視線情報と操作ログを用いた アノテーション作業における回答信頼度推定手法の検討

A Preliminary Study on Estimating Annotation Reliability Using Gaze Data and Interaction Logs

中村 悠一 † 渡邉 洸 ‡ 松田 裕貴 †
Yuichi Nakamura Ko Watanabe Yuki Matsuda

1 はじめに

近年，機械学習に用いる教師データを作成するため，画像や動画をはじめとするさまざまなデータにラベルを付与するアノテーション作業が広く行われている．機械学習モデルの性能は教師データの品質に影響されるため，誤ったアノテーションを把握し，適切に修正することが重要である．しかし，人手によるアノテーションでは，すべての回答が同程度の信頼性を持つとは限らない．そのため，得られた回答のうち，確認や修正が必要な回答を効率的に抽出する方法が求められる．

アノテーション結果を確認する方法として，複数の回答者による結果を比較する方法や，回答者に自信度および迷いの有無を申告させる方法が考えられる．複数の回答者に同一のデータを割り当てる方法では，回答の一致度を確認できる一方，必要となる作業量や費用が増加する．また，自己申告による自信度や迷いには回答者ごとの基準差があり，判断に不確実性があっても迷いを申告しない場合や，誤ったラベルを正しいと確信して選択する場合がある．したがって，回答ラベルや自己申告情報から個々の回答の正誤を十分に判断することは困難である．

そこで本研究では，アノテーション作業中に生じる視線行動およびマウス操作に着目する．アノテーションの判断過程は，反応時間，視線移動，マウス軌跡などの行動に表れる可能性がある．これらの情報を利用することで，回答者の自己申告だけでは捉えにくい，回答の正誤に関する情報を取得できると考えられる．本稿では回答信頼度として，回答が主ラベルまたは許容副ラベルと整合する程度を表す正誤判定値を定義し，タスク遂行中の視線およびマウスの時系列データから，参加者が付与したラベルの正誤を推定するモデルを構築する．(図 1)

評価では，画像アノテーションタスクとして点字ブロック画像の状態分類 [1] を対象とし，点字ブロックの汚れ，割れ，溶け，摩耗，色褪せなど，視覚的特徴が類似する状態や，複数の状態が同時に現れる場合をラベル付けする実験を行った．構築したモデルは参加者単位の交差検証を行い，不正解検出性能と，確認対象件数を制限した条件における不正解発見性能の両面から評価する．これ

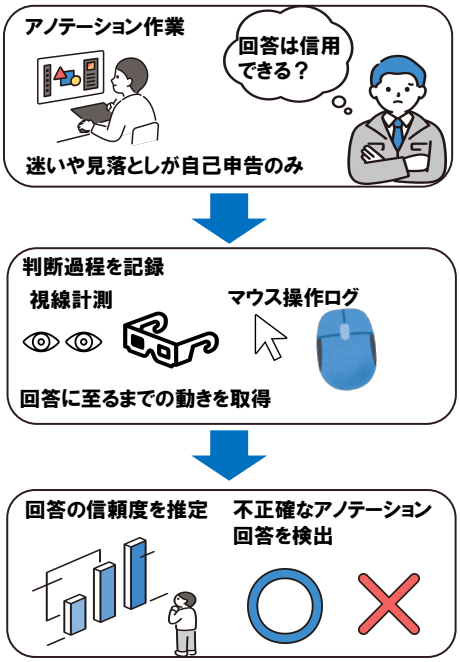

図 1: 研究の全体像

により，不正解の可能性があるアノテーションを抽出し，教師データの品質確認を支援できるかを検討する．

2 関連研究

2.1 アノテーションの品質管理

アノテーションの品質を確保する代表的な方法として，同一のデータを複数の回答者に割り当て，得られたラベルを統合する方法がある．Dawid らは，正解ラベルが未知の場合でも，複数の回答結果から回答者ごとの誤り率と正解ラベルを推定する手法を提案した [2]．Snow らは，自然言語処理に関する複数の課題を対象として，クラウドソーシングによる非専門家の回答を専門家のラベルと比較し，複数の回答を統合することの有効性を示した [3]．また，Sheng らは，誤りを含むラベルを扱う場合，同一データに対して複数のラベルを取得することで，教師データと学習モデルの品質を改善できる場合があることを示した [4]．

これらの方法は複数の回答を用いて品質を高める一方で，同一のデータを複数の回答者に割り当てる必要があるため，アノテーションに必要な作業量と費用が増加す

† 岡山大学，Okayama University
‡ ドイツ人工知能研究センター，DFKI

るという課題がある．

回答者が申告する自信度を品質管理に利用する研究も行われている．Oyama らは，クラウドソーシングで得られたラベルと自己申告自信度を用いて，複数のラベルを統合する確率モデルを提案した [5]．自己申告自信度は回答品質を判断するための情報となり得る一方，回答者によって過信や自信不足が生じるため，申告基準の違いを考慮する必要がある．

