日本語フォント文字の印象評価と感情字幕生成への応用 *

☆中村史也 (神戸大), 相原龍 (三菱電機), 高島遼一, 滝口哲也 (神戸大), △伊谷裕介 (三菱電機)

1 はじめに

自動音声認識 (automatic speech recognition; ASR) を用いた自動字幕システムは，発話内容の理解を文字という視覚的な情報で支援できる技術として期待されている。しかし，人間はコミュニケーションにおいて，話者の表情や声の調子から推定される感情情報などのパラ言語情報を利用していることが知られており [1]，ASR で得られる言語情報だけでは発話を理解するには不十分な場合がある。そのような場合，テレビや YouTube などの動画コンテンツでは，その場の状況や出演者の感情を様々なフォントや色彩として反映した字幕（テロップ）がしばしば用いられる。

話者の感情を反映したテロップを自動生成するシステムとして,「感情表現字幕システム」（NHK テクノロジーズ，DNP 社）[2] が提案されている。従来システムでは，話者の表情を画像認識によって感情クラスに分類し，感情クラスごとにあらかじめ定義したフォントを用いてテロップを生成する。しかし，従来手法はクラス分類の結果からルールベースでフォントを決定する以上,「激しい怒り」と「弱い怒り」,「悲しみを含んだ怒り」のような細かな感情の差異を表現することができないという問題がある。

そこで我々は以前，話者の感情（感情ごとの確率）に基づき，感情クラスごとに対応するフォント同士を画像的に補間することで複雑な感情を反映したフォントを生成する手法を提案した [3]。しかし，フォントを画像として補間することで，線の太さや丸みなどの感情に関わると考えられる特徴が薄れてしまうことがあり，感情を適切に反映できているとは言いがたい結果であった。

そこで本研究では，感情の機微を反映したフォント字幕の生成を行うため，日本語フォント文字の印象評価を行い，感情次元空間上におけるフォントの分布を調査する。さらに，得られたフォントと感情の対応データを用いて感情次元からフォントを生成するモデルを提案し，連続的な感情変化に対するフォント文字の生成を試みる。

Fig. 1: Examples of fonts selected for impression evaluation

2 日本語フォント文字の印象評価

2.1 評価方法

感情の差異をフォントに反映させるためには，まず感情とフォントの対応関係を調査する必要がある。そこで我々は，日本語のフリーフォントを 20 個収集し，9 名の日本語母語話者を対象に感情に関する印象の評価を行った。収集したフォントのサンプルを Fig. 1 に示す。

感情の評価尺度には，心理学 [4, 5, 6] や音声感情認識 (speech emotion recognition; SER)[8] の分野で一般的に用いられている「感情カテゴリ」と「感情次元」の 2 種類を採用した。

感情カテゴリは,「怒り」や「悲しみ」のような基本感情に感情を分類する評価手法である。本研究では SER の研究でよく用いられる「喜び」「怒り」「悲しみ」「平静」の 4 感情に「その他」を追加した計 5 つのカテゴリから最も近い感情を選択する形式とした。

感情次元は，快–不快などの軸上の値として感情を連続的に評価することで，感情を次元空間上で表現する手法である。本研究では，Russell [7] に倣って快–不快軸 (valence)，覚醒–睡眠軸 (arousal) の 2 軸を感情次元として採用し，それぞれの軸で 1（不快/睡眠）から 5（快/覚醒）の 5 段階で評価を行った。

評価者に提示するフォントのサンプル文字列は，次の理由から「いろはにほへとちりぬるをワカヨタレソツネナラム」の 23 文字とした。

---

* Impression evaluation of Japanese font characters and its application to emotional subtitle generation. by Fumiya Nakamura (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), Yusuke Itani (Mitsubishi Electric Corporation)

Table 1: Emotional category evaluation of fonts

Happiness	Anger	Sadness	Neutral	Others
6	4	2	7	1

Table 2: Number of votes for each emotional category

Happiness	Anger	Sadness	Neutral	Others
43	37	25	57	18

- 文字の重複がない。
- 文字列の意味によって評価者の感じる感情が左右されづらいと考えられる。
- 同じフォントでもひらがなとカタカナで印象が変わる可能性があるため，ひらがなとカタカナの両方を含む文字列であることが望ましい。

