近赤外画像とサーマル画像を用いた接触検知における マルチバンド画像処理の基礎検討

野口 憲人[1,a)]　岸野 泰恵[1,2]　白井 良成[2]　村田 正幸[1]

概要

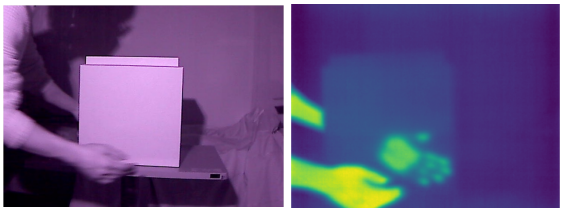
近年，深層学習を用いた画像処理技術は飛躍的に発展し，特に複数のバンド帯の動画像を用いたマルチバンド画像処理により，従来の単一バンドの画像処理では得られなかったさまざまな応用が実現されつつある．マルチバンド画像処理の活用例としては，人が物体表面に触れた後に残る熱痕跡をサーマル画像を用いて識別する接触検知が挙げられ，感染症対策に有効である．本稿では，近赤外画像とサーマル画像を組み合わせた深層学習による接触検知を対象として，時間方向の畳み込みの有無，入力する動画像の種類やサイズが認識精度にどのような影響を与えるか基礎的な検討を行った結果を報告する．

1. はじめに

近年，深層学習を用いた画像・動画認識の研究が盛んに行われている．特に画像の認識精度は年々大きく高まっており，現在では製造業や物流業をはじめとしたさまざまな産業で省人化を目的とした画像認識 AI の導入が進められている．また，最近では可視光カメラだけでなくサーマルカメラや近赤外カメラが安価に手に入るようになったため，機械の異常や野菜の目に見えない不良を検知するなど，目的に応じてセンサを選択することで課題解決の幅を広げている．

一方，2019 年末以降新型コロナウイルス感染拡大に伴い，感染症対策への関心が大きく高まっている．感染経路の一つである接触感染は，感染者が接触した箇所に残ったウイルスに非感染者が触れ，体内にウイルスが侵入することにより引き起こされる．これに対して，人の手が触れた接触箇所を可視化することにより，接触の回避や重点的な消毒行為の促進といった意識改善が期待される．

人が物体表面に触れると体温から伝わった熱が一定時間残留し，これを熱痕跡と呼ぶ．図 1 にサーマル画像に写る熱痕跡およびこの時の近赤外線の画像を示す．この画像は



(a) 近赤外画像　　(b) サーマル画像と熱痕跡

図 1: 接触時の各バンド画像

人が箱を触れた直後の画像であり，(a) の近赤外画像では人の手のみ写っているが (b) のサーマル画像では人が箱を触れていた熱痕跡が残っている．本研究では，熱痕跡を検出することで物体表面への接触を検出しようとしている．

図 1 で示したように，人はサーマル画像にも可視光画像や近赤外画像にも出現するが，熱痕跡はサーマル画像にしか出現しないため，サーマル画像単体を用いるよりも可視光画像や近赤外画像と組み合わせて画像処理を行うことで熱痕跡の認識精度の向上が図れる．また，熱痕跡は人が画面内に出現した後にのみ発生し，発生から時間経過とともに収縮する特性をもつ．そのため，時系列的に連続する複数の画像に対して同時に画像処理を行うことでも熱痕跡の認識精度の向上が期待できる．

以上を踏まえ，本研究では，マルチバンド画像処理による実世界理解に向け，近赤外画像とサーマル画像を用いた接触検知を対象として，マルチバンド画像処理に適した深層学習モデルの検討や，画像サイズや時間の影響を調査することを目的とする．本稿では，シンプルに検証が可能なモデルを設計し，時間方向の畳み込みの有無，入力する動画像の種類やサイズによる認識精度の変化について検証した結果を報告する．

