学位論文 博士（工学）

広範囲空間における頭部姿勢変動に頑健な キャリブレーションフリー視線推定

2017 年 7 月

慶應義塾大学大学院理工学研究科

田 村　仁 優

目次

図目次

表目次

第 1 章

序論

本章では，まず画像を入力とした視線推定の社会における応用例を紹介し，その必要性や現状について説明する．次に，関連研究の特性や課題点について言及し，提案手法の位置づけや目的を明確にする．

1.1 研究背景

近年，コンピュータの性能向上とイメージセンサの低価格化に伴い，カメラが様々なデバイスに内蔵され，日常生活における応用が広がっている．ノート PC やタブレットに内蔵されたカメラは，写真撮影のためではなく，ビデオ通話やユーザ認証などに用いられている．ビデオゲームの周辺機器としてのカメラは，体全体を使った次世代コントローラとして親しまれている．これに加えて，ショッピングモールや駅構内にある次世代型自動販売機やデジタルサイネージ（電子広告板）では，内蔵のカメラによって来客の特性やジェスチャを認識し，インタラクションやマーケティング応用に役立てている．このようにカメラの搭載された機器は今後とも増える見込みである．

一方，「目は口ほどにものを言う」と言われるように，人の視線情報は意図や興味の推定において重要である．そのため，ノート PC やデジタルサイネージなどの機器に搭載されたカメラに，特別なハードウェアの追加する事無くソフトウェア処理のみによって視線推定を実現できれば，商業応用に向けた価値が高いと言える．しかしながら従来の視線推定手法は，専用の装置（赤外線 LED・カメラのモジュールや装着型デバイス）を必要としているため普及の大きな妨げとなっていた．そこで，既存のインフラとも言えるデバイス内臓の単眼カメラのみを用いて，ユーザに金銭・物理的な負担を強いることなく，次世代のユーザーインターフェースとして新しい機能を提供可能であれば，これまでより多く

のユーザに受け入れてもらえ，普及のための鍵となりうる．それと同時に，マーケティングや興味推定，人の理解のためのより高度な情報を，デジタルサイネージ・家庭内のテレビ・店舗など社会全体から収集可能となり，商業展開可能性も高い．例えば，JR 駅構内に設置されている図 1.1 の自動販売機 [1] では，内蔵のカメラによって利用者の年齢・性別を判定し，おすすめの飲料を画面上に提示している．仮にこの機器上に視線推定機能を搭載すれば，実際にどの飲料が購入されたのかのデータに加え，視線推移から購入まで至らずとも利用者が興味を惹かれた飲料を推定可能となる．この情報を蓄積する事で，レイアウトや販売計画の見直しなど，マーケティングに役立てる事が可能となる．

広告効果測定への利用も期待されている．従来手法 [2] では，広告看板に内蔵されたカメラによって広告の前に人物の人数をカウントし，頭部姿勢推定によって広告へ顔を向けているか判定し，広告に興味を持っている人物の割合を集計していた．しかし，顔を向けずに目だけで広告を見る場合もあれば，逆の場合もあり，頭部姿勢情報のみでは不十分である．頭部姿勢に加えて，視線情報を得る事でより精度の高いマーケティング情報を蓄積することが可能となる．

テレビ視聴者に関する情報収集も関心が高まっている．従来は，テレビに接続された専用のセットトップボックスにより，テレビの電源状態や視聴されているチャンネルを収集し，「視聴率」として番組の価値を図る指標として用いられてきた．しかしながら，実際の利用状態においては，テレビの電源が入っていても必ずしもテレビに興味を示しているわけではない．そこで，テレビに内蔵のカメラ，もしくは，Kinect® などゲーム用カメラによってテレビ視聴者を観測する事で，視聴者がどのような状態にあるのか，画面が本当に見られているのか，更には画面上のどの領域が見られているかといった「視聴質」という新しい指標が研究されている．

エンターテイメントやマーケティング用途に加えて，自動車における先進予防安全技術としての応用も期待されている．図 1.2 のように，ドライバーの注意度や覚醒度を常時監視することで，事故原因として最も多い漫然運転や脇見行為を防止し，予防安全に役立てる事が可能である．そのため，従来から予防安全のためのドライバー視線推定は広く行われてきた [3–7]．しかし，個人キャリブレーションの煩雑さや環境光へのノイズ耐性が問題であった．これに加えて，昨今急速に技術開発が進んでいる自動運転のためにも，視線推定技術が必要不可欠である．例えば自動運転から手動運転へのハンドオーバーが必要なシチュエーションにおいて，ドライバーが運転可能な状態か判断する必要がある．また，歩行者や他車ドライバーの視線・認知状態を推定など，自動運転のための周辺環境センシングの応用のためにも視線推定技術が必要である..

