修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/26/2026 (MM/DD/YYYY)

専攻名(専門分野) Department	情報理工・ 情報通信専攻	氏 名 Name	細郷 壮希	指 導 教 員 Advisor	渡辺 裕 印 Seal
研究指導名 Research guidance	オーディオビジュアル情報処理研究	学籍番号 Student ID number	CD 5124F048-2		
研究題目 Title	文脈情報に基づくフレーム選択による動画内人物に焦点を当てた連続感情認識 Continuous Emotion Recognition via Frame Selection Based on Contextual Information Focusing on Characters in Video				

1. まえがき

映像からの連続感情認識は，監視や公共案内などの分野で重要性を増している．社会インフラの高度化や人中心 AI の発展により，人の内的状態を時間的に把握する技術への需要が高まっており，感情の時間的推移を扱う連続感情認識は，静的な感情分類を超える基盤技術として注目されている．特に第三者視点においては，音声や明示的な対話に頼らず，視覚的シーンに埋め込まれた文脈的手がかりを活用して感情状態を推定する必要がある．従来の連続感情認識手法は，固定間隔フレームあるいは全フレームをモデル入力として使用している．しかし，これらのアプローチは連続フレーム間の高い視覚的・意味的冗長性により，重要な感情変化や状況変化に対する感度が低下するという問題がある．

この課題に対処するために，本研究ではフレーム間の視覚的な非類似性に着目した新たな感情認識手法を提案する．具体的には，ターゲットフレームの前後から視覚的に異なるフレームを動的に選択し，それらを時系列順に構成した入力系列とする．この選択により，ターゲットの感情状態に関連するフレームが強調される．時間的連続性を重視する従来手法とは異なり，非類似性ベースの文脈選択は感情変化を捉えることができ，モデルの解釈性と汎化性を向上させる．

2. 関連手法

2.1 感情認識手法

連続感情認識では，CNN により各フレームの視覚特徴を抽出し，LSTM などの時系列モデルによって時間的変化をモデリングする手法が一般的である．近年では，自己注意機構に基づく Transformer モデルが長距離の時間的依存関係を捉える能力から注目を集めている．特に Vision Transformer(ViT)[1]は，時間的に離れたフレーム間の関係性を直接モデル化できる点で，文脈的感情変化の表現に適している．しかし，これらの手法においても，入力フレーム集合の構成は均一なサンプリングや全フレームを一様に入力する設計に依存している．

2.2 感情認識データセット

実世界の感情行動を捉えたデータセットである Aff-Wild[2]は広く使用されており，主に顔情報に焦点が当てられている．しかしながら，これらのデータセットは文脈情報や背景情報が限定的である．一方，VEATIC[3]データセットでは動画に対して二次元感情空間を構成する Valence(快－不快)および Arousal(覚醒度)が連続的に注釈付けされており，顔と環境の両方の文脈を含んでいる．これにより，人物の動作や背景変化を含む映像全体の情報に基づく感情変化の評価が可能となる．

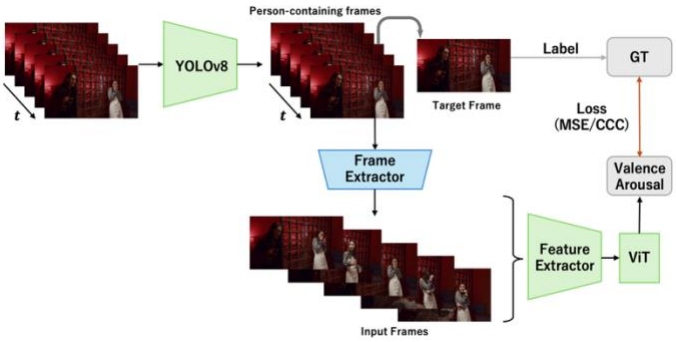

図1 提案手法の概要図(本図に含まれる例示画像は，https://www.pexels.com より取得した著作権フリー素材を使用している)

3. 提案手法

本研究では，第三者視点動画を対象として (1) 人物検出によるフレーム候補の制約，(2) 視覚的非類似度に基づくフレーム選択，(3) 選択フレーム集合を用いた感情推定モデルから構成される連続感情認識手法を提案する．

まず入力動画から得られるフレーム列に対して YOLOv8n[4]を用いた人物検出を行い，人物が含まれるフレームのみを後続のフレーム選択の候補集合として抽出する．第三者視点動画には人物不在フレームや背景中心フレームが多く含まれるため，候補を人物中心に制約することで，感情推定に寄与しにくい視覚的ノイズの混入を抑制する．推論時には，人物不在フレームを含めて連続的な推定系列を得る．

次に，人物を含む候補フレーム集合に対し，従来手法[3]に則り計 5 枚のフレームを選択する．具体的には図2のようにターゲットフレームを基準として過去・未来方向から視覚的に異なるフレームを探索する．フレームI_1とI_2の間の類似度は画素差分に基づき次式で定義する．

図2　入力フレームの選択方法（本図に含まれる例示画像は，https://www.pexels.com より取得した著作権フリー素材を使用している）

$$S(I_1, I_2) = 1 - \frac{1}{255} \cdot mean(|I_1 - I_2|) \quad (1)$$

Sは値が大きいほど類似度が高く，類似度が閾値τを下回るフレームを「非類似」と判定し選択する．本研究では，ターゲット 1 枚に対して過去 2 枚・未来 2 枚を選択し，計 5 枚を時間順に整列して入力フレーム集合を構成する．この入力設計により，均一サンプリングでは捉えにくい時間的・文脈的遷移（動作変化，人物間相互作用など）を効率的にモデルへ与えることができる．

最後に，選択された 5 フレームを感情推定モデルに入力し，Valence/Arousal を連続値として推定する．各フレームから ResNet-50[5]で空間特徴を抽出し，ViT によりフレーム間依存を考慮した特徴表現へ変換して回帰する．学習には，平均二乗誤差（MSE）と Concordance Correlation Coefficient（CCC）を組み合わせた損失関数を用い，以下で定義する．

$$\mathcal{L} = \mathcal{L}_{CCC} + \lambda \cdot \mathcal{L}_{MSE} \quad (2)$$

ここに，λの設定は従来手法に従い 0.1 とする．

4.　実験

4.1 データセット

映画やドキュメンタリー等から構成され，各フレームに Valence/Arousal の連続アノテーションが付与された VEATIC データセットを用いた．音声情報は除去されており，視覚情報のみで文脈的感情理解を評価できる．本研究では従来手法に従い，各動画を時間方向に前半 70%（学習）・後半 30%（評価）に分割した．

4.2 評価指標

VEATIC モデルに従い，CCC，PCC，RMSE，SAGR を用いた．特に CCC は相関に加えて平均・分散の一致度も評価できるため主要指標として扱う．

4.3 比較手法と実験設定

感情推定モデル構成は共通とし，入力フレーム選択のみを変更して比較した．ベースラインは VEATIC の均一サンプリング設定に準拠し，固定間隔$k = 5, 25, 50$で 5 枚の連続フレームを選択する．一方，提案手法ではターゲットフレームを中心に非類似フレームを選択し 5 枚を構成する．類似度指標として，画素差分に基づく Diff@τ，人物フレームのみに対する画素差分に基づく PeopleDiff@τ，SSIM に基づく SSIM@τを比較した．

4.4 結果と考察

実験結果を表1に示す．画素差分に基づく手法は，均一サンプリングを一貫して上回る性能を示した．特に PeopleDiff@0.80 は両次元で最も高い CCC を達成した．これは，人物不在フレームや背景中心フレームを除外することで，感情推定に寄与しにくい視覚的ノイズが抑制され，対象人物の状態変化に基づく学習が安定したためと考えられる．一方，SSIM に基づく手法は一定の改善は示すものの，変化点抽出という本課題の目的に対しては画素差分の方が適合した．