2. 関連研究

2.1 マルチバンド画像処理および熱痕跡検出に関する研究

近年，複数のバンドの画像を用いた物体認識手法が提案されている．一般的な画像処理と同様，画像の畳み込みを活用した CNN (Convolutional Neural Networks) による手法が主流である．例えば，熱伝導率が低く可視光を透過す

[1] 大阪大学大学院情報科学研究科
[2] NTT コミュニケーション科学基礎研究所
[a)] k-noguchi@ist.osaka-u.ac.jp

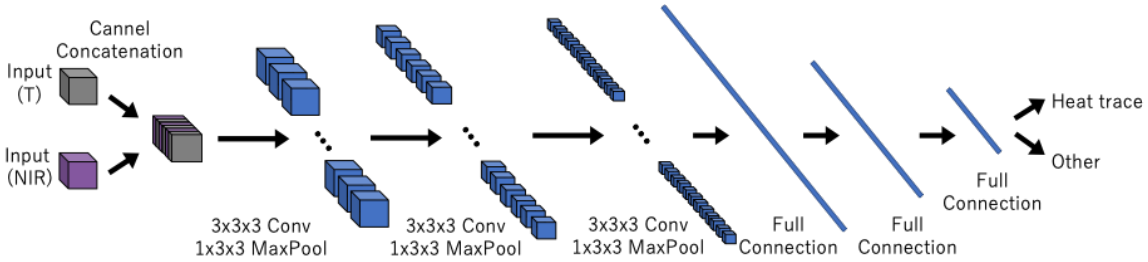


図 2: ネットワーク構造

るという特性をもつガラスは，可視光画像とサーマル画像を同時に CNN に入力することで，単体での入力よりも高い精度で検出できることが報告されている [1]．また接触検知に関する研究としては，U-Net[2] を用いることでフレーム毎に熱痕跡・人物・背景にセグメンテーションを行い，熱痕跡に対しては時間方向の累積和を取ることでサーマル動画における接触検知の手法が提案されている [3]．このように，複数のバンドの画像を用いて画像処理を行うことで，従来の可視光の画像処理のみでは検出できなかった応用が可能になりつつある．

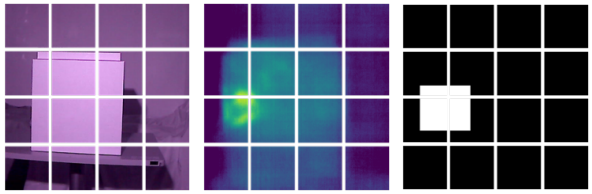
筆者らは，近赤外画像とサーマル画像を用いて画像処理することにより熱痕跡を可視化するシステムを開発した [4]．マルチバンドの画像を用いることで人の目に見えない温度差の情報を活用し，新型コロナウイルスの感染拡大防止に役立てようとしている．本稿では，マルチバンド画像処理によって温度の情報を活用しなければ認識が行えず，応用としての有用性も高い例として，接触検知を対象として検討進めることとした．

2.2 3DCNN

行動認識の分野では近年，画像認識に用いる CNN を時間方向に拡張した 3DCNN (3 Dimentional CNN) が注目されている．CNN は画像を一枚ずつ入力するものであったが，3DCNN では複数フレームを同時に入力に用い時間方向にも畳み込みを行うため，時空間的な変化を考慮した画像処理を行えるようになり，精度向上に効果がある．2013 年に初めて複数フレームを同時に入力する CNN が提案された [5]．以降，時間軸を加えた 3 次元で畳み込む C3D[6] および，オプティカルフローと呼ばれる時間方向の移動差分情報を用いた Two-Stream ConvNet[7] の 2 種類をベースとして，現在に至るまで様々なモデルが提案されている [8][9]．熱痕跡も人の行動と同様に時間方向の特性をもつため，3DCNN が有効ではないかと考えている．