2.2 評価結果

フォントごとに最も得票数の多かった感情カテゴリの個数を Table 1 に，全フォントを通した感情カテゴリの得票数を Table 2 に示す。いずれの結果でも，評価対象のフォントの印象として「平静」の感情が最も多く,「悲しみ」の感情は最も少なかった。また，9 名による感情カテゴリの評価から Fleiss の kappa 係数を計算したところ，$\kappa = 0.44$ であった。これは評価者間で適度な一致が見られると解釈できる。

さらに，20 個のフォントの感情次元平面における分布を Fig. 2 に示す。図中の点の色と形はそのフォントで最も得票数の多かった感情カテゴリを示している。Fig. 3 は Russell の感情円環モデル [7] における感情カテゴリと感情次元の関係を示しており，Fig. 2 における感情カテゴリの位置関係（「怒り」と評価されたフォントは左上に,「喜び」のフォントは右上に位置している，など）と符合していることが確認できる。

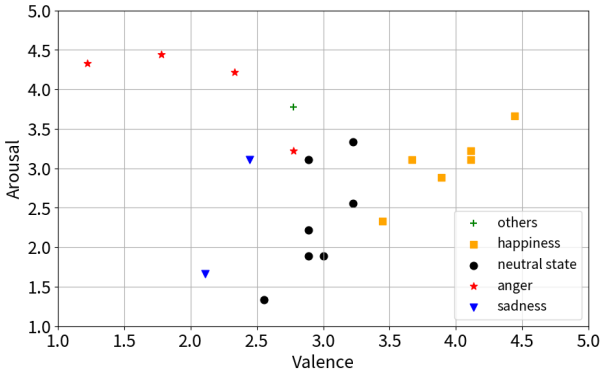


Fig. 2: Distribution of fonts in the emotional dimensions space

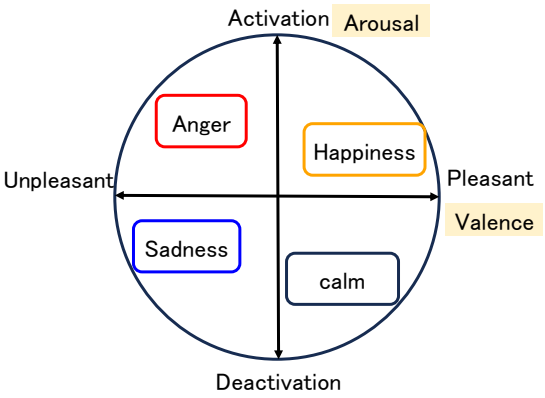


Fig. 3: Relationship between emotional categories and Russell's circumplex models [7]

3 印象評価に基づくフォント画像生成手法

3.1 従来手法：zi2zi によるフォントの補間

以前の研究 [3] では，フォント画像の生成に zi2zi [9] を使用した。zi2zi は，GAN に基づく Image-to-Image モデルの一種である pix2pix [10] を拡張し，フォントごとに異なるガウシアンノイズをカテゴリ埋め込みとして使用することで一対多のフォント文字画像の生成に適用したモデルである。zi2zi では，このカテゴリ埋め込みを操作することによってフォントを画像的に補間することができる。N 個のフォントのカテゴリ埋め込みを z_i $(i = 1, \ldots, N)$，各フォントの重みを w_i としたとき，生成時のカテゴリ埋め込み $\hat{z}$ はそれらの線形結合によって作成する。

$$\hat{z} = \sum_{i=1}^{N} w_i z_i \tag{1}$$

$$\sum_{i=1}^{N} w_i = 1 \tag{2}$$

このようなフォントを補間する手法を用いて与えられた感情次元 y に対応したフォントを生成するには，次の手順が考えられる。まず，フォント f の感情次元 x_f と y の距離 d_f をそれぞれ計算し，距離の近い 2 個のフォントを α，β とする。

$$d_f = \|x_f - y\|_2 \tag{3}$$

次に，フォント α と β を距離 d_α と d_β に基づく重み w_α と w_β で式 (1) に従って補間することで，感情次元 y に対してフォントを生成する。

