CJ-005

# モダリティ融合に基づく感情認識・潜在表現単一化による感情の数理的表現: モダリティ疑似欠損下での感情空間の表現力

# Mathematical Representation of Emotion by Emotion Recognition and Latent Space Unification Based on Multimodal Fusion: Representational Power of Emotional Space Under Modality Ablation

原田 誠一[1)]　佐久間 拓人[1)]　加藤 昇平[1)2)]
Seiichi Harata　Takuto Sakuma　Shohei Kato

## 1　はじめに

Human-Computer Interaction においては，計算機に感情的な振る舞いをさせることでユーザに楽しみや安らぎ等の精神的効果を与える感性対話エージェントの実現が期待できる．感性対話は，感情認識，感情生成，感情表出の 3 つの要素に分けられ，それらを通じてユーザやエージェントの感情状態を計算機内で扱い制御する．これを実現するために，エージェントの内部状態として感情を数理モデル化する，統一的な表現が必要と考える．

Russell[1] は，連続的な感情の表現方法として覚醒度（Arousal）と感情価（Pleasure，または Valence）による 2 次元空間上に感情が円環状に布置するモデルを作成した．Gotoh ら [2] および Hieida ら [3] は，感情コミュケーションモデル内で 2 次元の連続空間よって感情や情動を扱い，一定の感性インタラクション能力の実現可能性を示したが，多様で複雑な感情を表現する空間として適切な次元数について議論の余地がある．

本研究では，Deep Neural Networks（DNN）が入力刺激の特徴を低次元空間に写像できる能力に着目する．エージェントが相手の感情を知覚するプロセスである感情認識に注目して，音声や表情などの実際に表出されたデータから，感情を数理的に表現する連続値のベクトル空間（感情空間）への写像の獲得を目指す．Kervadec & Vielzeuf ら [4] は，顔画像から感情を識別する DNN を学習し，感情空間の獲得を試みた．そして，感情を連続空間で表現するための適切な次元数について考察したが，この手法では顔画像データのみを用いて DNN を学習するため，顔画像モダリティに特化した感情空間が学習され，多様なモダリティのコミュケーション手段をもつエージェントの内部状態としては適さないと考える．

そこで本研究では，人間が複数モダリティからの知覚を融合すると同時に，特定のモダリティが欠損しても一定の精度で感情を認識できることに着目し，DNN を用いて複数のモダリティに共通する感情の表現を獲得する手法を提案する．提案モデルにより，モダリティに依存しない純粋な感情の情報を感情空間で表現することで，人間の感情認知のモデル化を目指す．筆者ら [5] は先行研究において，単一のモダリティで学習した感情空間がモダリティ間で類似しないことを示し，モダリティを融合した DNN によって感情の多様性を維持しつつ音声と顔画像モダリティ間で共通の感情空間が学習可能であることを示唆した．本稿では感情生成や感情表出等の下流タスクへの応用のために，感情認識性能に加え異なるモダリティから得られる感情空間の一致性の観点から感情を数理的に表現するのに有効な空間とその次元数について考察する．

1）　名古屋工業大学　大学院工学研究科　工学専攻
Dept. of Engineering, Graduate School of Engineering, Nagoya Institute of Technology
2）　名古屋工業大学　情報科学フロンティア研究院
Frontier Research Institute for Information Science, Nagoya Institute of Technology

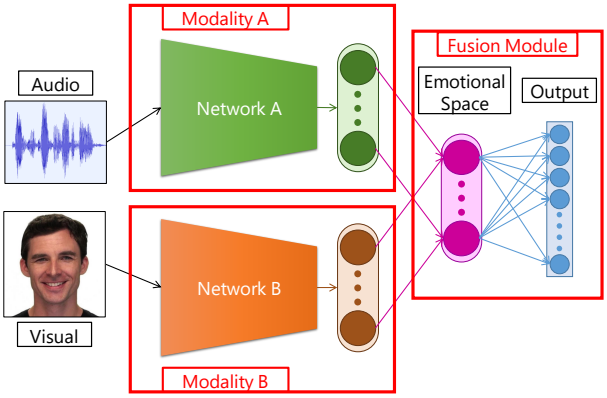



図 1: 提案モデルの概観．複数のモダリティを感情空間で融合する．

## 2　提案手法

### 2.1　提案モデル

図 1 に本研究で提案する複数のモダリティを融合した深層モデルの概観を示す．本研究では，モデルの出力層の直前の層を感情空間と定義し，この層が感情の数理的表現を獲得するための学習をする．モダリティ A 部には音声データを入力し，モダリティ B 部には顔画像データ入力し，各モダリティからそれぞれ感情空間を得る．