3. 動画像処理モデルの設計

本節は，本研究で用いる動画像処理の CNN の設計について述べる．本稿で目的としているのは，マルチバンド動画の入力および時間方向への確報の効果を確認することで



(a) 近赤外画像　(b) サーマル画像　(c) ラベル

図 3: グリッドに分割した画像の例

ある．そのためシンプルな構造の CNN を設計し，その効果を確認する基礎検討を行えるようにした．ネットワークの構造を 2 に示す．本稿で扱うマルチバンド画像処理では，近赤外線領域を撮影した RGB 画像と，LWIR (Longwave Infrared) 領域を撮影したサーマル画像を入力とする．近赤外画像と撮影領域とサイズを合わせることでサーマル画像を T とすれば，近赤外画像とサーマル画像は RGB 画像の R，G，B の各色を 3 チャンネルの入力とするのと同じように NIR と T の 2 チャンネルを結合して CNN の入力とできる．時間方向にフレーム数 f 続く画像を入力とし，時空間方向に $3 \times 3 \times 3$ のカーネルを用いて畳み込み，$1 \times 3 \times 3$ のプーリング処理を行うことを 3 回繰り返す．このようにして得られた特徴量を 3 層の全結合層で処理し，最終的に出力層では熱痕跡有とその他の 2 状態を推定する．またバッチサイズ b のバッチを用いて処理を進めた．

比較対象であるサーマル画像のみの場合や近赤外画像のみを入力とする場合は，入力データ量は異なるものの，同じモデルサイズで識別するものとした．時間方向の畳み込みを用いない 2DCNN の場合も入力データを $f = 1$ とし，同構造のモデルを用いて評価を行った．

入力動画の領域の広さによる認識性能への影響を調査するため，画像は縦 m 個，横 n 個のグリッドに分割し，その 1 つの画像のみ f フレーム連結させたものを入力動画とする．画像を分割する例を図 3 に示す．この例では 4×4 に画像を分割しているが，実験ではこの他に 2×2 に分割した画像でも性能を調査した．データの分割数を変更して推論結果を比較することで，推論を行う画像・動画内部に写っている情報（人や熱痕跡）によって識別結果がどの程度影響を受けるか確認する．

表 1: カメラスペック

Camera	Band	Image size [pixels]	Frame rate [fps]
FLIR Boson 640	NWIR	640 × 512	60
Gazo MCM-320	NIR*	640 × 480	30

*可視光カットフィルタで近赤外領域のみを撮影

4. 実験

4.1 データの収集

筆者らがこれまでに構築した熱痕跡可視化システム [4] と同様にして動画の撮影を行った．接触対象としては，紙でできた箱を用いることとした．紙は，金属やプラスチックと比較して熱伝導率が低く，手の熱が物体表面にとどまりやすいため，初期検討に適していると考えたためである．本実験で使用したカメラのスペックを表 1 に示す．これらのカメラはフレームレートが異なるため，より低いフレームレートである近赤外カメラの各フレーム撮影時刻に最も近いサーマル画像を対応する画像としている．

データセット作成時には，実世界では人がさまざまな行動をとることを想定し，背景のみ・対象物に接触・接触せずカメラ前を通過・接触せずカメラ前に滞在といった複数の行動と接触を組み合わせた状況を撮影した．動画は 2 本撮影し，片方を訓練データ，もう一方をテストデータとした．

4.2 動画の前処理

データセットの動画を再生し，手作業で熱痕跡が存在する領域を矩形で囲むことによりアノテーションを行った．現状熱痕跡は明確な定義がないため，熱痕跡は人の手が物体への接触後完全に熱痕跡と手が離れた時に出現し，熱痕跡と背景の差が目視で確認できなくなった時を消失とした．今回の条件下では，温度差はおおむね 0.1 度以下で熱痕跡が確認できなくなった．図 3(c) にアノテーションデータの例を示す．データセットのラベルは，動画内に熱痕跡を含む画像が 1 枚でもあれば熱痕跡ありと設定している．取得した近赤外動画およびサーマル動画はそれぞれ異なるカメラから取得しており位置ずれが発生しているため，本データセットでは手動で撮影領域が同じ部分を指定することにより位置合わせを行った．最終的に得られたデータセットおよびパラメータを表 2 に示す．