2.2 行動ログを用いたアノテーション品質の推定

回答結果だけでなく，作業中に記録される行動情報からアノテーション品質を推定する研究も行われている．Goyal らは，検索結果の関連性を判定するクラウドソーシング課題を対象として，クリック，マウス移動，スクロール，キー入力，ウィンドウのフォーカス変更，作業時間などを記録した [6]．これらの行動情報を用いて，個々のラベルの正誤と回答者ごとの正解率を予測し，行動履歴が回答品質の推定に利用できる可能性を示した．

Pei らは，個々のアノテーション作業で記録される操作履歴から細粒度の行動特徴を抽出し，回答品質の管理に利用する枠組みを提案した [7]．同研究では，正解が定められた課題における回答品質の予測，主観的な課題における不審な行動の検出，および行動傾向に基づく回答者の分類を行い，行動特徴の有効性を検証している．この結果は，回答者の過去の成績だけでなく，個々の回答に至るまでの操作過程からも回答品質を推定できる可能性を示している．

しかしながら，これらの研究で利用される情報は，主にマウス操作や Web ブラウザ上の操作履歴である．そのため，回答者が画像や選択肢のどの部分を確認して判断したかを直接表す視線情報は扱われていない．

本研究グループではこれまで，回答中に取得される行動情報を用いた回答品質の推定に取り組んできた．Gogami らは，スマートフォン上で記録される回答操作を用いて，オンライン調査における不注意回答を検出する手法を提案した [8]．Fukumitsu らは，PC 上のアノテーション作業を対象として，回答時間やカーソル移動などの操作ログから不注意回答を検出する手法を提案した [9]．また，Taira らは，スマートフォンの操作情報とセンサ情報を用いて，アノテーション作業中に生じる回答品質の低下を推定する手法を提案した [10]．

これらの研究では回答中の操作情報が利用されているが，回答者が対象をどのように確認したかを示す視線情報は扱われていない．

2.3 視線およびマウス操作と判断過程

視線行動は，視覚的な情報の探索過程や判断時の不確実性を捉える情報として利用されている．Brunyé らは，鮮明度の異なる顔画像および建物画像を分類する知覚判断課題を実施し，判断の確信度と視線行動との関係を調査した [11]．その結果，判断の確信度によって，注視の頻度と継続時間，サッカードの頻度や速度，瞳孔径などに変化が生じることを示した．

Quétard らは，実環境画像を用いた視覚探索課題において，視線とマウス軌跡を同時に記録し，視覚的な不確実性が判断過程に与える影響を調査した [12]．画像にノイズを加えて対象の判別を困難にした場合，判断時間や対象の確認時間が長くなり，マウスの移動速度にも変化が生じることを報告した．この結果は，視線行動とマウス操作が，情報の探索から選択に至る判断過程を異なる側面から捉える可能性を示している．

視線情報を用いて回答の正誤を推定する研究として，Bayazit らは，オンライン試験における学習者の視線指標とマウスクリック数を特徴量とし，Random Forest を用いて回答の正誤を推定した [13]．総注視時間，注視回数，領域への訪問時間，マウスクリック数などが，回答の正誤を予測する特徴となる可能性が示されている．

2.4 本研究の位置づけ

従来研究では，複数回答の統合や操作ログの分析によって，アノテーションの品質を管理する方法が検討されてきた．また，視線情報を用いて判断時の迷いや回答の正誤を分析する研究も行われている．しかし，画像アノテーション中の視線とマウス操作を同時に取得し，個々の回答の正誤推定に利用した研究は十分ではない．

そこで本研究では，アノテーション中の視線とマウスの時系列データを取得・分析し，回答の正誤を推定する手法を提案する．

3 提案手法

3.1 提案手法のコンセプトと全体像

本研究では，画像アノテーション中の視線情報とマウス操作ログを取得し，個々の回答の正誤を推定して，優先確認対象を抽出する手法を提案する．提案手法の全体構成を図 2 に示す．提案手法は，課題提示・回答記録部，視線・操作ログ取得部，時刻同期・試行抽出部，特徴量抽出部，正誤推定部から構成される．

課題提示・回答記録部では，Web ブラウザ上に分類対象画像と回答選択肢を提示し，選択ラベル，自信度，迷いの有無および各操作の時刻を記録する．具体的には，反応時間，最初のラベル選択までの時間，ラベル変更回数，選択肢間のホバー遷移，選択肢上のホバー時間およびマウス軌跡を取得する．また，選択ラベルを事前に定めた主ラベルおよび副ラベルと比較し，各回答の正誤状態を算出する．