表1　各手法の実験結果

Dimension	Method	CCC ↑	PCC ↑	RMSE ↓	SAGR ↑
Valence	VEATIC (k=5)	0.609	0.644	0.303	0.789
	VEATIC (k=25)	0.624	0.670	0.293	**0.798**
	VEATIC (k=50)	0.609	0.655	0.301	0.785
	Ours (Diff@0.75)	0.677	0.738	0.261	0.797
	Ours (Diff@0.80)	0.687	**0.750**	**0.258**	0.797
	Ours (PeopleDiff@0.80)	**0.723**	0.736	0.278	0.788
	Ours (SSIM@0.75)	0.606	0.688	0.285	0.769
	Ours (SSIM@0.80)	0.612	0.691	0.282	0.771
	Ours (SSIM@0.85)	0.599	0.679	0.288	0.766
Arousal	VEATIC (k=5)	0.630	0.668	0.210	0.779
	VEATIC (k=25)	0.641	0.684	0.202	0.768
	VEATIC (k=50)	0.622	0.653	0.214	0.764
	Ours (Diff@0.75)	0.670	0.733	0.190	0.803
	Ours (Diff@0.80)	0.685	**0.746**	**0.182**	**0.804**
	Ours (PeopleDiff@0.80)	**0.698**	0.721	0.201	0.798
	Ours (SSIM@0.75)	0.622	0.693	0.206	0.785
	Ours (SSIM@0.80)	0.608	0.692	0.205	0.772
	Ours (SSIM@0.85)	0.607	0.691	0.200	0.780

5.　結論

本研究では，ターゲットフレームの前後から視覚的に非類似なフレームを動的に選択する連続感情認識手法を提案した．VEATIC を用いた評価の結果，提案手法はすべての均一サンプリングベースラインを上回る性能を示した．この結果から，時間的・文脈的に多様なフレームを入力に反映することが連続感情認識の精度向上に有効であることが確認された．今後の課題として，複数人物の識別や背景情報などの文脈情報の統合が残されている．

参考文献

[1] A. Dosovitskiy *et al.,* “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” Int. Conf. Learning Representations (ICLR), 2021.
[2] S. Zafeiriou et al., “Aff-Wild: Valence and Arousal ‘In-the-Wild’ Challenge,” IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 34–41, Jul. 2017.
[3] Z. Ren *et al.,* “VEATIC: Video-Based Emotion and Affect Tracking in Context Dataset,” in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), pp. 2484–2494, Jan. 2024.
[4] G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by Ultralytics,” Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778, Jun. 2016.

2025 年度

早稲田大学大学院　基幹理工学研究科 情報理工・情報通信専攻　修士論文

文脈情報に基づくフレーム選択による動画内人物に焦点を当てた連続感情認識

Continuous Emotion Recognition via Frame Selection Based on Contextual Information Focusing on Characters in Video

細郷 壮希
(5124F048-2)

提出日：2026 年 1 月 26 日

指導教員：渡辺 裕 教授

研究指導名：オーディオビジュアル情報処理研究

論文要旨

近年，監視カメラやモバイル端末，動画配信サービスの普及に伴い，第三者視点から人間の行動や状態を理解する技術への関心が高まっている．中でも，動画を用いた連続感情認識は，セキュリティや公共空間における監視，ヒューマン・コンピュータ・インタラクションなど，対象者と直接的に対話することが困難な場面において重要な役割を果たすと期待されている．しかし，従来の多くの研究は顔表情に強く依存しており，周囲の状況や文脈に基づく感情変化を十分に捉えられていないという課題がある．特に第三者視点動画では，顔の遮蔽や人物不在フレームが頻繁に発生するため，表情情報のみに基づく感情推定は不安定になりやすい．

本研究では，第三者視点動画を対象とした連続感情認識において，入力フレームの選択方法に着目した新たな感情推定手法を提案する．提案手法では，まず人物検出モデルを用いて人物が含まれるフレームに処理対象を限定し，感情推定に寄与しにくい背景フレームの影響を抑制する．次に，人物が含まれるフレーム集合に対して，フレーム間の視覚的非類似度に基づく動的フレーム選択を行い，感情変化に関与する時間的・文脈的遷移を含む 5 枚のフレームを 1 セットとして感情推定モデルに入力する．この構成により，時間的冗長性を抑制しつつ，人物の状態変化や周囲環境を反映した文脈情報を効果的に活用することを可能とする．

評価には，第三者視点動画に対して Valence および Arousal の連続アノテーションが付与された VEATIC を用いた．均一サンプリングに基づく既存手法や，構造的類似度指標（SSIM）を用いたフレーム選択手法との比較実験を行った結果，人物抽出を行った上で画素差分に基づく非類似度を用いる提案手法が，Concordance Correlation Coefficient（CCC）において良好な性能を示すことを確認した．これらの結果から，第三者視点動画における連続感情認識では，人物を中心とした文脈に基づくフレーム選択が有効であることが示唆された．

キーワード

連続感情認識，第三者視点動画，文脈理解，フレーム選択，Valence–Arousal

目次

図目次

表目次

第 1 章

序論

1.1 研究背景

近年，動画配信サービスの普及，オンライン会議の常態化，監視カメラやモバイル端末による動画取得・共有の一般化に伴い，動画コンテンツの生成および利活用の機会は急速に拡大している．これにより，動画を通じて人間の行動や状態を理解する技術への関心が高まっており，特に人間の内的状態を推定する感情認識（Emotion Recognition）は，セキュリティ，公共誘導，ヒューマン・コンピュータ・インタラクション（Human–Computer Interaction: HCI）など多様な分野で重要性を増している．

動画を用いた感情認識では，時間方向に連続する情報を扱うことが可能であるため，静止画像と比較して，感情状態の時間的推移や変化過程を捉えられる点に特徴がある．一方で，動画は時間的冗長性を含む大規模データであり，感情推定に有効な情報と無関係な情報が混在している．このため，どのフレームや視覚的手がかりを用いて感情を推定するかという入力設計は，推定精度およびモデルの安定性に大きな影響を与える．

これまでの感情認識研究の多くは，顔表情に基づくアプローチを中心に発展してきた．畳み込みニューラルネットワーク（Convolutional Neural Network: CNN）や再帰型ニューラルネットワーク（Recurrent Neural Network: RNN），さらには Transformer モデルを用いて，表情の時間的変化から感情状態を推定する手法が数多く提案されている．しかし，表情情報のみに依存した感情推定には本質的な限界が存在する．すなわち，同一の表情であっても，文脈や状況によって異なる感情状態が対応し得るため，表情だけから感情を一意に定めることは困難である．例えば，涙を流す表情は，喜びに起因する場合もあれば，悲しみや悔しさに起因する場合もあり，表情単体では感情の違いを識別できない．

この問題は，第三者視点で撮影された動画において特に顕在化する．第三者視点動画では，顔が小さく映る，遮蔽される，あるいは画面内に人物が存在しないといった状況が頻繁に発生する．そのため，顔表情に依存した手法では安定した感情推定が難しく，人物の姿勢や行動，周囲の環境，人物間の関係性といった文脈情報を考慮する必要性が高まっている．

このような背景のもと，文脈情報を含む第三者視点動画を対象とした連続感情認識が注目されている．連続感情認識では，感情状態を Valence（快－不快）および Arousal（覚醒度）といった連続値として扱い，時間的に変化する感情の推移を捉えることを目的とする．文脈を重視したデータセットとして，Video-Based Emotion and Affect Tracking in Context （VEATIC） [1] が提案され，第三者視点における感情理解のベ

ンチマークとして用いられている．VEATIC は，公共空間や映像アーカイブなど，音声情報の取得が困難，あるいは利用が制限される状況を想定して設計されており，視覚情報のみに基づく感情推定が求められる．

しかしながら，VEATIC を含む第三者視点動画を対象とした従来手法では，評価の簡便性や手法間の公平な比較を目的として，時間軸上で一定間隔にフレームを抽出する均一サンプリングが採用されることが多い．このような方法では，視覚的に冗長なフレームや，感情推定に寄与しにくいフレームが入力に含まれる可能性が高く，感情変化に関与する重要な時間的遷移や文脈的差異が十分に活用されない．

以上より，第三者視点動画において，表情情報に過度に依存せず，人物およびその周囲の文脈を考慮しつつ，感情推定に有効なフレームを適切に選択する入力設計の在り方が，重要な研究課題として位置付けられる．