4.3 実験結果

このようにして得られたデータセットを用い，3 章で説明したモデルを訓練データを用いて学習し，テストデータで推論を行う実験を行った．2DCNN および 3DCNN によるテストデータセットに対する推論結果を表 3 に示す．評価指標として正解率，適合率，再現率，F 値を用いた．

表 2: データセットのパラメータおよびデータ数

				Train		Test	
CNN	$m \times n$	f	b	T	F	T	F
2D	2×2	1	300	9332	13468	5777	7123
	4×4	1	300	26532	65568	12415	39185
3D	2×2	10	10	1909	2691	598	692
	2×2	30	10	396	524	210	210
	4×4	10	10	5430	12980	1287	3873
	4×4	30	10	1128	2552	456	1254

T：熱痕跡有，F：熱痕跡無

5. 考察

5.1 使用するバンドによる影響

当初，サーマル動画または近赤外動画を単体で用いるよりも両方用いた方が精度が高くなると予想していた．しかしサーマル動画のみを用いた場合は正解率および適合率が高くなり，近赤外動画のみを用いた場合は再現率および F 値が高くなり，両方を用いた場合では双方のスコアの間の値を取る結果となった．今回の実験では対象となる物体が一通りだけであったため，サーマル画像と 2 種類の画像での認識精度に差があまり出ず，一方で近赤外画像では箱が写っている領域に熱痕跡が出ると予測するだけでも一定の精度が出たと考えられる．箱だけではなく壁や机などバリエーションに富んだデータセットで追試を行うことで当初の予想が正しいかを今後確認する予定である．

5.2 時間方向の特徴による影響

熱痕跡は時間経過に伴い消失していく特性があるため，フレーム数が長ければ長いほど精度が高くなると想定していたが，フレーム数が 10 の時に最もスコアが高くなる結果となった．これは，フレーム数が増えるに従って時間変化を学習する一方，フレーム数が多いと動画内での熱痕跡が出現するタイミングのバリエーションが増えるため，限られたデータセットでは学習しきれなかったと考えられる．今後，より具体的な時間方向の影響を検証するため，画像を間引いて 30fps よりも低いフレームレートに変換したデータセットを用いて，入力動画の時間長が学習に及ぼす影響を確認する予定である．

5.3 分割数による影響

全体的な傾向として 2×2 よりも 4×4 に分割した方が低いスコアとなったことから，熱痕跡を識別する上でより広範囲の状況を利用した方が認識性能が高くなると考えられる．従って，より狭いグリッド単位で熱痕跡の有無を識別するような状況においても，全体の画像の特徴をうまく利用できるモデルが望ましい．

表 3: 推論結果

		2 × 2			4 × 4		
		T	NIR	T + NIR	T	NIR	T + NIR
CNN	Accuracy	0.5307	0.5225	0.5471	0.7327	0.6254	0.7193
	Precision	0,2289	0.4628	0.4712	0.3571	0.3038	0.2721
f=1,	Recall	0.0137	0.4947	0.2337	0.093	0.4314	0.1065
b=300	F1	0.0238	0.4738	0.2821	0.1321	0.3553	0.1343
3DCNN	Accuracy	0.5751	0.5554	0.5299	0.6918	0.6901	0.7376
	Precision	0.7675	0.5214	0.4936	0.3791	0.3400	0.4627
f=10,	Recall	0.1475	0.5037	0.2763	0.1942	0.2357	0.1493
b=10	F1	0.2319	0.4973	0.3294	0.1958	0.2666	0.2053
3DCNN	Accuracy	0.4903	0.5185	0.5192	0.703	0.6317	0.6864
	Precision	0.489	0.5043	0.425	0.3537	0.3121	0.3455
f=30,	Recall	0.205	0.479	0.4042	0.1247	0.3146	0.2022
b=10	F1	0.2398	0.4664	0.4102	0.1421	0.2927	0.2276