ここで感情空間を構成する際に，ユークリッド空間上で感情空間を定義するモデルと，半超球面上で感情空間を定義するモデルの 2 種類を考える．ユークリッド空間上で感情空間を定義する場合，図 1 の融合部での活性化関数にはベクトル要素ごとの tanh 関数を用い，感情空間に写像された 2 データ間の距離をユークリッド距離で定義する．一方，半超球面上で感情空間を定義する場合，2 データ間の距離を arccos 関数で定義する．以下の 2.2 節では，特に半超球面上に感情空間を定義する手法について述べる．

### 2.2　$D+1$ 次元半超球面上への写像

感情は多様で複雑な構造をしていると考えられており，Plutchik[6] は感情を円盤状又は円錐形の空間としてモデル化し，感情の種類だけでなく，感情の強度や混合感情についてもその空間で表現している．そのため，本研究における DNN によってモダリティに共通した感情空間を獲得する目的においても，そのような表現力の高い空間で感情を数理的に表現することが有効だと考える．また深層学習分野において，埋め込み空間を超球で

表現する手法は，距離学習 [7] や生成モデル [8] において成果を残している．そこで本稿では，$D$ 次元の感情空間を，$D+1$ 次元の半超球面上で表現する手法を提案する．

図 1 の融合部では，まず以下の活性化関数 $h(\cdot)$ により，各モダリティのモデル毎に得られた $D$ 次元ベクトル $\boldsymbol{x}$ を，それぞれ $D+1$ 次元の半超球面上に写像する．

$$h(\boldsymbol{x}) = \left[a(\boldsymbol{x}), \sqrt{1-(a(\boldsymbol{x})^T a(\boldsymbol{x}))}\right], \quad (1)$$

$$a(\boldsymbol{x}) = \frac{\tanh\sqrt{\boldsymbol{x}^T\boldsymbol{x}}}{\sqrt{\boldsymbol{x}^T\boldsymbol{x}}} \cdot \boldsymbol{x}. \quad (2)$$

まず $a(\boldsymbol{x})$ によって，モダリティ毎のモデルから得られる $D$ 次元ユークリッド空間のベクトル $\boldsymbol{x}$ を $D$ 次元超球内に写像する．その後 $h(\cdot)$ によって，$D+1$ 次元空間上で半超球面上に射影されるように新たな 1 次元を加えることで，半超球面として感情空間を表現する．$D+1$ 次元半超球面上に感情空間を定義した場合，感情空間上の 1 点は $D$ 自由度で表現されるため，これを $D$ 次元感情空間として扱う．

異なるモダリティからの感情空間を融合する手法は，単に各モダリティの感情空間を結合するのではなく，複合した感覚の情報を線形加算融合によって統合する知見 [9, 10] に基づき，各モダリティ毎のモデルにより得られた感情空間同士を加算平均に基づく計算により融合する．本稿では，各モダリティごとの $D+1$ 次元半超球面を加算平均し，半超球面上に布置されるよう正規化することによって感情空間を融合する．

### 2.3 認識タスクと単一化タスクの複合

本研究目的においては，似た感情を表出しているデータの組は感情空間上の距離が小さい位置に写像され，似ていない感情のデータの組は感情空間では距離が遠い位置に写像されるべきである．深層学習におけるクラス識別問題に用いられる Softmax-CrossEntropy 損失関数は，潜在空間において異なるクラス間のデータ同士の距離を大きくさせる役割を持つと考えられており，提案モデルの学習には，この関数を用いた感情認識タスクをする．

また本研究では，人間が相手の表情や声などの情報から単一の感情を知覚していると仮定して，同時に表出された複数のモダリティから得られる感情空間の共通部分を抽出することを試みる．そこで，マルチモーダルなネットワークを統合する際に，それぞれのモダリティ用のネットワークが同一の空間を学習するような単一化タスクを課し，モダリティに共通な感情空間が獲得されることを目指す．