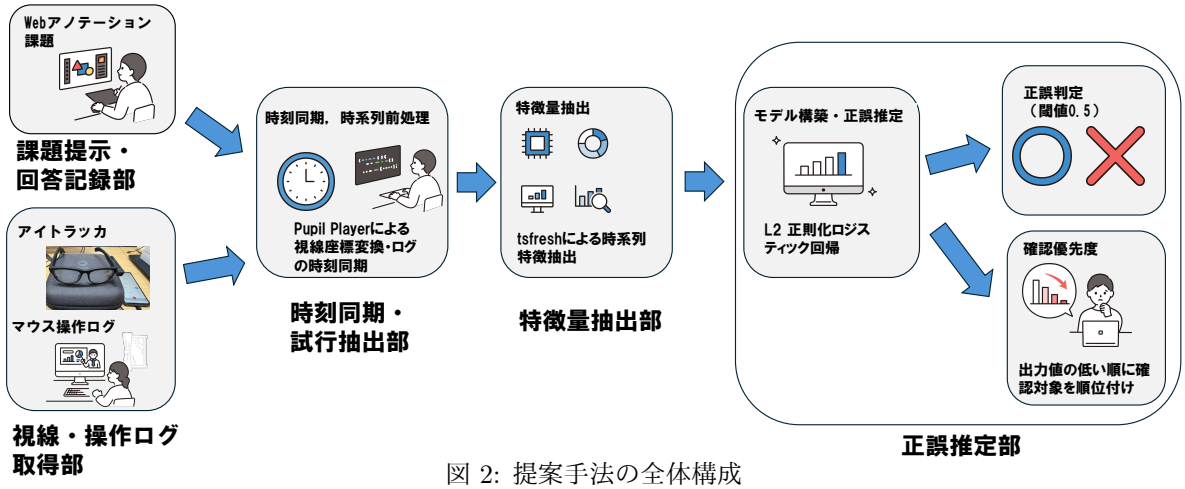

図 2: 提案手法の全体構成

視線・操作ログ取得部では，アイトラッカが計測した視線位置と，ブラウザ上のマウス位置を時系列データとして記録する．視線位置は画面内の正規化座標へ変換し，参加者が対象画像や選択肢のどこを見ていたかを分析可能にする．

時刻同期・試行抽出部では，Web アノテーションシステムとアイトラッカの記録時刻を対応付ける．その後，各回答の開始から確定までの区間に含まれる視線およびマウス操作ログを，試行ごとに切り出す．

特徴量抽出部では，試行ごとに切り出した視線およびマウスの時系列データから特徴量を生成する．

正誤推定部では，生成した特徴量を入力として，各回答に対する 0 から 1 の正誤判定値を出力する．この値は，大きいほど正解寄り，小さいほど不正解寄りであるとモデルが判断したことを表す．実際のアノテーション作業では，すべての回答を人手で確認すると大きな作業負担が生じる．そこで本手法では，正誤判定値の小さい回答から順に並べ，不正解の可能性が高い回答を優先確認対象として提示する．これにより，回答を自動的に修正するのではなく，限られた確認作業の中で不正解を効率的に発見することを支援する．

3.2 時刻同期と試行単位のデータ抽出

アノテーション作業開始前に，システム上で同期基準時刻を記録する．また，アイトラッカの録画開始から同期操作までの時間を計測する．サイトログ上の試行開始時刻を t_{trial}，同期基準時刻を t_{sync}，録画開始から同期操作までの時間を d とすると，アイトラッカ録画内の試行開始時刻 T_{start} を式 (1) によって求める．

$$T_{\mathrm{start}} = d + \frac{t_{\mathrm{trial}} - t_{\mathrm{sync}}}{1000} \tag{1}$$

試行終了時刻についても同様に算出し，試行開始から終了までの視線データを回答単位で切り出す．解析には，モニタ画面内で計測され，アイトラッカの計測信頼度を表す `confidence` が 0.6 以上である視線データのみを使用した．視線位置と Web 画面上の位置を対応付けるため，両者で異なる Y 軸の向きを統一した．また，マウスの移動中は位置を 50 ms 間隔で記録し，各試行の開始から経過した時間と画面上の座標を保存した．