これらの日常的な環境において社会で幅広く使われるための視線推定の要件は，安価な単眼カメラのみで動作し，ユーザ個別のキャリブレーションを必要とせず，解像度や環境光での悪条件に頑健な事である．以上の条件を満たすシステムは実利用に適しており，幅

図 1.1　次世代型自動販売機 [1]

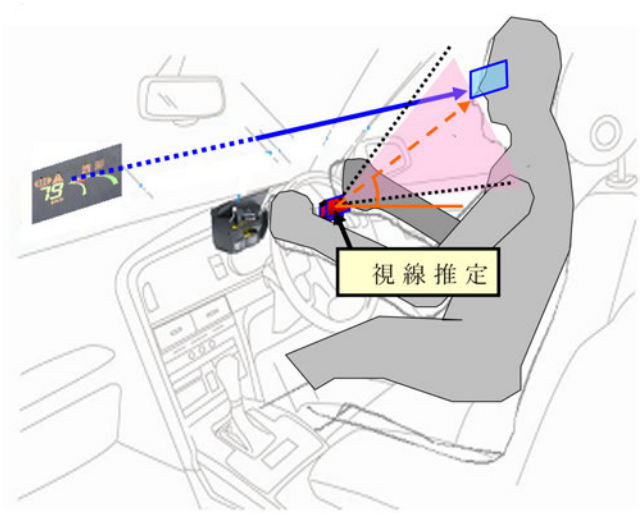


図 1.2　車載環境での応用

広い展開可能性が見込まれる．これまでに様々な視線推定手法が提案されてきたが，これらの条件を全て満たす手法は限られている．1.2 節で視線推定に関する研究をまとめつつ，我々の研究の立ち位置を明確にする．

1.2　関連研究

視線推定技術は様々な環境設定や目標のもと幅広く研究されてきた．それらは大きく「赤外線カメラを使用する手法」と「可視光カメラによる手法」に大別される [8].

1.2.1　赤外線カメラを使用する手法

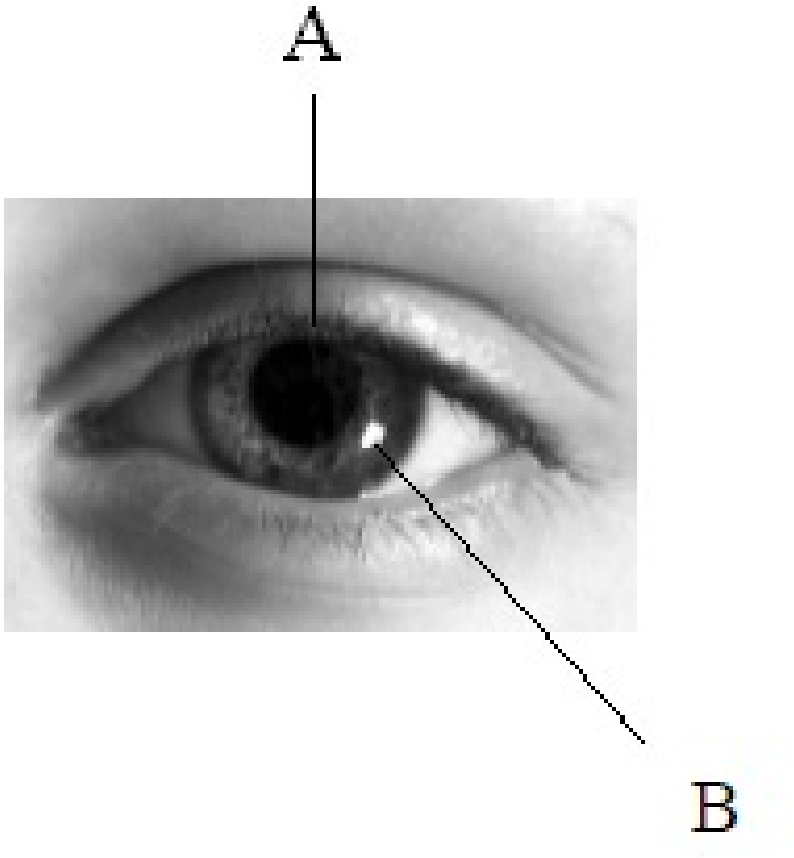
近赤外線を利用する手法 [9–16] では，赤外線 LED からユーザの角膜に近赤外光を投光し，その角膜上の反射点 (角膜反射点もしくは glint と呼ばれる, 図 1.3 A) と瞳孔中心位置 (図 1.3 B) から視線推定を行う．この方式のメリットは 2 つある．1 つ目は，眼球上の角膜はほぼ球体表面に近似できるため，視線方向が変わっても反射点の位置が変化せず，基準点として適している点である．2 つ目は，虹彩は赤外波長をよく反射するため明るく映り，瞳孔のみを観測しやすい点である．可視光線においては虹彩と瞳孔の区別を付けることは非常に難しいが，赤外画像では虹彩が顕著になるため，瞳孔のみを検出しやすい．虹彩と比較し瞳孔はその半径が小さく，まぶたによる遮蔽が少ないため，瞳孔中心位置の推定が容易である．

角膜反射点から瞳孔中心までのベクトル (pupil-glint ベクトル) を入力情報とし，キャリブレーション作業（既知のマーカを順次注視など）を経て入力ベクトルと注視点を結びつける写像を学習し，視線推定を実現する．瞳孔中心点は，瞳孔と虹彩の境界エッジに対し楕円フィッティングを施し，その中心位置として決定される．一般に計算コストは低く，200fps を超える超高速撮影や計算量の限られたモバイルデバイスでも利用可能である．

製品への応用例として，Tobii 社のアイトラッキング技術を搭載したタブレット端末 ibeam® があげられる (図 1.4)．このタブレットは波長の異なる 2 つの赤外線 LED および 2 つのカメラを搭載し，目線だけでアイコンの選択やページ送りができる機能が実装されている．