$$w_\alpha = \frac{d_\beta}{d_\alpha + d_\beta} \tag{4}$$

$$w_\beta = \frac{d_\alpha}{d_\alpha + d_\beta} \tag{5}$$

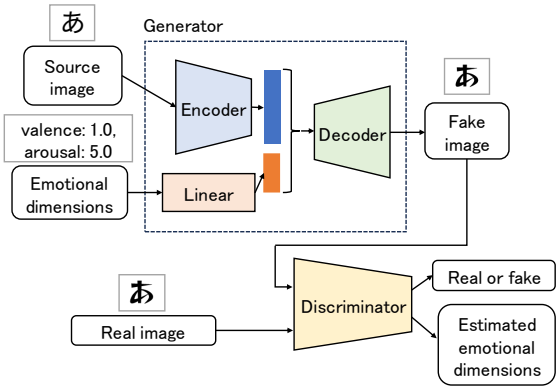


Fig. 4: Model structure of the proposed method

3.2 提案手法：感情次元ベクトルからのフォント生成

本稿では，2 章で得られた感情次元とフォントの対応データを用いて，感情次元の 2 次元ベクトル（快–不快軸，覚醒–睡眠軸）からフォントを生成する手法を提案する。

提案モデルの構造を Fig. 4 に示す。Generator 部では，zi2zi や pix2pix と同様の U-Net 構造 [11] の Encoder-Decoder モデルに加えて，zi2zi のカテゴリ埋め込みの代わりに感情次元ベクトルを線形層で 128 次元ベクトルに射影した出力を Encoder 出力と結合して Decoder に入力する。さらに Discriminator 部では，ACGAN[12] を参考にして画像の真贋を推定する他にフォント画像から感情次元を推定するための線形層を設けている。

損失関数は，GAN の敵対的損失に加えて，Generator の L1 損失，コンスタント損失 [13] ，Discriminator の感情次元推定損失を採用している。感情次元推定損失 L_{aff} は以下の式で計算される。

$$L_{aff} = \frac{1}{2}\sum_{i=1}^{2}(\hat{y}_i - y_i)^2 \tag{6}$$

ただし，y および $\hat{y}$ はそれぞれ感情次元ベクトルの正解値および推定値，y_i はベクトル y の i 次元目の値である。なお，感情次元推定損失は敵対的損失とは異なり，Generator と Discriminator の両方で最小化するように学習する。

4 評価実験

4.1 実験設定

ソース画像には一般的なゴシック体フォントの一種である「源真ゴシック Regular」を使用し，ターゲット画像には 2 章で評価を行った 20 個のフォントを使用した。学習にはフォントごとに日本語ひらがなカタカナ 160 字分から，一部のフォントで実装されていない文字を除いた計 3,197 枚の文字画像を用いた。文字画像の大きさは 256×256 で，3 チャネルのカラー画像として生成を行った。データ拡張のため，学習時に一定の範囲内でランダムに文字の回転処理を施した。モデルの L1 損失の重みは 100，コンスタント損失の重みは 15，感情次元推定損失の重みは 10 に設定した。オプティマイザは Adam を使用し，1,200 エポック学習時点のモデルでフォント文字画像生成の評価を行った。

Table 3: MSE, MAE and MS-SSIM between generated image and ground truth

Method	MSE ($\times 10^{-2}$)	MAE ($\times 10^{-2}$)	MS-SSIM
zi2zi	1.86	2.45	0.868
ours	1.82	2.39	0.864

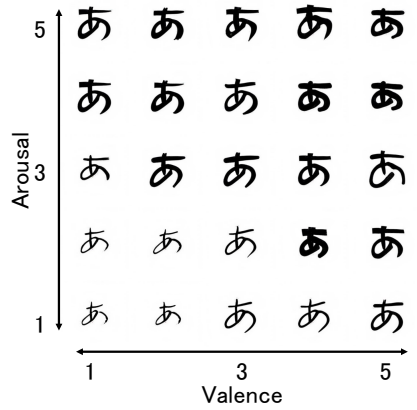
4.2 フォントの生成品質の客観評価

学習データに存在するフォントを生成した際の正解フォントとの MSE (Mean Squared Error)，MAE (Mean Absolute Error)，MS-SSIM の値を Table 3 に示す。ただし，MSE および MAE は $[0, 255]$ の画素値を $[0, 1]$ に正規化して計算を行った。この結果から，提案手法は zi2zi を用いた従来手法と同程度の品質でフォント画像を生成できていることが確認できる。ここで，zi2zi はフォントラベルを入力して文字画像の生成を行っているのに対して，提案手法ではフォントの感情次元評価値を入力していることに注意が必要である。