1.2 研究目的

本研究の目的は，第三者視点で撮影された動画を対象とした連続感情認識において，以下の三点を満たす感情推定手法を確立することである．

（i）表情情報に過度に依存せず，人物およびその周囲の文脈情報を活用可能な入力構成を設計すること，

（ii）時間方向に冗長なフレームを抑制しつつ，感情変化に寄与する視覚的遷移を効果的に捉えること，

（iii）入力フレーム数を固定した条件下において，推定精度の向上を図ることである．

本研究では，フレーム選択に先立って人物が含まれるフレームに処理対象を限定し，感情推定に寄与しにくい背景情報の影響を抑制する．その上で，視覚的非類似性に基づく動的フレーム選択を行うことで，人物の状態変化や行動，周囲の状況といった文脈情報を反映した連続感情推定の実現を目指す．

1.3 本論文の貢献

本論文の主な貢献は以下のとおりである．

- 連続感情認識を対象とし，フレーム選択に先立って人物が含まれるフレームに処理対象を限定する入力設計を導入することで，表情情報に依存しない，人物中心の文脈情報に基づく感情推定の枠組みを提示した．
- 視覚的非類似度に基づく動的フレーム選択手法を提案し，入力フレーム数を固定した条件下においても，感情変化に寄与する時間的・文脈的遷移を効果的に捉えられることを示した．
- 文脈情報を含む第三者視点動画を対象とした VEATIC データセットを用いた実験を通じて，提案手法が均一サンプリングに基づく既存手法と比較して，連続感情推定の精度向上に有効であることを定量的に検証した．

1.4 本論文の構成

本論文の構成を以下に示す．第 1 章は本章であり，本論文の研究背景，研究目的，および貢献について述べる．第 2 章では，本研究で用いる従来の感情推定手法および関連技術について述べる．第 3 章では，本研究で

提案する手法について述べる．第 4 章では，本研究の実験内容およびその結果と考察を述べる．第 5 章では，本研究の結論と今後の課題について述べる．

第 2 章

関連研究

2.1　まえがき

本章では，連続感情認識に関する従来研究を概観し，本研究の位置付けを明確にする．特に，感情認識手法を「表情ベースの手法」と「文脈ベースの手法」に大別し，それぞれの特徴と課題を整理する．その上で，第三者視点動画における連続感情認識において未解決となっている問題点を明らかにし，本研究が取り組む課題への導入とする．

2.2　表情ベースの感情認識

感情認識に関する初期の研究では，顔表情が感情状態を直接的に反映する重要な視覚的手がかりであると考えられ，表情情報に基づく手法が中心的に研究されてきた．従来手法では，顔画像から抽出した特徴量を用いて感情を分類する枠組みが主流であり，連続感情認識においても，表情の時間的変化を捉えるために再帰型ニューラルネットワーク（RNN）や畳み込みニューラルネットワーク（CNN）が広く用いられてきた [2]．例えば，CNN と LSTM を組み合わせた手法 [3] では，各フレームから抽出した表情特徴を時系列として処理することで，感情状態の連続的推移を捉えることが試みられている．また，時間方向の情報を空間情報と同時に扱う手法として，3D 畳み込みニューラルネットワーク（3D-CNN） [4] が提案されている．3D-CNN は，複数フレームを一括して入力とすることで，空間的特徴と時間的変化を統合的に学習できる点に特徴がある．しかし，計算コストが高く，時間的に冗長な情報をそのまま処理してしまうという課題を有する．

近年では，長距離の時間的依存関係を効果的に捉えられる Transformer モデルが注目されている．Vision Transformer（ViT） [5] をフレーム単位の特徴抽出器として用い，その出力系列を自己注意機構により統合する手法では，時間的に離れたフレーム間の関係性を考慮した感情推定が可能であることが示されている．さらに，注意機構を明示的に導入した Transformer ベースの手法 [6] では，微細な表情変化に着目することで高精度な感情推定が実現されている．

しかし，これらの表情ベース手法は顔表情が明瞭に観測可能であることを前提としており，第三者視点映像のように顔の遮蔽や人物不在フレームが頻発する条件下では安定した感情推定を行うことが構造的に困難である．

2.3 文脈ベースの感情認識

表情ベース手法の限界を背景として，近年では人物の姿勢や行動，周囲の環境，人物間の関係性といった文脈情報を考慮した感情認識手法が注目されている．文脈ベースの感情認識では，顔表情を感情の唯一の手がかりとするのではなく，全身動作や背景，社会的相互作用を含めたより包括的な情報に基づいて感情状態を推定することを目指す．

このような研究を支援するデータセットとして，AFEW-VA [7] や Aff-Wild，Aff-Wild2 [8] が提案されてきた．これらのデータセットは，自然環境下で撮影された動画に valence および arousal の連続アノテーションを付与している点で重要な貢献を果たしている．しかし，動画の多くは顔領域が中心であり，背景や人物間の関係性といった広範な文脈情報を十分に含んでいるとは言い難い．

これに対し，VEATIC は，映画やドキュメンタリーから抽出された第三者視点の全身映像を対象とし，valence および arousal の連続的アノテーションを提供している．VEATIC は，身体姿勢，行動，背景，社会的相互作用といった多様な文脈情報を含む点に特徴があり，表情に依存しない感情理解を評価するための代表的なベンチマークとして位置付けられる．

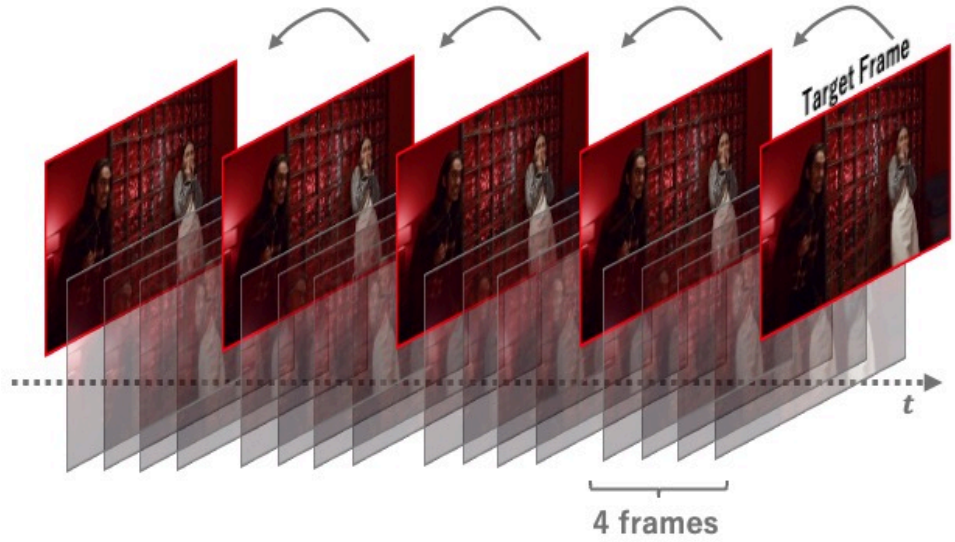

図 2.1 VEATIC ベンチマークにおける均一サンプリングに基づくフレーム選択 (本図に含まれる例示画像は，https://www.pexels.com より取得した著作権フリー素材を使用している)

2.4 従来研究の問題点

表情ベースの感情認識手法は，顔表情の詳細な変化を捉えられる一方で，同一の表情が異なる感情状態に対応し得るという本質的な問題を抱えている．例えば，涙を流す表情は喜びに起因する場合もあれば，悲しみや悔しさに起因する場合もあり，表情情報のみから感情を一意に定めることは困難である．この問題は，第三者視点動画において特に顕在化し，顔の視認性が低下する状況では表情ベース手法の適用が大きく制限される．

一方，文脈ベースの感情認識手法においても課題は残されている．VEATIC ベンチマークを含む従来手法

では，図 2.1 のように時間軸上で一定間隔にフレームを抽出する均一サンプリングが採用されており，視覚的に冗長なフレームや，感情推定に寄与しにくいフレームが入力に含まれる可能性がある．処理効率や冗長性低減を目的としたフレーム選択に関する研究も存在するが，それらは離散的な感情分類を対象としており，連続感情認識への応用は限定的である．

