視線ヒートマップを用いた行為意図の識別とその判断根拠の解明

岡田 来波生[1,a)]　金野 武司[2]

概要：視線を用いたインタフェースは，人間と機械との間の有力なインタラクション手段として期待されている．しかし，その多くは注視点の滞在場所や時間などの静的情報から入力を判定するため，単に対象を見ているだけの状態でも操作が発生しうるという課題を抱える．このような意図しない入力を抑え，直感的で誤動作の少ないインタフェースを実現するには，視線の動的情報から行為意図を推定する必要がある．そこで本研究では，視線の移動速度と方向を 2 次元ヒートマップとして画像化し，畳み込みニューラルネットワーク（CNN）に学習させることで，ユーザの「探索」と「操作」という異なる行為意図を識別する手法を検証した．その結果，97%という高い識別精度が得られたが，CNN の判断根拠は未解明であった．本稿では，識別モデルのブラックボックス性を解消するため，差分ヒートマップおよび Grad-CAM++ を用いた可視化分析を行った．解析の結果，操作モードは高速度域への集中，探索モードは中速度域への拡散という特徴が明らかになり，CNN はこれらヒートマップ上の微細なテクスチャなどの複合的特徴を捉えて判断していることが示唆された．

1. はじめに

視線計測技術の発展に伴い，視線は単なる眼球運動の記録ツールから，人間と機械とをつなぐ重要なインタフェースへと進化を遂げている [1]．特に，筋萎縮性側索硬化症（ALS）や脳性まひなどの重度運動障害を持つ人々にとって，視線は残された数少ない身体運動の一つであり，これを用いた補助代替コミュニケーション（AAC）機器の実用化が進んでいる [2]．代表的な製品として，屋外対応アイトラッカを組み込んだ Tobii Dynavox 社の TD I-Series [3] などが挙げられ，視線のみによる PC 操作や環境制御を可能にしている．また，車いすやロボットのナビゲーションなど，視線を入力とした半自律的な制御システムの開発も進められている [4]．

しかし，こうした視線インタフェースの多くは，注視点の位置や滞留時間といった静的な情報に基づいてユーザの関心を推定する手法が主流である．人間の視線には，外界の情報を知覚するための「探索」としての機能と，他者へ意図を伝える「表出」としての機能（ノンバーバル・コミュニケーション）が混在している [5]．従来のシステムではこの区別が困難であり，単に対象を見ただけで操作が発生してしまう「Midas Touch 問題」[6] が古くからの課題として指摘されてきた．

より直感的で誤動作の少ないインタフェースを実現するためには，視線動作に内在するこれらの異なる行為意図を，動的な振る舞いから読み解く必要がある．我々は先行研究 [7] において，異なる意図が混在する視線データを取得するための実験課題を構築し，その動作特徴の分析と識別を試みた．その結果，視線の移動速度や方向といった単純な統計量においては，モード間で人間が判別できるほどの明確な差異は見出せなかった．他方，それらの分布をヒートマップ画像化し，畳み込みニューラルネットワーク（CNN）を用いて学習させた場合には，極めて高い精度で行為の意図を識別できることが確認されている．

しかし，先行研究においては，CNN が高い識別性能を示した要因，すなわち「視線動作の具体的にどのような特徴を捉えて行為の意図を判別しているのか」までは明らかにされていなかった．人間には知覚できない微細な特徴や複雑なパターンを CNN が捉えていることは推測されるが，その判断根拠が不明なままでは，視線動作の生理学的・認知的な理解につながらず，より堅牢なインタフェース設計への応用も困難である．

そこで本研究では，先行研究にて収集した視線データセットに対し，説明可能 AI（XAI）[8] のアプローチを用いた詳細な分析を行う．具体的には，行為の意図（モード）の間の差異を強調する差分ヒートマップや CNN モデルの