一般的に赤外線を用いる手法では非常に高精度な結果が得られる．その反面，赤外線ライトと赤外線カメラという特別な機械が必要であり，普及価格帯の製品に搭載する上では大きな障壁となる．また，使用前のキャリブレーションが必須であり，頭部位置変動の影響を受けやすい．Arantxa らはキャリブレーション負荷を低減するため，図 1.5 のようにモニタ下部に設置した赤外線 LED とカメラをカメラを組み合わせ one-キャリブレーションを実現した [12]．しかしながら，複数の LED の必要性はシステムの複雑性を増し，コストや設置環境を大きく限定する．

赤外線を用いる手法では，瞳孔エッジ点への楕円フィッティングが精度を左右する．楕円形状を完全に復元するためには，円弧の半分以上のエッジ点を捉える必要があり，必然



A: 瞳孔中心 (pupil) C: 角膜反射点 (glint)

図 1.3　瞳孔と角膜反射点 [13]

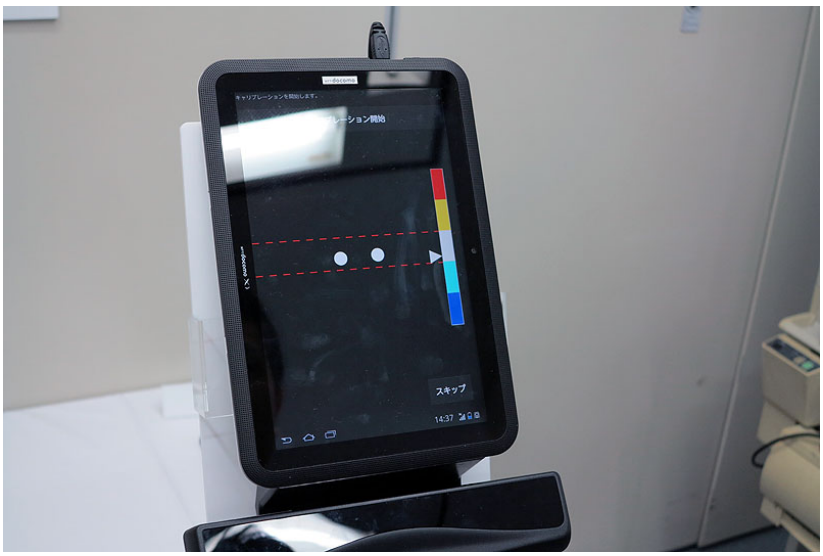
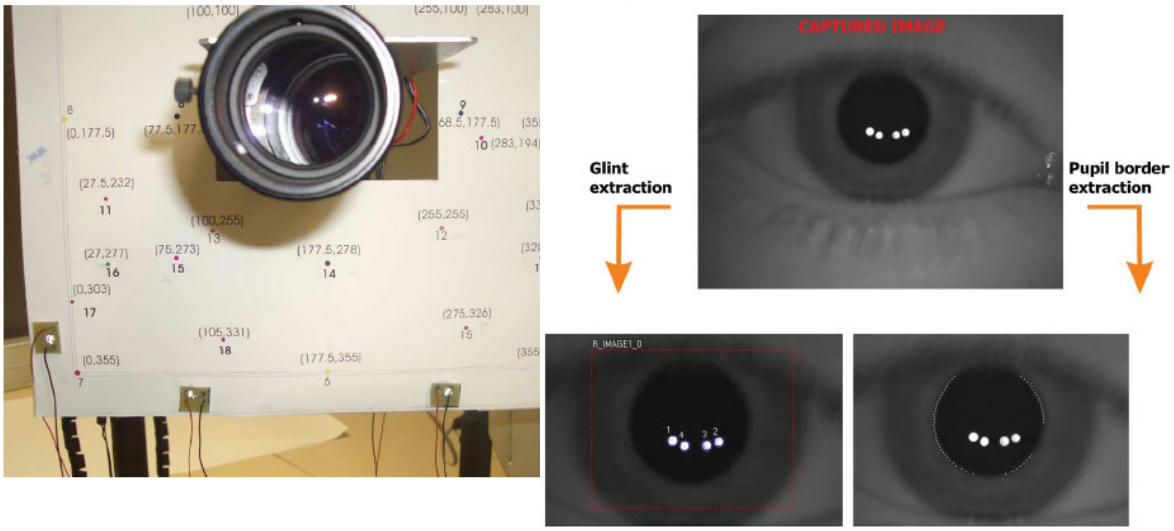


図 1.4　ibeam: 下部ユニット内に視線推定用のカメラと赤外線 LED を内蔵

的に高解像度の目画像が要求される．そして，LED から発せられた赤外線が角膜上で強く反射する必要があるため，カメラの近傍に使用者の位置が限られるため，使用環境が限定されるといった問題がある．



(a) カメラ下部に設置された LED　　(b)4 つの glint の検出と瞳孔境界抽出例

図 1.5　複数 LED によるキャリブレーション負荷の低減 [12]

1.2.2　可視光カメラによる手法

可視光カメラを使用する手法は，安価で広く普及しているデバイスで動作可能である点が大きなメリットである．可視光カメラ手法はアピアランスベース手法とモデルベース手法に分けられる．アピアランスベース手法は，被験者の目画像そのものを入力情報とし，機械学習によって注視点と組み合わせを学習し，新規目画像に対して注視点を推定する手法である．対してモデルベース手法は，眼球や顔のモデルを利用して視線方向を推定する．