2.5 むすび

本章では，連続感情認識に関する従来研究を，表情ベース手法と文脈ベース手法の観点から整理した．その結果，第三者視点動画における連続感情認識では，表情情報に依存しない文脈重視の設計と，感情変化に寄与するフレームを適切に選択する入力設計が依然として重要な課題として残されていることを示した．

第 3 章

提案手法

3.1　まえがき

本研究では，第三者視点動画を対象とした連続感情認識において，入力フレームの選択方法に着目した感情推定手法を提案する．提案手法は，(1) 人物検出に基づくフレーム候補の制約，(2) 視覚的な非類似度に基づく動的フレーム選択，(3) 選択されたフレーム集合を入力とする感情推定モデルという三段階の処理構成を有する．

本章では，まず提案手法の全体構成を示した後，各処理段階について順に説明する．

3.2　提案手法

3.2.1　提案手法の全体構成

提案手法の全体構成を図 3.1 に示す．本手法では，入力動画から得られるフレーム列に対して，まず人物検出を行い，人物が含まれるフレームのみを処理対象として抽出する．次に，抽出されたフレーム集合に対して，視覚的な非類似度に基づく動的フレーム選択を行い，感情変化に寄与する時間的・文脈的遷移を含む 5 枚のフレーム集合を構成する．最後に，選択されたフレーム集合を感情推定モデルに入力し，Valence および Arousal の連続値を推定する．

3.2.2　人物検出によるフレーム候補の制約

第三者視点動画には，人物が映っていないフレームや，感情推定に寄与しにくい背景のみのフレームが多く含まれる．このようなフレームが入力に含まれると，人物の状態変化や行動といった感情に関連する手がかりが希薄となり，感情推定の安定性が低下する可能性がある．

そこで本研究では，フレーム選択に先立ち，物体検出モデル YOLOv8 [9] を用いた人物検出を行い，人物が含まれるフレームのみを後続の処理対象とする．この前処理により，入力候補を人物中心のフレームに限定し，感情推定に有効な視覚情報に基づくフレーム選択を可能とする．なお，テスト時の推論は，人物が検出されたフレームと検出されないフレームで処理を分けて行う．目的は，学習時と同様に「人物が写るフレーム」

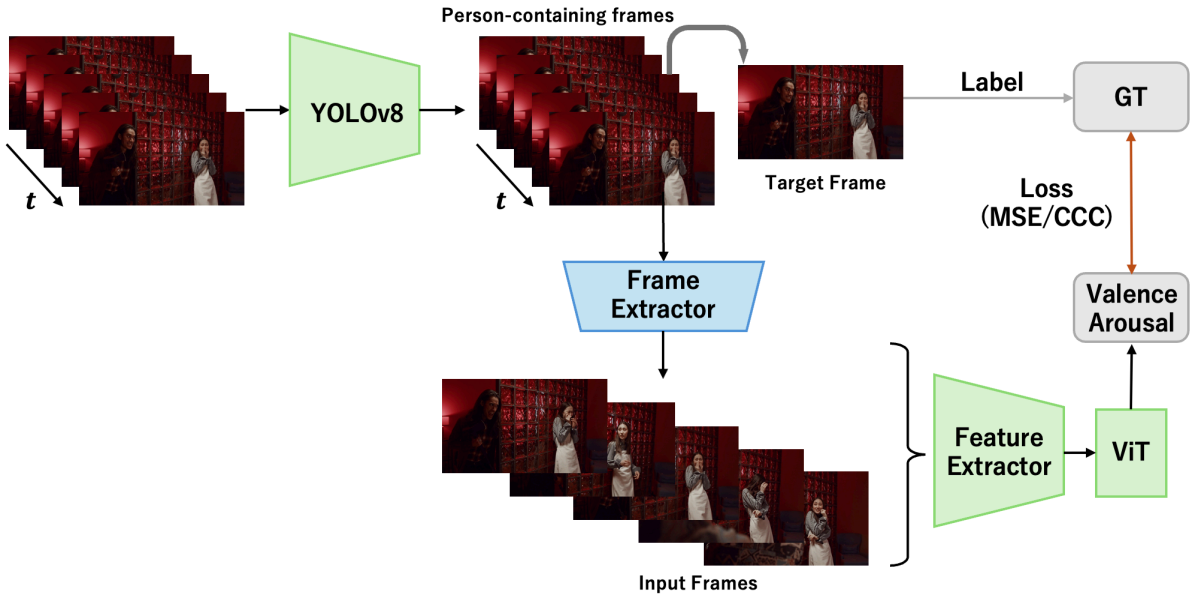

図 3.1 提案手法の全体構成 (本図に含まれる例示画像は，https://www.pexels.com より取得した著作権フリー素材を使用している)

を主に用いて推定しつつ，人物不在フレームに対しても動画全体として連続的な予測系列を得ることである．推論手順を以下に示す．

- **人物検出フレーム：** YOLOv8 により人物が検出されたフレームに対して，訓練済みモデルを用いて valence / arousal を直接推定する．
- **人物不在フレーム：** YOLOv8 により人物が検出されなかったフレームに対しては，同一動画内の直近の「推定済み人物フレーム」を参照し，推定値を補間する．

人物不在フレームのフレーム番号を t_i とし，その直前および直後に存在する推定済み人物フレームのフレーム番号をそれぞれ t_p，t_n とする．また，それらの推定値を $\hat{y}_p$，$\hat{y}_n$ とする．t_p と t_n の両方が存在する場合は線形補間を用い，補間値 $\hat{y}_i$ を次式で計算する．

$$\hat{y}_i = (1-\alpha)\hat{y}_p + \alpha\hat{y}_n \tag{3.1}$$

$$\alpha = \frac{t_i - t_p}{t_n - t_p} \tag{3.2}$$

直前の推定済み人物フレームのみが存在する場合は $\hat{y}_i = \hat{y}_p$，直後のみが存在する場合は $\hat{y}_i = \hat{y}_n$ を用いる．この処理により，人物が検出されない区間に対しても，前後の推定結果に整合した連続的な推定系列を得ることができる．

3.2.3 視覚的な非類似度に基づく動的フレーム選択

人物が含まれるフレーム集合に対して，フレーム間の視覚的な非類似度に基づく動的フレーム選択を行う．従来の均一サンプリングでは，時間的に近接した視覚的に類似したフレームが連続して選択される可能性があ

図 3.2 視覚的な非類似度に基づくフレーム選択の概略 (本図に含まれる例示画像は，https://www.pexels.com より取得した著作権フリー素材を使用している)

り，感情変化に関与する文脈的な遷移が十分に反映されない場合がある．

本研究では，ある基準フレーム（ターゲットフレーム）に対し，過去方向および未来方向へ探索を行い，視覚的に十分異なるフレームを選択する．フレーム探索手順の概略を図 3.2 に示す．RGB フレーム I_1 および I_2 の類似度 $S(I_1, I_2)$ は，画素差分に基づき次式で定義する．

$$S(I_1, I_2) = 1 - \frac{1}{255} \cdot \mathrm{mean}(|I_1 - I_2|) \tag{3.3}$$

なお，本研究で用いる類似度指標 $S(\cdot, \cdot)$ は，いずれも値が大きいほどフレーム間の類似性が高くなるように定義している．画素差分に基づく類似度 $S(I_1, I_2)$ は $[0, 1]$ の値域をとり，1 に近いほど両フレームが視覚的に類似していることを示す．フレーム選択においては，類似度があらかじめ定めた閾値 τ を下回った場合に，当該フレームを視覚的に非類似なフレームとして選択する．

ターゲットフレームに対して，過去方向および未来方向からそれぞれ 2 枚ずつフレームを選択し，合計 5 枚のフレームを時間順に並べて入力フレーム集合を構成する．なお，本研究では VEATIC データセットの評価設定に準じて，入力フレーム数を 5 枚に固定している．

3.2.4 感情推定モデル

視覚的な非類似度に基づいて選択されたフレーム集合を入力として，連続的な感情状態を推定する感情推定モデルを用いる．本研究では，各入力フレームから空間的特徴を抽出するバックボーンネットワークと，フレーム系列の時間的依存関係をモデル化する時系列モデルから構成されるモデルを採用する．