3.3 時系列の正規化と特徴抽出

時系列モデルにおいて，回答時間の長さがサンプル数の違いとして直接推定に利用されることを防ぐため，各試行の開始時刻を 0，終了時刻を 1 とする試行内進行率へ変換する．その上で，視線の X 座標，Y 座標および視線速度の 3 系列を 128 点に，マウスの X 座標，Y 座標および移動速度の 3 系列を 64 点に線形補間する．視線座標には Surface 上の正規化座標を用い，マウス座標は画面の幅と高さによって 0 から 1 の範囲へ正規化する．回答時間そのものは時系列とは分離し，反応時間モデルの入力として用いる．

正規化した各時系列から，時系列特徴量抽出ライブラリ tsfresh [14] の `EfficientFCParameters` を用いて特徴候補を抽出する．これにより，座標や速度の変化量，分布，自己相関などを表す特徴量を生成し，視線とマウスからそれぞれ 2,331 特徴を得る．

正誤推定では，欠損値の中央値補完，分散が 0 の特徴の除去，単変量特徴選択，標準化および L2 正則化ロジスティック回帰を一つの `Pipeline` として構成する．ロジスティック回帰では，選択された特徴量ベクトル $\boldsymbol{x}$ から線形結合値 z を求め，シグモイド関数によって 0 から 1 の値 s へ変換する．

$$z = \boldsymbol{w}^{\top}\boldsymbol{x} + b, \qquad s = \frac{1}{1 + \exp(-z)} \tag{2}$$

ここで，$\boldsymbol{w}$ と b は学習によって決定されるパラメータである．s は正解クラスに対するモデル出力値であり，1 に近いほど正解寄り，0 に近いほど不正解寄りであることを表す．閾値による正誤分類では s を 0.5 で二値化し，

表 1: 提案手法で構築する正誤推定モデル

モデル	使用する情報
反応時間モデル	回答までの時間
自己申告モデル	自信度，迷いの有無
視線時系列モデル	視線 X・Y 座標，視線速度
マウス時系列モデル	マウス X・Y 座標，移動速度
統合モデル	視線時系列，マウス時系列
均衡化統合モデル	統合モデル，正誤数の均衡化

確認対象の優先順位付けでは二値化前の s を用いる．また，前処理，特徴選択およびモデル学習は交差検証の学習データ内だけで実施し，評価対象データの情報が学習へ混入することを防ぐ．

3.4 正誤推定モデル群

取得する情報の違いが正誤推定性能に与える影響を評価するため，表 1 に示す複数のモデルを構築する．各モデルでは同じ L2 正則化ロジスティック回帰を用い，モデルへ入力する情報のみを変更する．

比較対象として，回答までの時間を用いる反応時間モデルと，自信度および迷いの有無を用いる自己申告モデルを構築する．さらに，提案手法によって取得する情報を用いたモデルとして，視線時系列モデル，マウス時系列モデル，および両者を用いる統合モデルを構築する．

本研究の評価データでは，不正解よりも正解が多い．このような正誤数の偏りが不正解検出に与える影響を確認するため，モデル学習時の正解を不正解と同数まで無作為に減らすランダムアンダーサンプリングを統合モデルへ適用し，均衡化統合モデルを構築する．この処理は評価対象の回答を含まない学習データ内だけで実施する．

各モデルについて，前節で定義したモデル出力値 s を用いて，閾値による分類性能と順位付け性能を評価する．閾値による正誤分類では，s が 0.5 未満の回答を不正解，0.5 以上の回答を正解と判定する．順位付け性能は s を二値化せずに評価し，値が低い回答ほど不正解寄りに順位付けされたものとして扱う．

4 評価実験

提案手法の実現可能性および不正解アノテーション検出への適用可能性を確認するための評価実験を行った．対象としては，著者らのグループで研究を行っている点字ブロック状態分類 [1] をタスクとして取り上げた．

本実験では，第一に，アノテーション作業中の視線情報とマウス操作ログを取得し，時刻同期によって回答単位に対応付けられることを確認する．第二に，反応時間，自己申告，視線時系列，マウス時系列およびこれらを組み合わせたモデルの正誤推定性能を比較する．第三に，モデルが算出した正誤判定値を用いることで，不正解を含む可能性が高い回答を優先的に抽出できるかを評価する．