これらを踏まえ，以下の損失関数により感情認識と感情空間単一化のマルチタスクでモデルを訓練する．$Loss^{recog}$ は認識タスクの損失を表し，$Loss^{unif}$ は単一化タスクの損失を表す．

$$Loss = Loss^{recog} + Loss^{unif}, \quad (3)$$

$$Loss^{recog} = \frac{1}{N}\sum_{i=1}^{N} \frac{N}{N_{y_i}\cdot|Label|} \cdot CrossEntropy(\hat{y_i}, y_i),$$

$$Loss^{unif} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D}\left(ES_{i,j}^{audio} - ES_{i,j}^{visual}\right)^2 \\ \qquad (euclidean). \\ \frac{1}{N}\sum_{i=1}^{N}\arccos\left((ES_i^{audio})^T(ES_i^{visual})\right) \\ \qquad (hypersphere). \end{cases}$$

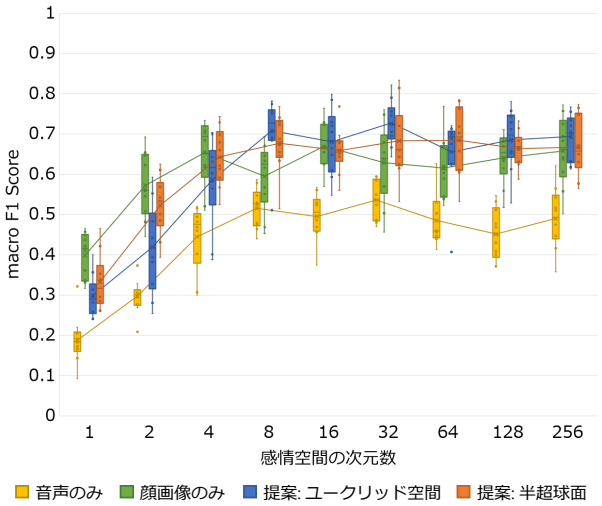



図 2: 感情空間の各次元数における *macro F1 Score*．ユークリッド空間上に感情空間を定義した音声のみ（黄）・顔画像のみ（緑）と提案モデル（青），半超球面上に感情空間を定義した提案モデル（橙）

ここで，$N$ は全データ数，$Label$ は感情ラベル集合，$\hat{y_i}$ はデータ $i$ の予測ラベルの確率，$y_i(\in Label)$ は $i$ の正解ラベル，$N_{y_i}$ はラベル $y_i$ のデータ数，$ES^{\{audio,visual\}}$ は音声と顔画像それぞれから，ユークリッド空間で表現する場合は tanh 関数，半超球面上で表現する場合は式 1 によって得られた感情空間を表す．$Loss^{recog}$ は重み付け CrossEntropy 関数で，ラベル $y_i$ 毎にデータ数の不均衡を補正するための重みを乗じて学習する．$Loss^{unif}$ は，対応するデータ $i$ の $D$ 次元ユークリッド空間上の表現のモダリティ間の二乗平均誤差または，$D+1$ 次元半超球面上の表現がモダリティごとになす角のデータ数による平均とする．マルチタスクな損失関数により，単純に認識性能を高めるのではなく，高い認識性能を保ちつつ，モダリティ間の感情空間の一致度を高めることを目指す．

### 2.4 データセット

本稿では，音声と顔表情の両方のモダリティが含まれる映像データセットである，RAVDESS[11] を用いる．このデータセットは，北アメリカ英語の演技発話データからなり，21–33 歳の本職の俳優 24 名（男女各 12 名）分のデータがある．本研究では，Actor 1,2 をテストデータ，Actor 3,4 を検証データ，残りの 20 名を訓練データとして用い，7 つ（Neutral, Happy, Surprise, Fear, Anger, Disgust, and Sad）の感情ラベルによる感情認識タスクと単一化タスクでモデルを学習・評価する．音声モダリティの入力としては対数メルスペクトログラムを用い，顔画像モダリティの入力としては，OpenCV [12] により顔検知した領域の画像を用いる．

## 3 実験

### 3.1 感情空間の次元数ごとの感情認識性能の比較

音声と顔画像モダリティを融合し，感情認識タスクに加え異なるモダリティから同一の感情空間を得るような単一化タスクで学習した提案モデルによって得られる感情空間が，どの程度感情の多様性を保有しているかを確認する．ここでは，ユークリッド空間で感情空間を定義し，音声モダリティのみ，顔画像モダリティのみそれぞれで学習したモデルの感情認識性能と，ユークリッド空間，半超球面それぞれで感情空間を定義した提案モデル