空間特徴抽出器として ResNet-50 [2] を用い，各フレームを独立に処理することで高次の視覚特徴を得る．得られたフレーム特徴系列は，Vision Transformer（ViT）に入力され，自己注意機構によりフレーム間の時間的関係性を考慮した特徴表現へと変換される．ViT は，時間的に離れたフレーム間の依存関係も柔軟に捉えることが可能であり，第三者視点動画における文脈的な感情変化のモデリングに適している．

なお，本研究では，感情推定モデルの基本構成自体は先行研究に準じており，入力として与えるフレーム集合の構成方法に主眼を置く．

3.2.5 損失関数と学習方法

感情推定モデルの学習には，VEATIC において採用されている損失関数および学習設定に準拠した構成を用いる．予測値と正解値の局所的な誤差と，系列全体としての統計的一致性を同時に考慮するため，Mean Squared Error（MSE）損失と Concordance Correlation Coefficient（CCC）損失を組み合わせた損失関数

を用いる．

予測値 x と正解値 y に対する CCC ρ_c は次式で定義される．

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{3.4}$$

ここで，s_x^2 および s_y^2 はそれぞれ分散，$\bar{x}$ および $\bar{y}$ は平均値，s_{xy} は共分散を表す．

Valence および Arousal に対する CCC をそれぞれ ρ_v，ρ_a とすると，CCC 損失 $\mathcal{L}_{\mathrm{CCC}}$ は次式で定義される．

$$\mathcal{L}_{\mathrm{CCC}} = 1 - \frac{\rho_a + \rho_v}{2} \tag{3.5}$$

一方，系列長を T，時刻 t における予測値および正解値をそれぞれ x_t，y_t とすると，MSE 損失 $\mathcal{L}_{\mathrm{MSE}}$ は次式で定義される．

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{T}\sum_{t=1}^{T}(x_t - y_t)^2 \tag{3.6}$$

最終的な学習損失 $\mathcal{L}$ は，次式で定義される．

$$\mathcal{L} = \mathcal{L}_{\mathrm{CCC}} + \lambda\mathcal{L}_{\mathrm{MSE}} \tag{3.7}$$

ここで，λ は損失の重み係数であり，VEATIC の設定に従い $\lambda = 0.1$ に固定した．

3.3 むすび

本章では，人物検出および視覚的な非類似度に基づくフレーム選択を用いた連続感情認識手法について，その処理構成および各段階の設計方針を述べた．従来の均一サンプリングに対し，入力フレームの選択方法を工夫することで，感情推定に有効な時間的・文脈的情報をより適切に反映できる点を示した．

第 4 章

実験

4.1　まえがき

本章では，連続感情認識において，「どのようなフレームを入力として選択すべきか」という問いに対し，視覚的な非類似度および人物制約の有効性を実験的に検証する．さらに，入力設計の違いおよび学習設定の違いが，感情推定性能に与える影響を定量的に評価する．

4.2　実験

4.2.1　データセット

本研究では, 第三者視点動画における連続感情認識を対象とした Video-based Emotion and Affect Tracking in Context Dataset（VEATIC）を用いた．VEATIC は，映画，ドキュメンタリーなどから構成される文脈重視型の連続感情認識ベンチマークであり，各フレームに対して valence（快・不快度）および arousal（覚醒度）の連続値アノテーションが付与されている．

本データセットの特徴は，顔表情に依存した感情推定ではなく，身体動作，背景情報，人物間の関係性といった視覚的文脈情報に基づく感情理解を評価可能である点にある．実際の社会的場面では，対象人物の顔が明確に観測できない状況や，文脈情報が感情判断に大きく影響する場面が多く存在する．VEATIC は，こうした状況を多く含む動画を収録することで，表情情報に過度に依存しない感情推定手法の検証を可能としている．

各動画ごとに感情を追跡する対象人物が事前に指定されており，アノテータはその人物の感情状態を動画の時間経過に沿って連続的に評価する．アノテーションは，図 4.1 に示す専用のユーザインタフェースを用いて行われ，valence–arousal の 2 次元感情空間上でマウス操作によりリアルタイムに感情状態を追跡する形式が採用されている．また，VEATIC では視覚的文脈情報の影響を明確に評価する目的から音声情報は除去されている．このため，本研究はマルチモーダル感情認識を対象とするものではなく，視覚情報のみに基づく文脈的感情推定を主眼としている．なお，本研究では VEATIC のベンチマーク設定に従い，各動画を時間方向に前半 70%（学習区間）と後半 30%（テスト区間）に分割した．すなわち，分割は動画単位ではなく，各動画内の時間区間に基づいて行う．以降の実験は，この分割に基づいて学習および評価を行う．

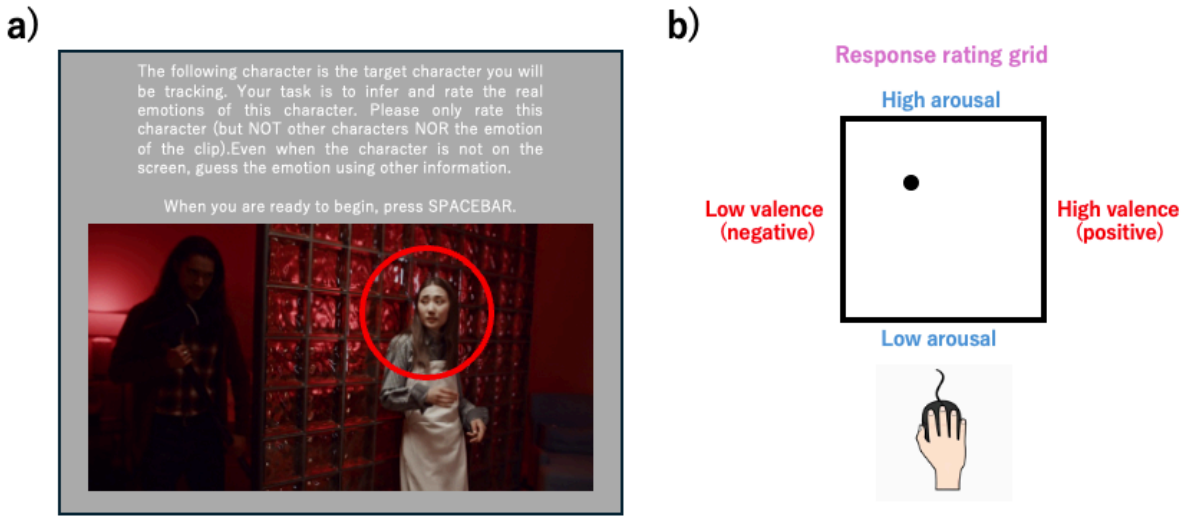

図 4.1 VEATIC における感情アノテーション用のユーザインタフェース．(a) 各動画の開始前に，アノテーターに提示される感情推定対象となるキャラクターおよび課題．(b) 映像再生中に表示される valence–arousal の 2 次元評価グリッド．(本図に含まれる例示画像は，https://www.pexels.com/より取得した著作権フリー素材を使用している)

4.2.2 評価指標

感情推定性能の評価には，VEATIC のベンチマーク設定に従い，Concordance Correlation Coefficient（CCC），Pearson Correlation Coefficient（PCC），Root Mean Squared Error（RMSE），Sign Agreement Metric（SAGR）を用いた．

これらのうち，CCC は相関に加えて平均値および分散の一致度を同時に評価できる指標であり，連続感情認識において最も重要な評価指標として広く用いられている．そのため，本研究では CCC を主要指標として結果を解釈し，PCC，RMSE，SAGR は補助的指標として併せて報告する．CCC の定義式については第 3 章で述べたため，本節では CCC 以外の評価指標について説明する．

Pearson Correlation Coefficient（PCC）

PCC は，予測値と正解値の線形相関の強さを評価する指標である．テストサンプル数を N，i 番目のサンプルに対する予測値と正解値をそれぞれ $\hat{y}_i$，y_i とすると，PCC は次式で定義される．