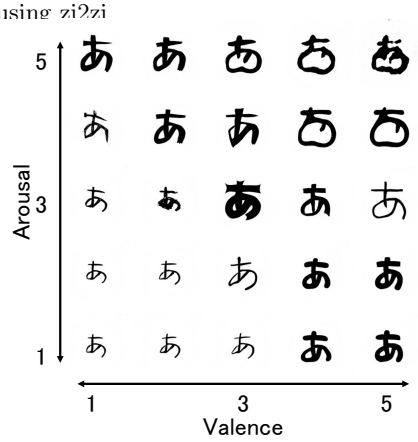
4.3 感情空間上における生成

3.1 節および 3.2 節で述べた手法を用いて感情次元平面上の点からフォント文字「あ」を生成した結果をそれぞれ Fig. 5(a)，Fig. 5(b) に示す。これらは，Fig. 2 の座標平面上の格子点にあたる感情次元を用いて生成を行った結果である。従来手法（Fig. 5(a)）では感情次元に対して生成画像が滑らかに変化している一方，補間によってそれぞれのフォントの感情的特徴（線の太さや丸みなど）が薄れており，生成されるフォントが多様性に乏しい。対して，提案手法（Fig. 5(b)）では (valence, arousal)=(1, 4) や (3, 4) などの感情次元で特徴的なフォントの生成が行えていることから，従来手法よりも感情的特徴を保持されていると考えられる。さらに，提案手法では不快方向では先端のとがった文字が，快方向では丸みを帯びた文字が生成されており，不快・睡眠の方向に進むにつれて線が細くなっていることが確認できる。このことから，提案手法は従来手法よりも感情とフォントの対応関係を学習できていると考えられる。

提案手法の課題として，(5, 5) など右上の感情次元に対する生成が乱れている点が挙げられる。これは学習データの中にそのような極端な評価値を持つフォントが存在しなかったためであると考えられ，今後よ



(a) Generated images by interpolation using zi2zi



(b) Generated images by the proposed method

Fig. 5: Samples of generated font images from emotional dimension values

Fig. 6: Changes in generated image with increasing valence (arousal=3.0)

り多くのフォントで印象評価を実施して学習データを増やすことで改善すると期待できる。また，提案手法において，arousal を 3.0 に固定して valence を 1 から 5 まで 0.25 ずつ増加させた場合の生成結果を Fig. 6 に示す。図中の valence が 4.0 から 4.25 の点など，感情次元の変化に対してフォントが突然変化する箇所が存在するため，感情的特徴を保ったまま緩やかにフォントが変化していくようにモデルを改善する必要がある。

5 おわりに

本研究では，感情の細かな差異を表現するフォント字幕生成システムの構築を目標に，日本語フォント文字の印象評価を行い，フォントと感情の関連を明らかにした。そして，得られたフォントの印象データを用いて，感情次元からフォント文字画像を生成する手法を提案した。評価実験により，提案手法は従来のフォント生成手法と同程度の品質で感情次元からフォントを生成できることを示し，従来手法よりも感情的特徴を保っていることを確認した。今後の課題としては，学習データ不足による生成の乱れを解消することと，感情次元に対するフォントの急激な変化を緩やかにすることが挙げられる。

参考文献

[1] R. Cowie *et al.*, “Emotion recognition in human-computer interaction,” IEEE Signal processing magazine, 18 (1), 32–80, 2001.

[2] https://www.dnp.co.jp/news/detail/10158470_1587.html

[3] 中村ら, “発話音声の感情を反映したテロップ画像の自動生成手法の検討,” 音講論（春）, 887-890, 2023.

[4] R. Plutchik, “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, 89 (4), 344-350, 2001.

[5] A. Mehrabian, “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” Current Psychology, 14, 261-292, 1996.

[6] I. Bakker *et al.*, “Pleasure, arousal, dominance: Mehrabian and Russell revisited,” Current Psychology, 33, 405-421, 2014.

[7] J. A. Russell, “A circumplex model of affect,” Journal of personality and social psychology, 39(6), 1161-1178, 1980.

[8] 赤木, “音声に含まれる感情情報の認識：感情空間をどのように表現するか,” 音響学会誌, 66 (8), 393-398, 2010.

[9] https://kaonashi-tyc.github.io/2017/04/06/zi2zi.html

[10] P. Isola *et al.*, “Image-to-Image Translation with Conditional Adversarial Networks,” arXiv:1611.07004, 2017.

[11] O. Ronneberger *et al.*, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” arXiv:1505.04597, 2015.

[12] A. Odena *et al.*, “Conditional Image Synthesis With Auxiliary Classifier GANs,” arXiv:1610.09585, 2016.

[13] Y. Taigman *et al.*, “Unsupervised Cross-Domain Image Generation,” arXiv:1611.02200, 2016.