アピアランスベース手法

アピアランスベース手法は，目の画像そのものを入力として，機械学習により注視点を推定する手法である．学習手法は，適応的線形回帰 [17], ガウス過程回帰 [18] などがあり，近年は畳み込みニューラルネットワーク (Convolutional Neural Networks, CNN) [19, 20] による手法が報告されている．

一般的にモデルベース手法と比較し低解像度画像に頑健と言われている [20]．その半面，早期の研究では，環境光変化や頭部姿勢変動の影響を受けやすいという性質があった [21–24]．しかし，近年の報告では，顔モデルにもとづき推定した 3 次元頭部姿勢追跡と目画像の混合モデルにて学習するなど，モデルベース手法のアイデアを組み合わせる事で頭部姿勢拘束条件を緩めた [25–27]．Sugano ら [28] は，ノート PC 操作環境においてユーザがマウスでクリックする座標を正解注視点と仮定し，機械学習によって未知の目画像から注視点を推定する手法を提案した．また，映像中の視覚的顕著性の高い領域を正解注視領域とみなす事で，ユーザに特別な操作を強いることの無い可視光カメラのみによる視線推定手法を実現した [29]．しかし，良い精度が得られるまでに個人別で数百枚の正解サンプルが必要であり，未知のユーザには適応できない．一般にアピアランスベース手法はモデルベース手法よりも多くの個人別学習用データが必要と言われており，未知の被験者に対する有効性，言い換えればキャリブレーションフリー性は不明確であった．これに対し，Zhang らは環境光変化の激しい条件下 (in-the-wild) で MPII gaze dataset と呼ばれる大規模データセットを作成し，CNN による学習を通して，ユーザー個別学習の必要無い視線推定手法を報告した [20]．しかしながら，これまで提案されたデータセット [20, 30] は，ノート PC やタブレットの使用者を対象としたものであり，被験者の位置がカメラの近傍に限られており，デジタルサイネージやテレビなど，頭部位置がカメラから遠い環境での検証は不十分であった．

モデルベース手法

モデルベース手法は，顔や目のモデルを入力画像にフィッティングする事で視線を推定する手法である．この手法では顔や眼球などの人間の解剖学的特徴を手がかりとするため，シンプルなモデルで記述が可能であり，アピアランスベース手法のように個人毎の大量の訓練データを必要としない．モデルベース手法の初期の研究 [31, 32] では，虹彩の形状から視線方向を推定し，これは”circle アルゴリズム” と呼ばれている．これらの手法の入力情報は目画像のみであるものの，高解像度の目画像を必要であった．後に，モデルベースのアプローチは 3D 眼球モデルを使用する [3, 33]．これらのアプローチでは，視線方向は眼球中心から虹彩中心までのベクトルと定義される．Kitagawa らの手法 [34] は放物線で近似したまぶた形状を含む眼球モデルを使用し，単眼カメラのみで視線推定を実現した．しかし．まぶたの形状は多種多様であり放物線で近似できない場合もある．加えて，使用開始前に目尻目頭の 4 点を手動で指定する必要があり，全自動のシステムとはなっていない．手法 [35] では 3 次元視線ベクトルを推定するために顔特徴点を利用したが，目の位置の正確な検出が前提であり，加えて 1 回のキャリブレーションが必要であった．

モデルベース手法の利点は，追跡の際に顔画像から得られる頭部位置情報や方向情報を利用する事が多いため，アピアランスベースの手法より効果的に頭部姿勢変化を考慮でき

る事である [36]．手法 [37] は，３次元頭部姿勢，唇，眉毛，虹彩の追跡手法を提案した．手法 [38] では頭部姿勢情報と目の位置情報を組み合わせる方法を提案した．しかしながら，既知のターゲットを見るというキャリブレーションフェーズが依然として必要であった．モデルベース手法の欠点の一つは，頭部内の正確な眼球位置など個人間で異なるパラメータを得るために，使用開始前に個人ごとのキャリブレーションプロセスが必要である事である．この必要性は多くの使用環境を制限してしまう．Yamazoe ら [39] はユーザの負担を軽減するための自動キャリブレーション手法を提案した．隠れキャリブレーションとして，画像とモデルの投影誤差を最小にするように顔モデルと眼球モデルのパラメータを最適化する．この手法は特別なキャリブレーション動作を必要としない点で注目に値する．しかしながら，例えば短時間であっても，キャリブレーションにかかる時間はその適応性を制限する． さらに，頭部姿勢の変化量に対する頑健性は明確にされていない．Cazzato ら [40] は RGB-D センサを用い，また，眼球中心位置を眼球表面から 12 mm と一般化することにより，キャリブレーション不要かつ即座に使用可能な視線推定手法を提案した．これはいくつかの状況には適しているものの，デプスセンサの必要性は利用可能なシーンを制限する．実験は RGB-D センサから 70 cm の場所で行われたが，遠方の人物の視線推定については検証されていない．

一般的に，モデルベース手法では顔や目の特徴点の検出位置精度が重要であるため，高解像度を要求する [20]．そのため，頭部位置がカメラから遠く離れた場合 (1 m 以上) など，顔や目領域の画像解像度が低い場合の推定精度が課題であった．カメラに対する人物の移動の自由度，つまり追跡可能な幅および深度の頭部位置範囲の広さは，Human Computer Interaction (HCI)，デジタルサイネージ，TV などの用途にとって重要である．しかしながら，上記の視線推定手法は，対象人物の位置をカメラの近傍 (1 m 未満) に制限するか，椅子などを利用して人物の位置を規定の位置に制限していた．