Epoch number: 30, Learning rate: $1e^{-4}$

6. 議論

6.1 アノテーションおよび前処理

熱痕跡は現状明確な定義が存在せず，本実験や関連研究 [3] ではラベル付けが人手で行われているため，ラベルをつける人や動画によって判定が揺らぐ可能性が高い．そこで，ラベリングツールに一定の温度閾値以上の部分を熱痕跡の候補として提示する機能を付加するなどラベリングの安定化を図ることで精度向上に効果があると考えられる．また，熱痕跡の形は場所や時間経過によって形が不明瞭になりやすく，人がカメラと熱痕跡の間を横切ることも多い．そこで，単なる熱痕跡の有無の判別モデルではなく，事前に既存のセグメンテーション技術により人の領域を除去したり，前の時刻の識別結果を予測に活かすアルゴリズムを組み合わせることでも精度を向上できると考えている．

6.2 モデル設計の再検討

本実験では近赤外画像とサーマル画像を 1 つに結合することで 2 種類のバンドの画像を入力としたものの，各バンドの特徴を活用しきれていないことが窺える．理由のひとつとして，各バンドの動画をまとめて畳み込むため想定する特徴が抽出できていないのではないかと考えられる．そこで，各バンドに対してエンコーダを設計し，抽出された特徴を結合したものを入力とするモデルを設計するなど，特徴を活用できる方法についても今後検証したい．

7. まとめ

本稿では，感染症対策にも有効な人の物体表面への接触検知を対象として CNN および 3DCNN を用いた熱痕跡の識別を行い，入力画像のバンドや時間方向の畳み込みの有無，画像サイズが識別結果に与える影響を調査した．実験の結果，時間方向の畳み込みやマルチバンドの画像の入力が識別性能向上に有効であることが示唆された．一方で，今回の実験はデータセットも限られ基礎的な検討にとどまった．今後入力画像のバリエーションを増やすことでさらに検討を進める予定である．また，時間方向やバンド間のデータの差から得られる特徴を反映させやすい識別モデルにより認識性能を向上させたい．さらに，熱痕跡識別による接触検知以外の応用についても検討を進め，マルチバンド画像処理技術の発展に貢献していきたい．

参考文献

[1] D. Huo, J. Wang, Y. Qian and Y. H. Yang, “Glass Segmentation with RGB-Thermal Image Pairs,” arXiv preprint arXiv:2204.05453, 2022.

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” MICCAI, pp. 234-241, 2015.

[3] G. Ma, W. Ross, M. Tucker, P. -C. Hsu, D. M. Buckland and P. J. Codd, “Touch-Point Detection Using Thermal Video With Applications to Prevent Indirect Virus Spread,” IEEE Journal of Translational Engineering in Health and Medicine, Vol. 9, pp. 1-11, 2021.

[4] 白井 良成，岸野 泰恵，柳沢 豊，尾原 和也，水谷 伸，須山 敬之, “Alertable Surfaces: ウイルスの付着を警告可能な実環境,” WISS 2020, 2020.

[5] S. Ji, W. Xu, M. Yang and K. Yu, “3D Convolutional Neural Networks for Human Action Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 1, pp. 221-231, 2013.

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” ICCV, pp. 4489-4497, 2015.

[7] K. Simonyan, A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos,” NIPS, Vol. 27, 2014.

[8] M. Zolfaghari, K. Singh, T. Brox, “ECO: Efficient Convolutional Network for online Video Understanding,”ECCV, pp. 695-712, 2018.

[9] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić and C. Schmid, “ViViT: A Video Vision Transformer,” ICCV, pp. 6816-6826, 2021.