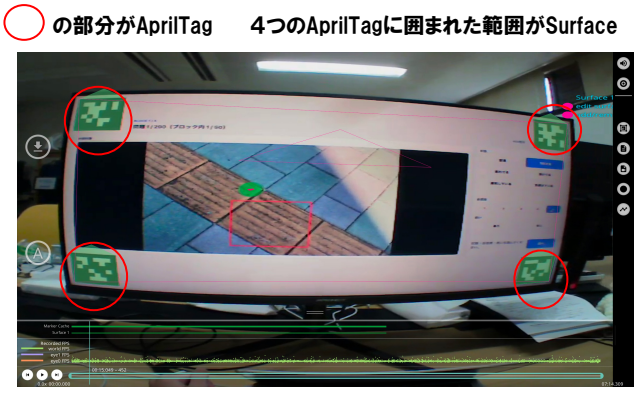

図 3: Pupil Player による画面領域と視線位置の対応付け

4.1 参加者と実験課題

21 歳から 22 歳の大学生 6 名を対象に，点字ブロック画像の状態を分類する画像アノテーション課題を実施した．各参加者は 200 問に回答し，総試行数は 1,200 試行であった．実験課題は 50 問ずつの 4 ブロックに分け，各ブロック間に休憩を設けた．

各試行では，分類対象となる点字ブロック領域を赤枠で示した画像を画面左側に提示し，画面右側に 6 種類の状態ラベルを配置した．参加者は「普通」,「汚れている」,「割れている」,「溶けている」,「摩耗している」,「色褪せている」から 1 つを選択した．続いて，回答への自信度を 1 から 5 の 5 段階で回答し，判断に迷いがあったかを「あり」,「なし」で回答した．200 問の内訳は，普通 50 問，汚れている 68 問，割れている 43 問，溶けている 6 問，摩耗している 22 問，色褪せている 11 問であった．全参加者に同一の問題集合を同一順序で提示した．

4.2 実験環境と手順

実験画面は，31.5 インチ，解像度 1,920 × 1,080 画素のモニタに全画面表示した．参加者とモニタとの距離は約 50–60 cm とした．視線計測には Pupil Invisible[*1]（Pupil Labs 社製）を用いた．機器の装着状態と視線データが取得されていることを確認した後に録画を開始した．画面四隅に配置した AprilTag を基準として，Pupil Player の Surface Tracker によりモニタの表示領域を検出した．検出した領域を視線位置の対応付け対象（Surface）として定義した (図 3)．

実験画面の流れおよび実験環境を図 4 および図 5 に示す．参加者には最初に 6 問の練習課題を行わせ，分類ラベルと操作方法を確認させた．練習課題の回答は評価対象に含めなかった．その後，アイトラッカの録画を開始して同期操作を行い，50 問の本番課題を実施した．各ブ

[*1] https://pupil-labs.com/products/invisible

図 4: 実験画面

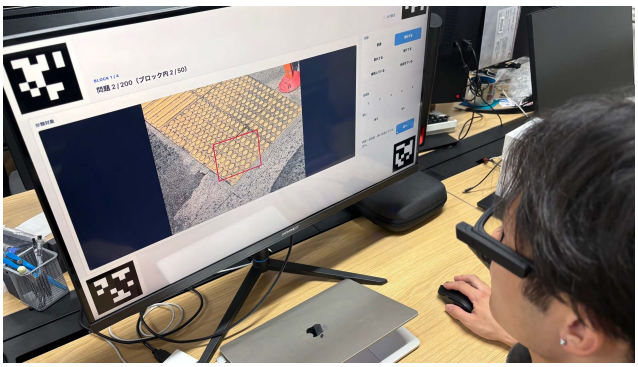

図 5: 実験環境

ロック終了後にサイトログを保存して録画を停止し，参加者が休憩した後に次のブロックを開始した．この手順を 4 ブロック繰り返し，参加者 1 名につき 200 試行を取得した．

4.3 正解ラベルと評価データ

画像の正解ラベルは，第一著者および点字ブロックに関する研究に従事する研究者 1 名の計 2 名がそれぞれ確認した後，両者の協議によって決定した．また，協議の結果，複数の状態として妥当に解釈できると判断した 19 問には副ラベルを設定した．主ラベルと一致した回答を完全正解，副ラベルと一致した回答を部分正解，それ以外を不正解とした．正誤推定モデルでは，完全正解と部分正解をまとめて正解（1），それ以外を不正解（0）として扱った．

提案手法によって取得した 1,200 試行を評価データとして使用し，表 1 に示した各モデルで評価した．

4.4 モデルの評価方法

未知の参加者に対する性能を評価するため，参加者単位のネスト交差検証を行った．外側では 6 名のうち 1 名を評価用，残り 5 名を学習用とし，評価対象者を入れ替えながら計 6 回評価した．内側では学習用データのみを用いて，特徴量数と正則化パラメータを選択した．特徴選択，標準化およびモデル学習は，評価用データを用いず，各交差検証の学習用データ内で実施した．また，均衡化モデルのアンダーサンプリングも学習用データ内でのみ実施した．