遠距離にいる人物を対象とした視線推定

前項で述べたように，カメラに対して広範囲 (左右方向および深度方向) を移動する人物の視線推定は困難であった．手法 [41,42] では，視線推定のために，虹彩追跡をする代わりに頭部と体の姿勢情報を用いた．手法 [43] では，固定した広角カメラから顔を検出する事で大まかな目領域の位置を推定し，別のパンチルトズームカメラで目の領域にズームする２カメラシステムを提案した．この手法は，遠方の人物 (カメラから 4 m) の注視推定を達成した．しかしながら，パンチルトズームカメラでは構造上複数の人物を同時に追跡できない．さらに，キャリブレーションのために，無意識的であれ対象人物が一度カメラを直視する動作が必要であるが，実際のアプリケーションでこのような動作が発生するかは不明である．手法 [44,45] では，遠くに離れた人物の頭部姿勢変動に対応した視線推定のために RGB-D センサを用いた．[44] では，特徴点追跡において高解像度画像が必要と

なってしまう幾何的アプローチを避けるために，生成モデルを提案した．Cazzato ら [45] は，頭部姿勢が視線方向と相関があるという前提のもとで，頭部姿勢情報のみに基づく視線推定手法を提案した．彼らはカメラからいくつかの距離 (70 cm, 150 cm, 250 cm) で検証したが，カメラに対し左右方向の移動自由度については明確にされていない．

1.2.3 顔特徴点追跡および頭部姿勢推定手法

顔特徴点追跡および頭部姿勢推定は，可視光カメラによる視線推定において重要な要素となる．近年の視線推定手法は，目の検出のための前処理として，そして頭部姿勢変動への対応のため，頭部姿勢推定を処理の第一段階としている．

コンピュータビジョンの分野においては，過去 20 年以上に渡ってこの分野の研究がされてきた [8, 46–50]．代表的なアプローチは，アクティブシェイプモデル (ASM) やアクティブアピアランスモデル (AAM)，およびそれらの組み合わせがある．アクティブシェイプモデル手法は，顔形状のバリエーションを事前に学習し，固有空間内の線形結合としてモデル化する．アクティブアピアランス手法では，顔の外観をモデリング化する．これらを組み合わせ，入力画像にフィッティングする事により，顔の形状を得る．この形状はあくまで画像上の 2 次元情報であるが，これらの相対位置から顔姿勢を推論する事が可能である．一つの方法は，単純な線形回帰を使用し，顔姿勢から頭部姿勢を得る [51] が，別の手法では，3 次元アクティブシェイプモデルモデルを 2 次元フィッティングの制約に用いる [52]．これは 3 次元モデルのパラメータから頭部姿勢を直接推定する事を可能とする．Saragih らは [53]，目尻や鼻の穴といった局所特徴パッチを 3 次元顔形状モデルに基いて画像上に散布し全体の整合性を保ちつつフィッティングする Constrained Local Model (CLM) を提案した．

近年はニューラルネットワークによる手法に焦点があたっている．Sun ら [54] は，3 階層カスケード CNN を提案し，両目，鼻，左右口角の 5 点顔特徴点検出手法を提案した．Baltursaitis ら [55] は，CLM を拡張し，局所特徴パッチを 3 層のニューラルネットワークで学習した Constrained Local Neural Fields (CLNF) 特徴に基づく顔姿勢推定，視線推定，および Action Unit 推定のフレームワークを提案し，OpenFace としてソースコードを公開している．この手法ではまず，入力画像から局所領域を抽出し，ニューラルネットワークにより特徴点の応答マップを得る．続いて，応答マップと各特徴点毎の信頼度を考慮し最適な特徴点位置を決定する．Baltursaitis らはこの手法が IBUG Dataset および AFW Dataset において最も高い精度を有したと主張している．また，視線推定は，MPII gaze dataset において [20] を上回り，最先端の精度を実現している．以後この手法を OpenFace と呼称し，本稿の比較対象とする．

1.2.4　関連研究のまとめ

ここでは，これまでに述べてきた関連手法についてまとめる．単眼カメラによる視線推定手法は，赤外線カメラを用いる手法と可視光カメラを用いる手法に大別される．赤外線カメラを用いる手法は，非常に高い精度が得られる反面，専用の機器の設置コストや，頭部姿勢変動の影響を受けやすさ，個人別学習の必要性といった理由により，社会で幅広く使われる手法としては適していない．可視光カメラを用いる手法は，更に，アピアランスベース手法とモデルベース手法に分けられる．アピアランスベース手法は，目領域そのものを入力とし，機械学習によって注視点を推定する手法である．環境光変化に頑健である反面，頭部姿勢変動の影響を受けやすく，大量の個人別学習画像が必要となる性質を持っている．近年の報告では，頭部姿勢推定情報とともに大規模データセットによる学習によって，被験者に依存しない汎用性を獲得している．しかしながら，検証はカメラにごく近い領域 (カメラから 1 m 以内) でのみ行われており，実社会での応用に即した広い空間 (カメラから 2 m 以上) での精度は明確にされていない．モデルベース手法は，大規模な学習をせずとも視線推定が可能であり，頭部姿勢変動にも頑健である．しかし，高い精度のためには一般に高解像度画像を要求しており，被験者がカメラから離れるなどした場合，解像度低下の影響を受けやすい．