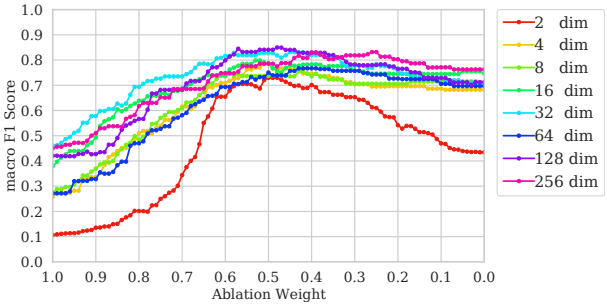



図 3: ユークリッド空間上に感情空間を定義し単一化タスクを用いずに学習した場合の，モダリティ疑似欠損実験における *macro F1 Score*.

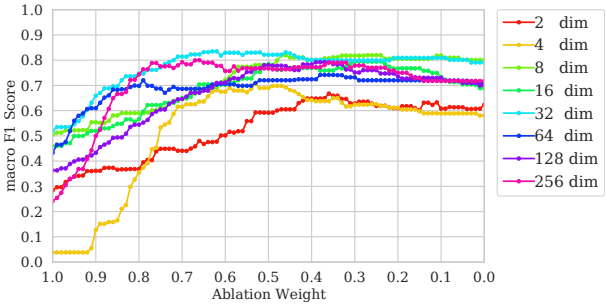



図 4: ユークリッド空間上に感情空間を定義した場合の，モダリティ疑似欠損実験における *macro F1 Score*.

の感情認識性能を比較する．パラメータの初期値等の影響を考慮に入れるため，異なる初期パラメータで 10 回の学習を試行し，感情空間の次元数ごとに比較モデルを訓練した．感情認識タスクの評価尺度としては *macro F1 Score* を用いた．

図 2 に各次元数における比較モデルの感情認識スコアを箱ひげ図で示す．この結果より，いずれの空間の定義においても提案モデルは音声モダリティのみで学習したモデルより感情認識性能が優れており，8 次元以上において顔画像モダリティのみで学習したモデルよりもわずかに優れていることがわかる．このため，感情認識とは異なるタスクである単一化タスクを加えることによる感情認識性能が劣化することは確認されず，複数のモダリティを融合して学習することによってより多様な感情を表現する感情空間が学習されたと考える．

### 3.2 疑似モダリティ欠損による感情認識への影響

本研究のモダリティ融合手法は，各モダリティ毎の感情空間を加算平均に基づく計算で融合している．そのため，人間による感情認識と同様に片方のモダリティが欠損した場合でも得られたモダリティから補完して感情を推定できる．モダリティを補完した感情認識の性能を評価することにより，感情空間上での感情の多様性とモダリティごとに得られる感情空間の一致度の両方を評価できる．そのため，得られるモダリティを擬似的に欠損させ，モダリティ補完による感情認識の性能を観察し，ユークリッド空間での感情空間と半超球面上での感情空間を比較する．

擬似的なモダリティ欠損として以下の式で融合した感情空間（$ES^{fused}$）を，各モデルの出力層に入力することによって推定感情を得る．

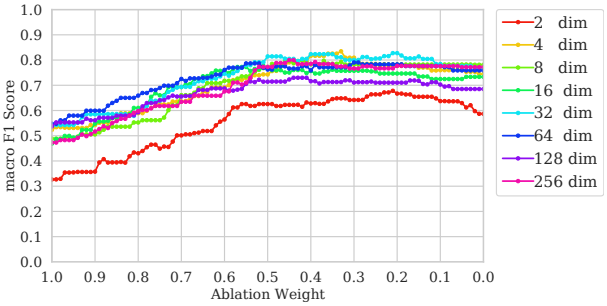



図 5: 半超球面上に感情空間を定義した場合の，モダリティ疑似欠損実験における *macro F1 Score*.