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{4.1}$$

ここで，$\bar{\hat{y}}$ および $\bar{y}$ はそれぞれ予測値と正解値の平均を表す．PCC は $[-1, 1]$ の範囲をとり，値が 1 に近いほど予測値が正解値と同様の変動傾向を持つことを意味する．ただし，平均値や分散の一致度は考慮しないため，予測値に一定のバイアスが存在しても高い値を示す場合がある．

Root Mean Squared Error（RMSE）

RMSE は，予測値と正解値の絶対的な誤差量を評価する指標である．RMSE は次式で定義される．

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{4.2}$$

RMSE は 0 に近いほど予測誤差が小さいことを示し，予測値が正解値からどの程度乖離しているかを元の感情スケール上で直感的に解釈できる利点を持つ．一方で，大きな誤差に対して二乗による強いペナルティが課されるため，外れ値の影響を受けやすいという性質がある．

Sign Agreement Metric（SAGR）

SAGR は，予測値と正解値の符号が一致している割合を評価する指標であり，感情の増減方向が正しく推定できているかを確認するために用いられる．SAGR は次式で定義される．

$$\mathrm{SAGR} = \frac{1}{N}\sum_{i=1}^{N}\delta\Big(\mathrm{sign}(\hat{y}_i),\, \mathrm{sign}(y_i)\Big) \tag{4.3}$$

ここで，$\mathrm{sign}(\cdot)$ は符号関数，$\delta(\cdot,\cdot)$ は Kronecker のデルタ関数であり，予測値と正解値の符号が一致する場合に 1，一致しない場合に 0 を返す．SAGR は $[0,1]$ の範囲をとり，値が大きいほど感情変化の方向性を正しく捉えられていることを示す．ただし，誤差の大きさ自体は反映しないため，RMSE や CCC と併せて解釈する必要がある．

4.2.3 比較手法

本研究では，2 枚のフレーム間の視覚的類似度を表す関数を $S(\cdot,\cdot)$ と表記する．$S(\cdot,\cdot)$ は用いる手法に応じて定義が異なり，画素差分に基づく類似度（Diff，PeopleDiff）または Structural Similarity Index Measure（SSIM） [10] に基づく類似度（SSIM）を用いる．いずれの類似度指標においても，$S(\cdot,\cdot)$ は値が大きいほどフレーム間の視覚的類似性が高くなるように定義されている．フレーム探索では，類似度が閾値 τ を下回った場合に，当該フレームを「視覚的に非類似」と判定し，選択する．すべての手法において感情推定モデルの構成は共通とし，入力フレームの選択方法のみを変更して比較を行った．比較手法として，VEATIC において採用されている均一ダウンサンプリングに基づくフレーム選択手法をベースラインとした．具体的には，固定間隔 $k = 5, 25, 50$ で 5 枚の連続フレームを選択する手法を用いる．提案手法では，ターゲットフレームを中心に視覚的非類似度に基づいてフレームを選択する．類似度指標として，以下の手法を比較した．

- **Diff@τ**：画素差分ベース
- **PeopleDiff@τ**：人物フレーム事前抽出 + 画素差分ベース
- **SSIM@τ**：SSIM ベース

ここで τ は類似度に対する閾値を表す．

Diff@τ

画素差分に基づく視覚的類似度を用いたフレーム選択手法である．類似度の定義は第 3 章に示した式に従い，ターゲットフレームとの類似度が閾値 τ を下回るフレームを非類似フレームとして選択する．

PeopleDiff@τ

YOLOv8n により人物が検出されたフレームのみを候補集合として抽出した上で，Diff@τ と同一の画素差分に基づく類似度を用いてフレーム探索を行う手法である．人物が含まれない背景中心フレームの影響を抑制することを目的とする．

SSIM@τ

SSIM は，2 枚の画像間の類似度を，画素値の単純な差分ではなく，人間の視覚特性を考慮した輝度，コントラスト，構造情報の 3 要素に基づいて評価する指標である．画像 I_1 と I_2 に対する SSIM は，次式で定義される．

$$\mathrm{SSIM}(I_1, I_2) = \frac{(2\mu_{I_1}\mu_{I_2} + C_1)(2\sigma_{I_1 I_2} + C_2)}{(\mu_{I_1}^2 + \mu_{I_2}^2 + C_1)(\sigma_{I_1}^2 + \sigma_{I_2}^2 + C_2)} \tag{4.4}$$

ここで，μ_{I_1}, μ_{I_2} はそれぞれ画像 I_1, I_2 の平均輝度，$\sigma_{I_1}^2, \sigma_{I_2}^2$ は分散，$\sigma_{I_1 I_2}$ は共分散を表す．C_1, C_2 は分母が 0 になることを防ぐための定数である．SSIM の値域は $[0, 1]$ であり，値が大きいほど 2 枚の画像が構造的に類似していることを示す．

入力フレーム集合の事前生成

入力フレーム集合（5 フレーム組）は，学習・推論の都度に生成するのではなく，各手法のフレーム探索を事前に実行し，ターゲットフレームと対応する 5 フレーム（B2, B1, Target, F1, F2）の組を CSV ファイルとして保存する．学習時および推論時は，この CSV を参照して入力フレーム集合を取得する．

ここで，フレーム探索はターゲットフレームが属する区間（学習区間またはテスト区間）ごとに独立に実施し，各ターゲットに対する探索は同一動画内の同一区間に限定した．したがって，学習区間のターゲットを生成する際にテスト区間の情報を参照すること，ならびにテスト区間のターゲットを生成する際に学習区間の情報を参照することはない．

なお，動画端などの理由で所定枚数の前後フレームが得られない場合は，取得可能な範囲で探索を行い，不足分は最も近傍の選択済みフレームで補完した．

4.2.4 実験設定

各手法における類似度の閾値は，以下のように設定した．

- Diff@τ：$\tau \in \{0.75, 0.80\}$
- PeopleDiff@τ：$\tau = 0.80$
- SSIM@τ：$\tau \in \{0.75, 0.80, 0.85\}$

また，人物検出には YOLOv8n を用い，検出信頼度の閾値（confidence threshold）を 0.25 に設定した．信頼度が 0.25 以上の検出結果のみを有効な人物検出として採用した．推論時の入力画像サイズは 640×640 ピクセルに統一した．検出対象は COCO データセットにおける person クラス（クラス ID 0）である．

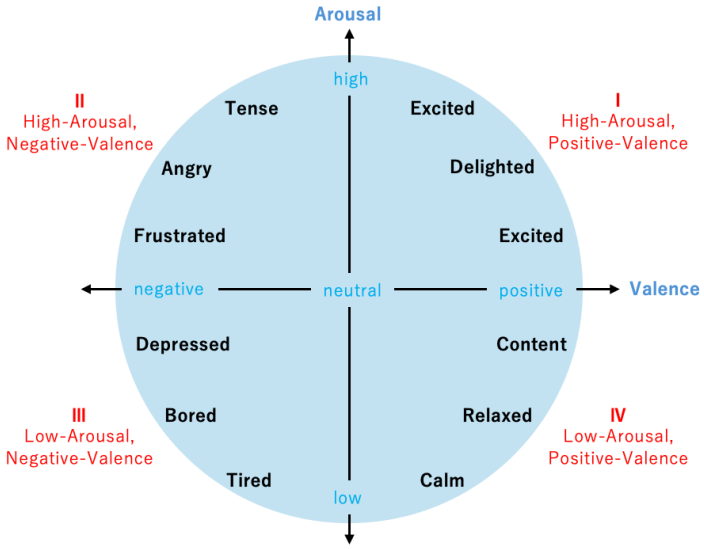

図 4.2 ラッセルの感情円環モデル

4.2.5 別学習と同時学習の比較

Arousal と Valence を同時に学習する設定と，それぞれを個別に学習する設定の比較を行った．

人間の感情状態は図 4.2 のように，Arousal（覚醒度）と Valence（快・不快度）からなる 2 次元の連続空間で表現されることが一般的である．これら 2 つの次元は定義上は独立した指標であり，感情推定においても，それぞれを独立にモデル化する方が適切である可能性が考えられる．

そこで本研究では，Diff@0.80 により選択された同一の入力フレーム集合を用いた上で，Arousal と Valence を別々のモデルで学習する場合と，単一のモデルで同時に学習する場合を比較することで，学習設定の違いが感情推定性能に与える影響を検証する．

