Information Processing Society of Japan

5. おわりに

ビデオ会議における参加者表情から深層学習を用いて感情分類を行い利用者にフィードバックするシステムについて述べた．分類結果が誤っていると感じた場合に簡単に修正し，それを新たなデータセットとして利用できるようにすることで，大幅な精度向上が実現できる見込みが立った．また，会議ごとに異なる学習モデルを切りかえて表情分類をすることができるため，それぞれの会議の参加者で作成したモデルを利用することで，表情分類の精度が高くなる．分類結果の表示方法のバリエーションを考察した．

謝辞 本研究は，東洋大学井上円了記念研究助成プログラムにより助成を受けたものです．同助成に感謝いたします．

参考文献

[1] 神場： オンライン会議における感情推定と参加者へのフィードバック方法，情報処理学会インタラクション 2022, 6D18
[2] Ben Weber, “People Analytics（邦訳：職場の人間科学），2013.
[3] 矢野：データの見えざる手　ウエアラブルセンサーが明かす人間・組織・社会の法則, 2014.
[4] S.Samrose, et al.: MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings, Proceedings of the 2021 CHI, Article No. 252, pp.1-13, 2021
[5] P. Murali et al: AffectiveSpotlight: Facilitating the Communication of Affective Responses from Audience Members during Online Presentations, Proceedings of the 2021 CHI, Article No. 247, pp.1-13, 2021
[6] S. Setty, et al. "Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations", NCVPRIPG 2013. https://cvit.iiit.ac.in/projects/IMFDB/
[7] Qiita：Chrome 拡張の作り方（2020 年 12 月 25 日更新）https://qiita.com/RyBB/items/32b2a7b879f21b3edefc
[8] Qiita: FaceNet の顔認証をお手軽に試す（2020 年 12 月 06 日更新）https://qiita.com/Takuya-Shuto-engineer/items/4dcbadbd16e16c3b1677
[9] P. Ekman and W.V.Friesen: Universals and cultural differences in the judgements of facial expressions of emotions, Journal of Personality and Social Psychology, Vol. 53, No.4, pp.712-717, 1987.
[10] VTuber とは：https://media.mobile.rakuten.co.jp/contents/articles/2022/00067/
[11] 南，深津：超高臨場感通信技術「Kirari!」，映像情報メディア学会誌, VO.73, No.5. pp.854-859 (2019).

© 2023 Information Processing Society of Japan