$$ES^{fused} = f(w \cdot ES^{audio} + (1-w) \cdot ES^{visual}), \tag{4}$$

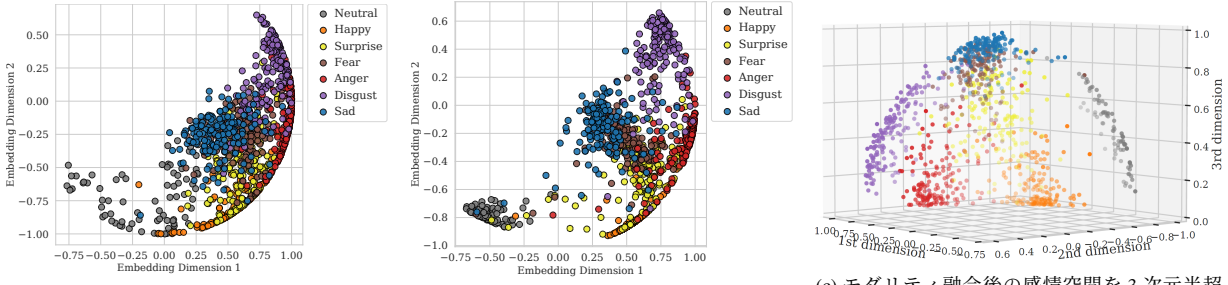
$$f(z) = \begin{cases} z & (euclidean). \\ \dfrac{z}{\sqrt{z^T z}} & (hypersphere). \end{cases} \tag{5}$$

ここで，$w$ はモダリティを融合する重みで，モデルの学習時には $w = 0.5$ としている．$w = 1$ は音声モダリティのみ与えられた場合に対応し，$w = 0$ は顔画像モダリティのみが与えられた場合に対応する．また $f(\cdot)$ はベクトルの正規化関数で，本稿の提案手法では融合後の感情空間も半超球面上に布置されるようにする．

図 3,4,5 に感情空間の各次元数におけるモダリティ疑似欠損下での感情認識スコアを示す．図 3,4 はユークリッド空間上に感情空間を定義し，図 3 は単一化タスクを用いず認識タスクのみで学習したモデルの結果，図 4 は認識・単一化タスクで学習した提案モデルの結果，図 5 は半超球面上に感情空間を定義して認識・単一化タスクで学習した提案モデルの結果を示している．図 3,4 を比較すると，複数モダリティが得られる際の感情認識性能は保ちつつ単一化タスクによってモダリティ補完能力は高いことがわかるが，特に $w = 0$ での音声のみ，$w = 1$ での顔画像のみ得られる条件において依然として著しい性能低下が見られ，モダリティ欠損時の補完能力は低いといえる．一方，図 5 においては，他 2 モデルと比較してモダリティ欠損に対する性能低下が小さいため，異なるモダリティから一致した感情の情報が抽出できていると考えられ，半超球面上の感情空間を定義する手法は，ユークリッド空間で定義するよりモダリティ非依存な感情の数理モデル化に適していると考えられる．

また，図 5 における $w = 0, w = 1$ での片方のモダリティが完全に欠損した条件での感情認識性能は，図 2 における単一モダリティで学習したモデルの感情認識性能と比較しても優れている．このため，異なるモダリティを融合して学習し，半超球面上で感情空間を定義することによって，それぞれのモダリティからより効率的に感情の情報を抽出できていることが考えられる．

加えて図 5 より，半超球面上で感情空間を定義した場合，低次元空間（4 次元）においても高次元空間（256 次元）で感情空間を学習したモデルと同等の感情認識性能・モダリティ補完性能を示している．この結果より，半超球面上に感情空間を定義する提案モデルによって，低次元空間で感情の数理モデル化ができることが示唆され，本稿の実験条件においては 4 次元程度で十分な表現能力を有していると考える．




(a) 音声モダリティから得られる 2 次元超球内での感情空間

(b) 顔画像モダリティから得られる 2 次元超球内での感情空間

(c) モダリティ融合後の感情空間を 3 次元半超球面上に射影

図 6: 半超球面で定義した 2 次元感情空間の可視化

### 3.3 感情空間の可視化

図 6 に半超球面上の感情空間の次元数を 2 次元として学習した際の訓練データのプロットを示す．図 6(a),6(b) はそれぞれ音声と顔画像から得られる 2 次元超球内へのプロット，図 6(c) はこれらを融合した感情空間の 3 次元半球面上での表現を示す．これらの可視化からも，異なるモダリティから類似度の高い感情空間が得られていることがわかり，異なる感情は離れて分布していることもわかる．また，この可視化を見ると，Happy は Surprise の近くに布置されており，Disgust は Sad や Angry の近くに布置されている．これらの感情のペアは，RAVDESS データセットの文献 [11] 内で行われている人間の評定者による感情評定実験における結果において，人間が誤認識した感情のペアとなっている．このため，人間が混同しやすい感情同士を感情空間上で近くに布置するように感情空間が学習できたといえ，提案手法によって人間の感情間の”距離”を表現する数理モデルが構築できる可能性が示唆された．