4.3 実験結果と考察

各手法の定量的評価結果を表 4.1 に示す．なお，各手法に付随する数値は，フレーム間の類似度に対して設定した閾値を表している．画素差分に基づく提案手法は，従来の VEATIC ベースラインと比較して，valence および arousal の両次元において一貫した性能向上を示した．特に Diff@0.80 および PeopleDiff@0.80 は，最重要指標である CCC において顕著な改善を達成している．

CCC は，予測値と正解値の相関に加えて平均値および分散の一致度を同時に評価する指標であり，連続感情系列としての時間的一貫性や振幅の再現性が反映されやすい．本研究において CCC が向上した主な要因は，感情推定に寄与しにくいフレームの混入を抑制し，感情変化に関係するフレームを重点的に入力として選択したことで，予測系列全体の分布が正解系列とより整合したためであると考えられる．すなわち，時間的な変化点が強調され，系列全体としての整合性が高まったことが CCC の改善として現れたと解釈できる．

一方で，PCC，RMSE，SAGR といった CCC 以外の指標では，改善幅が限定的，あるいは一部で低下する傾向も観測された．これは，フレーム選択によって重要な時間区間が強調される一方で，局所的な予測値の

表 4.1 VEATIC ベンチマークにおける，ベースライン手法および提案手法の定量的性能比較

Dimension	Method	CCC ↑	PCC ↑	RMSE ↓	SAGR ↑
Valence	VEATIC (k=5)	0.609	0.644	0.303	0.789
	VEATIC (k=25)	0.624	0.670	0.293	**0.798**
	VEATIC (k=50)	0.609	0.655	0.301	0.785
	Ours (Diff@0.75)	0.677	<u>0.738</u>	<u>0.261</u>	0.797
	Ours (Diff@0.80)	<u>0.687</u>	**0.750**	**0.258**	<u>0.797</u>
	Ours (PeopleDiff@0.80)	**0.723**	0.736	0.278	0.788
	Ours (SSIM@0.75)	0.606	0.688	0.285	0.769
	Ours (SSIM@0.80)	0.612	0.691	0.282	0.771
	Ours (SSIM@0.85)	0.599	0.679	0.288	0.766
Arousal	VEATIC (k=5)	0.630	0.668	0.210	0.779
	VEATIC (k=25)	0.641	0.684	0.202	0.768
	VEATIC (k=50)	0.622	0.653	0.214	0.764
	Ours (Diff@0.75)	0.670	<u>0.733</u>	<u>0.190</u>	<u>0.803</u>
	Ours (Diff@0.80)	<u>0.685</u>	**0.746**	**0.182**	**0.804**
	Ours (PeopleDiff@0.80)	**0.698**	0.721	0.201	0.798
	Ours (SSIM@0.75)	0.622	0.693	0.206	0.785
	Ours (SSIM@0.80)	0.608	0.692	0.205	0.772
	Ours (SSIM@0.85)	0.607	0.691	0.200	0.780

揺らぎや振幅の差が増大し，フレーム単位での絶対誤差（RMSE）や線形相関（PCC）が必ずしも最適化されなかった可能性によるものと考えられる．また，SAGR は符号の一致のみを評価する指標であるため，系列全体の分布整合が改善しても，必ずしも単調に向上しない場合がある．以上より，本研究の設定では，フレーム単位の精度よりも系列全体としての整合性を評価する CCC が，提案手法の有効性を最も適切に反映した指標であるといえる．

特に PeopleDiff@0.80 は，valence および arousal の両次元において最も高い CCC を達成した．この性能向上は，人物検出により背景のみのフレームを入力候補から除外したことにより，感情推定に寄与しにくい視覚的変動が抑制されたことが主な要因であると考えられる．これらの効果により，モデルは人物中心の視覚情報とその時間的推移により強く依存した表現を安定して学習できたと推察される．この結果は，第三者視点動画における連続感情認識において，人物制約を導入したフレーム選択が有効であることを示している．

SSIM ベースのフレーム選択手法は，ベースラインに対して一定の性能向上を示したものの，画素差分に基づく手法には及ばなかった．SSIM は構造的類似性を重視する指標であるため，ターゲットフレームと視覚的に類似したフレームを選択しやすい傾向を持つ．その結果，感情変化に関与する動的な視覚変化や時間的に重要な変化点が十分に反映されず，非類似なフレームを積極的に選択するという観点では画素差分に比べて有利に働かなかった可能性がある．このことから，本課題のように時間的変化点の抽出が重要となる設定では，構造的類似性よりも視覚的変化量に基づく非類似度指標が有効である場合があることが示唆される．

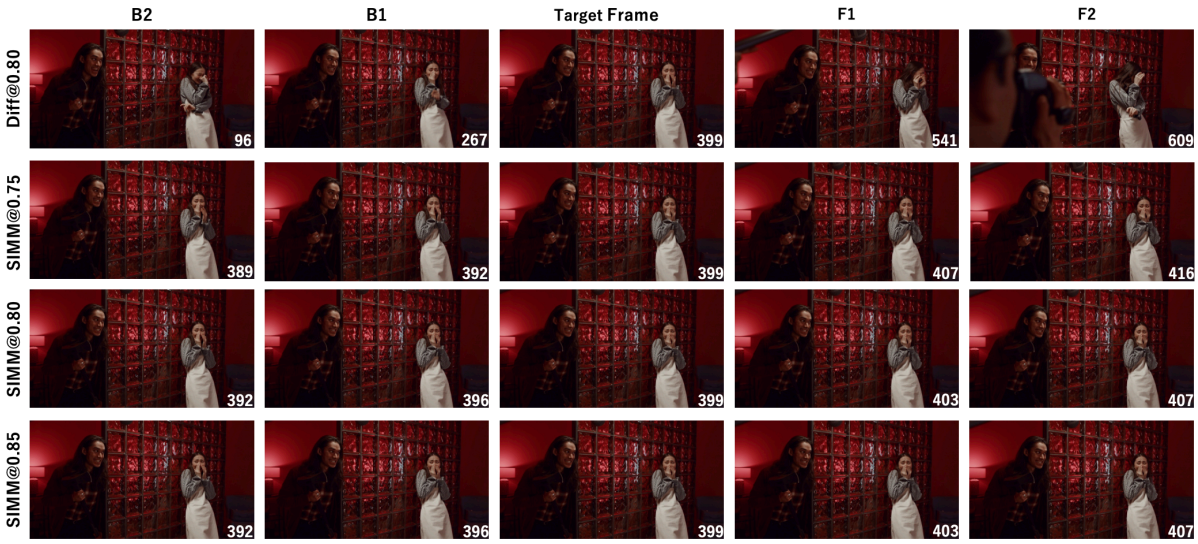

図 4.3 各手法の入力フレームのイメージ図．各行は上から順に Diff@0.80，SSIM@0.75，SSIM@0.80，SSIM@0.85 を示し，各列は左から順に B2，B1，ターゲットフレーム，F1，F2 に対応する．また，各フレームには対応するフレーム番号を重ねて表示している．本図は手法間のフレーム選択傾向の違いを直感的に示すための説明用イメージであり，実験に用いた VEATIC データセットに基づく評価結果そのものではない．(著作権への配慮から，https://www.pexels.com より取得した著作権フリー素材を使用している)

図 4.3 に，異なるフレーム選択手法による入力フレームのイメージを示す．画素差分ベースの Diff@0.80 では，ターゲットフレームの前後から時間的にも視覚的にも多様なフレームが選択されていることが確認できる．構図や人物の姿勢，動作などに明確な変化が見られるフレームが選択されており，感情変化に関連する時間的な変化点を捉えやすい傾向を持つことが分かる．

一方，SSIM ベースのフレーム選択手法では，ターゲットフレームと極めて類似したフレームが前後フレームとして選ばれる傾向が確認できる．その結果，視覚的には安定した入力集合が構成される一方で，感情変化に関与する動的な視覚情報や時間的に重要な変化点が十分に反映されにくい可能性がある．