[1] 金沢工業大学 情報工学科
[2] 金沢工業大学 知能情報システム学科
[a)] c1204100@st.kanazawa-it.ac.jp

図 1　実験で提示されたオブジェクト探索画像

判断根拠を可視化する Grad-CAM++ [9] を用いて，視線動作に内在する意図の差異を示す特徴を解明することを目的とする．

2.　視線データを取得する実験課題

本研究では，視線動作に内在する行為意図を分析するため，我々が先行研究において構築した，多数のオブジェクトが描かれた画像から特定の対象を探索する実験課題を用いた．実験課題には，画面上に **図 1** に示すような 41 種類のオブジェクトを計 1930 個配置した画像（1896 × 1030 pixel）を 5 枚用意し，実験参加者には，各画像につき指定された 5 種類のターゲット（人型ロボット，犬，カブトムシ，カエル，ペットボトル）を 3 分以内に探し出すことを求めた．

本課題では，参加者が能動的に画面を操作しようとする意図と，受動的に対象を探そうとする意図を明確に区別して発揮できるよう，手元のキーボードの Z キーを押下することで任意に切り替え可能な 2 つのモード（操作モードと探索モード）を実装した．「操作モード」は，視線によって画像の拡大・縮小および平行移動が可能な状態であり，視線の移動速度が一定の閾値以下になった際に画像の拡大（ズームイン）が，左目を閉じる動作によって縮小（ズームアウト）が実行されるほか，視線を画像の端へ移動させることでその方向へのスクロールが可能となっていた．他方，「探索モード」は画面操作が無効化された状態であり，参加者は静止した画像の中からターゲットとなる対象の探索が行えるようになっていた．

視線計測には Tobii Technology 社製のアイトラッカ X2-30 Compact を用い，これを PC（iMac 4.5K Retina ディスプレイモデル）のモニタ下部に設置して計測を行った（**図 2**）．実験には金沢工業大学の男子学生 10 名（平均年齢 21.6 歳，$SD = 0.966$）が参加した．本実験は所属機関の倫理審査委員会の承認を得て実施された．

3.　視線データの分析

収集された視線データは，参加者 10 名に対し各 5 枚の

図 2　実験環境

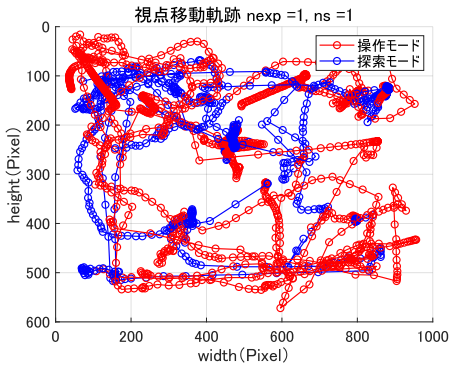

図 3　視点の画像上の移動軌跡．参加者 1 の 1 回目の試行

画像を提示した計 50 試行分であり，総計測時間は各試行 3 分間の計 150 分に及ぶ．一例として，画面上を移動した視点の軌道を**図 3** に示す．先行研究では，まずこれらのデータから「探索モード」と「操作モード」を識別するための定量的な特徴を抽出することを試みた．

具体的には，視線の停留点ではなく動きそのものに着目し，各モードにおける視線移動速度および移動方向のデータを確率密度分布として可視化した（**図 4**）．その結果，速度分布においては探索モードの方が高速度域での頻度がわずかに高い傾向が見られ，方向分布においては両モードともに十字方向（0, 90, 180, -90 度）付近にピークが確認された．

しかし，これらの統計的な分布形状は両モード間で酷似しており，試行ごとの変動も大きいため，単純な閾値処理や統計量の比較によってモードを明確に分離することは困難であるという結論に至った．そこで，統計的手法では捉えきれない視線動作の非線形かつ微細な特徴を抽出するため，視線の動作量を画像化し，機械学習である畳み込みニューラルネットワーク（CNN）による識別を試みた．

