顔表情認識におけるネガティブ感情の精度向上の研究

Study on Improving Accuracy of Negative Emotion Recognition in Facial Expression Recognition

磯野　智洋
指導教員　青木　輝勝

東京工科大学大学院　バイオ・情報メディア研究科　コンピュータサイエンス専攻
コンピュータビジョン研究室

本研究は,顔表情認識におけるネガティブな感情（恐怖や嫌悪）の認識精度向上を目指す.従来の課題である精度低下を改善するため,言語誘導型コントラスト学習モデル CLIP の特徴を取り入れた表情認識向けモデルを使用し,精度向上に寄与することを目指す.

キーワード：深層学習，顔表情認識，ネガティブ感情，CLIP

1．　はじめに

現在コンピュータを用いて人間の顔表情を分類する研究が盛んに行われている.顔表情認識の研究が進むことによって， 医療分野における精神的な分析などの分野で応用が期待される.

顔表情認識は深層学習を用いた研究が一般的となっており， 深層学習の登場以前の研究と比べて非常に高い精度を実現している.

顔表情認識において,人間の感情は surprise , sad, happy, fear, disgust, angry などの 6 種類程度に分類されるのがこの分野では一般的である.現在この感情の内 fear や disgust などの一部のネガティブな感情に関しては他と比べて顔表情認識の精度が低くなってしまうことが既存研究から示されている.そのため,これらの感情に焦点を当てて精度向上を目指すことによって,結果的にシステムの精度向上が期待できる.

2．　関連研究と提案手法

今回,顔表情認識のモデルを作成するにあたり,LASTED(Language-guided SynThEsis Detection)を使用する.

LASTED は,AI の生成画像の検出を主な目的とし

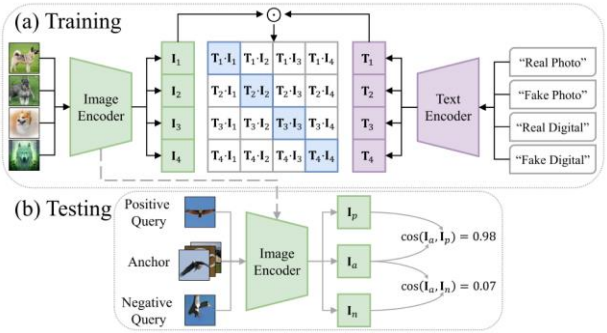

図 1. LASTED の構造

て開発されたアーキテクチャであり， 言語と画像のマルチモーダルモデルの一つである CLIP[2]をベースに、特定のテキストラベルを使用してより専門的なタスクに対応している.

LASTED の特徴は,言語誘導型コントラスト学習を利用している点にある.このアプローチでは,画像とテキストのペアを用いて学習を行い,画像の特徴を高精度に抽出する.また,LASTED は CLIP と同様のコントラスト学習手法を用いており,視覚と言語の両方の情報を統合することで,より一般化された特徴を抽出する.

LASTED は合成画像の検出を主な目的としているが,同じクラス間の距離が微妙な顔表情認識にも適用できると考えられる.顔表情認識においては,微細な表情の違いを高い精度で認識することが求められる.LASTED の言語誘導型コントラスト学習は,このような微妙な違いを捉える能力を持って

おり,顔表情認識モデルの性能向上に寄与することが期待される.

3. 実験

提案手法で紹介した LASTED を使用し,顔表情認識の実験を行う.比較対象として, LASTED のバックボーンとして使われている ResNet50x60 と比較する.

3．1. データセット

FER2013 は,Kaggle で開催された FER チャレンジで導入されたデータセットである.このデータセットは,Google の画像検索 API を使用して収集されたもので,約 36,000 枚のサンプル画像が含まれている.これらのサンプル画像は 1 人のタグ付け者によって surprise , sad, happy, fear, disgust, angry, neutral の 7 つの基本的な表情ラベルが付けられている.

3．2. 実験結果

実験結果として、それぞれの Confusion Mtrix を提示する.

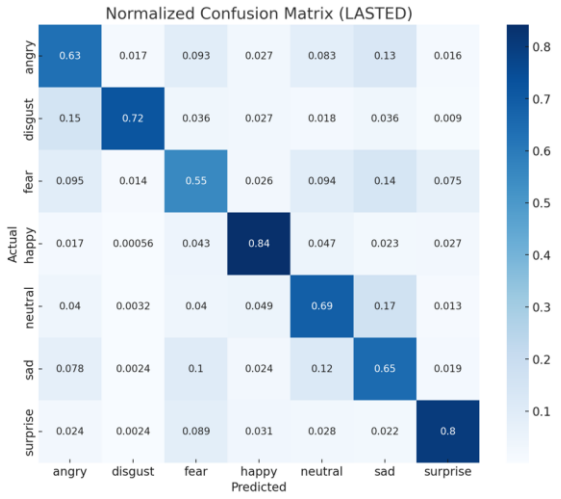

図 2. 実験結果(LASTED)

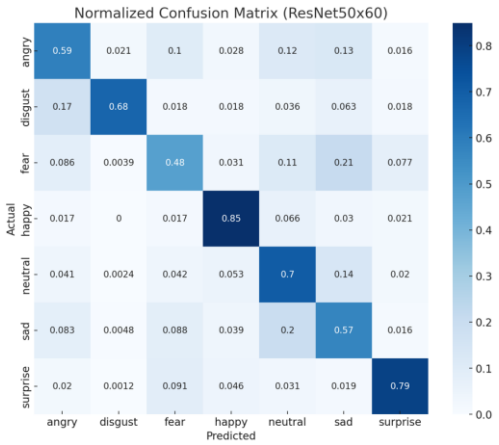

図 3. 実験結果(ResNet50x60)

表 1.平均 accuracy

	average acc
LASTED	0.697
RESNET50*60	0.665

4. 今後の展望

現状バックボーンと比べて精度が向上し,ネガティブな感情の精度も向上したことは目標通りであるが、既存の顔表情モデルと比べると精度が低いのが課題点である.

今後行う予定として,CLIP 内の画像エンコーダを別のネットワークに変更し,精度向上を目指す予定である.

LASTED に内蔵されている CLIP は画像エンコーダと言語エンコーダの 2 種類が存在し, 画像エンコーダのアーキテクチャは ResNet50 か Vision Transformer のいずれかを選択して使用する.これらは一般的なアーキテクチャであるため,顔表情認識向けのアーキテクチャに変更すれば,既存の顔表情認識モデルの精度を上回ることが期待できる.

参考文献

[1] Wu, H.; Zhou, J.; and Zhang, S. 2023. Generalizable synthetic image detection via language-guided contrastive learning. arXiv preprint arXiv:2305.13800.