以上の特徴を表 1.1 にまとめた．本論文で説明する提案手法では，低解像度目画像に頑健な虹彩追跡手法を軸に，耐環境性の高い顔特徴点追跡，および独自のデータセットによる使用可能頭部位置の拡張により，実用的な視線推定手法を目指す．

表 1.1　視線推定手法の種類と特徴

	赤外線カメラ	可視光カメラ		
		アピアランスベース	モデルベース	提案手法
精度	◎	△	○	○
低解像度頑健	×	△	×	○
頭部姿勢変動対応	△	△	○	○
耐環境光	×	○	×	△
キャリブレーションフリー	×	○	△	○
使用コスト	×	○	○	○
使用可能エリア	×	×	△	○

1.3　研究目的

本論文では，社会で広く使われるための視線推定手法を目指し，以下の３つを研究の要件として設定した．

１．安価で普及している単眼カメラのみで動作

２．未知のユーザに対してもキャリブレーションフリーで推定

３．カメラの近傍に限定しない広いエリアで推定可能

従来の手法のうち，距離センサやステレオカメラ，赤外線カメラを用いるものは，確かに高精度であるものの，専用のデバイスを一般用との製品に搭載する事のコストは非常に大きいため，社会で広く使われる上での障壁となる．よって，可視光単眼カメラのみで動作する事を一つ目の要件とした．

キャリブレーションプロセスについては，明示的に行われる場合，ユーザにとって負担となり，視線推定の応用展開性を大きく妨げることになる．また，非明示的であったとしても，キャリブレーションプロセスの間，視線推定開始までの時間のデータは得られない．例えば，街角での広告効果測定などの応用では，人物が対象に興味を示す時間はわずかであり，キャリブレーションが完了する前にその人物が去ってしまうケースが容易に考えられる．よって，これも使用シーンを限定してしまう．我々は，ワンショット，つまり１フレーム目から推定を開始する事が重要と考えている．これを二つ目の要件と定めた．

また，使用可能な範囲の広さも重要であると考えられる．従来手法では，パーソナルデバイスを対処うとし，タブレットや PC を使う範囲，つまり 1m 以下の距離のユーザを対象としてきた．しかしながら，デジタルサイネージや歩行者，ロボットへの応用を考えた場合，この範囲は出来る限り拡張されるべきである．人物の位置が遠い場合，得られる目領域解像度が著しく低下し，また一般的には頭部位置の推定精度も低下するため，これらの課題を解決する必要がある．

1.4　提案手法の概要

1.3 節で述べた要件を満たすために，低解像度に頑健な虹彩追跡手法を軸とし，データセットに基づくキャリブレーションフリー視線推定手法を提案する．まず，頭部位置の非拘束かつキャリブレーションフリーを実現するため，また，検証のため，我々は室内に設置したテレビを注視する人物の顔写真を収集したデータセット「Room-Scale Gaze Dataset」を作成した．以後このデータセットを RSGD と呼称する．RSGD では，従来のデータセットと比較し，カメラからの人物の頭部位置遠い事が特性である．また，距離のみでなく，広角のカメラを用いていることで，左右方向の分布も広い．同じカメラ解像度

だとしても，広角カメラで遠い人物を捉えると，人物の顔領域の解像度が低下してしまう．同時に目領域解像度も低くなる．そのため，従来手法では大きく精度低下が発生してしまう事が課題となる．これを解決するため，我々は，低解像度に頑健な顔特徴点検出手法および虹彩追跡手法を提案する．

提案手法の処理の流れを，図 1.6 に示す．まず，入力画像に対して，頭部姿勢推定を行う．頭部姿勢は顔特徴点検出後，予め作成した 3 次元モデルをフィッティングさせる事で得られる回転行列および並進行列から推定される．並進行列からカメラを原点とした世界座標系内の 3 次元頭部位置を算出し $\boldsymbol{t}$ と表記する．$\boldsymbol{t}$ は全顔特徴点の重心位置に相当する．また，回転行列から回転角を算出し，頭部中心から顔の正面方向に向いたベクトルとして頭部方向を $\boldsymbol{r}$ と表す．続いて，頭部位置 $\boldsymbol{t}$，頭部方向 $\boldsymbol{r}$ から画像中の眼球中心座標 $\boldsymbol{e}_l$，$\boldsymbol{e}_r$ を推定する．この位置をもとに，3 次元眼球モデルに基づく虹彩検出を行い，視線方向 $\boldsymbol{g}_l, \boldsymbol{g}_r$ を算出する．ここまではモデルベース手法に基づいた流れであるが，広い空間内での非拘束視線推定精度の改善のため，RSGD によって学習された回帰モデルによって $\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{g}_l, \boldsymbol{g}_r$ から画面上の注視点 $\boldsymbol{p}$ を推定する．

1.5　本論文の構成

本論文の構成を以下に示す．第 1 章では，研究背景として，社会における視線推定の重要性や応用例を紹介し，その現状について述べる．また，関連する手法を挙げて，本研究の位置付けや研究目的を明らかにする．

第 2 章では，顔画像からの顔特徴点検出手法について詳説する．顔特徴点検出では，自然環境下で撮影された様々な人種，顔向き画像を含むデータセットを学習した CNN を用いることで，環境光変化や頭部姿勢変動に頑健かつ高精度な顔特徴点検出を行う．さらに，視線推定のための前処理として，目領域のみを学習したネットワークにより詳細な目形状の推定について述べる．また，検証のため，顔特徴点検出のために一般的に用いられるデータセットを用いて精度比較実験を行う．