各モデルは，回答ごとに 0 から 1 の正誤判定値を出力する．値が低いほど不正解，高いほど正解とみなす．モデルの評価には，不正解に低い値，正解に高い値を付けられたかを表す ROC-AUC を用いた．また，正誤判定値 0.5 を基準に正誤を分類し，実際の不正解を検出できた割合である不正解 Recall と，正解および不正解の判定性能を均等に評価する Balanced Accuracy を算出した．なお，均衡化モデルの出力値は確率ではなく，正誤を判定するための値として扱った．

確認支援への適用可能性を調べるため，統合モデルの正誤判定値が低い方から，各参加者 20 試行を優先確認対象として抽出した．その中に含まれる不正解数と不正解率を算出し，同じ件数を無作為に抽出する処理を 10,000 回行った場合と比較した．

5 結果

5.1 データ取得と回答結果

参加者 6 名から各 200 試行，合計 1,200 試行を取得した．全 4 ブロックについてシステム操作ログと視線ログを記録し，同期基準時刻を用いて各回答区間に対応する視線およびマウスの時系列を試行単位で抽出した．これにより，全 1,200 試行を特徴抽出と正誤推定の入力データとして整形した．正解は 924 試行，不正解は 276 試行であり，正解率は 77.0%であった．参加者別の正解率は 67.5%から 84.5%，不正解数は 31 件から 65 件の範囲であった．

5.2 取得情報別の正誤推定結果

取得情報の種類と組合せごとの結果を表 2 に示す．各モデルは，回答ごとに 0 から 1 の正誤判定値を出力し，値が大きいほど正解，小さいほど不正解であると判定する．ROC-AUC を用いて，正解に高い値，不正解に低い値を付けられるかを評価した．ROC-AUC は 0.5 が無作為な判定に相当し，1 に近いほど正解と不正解を明確に区別できることを表す．また，正誤判定値を閾値 0.5 で

表 2: 取得情報別の正誤推定結果

使用情報	ROC-AUC	Balanced Accuracy	不正解 Recall
反応時間	0.623	0.526	0.062
自信度・迷い	0.540	0.514	0.065
視線時系列	0.634	0.523	0.058
マウス時系列	0.649	0.527	0.069
視線＋マウス	**0.664**	0.520	0.058
視線＋マウス（正誤数均衡化）	0.654	**0.601**	**0.554**

区切った場合の分類性能を，Balanced Accuracy と不正解 Recall によって評価した．

順位付け性能　自信度と迷いを用いた場合の ROC-AUC は 0.540 であった．これに対し，視線時系列では 0.634，マウス時系列では 0.649，両者を組み合わせた場合は 0.664 となった．本実験で比較した条件の中では，視線とマウスを組み合わせた条件が最も高い ROC-AUC を示した．この結果は，提案手法によって取得した視線およびマウスの時系列に，不正解である可能性が高い回答を優先的に抽出するための情報が含まれている可能性を示している．

閾値による不正解検出　正誤数を均衡化していない統合モデルでは，閾値 0.5 での不正解 Recall は 0.058 であり，多くの不正解を正解と誤判定した．一方，学習データ内の正解を不正解と同数にした均衡化統合モデルでは，ROC-AUC は 0.654，Balanced Accuracy は 0.601，不正解 Recall は 0.554 であった．このときの判定件数を図 6 に示す．実際の不正解 276 試行のうち 153 試行を検出した一方，正解 924 試行のうち 326 試行も不正解と判定した．したがって，正誤数の均衡化によって不正解を拾いやすくなる一方，確認対象となる正解も増えることが確認された．

5.3　確認対象の優先順位付け

次に，ROC-AUC が最も高かった視線とマウスの統合モデルについて，固定閾値による自動判定ではなく，正誤判定値を確認対象の優先順位付けに用いた．各参加者の正誤判定値下位 10%に当たる 20 試行を優先確認対象とした結果，全 1,200 試行のうち 120 試行が抽出された．このうち 60 試行が不正解であり，優先確認対象内の不正解率は 50.0%であった．また，発見した 60 試行は，全不正解 276 試行の 21.7%に相当した．

同じ 120 試行を無作為に抽出した場合に発見される不正解数は平均 27.6 試行であり，95%範囲は 19–36 試行であった．これに対し，正誤判定値による抽出では 60 試行を発見しており，無作為抽出の平均に対して 2.17 倍であった．モデルによる発見数 60 試行は無作為抽出より有意に多く，この差が偶然に生じる確率は 0.1%未満であった（$p < 0.001$）．