最後に，valence と arousal を別々に学習する場合と同時に学習する場合の比較結果を表 4.2 に示す．同時学習を行ったモデルは，すべての評価指標において別学習モデルを一貫して上回る性能を示した．この結果は，valence と arousal が定義上は独立した次元である一方で，実際の感情変化においては共通する時間的変動パターンや文脈的手がかりを共有しているためであると考えられる．同時学習により，両次元に共通する視覚的・時間的特徴が相互に補完され，より頑健な感情表現が獲得された結果，推定性能の向上につながったと

推察される．

表 4.2 画素差分に基づくフレーム選択（Diff@0.80）における，valence と arousal の別学習および同時学習の性能比較

Dimension	Training Type	CCC ↑	PCC ↑	RMSE ↓	SAGR ↑
Valence	Separate	0.367	0.485	0.351	0.697
	Joint	**0.687**	**0.750**	**0.258**	**0.797**
Arousal	Separate	0.403	0.558	0.254	0.729
	Joint	**0.685**	**0.746**	**0.182**	**0.804**

4.4 むすび

本章では，VEATIC データセットを用いた実験を通じて，提案手法の有効性を検証した．その結果，人物が含まれるフレームに着目した候補制約と，画素差分に基づく視覚的非類似度を組み合わせたフレーム選択手法が，連続感情認識において最も高い性能を示すことを確認した．特に，PeopleDiff@0.80 は最重要指標である CCC において最良の結果を達成し，第三者視点動画における感情推定に有効であることが示された．

第 5 章

結論と今後の課題

5.1 結論

本研究では，第三者視点動画を対象とした連続感情認識において，入力フレームの選択方法に着目した感情推定手法を提案した．従来の均一サンプリングに基づく手法に対し，人物が含まれるフレームのみを事前に抽出した上で，視覚的な非類似度に基づいてフレームを選択することで，時間的および文脈的に多様なフレーム集合を構成する．

提案手法では，YOLOv8 による人物検出を用いて感情推定に寄与しにくい背景フレームを除外し，その後，画素差分に基づく非類似度を用いてターゲットフレームの前後からフレームを探索する．これにより，急激な場面変化や文脈的な遷移を含むフレームが選択され，感情変化をより適切に捉えることが可能となる．

VEATIC データセットを用いた実験の結果，提案手法は従来の均一サンプリング手法を一貫して上回る性能を示した．特に，人物抽出と画素差分に基づくフレーム選択を組み合わせた PeopleDiff@0.80 は，連続感情認識において最も重要な評価指標である CCC において，valence および arousal の両次元で最良の結果を達成した．この結果は，人物領域に着目したフレーム候補の制約と，視覚的に十分異なるフレームを選択する戦略の有効性を示している．一方，SSIM に基づくフレーム選択手法は，一定の性能改善を示したものの，画素差分に基づく手法には及ばなかった．これは，SSIM が構造的に類似したフレームを選択しやすく，感情推定に必要な動的な文脈変化を十分に捉えられなかったためであると考えられる．

また，valence と arousal を同時に学習することで，両者に共通する時間的変動パターンが効果的にモデル化され，推定精度が向上することも確認された．

以上より，本研究は，第三者視点動画における連続感情認識において，入力フレームの選び方が推定性能に大きく影響することを示すとともに，人物抽出と視覚的な非類似度に基づくフレーム選択が有効なアプローチであることを明らかにした．さらに，本研究は，感情推定モデルの構造を変更することなく，入力設計の工夫のみで性能向上を達成できることを示しており，従来の連続感情認識手法に対して汎用的に適用可能な改善指針を与える点に意義がある．

5.2 今後の課題

今後の課題として，第一に，フレーム間類似度の定義および探索戦略の高度化が挙げられる．本研究では画素差分および SSIM を用いたが，画素差分は輝度変動やカメラ揺れに敏感であり，SSIM は構造類似性を重視する一方で変化点の抽出に不利となる場合がある．今後は，次のような手法を導入することで，感情変化により整合したフレーム選択が可能になると考えられる．

- 深層特徴（ResNet や ViT の中間特徴）に基づく類似度の導入による，意味的変化の反映
- 学習ベースの知覚類似度（LPIPS [11] など）を用いた，人間知覚に近い差異の評価
- 閾値固定ではなく，動画内の変化量分布に応じた適応的閾値設定や，探索範囲の動的制御

第二に，人物制約の精緻化が挙げられる．本研究では人物の有無に基づくフィルタリングに留まっているが，第三者視点動画では複数人物が同時に存在し得るため，「対象人物が十分に観測できているか」という観点が重要となる．今後は，ターゲット人物の位置・領域の推定や追跡を導入し，対象人物中心の視覚情報を安定して抽出することで，教師信号との対応関係をさらに強化できる可能性がある．また，人物間距離や対面配置などの相互作用の手がかりを明示的に取り込むことで，文脈理解の高度化につながる可能性もある．

第三に，推論時の補間処理の改善が挙げられる．本研究では，探索によって選択されなかったフレームに対し，前後推定結果の補間によって連続性を担保したが，この補間は感情変化が急激な区間において誤差を生じる可能性がある．一方で，本研究の目的が人物中心フレームに基づく感情推定性能の検証にあることを踏まえると，補間は影響範囲を限定した実用的な近似として位置づけられる．今後は，補間区間の不確実性を明示的に扱う推定や，時間モデル側で欠落フレームを吸収する設計などを検討する必要がある．

最後に，汎化性能の検証が挙げられる．本研究は VEATIC に基づく評価に留まっているため，他データセットや実環境動画への適用可能性を検証し，手法の一般性と限界を明確化する必要がある．

これらの課題はいずれも，対象人物と文脈の時間的変化をより正確に捉えることを目的としており，入力設計・人物理解・時間モデリングを統合的に高度化する方向性を示している．

以上の課題に取り組むことで，第三者視点動画における連続感情認識の性能向上と，人間とシステムの自然なインタラクション支援に資する感情理解技術の発展につながると期待される．

謝辞

本研究を行うにあたり，多くの方々より貴重な助言ならびに温かいご支援をいただいた．この場を借りて，深く感謝の意を表する．

はじめに，指導教員である渡辺裕教授には，研究テーマの検討段階から論文執筆に至るまで，常に的確なご指導を賜った．研究の進め方に関する助言に加え，課題設定の重要性や結果の解釈，考察の組み立て方など，研究する上で必要な姿勢について多くを学ばせていただいた．ここに心より御礼申し上げる．

また，渡辺研究室の皆様には，日々の議論やゼミを通じて数多くの示唆をいただいた．実験設計や実装に関する課題について意見交換を重ねる中で，多角的な視点を得ることができた．互いに刺激を受けながら研究に取り組めた環境は，非常に恵まれたものであり，本研究を進める上で大きな支えとなった．ここに感謝の意を表したい．

さらに，これまで常に支え，研究生活を見守ってくれた両親に深く感謝する．日常生活における支援のみならず，挑戦を後押ししてくれたことが，本研究を最後までやり遂げる原動力となった．

最後に，本研究に関わり，支えてくださったすべての方々に，心より感謝申し上げる．

参考文献

[1] Z. Ren, J. Ortega, Y. Wang, Z. Chen, Y. Guo, S. X. Yu, and D. Whitney, "Veatic: Video-based emotion and affect tracking in context dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4467–4477, Jan. 2024.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Jun. 2016.

[3] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*, pp. 1–10, Mar. 2016.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, Dec. 2015.

[5] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, May. 2021.

[6] H. Song, W. Zhang, Z. Lin, and Y. Liu, "Affective computing using attention-based transformer networks," *IEEE Transactions on Affective Computing*, 2023.

[7] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[8] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: valence and arousal'in-the-wild'challenge," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 34–41, Jul. 2017.

[9] G. Jocher, A. Chaurasia, and J. Qiu, "Yolo by ultralytics," 2023. YOLOv8 object detection model.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, Jun. 2018.

発表文献

[1] **Soki Saigo**, T. Hayami, and H. Watanabe, “Enhancing continuous emotion recognition via visually diverse frame selection,” in *IEEE Global Conference on Consumer Electronics (GCCE)*, pp. 1275–1278, Sep. 2025. DOI: 10.1109/GCCE65946.2025.11274966.