CNN に入力する画像は，まず時系列の視線データから 5 秒から 10 秒の時間窓でデータを切り出し，各サンプリング時点における「視線移動速度（pixel/s）」と「視線移動方向（rad）」を算出した．次に，これらを横軸が角度

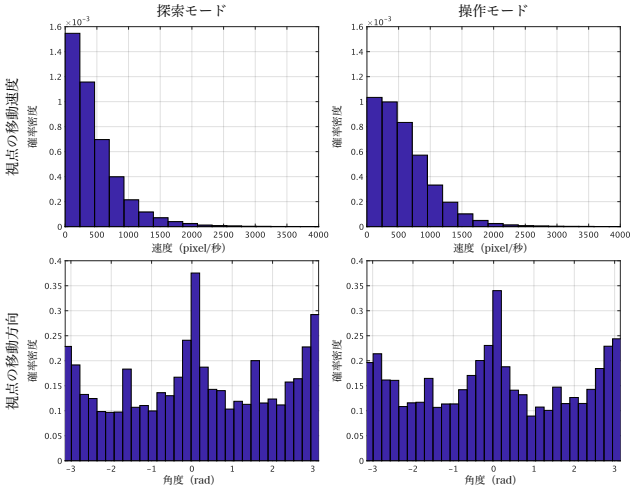

図 4 探索モード（左列）と操作モード（右列）における視点の移動速度（上段）と移動方向（下段）の確率密度分布

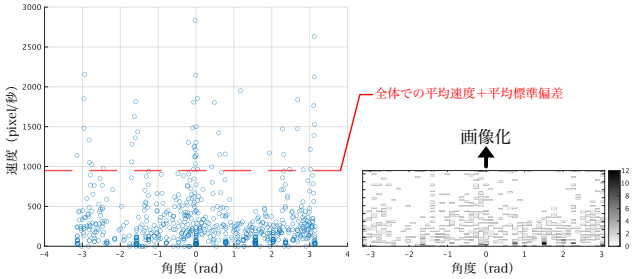

図 5 視点の移動方向と速度の散布図とその画像化．参加者 1 の 1 回目の試行

$(-\pi \leq \theta \leq \pi)$，縦軸が速度となる 2 次元平面上の点群として扱った（**図 5** 左）．この際，外れ値の影響を抑制するため，速度値については全データの平均値に標準偏差を加算した値を上限とするクリッピング処理（上限を超える値の置き換え）を施した．続いて，この速度-方向平面を 50×50 のグリッドに分割し，各グリッド内に含まれるデータ点の数を集計した．この頻度分布（2 次元ヒストグラム）に対し，最小値から最大値の範囲を 0 から 255 に線形スケーリングすることで，頻度が高い領域ほど輝度が高い 256 階調のグレースケール画像を生成した（図 5 右）．このようにして作成されたヒートマップは，単なる統計グラフではなく，ある時間区間における視線の「速度と方向の結合分布」を，空間的なテクスチャとして保持することになると考えられる．

この画像を，MathWorks 社の MATLAB で提供されているサンプルネットワーク [10] を利用し，3 層の畳み込み層と 2 層のプーリング層で構成する CNN に学習させたところ，検証用データセットに対して 0.97 という極めて高い識別率が得られた．また，連続する時系列データに対するモード推定においても，全体で 6 割以上，参加者によっては 9 割を超える一致率が確認された（**図 6**）．