## 4 おわりに

本稿では，Affective Computing において感情を数理的表現（感情空間）として扱うために，感情空間をユークリッド空間と超球上の空間として 2 種類定義し，マルチモーダルな DNN の感情認識・感情空間単一化のマルチタスク学習によって，複数モダリティから共通した感情空間を獲得する手法を提案した．音声と顔画像のデータによる実験で，各モダリティを擬似的に欠損させた条件における感情認識性能の頑健性を観察することにより，異なるモダリティに共通する感情の情報を低次元空間に埋め込むことができることを確認した．特に，エージェントが相手の感情を認知するプロセスのモデル化において，本稿の実験条件においては 4 次元程度の低次元の超球で感情の数理的表現が可能であることが示唆された．

今後は提案手法による感情の超球表現と心理学の感情モデル [1, 6] との関連性について考察・解釈する．また，提案モデルによる感情認識の混同傾向と人間の評定者による感情認識の混同傾向を比較し，人間と同様の感情知覚の獲得を目指す．加えて，音声や顔画像以外でエージェントが保有しうる対話モダリティも統合した感情空間の獲得を目指し，より一般的な感情認知プロセスのモデル化に取り組む．更に，Affective Computing への応用に向けて，感情空間を感情表出モデルと組合せることにより，感情を表出するモダリティが様々な場合でも一貫した感情を表出できるかを調査する．

### 謝辞

本研究は，一部，文部科学省科学研究費補助金（課題番号 JP19H01137，JP19H04025，および，JP20H04018）の助成により行われた．

### 参考文献

[1] Russell, J. A.: A Circumplex Model of Affect, *Journal of Personality and Social Psychology*, Vol. 39, pp. 1161–1178 (1980).

[2] Gotoh, M., Kanoh, M., Kato, S., Kunitachi, T. and Itoh, H.: Face Generator for Sensibility Robot based on Emotional Regions, in *Proceedings of the 36th International Symposium on Robotics*, Vol. 36 of *ISR 2005*, pp. WE31–6, Tokyo, Japan (2005), Citeseer.

[3] Hieida, C., Horii, T. and Nagai, T.: Emotion Differentiation based on Decision-Making in Emotion Model, in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 659–665 (2018).

[4] Kervadec, C., Vielzeuf, V., Pateux, S., Lechervy, A. and Jurie, F.: CAKE: Compact and Accurate K-dimensional representation of Emotion, in *Image Analysis for Human Facial and Activity Recognition (BMVC Workshop)*, Newcastle, United Kingdom (2018).

[5] 原田誠一, 佐久間拓人, 加藤昇平：モダリティを統合した認識モデルに基づく深層マルチタスク学習による感情の数理的表現, 電気学会論文誌 C（電子・情報・システム部門誌）, Vol. 140, No. 12, pp. 1343–1351 (2020).

[6] Plutchik, R.: The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, *American Scientist*, Vol. 89, No. 4, pp. 344–350 (2001).

[7] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B. and Song, L.: SphereFace: Deep Hypersphere Embedding for Face Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6738–6746 (2017).

[8] Davidson, T., Falorsi, L., De Cao, N., Kipf, T. and Tomczak, J.: Hyperspherical variational auto-encoders, in *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 856–865 (2018).

[9] Landy, M. S., Maloney, L. T., Johnston, E. B. and Young, M.: Measurement and modeling of depth cue combination: in defense of weak fusion, *Vision Research*, Vol. 35, No. 3, pp. 389 – 412 (1995).

[10] Ernst, M. O. and Banks, M. S.: Humans integrate visual and haptic information in a statistically optimal fashion, *Nature*, Vol. 415, No. 6870, pp. 429–433 (2002).

[11] Livingstone, S. R. and Russo, F. A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, *PLOS ONE*, Vol. 13, No. 5, pp. 1–35 (2018).

[12] Bradski, G.: The OpenCV Library, *Dr. Dobb's Journal of Software Tools* (2000).