第 3 章では，虹彩追跡について詳説する．従来手法では外乱エッジや低解像度環境が課題であったが，提案手法では，3 次元眼球モデルをもとに，虹彩エッジの強度および勾配方向に着目した尤度評価と密なサンプリングにより頑健な虹彩追跡を実現する．このように低解像度に強い追跡を実現する手法について述べる．

第 4 章では，注視点推定について詳説する．広い空間内で非拘束視線推定を実現するための独自視線データセットを作成する．本データセットにより学習した回帰モデルによって，画面上の注視点を推定する方法について述べる．

第 5 章では，提案手法の有効性を検証する．第 4 章で作成した視線データセットによる検証を通して，従来のデータセットより低解像度の条件における注視点推定精度の比較を

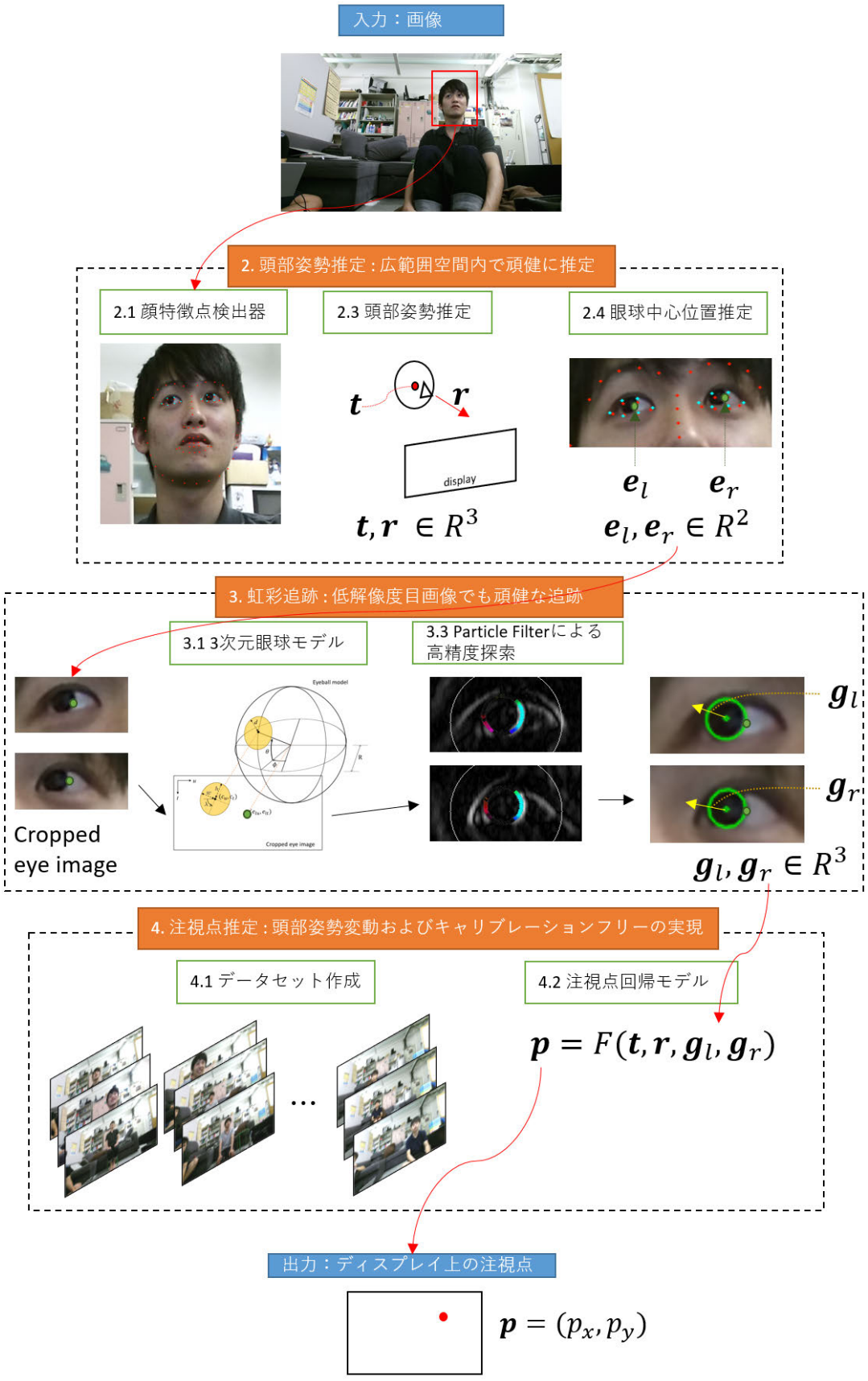

入力：画像
2. 頭部姿勢推定：広範囲空間内で頑健に推定
2.1 顔特徴点検出器
2.3 頭部姿勢推定
2.4 眼球中心位置推定
t
r
display
t, r ∈ R^3
e_l
e_r
e_l, e_r ∈ R^2
3. 虹彩追跡：低解像度目画像でも頑健な追跡
3.1 3次元眼球モデル
3.3 Particle Filterによる
高精度探索
Eyeball model
Cropped eye image
Cropped
eye image
g_l
g_r
g_l, g_r ∈ R^3
4. 注視点推定：頭部姿勢変動およびキャリブレーションフリーの実現
4.1 データセット作成
4.2 注視点回帰モデル
p = F(t, r, g_l, g_r)
出力：ディスプレイ上の注視点
p = (p_x, p_y)