以上の結果から，人間の目や単純な統計分析では差異を

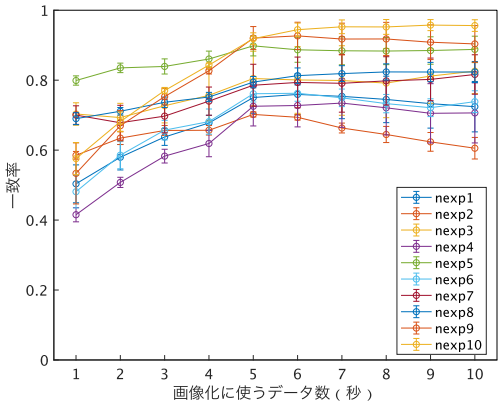

図 6 識別に使用するデータ時間幅に対するモード推定の平均一致率．エラーバーは標準誤差．

見出すことができなかった視線動作に対し，CNN はヒートマップ上の何らかの特徴パターンを捉えることで，両モードを明確に区別していることが明らかとなった．しかし，CNN が具体的にヒートマップ上のどの領域（速度帯や角度）に着目して判断を下したのかについては未解明のままであった．

そこで本研究では，この高い識別能力の根拠を明らかにするため，差分ヒートマップおよび説明可能 AI（XAI）のアプローチを用いた詳細な解析を行った．次節以降でそれらを説明する．

4. 視線移動速度と方向の分布におけるモード間の平均頻度差

CNN が高い識別性能を発揮した要因を解明するため，本研究ではデータセット全体におけるモード間の大局的な差異を可視化する「差分ヒートマップ」による分析を行った．

先行研究における統計的分析では，速度と方向をそれぞれ独立した 1 次元の頻度分布（ヒストグラム）として比較していたため，両者の相関関係に含まれる特徴が捨象されていた可能性がある．そこで，速度と方向の結合分布であるヒートマップにおいて，モード間にどのような構造的な差異が存在するかを可視化することを試みた．具体的には，全参加者・全試行のデータから「操作モード」とラベル付けされたヒートマップ群の平均画像と，「探索モード」とラベル付けされたヒートマップ群の平均画像をそれぞれ算出し，前者から後者を画素ごとに減算することで，モード間の頻度差を表す差分ヒートマップを作成した．

図 7 に生成された差分ヒートマップを示す．図中の赤色は操作モードにおいて出現頻度がより高い領域（正の差分），青色は探索モードにおいて出現頻度がより高い領域（負の差分）を表している． この結果を見ると，操作モードの特徴を示す赤色の領域は，特定の角度における高速度域に局所的な集中を見せている．これは，実験課題の仕様上，画面をスクロールさせるために視線を画面端へ素早く

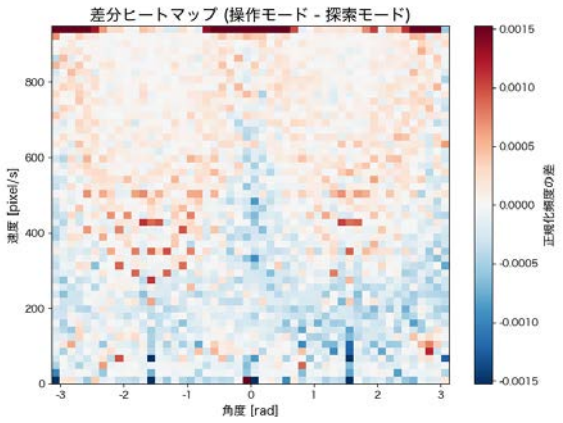

図 7 視線移動速度と方向分布におけるモード間差分

移動させる動作や，意図的に視線を飛ばすサッカード動作が頻繁に行われたことに起因すると考えられる．1 次元の速度ヒストグラムでは「操作モードの方がわずかに高速度成分が多い」程度の差として埋もれてしまっていた情報が，2 次元平面上で角度情報と組み合わせられることで，特定の方向への鋭い動きという明確なクラスタとして顕在化している．

対照的に，探索モードの特徴を示す青色の領域は，中速度域から低速度域にかけて広く拡散している様子が確認できる．これは，特定の方向や速度への偏りが少なく，画面全体を万遍なくスキャンする探索特有のランダムな視線移動が，ヒートマップ上では広範な分布として表れていると解釈できる． 以上の結果より，一見して類似しているように見えた 2 つのモードの視線動作には，2 次元の「速度-方向空間」において明確な構造的差異が存在することが確認された．CNN は，このようなヒートマップ上の局所的な高輝度領域や，分布の広がりを特徴量として捉えることで，高い精度での識別を実現したと推察される．