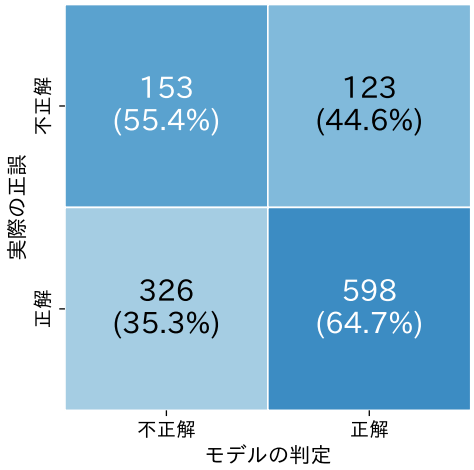

図 6: 均衡化統合モデルの混同行列（括弧内は実際のクラスごとの割合）

表 3: 統合モデルの参加者別評価結果

参加者	ROC-AUC	不正解数	優先確認 20 試行中の不正解数
P001	0.738	38	10
P002	0.745	45	9
P003	0.645	56	9
P004	0.640	41	11
P005	0.719	65	15
P006	0.660	31	6

5.4　参加者別の結果

統合モデルの参加者別結果を表 3 に示す．参加者別の ROC-AUC は 0.640–0.745 であり，参加者間で差が見られた．各参加者について，正誤判定値が低い 20 試行を確認したところ，そのうち 6–15 試行が不正解であった．

6　考察

6.1　提案手法の実現可能性

本研究では，Web アノテーションシステムとアイトラッカを用いて，回答，自信度，迷いの有無，マウス操作および視線情報を同時に記録した．さらに，サイトログと視線ログの時刻を同期し，各回答の開始から確定までのデータを切り出して，特徴抽出と正誤推定に用いる処理を構築した．評価実験では，6 名から取得した計 1,200 試行をこの処理によって解析した．以上より，画像アノテーション中の視線とマウス操作を回答単位で対応付け，不正解検出に利用する一連の手法について，技術的な実現可能性を確認した．

6.2　視線情報とマウス操作の利用可能性

自信度と迷いを用いた場合の ROC-AUC が 0.540 であったのに対し，視線時系列では 0.634，マウス時系列

では 0.649 となった．この結果は，自己申告だけでは表れない回答過程の違いが，視線移動やマウス操作に含まれていた可能性を示している．例えば，対象画像の確認方法，画像と選択肢の間の移動，回答確定までのカーソル操作などが，正誤と関係した可能性がある．

視線とマウスを組み合わせた場合の ROC-AUC は 0.664 であった．視線は画面上のどの情報を確認したか，マウスはどのように選択へ至ったかをそれぞれ反映するため，両者は回答過程の異なる側面を捉えたと考えられる．ただし，本研究では各条件間の差に対する統計検定を行っていない．したがって，視線またはマウスが自己申告より優れていることや，両者の組合せによって性能が向上することを断定することはできない．本実験の範囲では，これらの情報を正誤推定へ利用できる可能性が確認されたと位置付けるのが妥当である．

6.3 確認支援への適用可能性

5.3 節の結果では，統合モデルの正誤判定値が低い回答を抽出することで，無作為に抽出した場合よりも多くの不正解を確認対象に含められた．このことから，モデルの正誤判定値は，確認が必要な回答の優先順位付けに利用できる可能性がある．

一方，統合モデルによって優先確認対象とした 120 試行には，不正解 60 件が含まれていたが，全不正解 276 件のうち 216 件は対象外となった．この結果から，本手法は不正解を確認対象へ集中させられるものの，優先確認対象だけですべての不正解を発見することは困難である．

また，学習データの正解数を不正解数に揃えた均衡化統合モデルでは，不正解 276 件中 153 件を検出し，不正解 Recall は 55.4%となった．その一方で，正解 924 件のうち 326 件も不正解と判定され，確認対象は合計 479 件となった．すなわち，不正解の見逃しを減らすほど，人手で確認する正解回答も増加するというトレードオフが確認された．

以上より，本手法を回答の自動的な削除や修正に用いることは適切ではなく，人手による品質確認の対象を絞り込むための支援手段として利用することが妥当である．実際の運用では，確認に割ける作業量と不正解を見逃した場合の影響を考慮し，確認件数または判定閾値を設定する必要がある．

6.4 個人差と今後の検討課題

視線とマウスを用いた場合の参加者別 ROC-AUC は 0.640–0.745 であった．この差には，視線の動かし方，マウス操作，回答方略，点字ブロックに対する判断基準などの個人差が影響した可能性がある．

本実験の参加者は大学生 6 名に限られ，全員が同じ点字ブロック画像を同じ順序で分類した．また，評価参加者の回答は学習データに含まれないが，同一画像に対する他参加者の回答は学習データに含まれている．そのため，本評価は未知参加者に対する評価であり，未知画像に対する汎化性能までは確認できていない．今後は，参加者数の増加，提示順の無作為化，視線取得品質の評価に加え，点字ブロック以外の画像アノテーション課題への適用を通して，提案手法の汎用性を検証する必要がある．