5. 説明可能 AI を用いた解析

CNN が高い識別精度を実現している背景には，入力画像内の特定領域が判断に強く寄与していると考えられる．このような CNN の判断根拠を可視化する手法として，Class Activation Mapping (CAM) が知られている [11] ．CAM は，Global Average Pooling (GAP) 層の重みを用いることで，特定のクラス反応に貢献した画像領域を特定する手法であるが，特定のネットワーク構造を前提とするため，適用範囲に制約があった．

この課題を解決するために提案されたのが Grad-CAM (Gradient-weighted CAM) である [12]．Grad-CAM は，最終畳み込み層への勾配情報を用いることで，ネットワークの再学習や構造変更を必要とせず，汎用的な CNN モデルに適用可能である．しかし，Grad-CAM には，画像内に同じクラスの特徴が複数箇所に存在する場合や，広範囲にわたる特徴を捉える際に，最も顕著な一部分のみを強調してしまい，全体像を捉えきれない場合があるという課題が指摘されている．

本研究で扱う視線ヒートマップ画像は，単一の物体が写っている一般的な写真とは異なり，視線データの頻度分布が複雑なテクスチャとして画面全体に散在している可能性がある．したがって，局所的なピークだけでなく，微細な分布パターンを含めた全体の特徴を捉えることが不可欠である．そこで本研究では，Grad-CAM の拡張版である Grad-CAM++ を採用した [9]．Grad-CAM++は，勾配の重み付け計算において画素ごとの寄与度（高次微分）を考慮することで，対象が画像内に複数存在する場合や，より詳細な特徴領域の特定において，従来の Grad-CAM よりも優れた可視化性能を持つ．これにより，単純な速度のピークだけでなく，モード固有の複雑な分布形状を正確に捉えることを狙いとした．

前節の差分ヒートマップの分析では，操作モードにおいて高速度域や特定角度に，探索モードにおいて中速度域に特徴的な頻度分布が現れることが確認された．もし CNN が単純な統計量（頻度の高さ）のみを学習しているのであれば，Grad-CAM++の出力もこれらの「差分で優位だった領域」と一致するはずである．

しかし，実際の出力結果を確認すると，CNN は必ずしも差分ヒートマップで示された高頻度領域のみに注目しているわけではなく，単純な画素値の大小だけでは説明できない事例が確認されたのでそれを報告する．

具体的な事例として，真のラベルおよび予測結果が共に「操作モード」であった画像に対する Grad-CAM++の出力結果を**図 8** 左に示す．また，比較対象として同条件の「探索モード」の事例を図 8 右に示す． 差分ヒートマップ（図 7）の傾向に従えば，操作モードである図 8 左においては，本来頻度が高いとされる高速度域（画像上部）が注目されることが予想される．しかし実際には，CNN は画像中央部や左側に散在する局所的な特徴領域に強い反応（赤色・黄色の領域）を示しており，全体の頻度分布とは異なる箇所を根拠に判断を下していることが分かる．他方，探索モードである図 8 右においても，単純な分布の広がりではなく，画像下部の特定の領域に注目が集まっている様子が確認できる．

これらの結果は，CNN がヒートマップを単なる 2 次元ヒストグラムとしてではなく，空間的な構造を持った画像として処理していることを示唆している．すなわち，CNN は速度と方向の局所的な変化率や，ヒートマップ上の微細なテクスチャ（点の疎密パターンや形状の歪みなど）といった，より複合的かつ高次元な特徴に基づいて予測を行っていると考えられる．