さらに，現時点では取得したログを実験後に解析し，優先確認対象を算出している．今後は提案手法をアノテーションシステムへ組み込み，作業中または作業後に優先確認対象を提示した場合の不正解発見数，確認時間，確認後のラベル品質を評価する必要がある．

7 おわりに

本研究では，画像アノテーション作業中の回答，視線情報およびマウス操作ログを取得し，時刻同期，試行単位の時系列抽出，特徴量生成，正誤推定を行う一連の手法を提案した．さらに，反応時間，自己申告，視線，マウスおよびこれらの組合せを用いて，個々の回答の正誤判定値を算出し，不正解を含む可能性が高い回答を優先確認対象として抽出する仕組みを構築した．

提案手法の実現可能性を確認するため，大学生 6 名を対象として，点字ブロック画像の状態分類実験を実施した．合計 1,200 試行についてサイトログと視線ログを回答単位で対応付け，特徴抽出と正誤推定へ利用した．視線とマウスを組み合わせた場合の ROC-AUC は 0.664 であった．また，各参加者の正誤判定値下位 10%を優先確認対象とした場合，120 試行中 60 試行が不正解であり，無作為抽出と比較して 2.17 倍の不正解が含まれていた．以上から，アノテーション作業中の視線とマウス操作を取得し，不正解の可能性が高い回答の抽出へ利用できる可能性が示された．提案手法は回答を自動的に削除または修正するものではなく，人手による確認対象の優先順位付けを支援する仕組みとして位置付けられる．

今後は，提案した処理をアノテーションシステムへ組み込み，優先確認対象を人に提示して回答を確認･修正できる機能を実装し，不正解発見数，確認時間および確認後のラベル品質を評価する．また，参加者数とアノテーション課題の種類の増加により，参加者差および視線取得品質の影響と，他の課題に対する汎用性を検証する．

謝辞 本研究の一部は，JSPS 科研費（24K20763）および JST 共創の場形成支援プログラム（JPMJPF2115）の助成を受けて行われたものです．

参考文献

[1] Yuto Matsuda and Yuki Matsuda. Tactile Paving Detection and Classification Method Based on Cyclist-Participatory Road Image Sensing. In *27th International Conference on Distributed Computing and Networking (ICDCN '26 Companion)*, pp. 78–83, 2026.

[2] A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 20–28, 1979.

[3] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, 2008.

[4] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 614–622, 2008.

[5] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, IJCAI '13, pp. 2554–2560, 2013.

[6] Tanya Goyal, Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. Your behavior signals your reliability: Modeling crowd behavioral traces to ensure quality relevance annotations. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6, No. 1, pp. 41–49, 2018.

[7] Weiping Pei, Zhiju Yang, Monchu Chen, and Chuan Yue. Quality control in crowdsourcing based on fine-grained behavioral features. *Proc. ACM Hum. Comput. Interact.*, Vol. 5, No. CSCW2, pp. 1–28, 2021. Article No. 442.

[8] Masaki Gogami, Yuki Matsuda, Yutaka Arakawa, and Keiichi Yasumoto. Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone. *IEEE Access*, Vol. 9, pp. 53205–53218, 2021.

[9] Yoshinobu Fukumitsu, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Detecting Careless Responses in Dataset Annotation using Screen Operation Logs. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom '24)*, pp. 775–780, 2024.

[10] Shigeyuki Taira, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Detection of Response Quality Degradation During Crowdsourced Annotation Tasks Using a Conversational Interface. In *The 15th International Conference on Mobile Computing and Ubiquitous Networking (ICMU '25)*, pp. 1–6, 2025.

[11] Tad T. Brunyé and Aaron L. Gardony. Eye tracking measures of uncertainty during perceptual decision making. *International Journal of Psychophysiology*, Vol. 120, pp. 60–68, 2017.

[12] Boris Quétard, Jean Charles Quinton, Martial Mermillod, Laura Barca, Giovanni Pezzulo, Michèle Colomb, and Marie Izaute. Differential effects of visual uncertainty and contextual guidance on perceptual decisions: Evidence from eye and mouse tracking in visual search. *Journal of Vision*, Vol. 16, No. 11:28, pp. 1–21, 2016.

[13] Alper Bayazit, Petek Askar, and Erdal Cosgun. Predicting Learner Answers Correctness Through Eye Movements with Random Forest. In *Educational Data Mining: Applications and Trends*, pp. 203–226, 2014.

[14] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, Vol. 307, pp. 72–77, 2018.