6. 結論

先行研究における統計的な分析では，視線の移動速度や

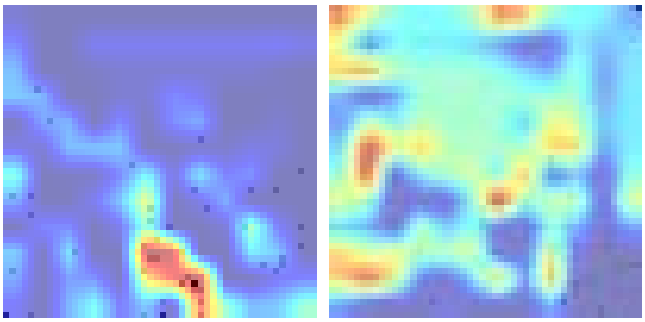

図 8 CNN が操作モードと予測した画像（左）と探索モードと予測した画像（右）の Grad-CAM++の出力

方向の 1 次元的な頻度分布において，モード間を明確に分離できる定量的な特徴は見出せなかった．しかし，視線動作の動的特性を 2 次元のヒートマップ画像として表現し，CNN を用いて学習させた結果，97%という極めて高い精度での識別が可能であることが確認された．

本研究では，この CNN が高い識別性能を発揮する要因を明らかにするため，差分ヒートマップおよび Grad-CAM++を用いた詳細な分析を行った．差分ヒートマップによる可視化の結果，操作モードでは高速度域や特定の角度に局所的なピークが存在し，探索モードでは中速度域に広く拡散する傾向があることが大局的な差異として確認された．一方で，Grad-CAM++を用いた判断根拠の特定においては，CNN が必ずしもこれら単純な高頻度領域のみに着目しているわけではなく，ヒートマップ上の微細なテクスチャや局所的な分布形状といった，より複合的かつ高次元な特徴を捉えて判断していることが示唆された．

これらの知見は，人間の視線動作には意識的な操作意図に伴う特有の運動パターンが含まれており，深層学習を用いることで，従来の統計指標では捉えきれなかった意図の差異を検出可能であることを示している．

参考文献

[1] 竹村憲太郎：視線推定技術とインタフェースへの応用，計測と制御，Vol.51，No.1，pp. 37–42 (2012).
[2] 伊藤史人：重度障害児・者のための視線入力装置活用と意思伝達支援，日本画像学会誌，Vol.58，No.5，pp. 506–513 (2019).
[3] Tobii Dynavox, “TD I-Series,” `https://www.tobiidynavox.com/products/td-i-series/`, (2025.7.6 アクセス)
[4] Subramanian, M., Park, S., Orlov, P., Shafti, A., & Faisal, A. A. : Gaze-contingent decoding of human navigation intention on an autonomous wheelchair platform. In 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 335–338 (2021).
[5] 髙木幸子：コミュニケーションにおける表情および身体動作の役割，早稲田大学大学院文学研究科紀要，Vol.51，pp. 25–36 (2005).
[6] Jacob, R.J.K.: The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look At Is What You Get, *ACM Transactions on Information Systems (TOIS)*, Vol.9, No.2, pp. 152–169 (1991).
[7] 岡田来波生，大谷一真，田中康平，金野武司：視線動作に内在する行為意図の動的特徴を捉える機械学習モデルの構築，日本認知科学会第 42 回大会発表論文集，P2-11 (2025).
[8] Gunning, D.: Explainable Artificial Intelligence (XAI), *Defense Advanced Research Projects Agency (DARPA), Web*, Vol.2, No.2 (2017).
[9] Chattopadhay, A., Sarkar, A., Howlader, P. and Balasubramanian, V.N.: Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks, *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 339–347 (2018).
[10] Mathworks, 分類用のシンプルな深層学習ニューラルネットワークの作成，`https://jp.mathworks.com/help/deeplearning/ug/create-simple-deep-learning-network-for-classification.html`, (2025.7.6 アクセス)．
[11] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A.: Learning Deep Features for Discriminative Localization, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929 (2016).
[12] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626 (2017).