図 1.6　提案手法の流れ

行う．

第 6 章では，本論文をまとめ，今後の課題と展望について述べる．

第 2 章

顔特徴点検出および頭部姿勢推定

本章では，視線推定のための前処理として，顔画像中から顔特徴点を検出する手法について説明する．提案手法では，二つの回帰モデルを事前に作成した．一つ目は，顔画像全体を入力とする顔特徴点検出器である．二つ目は，目領域画像を入力とする目形状点を推定する目形状検出器である．これらを直列に接続し，低解像度や遮蔽環境においても高精度な顔特徴点検出器の作成について述べる．続いて，予め作成した 3 次元顔モデルを得られた顔特徴点位置にフィッティングさせる事による頭部姿勢を推定と，得られた顔特徴点位置および頭部姿勢から画像中の眼球中心位置を求め方法について説明する．最後に，顔特徴点検出において一般的に使用されているデータセットにより本手法の評価を行う．

2.1　顔特徴点検出器

頭部姿勢変動に頑健な視線推定を実現するために，頭部位置及び回転の推定は極めて重要なベースとなる．提案手法では，まず顔特徴点位置を検出し，それに予め作成したモデルを当てはめる事で頭部姿勢を推定するアプローチを取る．そのための顔特徴点検出器について本節で説明する．

2.1.1　データセット

画像からの頭部姿勢検出は非常に研究が盛んな分野であり，実験室環境を対象としたデータセットでは精度が飽和状態にあった．そのため，自然環境内での顔特徴点検出が新たな目標となり，様々なアノテーション済みデータセットが公開された．しかしながら，

それらのデータセットは皆異なるフォーマット，特徴点数で作成されており，アノテーション精度にもばらつきがあったため，学習は検証，比較において問題が発生していた．

これを受けて，統一したフォーマット，特徴点数で作成されたのが Intelligent Behaviour Understanding Group (iBUG) の公開している iBug Facial Point Annotations (iBUG FPA) である [56–58]．iBUG FPA は 300W, AFW, LFPW, HELEN, IBUG, XM2VTS, FRGC Ver.2 のサブセットから構成される．我々は，このうち入手可能であった 300W, ALW, HELEN, IBUG, LFPW からなる 4437 枚のアノテーション済み顔画像のうち 9 割にあたる 3994 枚を学習用に，残りの 443 枚を検証用に用いた．図 2.1 にデータセット内の顔画像例を示す．また，図 2.2 にアノテーションに使われる 68 点を示す．

2.1.2　CNN による顔特徴点検出器

提案手法では，畳み込みニューラルネットワーク (CNN) により顔画像中から 68 点の顔特徴点全てを学習する．ネットワーク構造は図 2.3 に示す AlexNet である．AlexNet は Krizhevsky ら [59] の提案した 7 層の畳み込み層と 2 層の全結合層から構成されるネットワークであり，識別問題や回帰問題で広く使われている．近年，非常に多層のネットワーク [60, 61] が提案され，性能が飛躍的に向上しているが，視線推定という目的からリアルタイム性が重要であり，我々は比較的層の浅いこのネットワークを採用する．入力は 220x220 のサイズの顔画像とし，出力は 68 点の顔特徴点 × 画像座標 (x, y) の 2 次元の 136 次元ベクトルとする．過学習を抑制するため，全結合層において 0.6 の割合で重みを 0 にする dropout を行う．この比率は経験的に決定した．

ここでは，学習のために用いる顔画像群の作成について説明する．

iBUG FPA データセットは風景などを含む一般の画像であるため，アノテーションされた正解顔特徴点の位置を内包する矩形を切り抜いて顔画像を作成する．なお，画像中からの顔矩形検出は既存の顔検出器 [62] である Haarlike 特徴量 +Adaboost で学習済みのモデルと Dlib Face Detector を利用する．しかし，これらの顔検出器は必ずしも正確な顔矩形を検出するとは限らず，実際の顔位置よりもずれていたり，サイズが異なっていたりする．これらの位置ずれやスケール変化にロバストな検出のため，学習用データの顔画像の切り抜きにおいては，実際の顔矩形よりも大きく拡張した領域で学習する．拡張幅の上限と下限を padding_sup, padding_inf というパラメータとして学習時に指定し，その間でランダムに決定されたサイズで切り抜きを行った．

今回用いた学習用画像は約 4000 枚であるが，CNN のための学習には数万枚オーダーの画像が求められるとされており，十分なデータ数とは言えない．データ数の不足は過学習を招くが，顔特徴点のアノテーションは非常にコストのかかる作業であり，容易に枚数を増やせるものではない．そのため，仮想的に学習用画像を増やす Data Augmentation を

(a)　(b)

(c)　(d)

(e)　(f)

図 2.1　iBUG FPA データセット

行う．本論文では Data Augmentation のため，flipping, rotating, shifting を適用する．

flipping: 人間の顔の左右対称性を利用し，y 軸を基準に反転させた画像を作成し，別の顔画像とする．本来の右目は左目に，右耳は左耳となる．

shifting: 顔検出時のずれにロバストな追跡を目指し，cropping する矩形位置をランダムに並進させる shifting を行う．

rotating: 頭部姿勢変動に因る見えの変化を生成する．iBUG FPA データセットは 2 次元

図 2.2　アノテーションに使用される 68 点 [56]

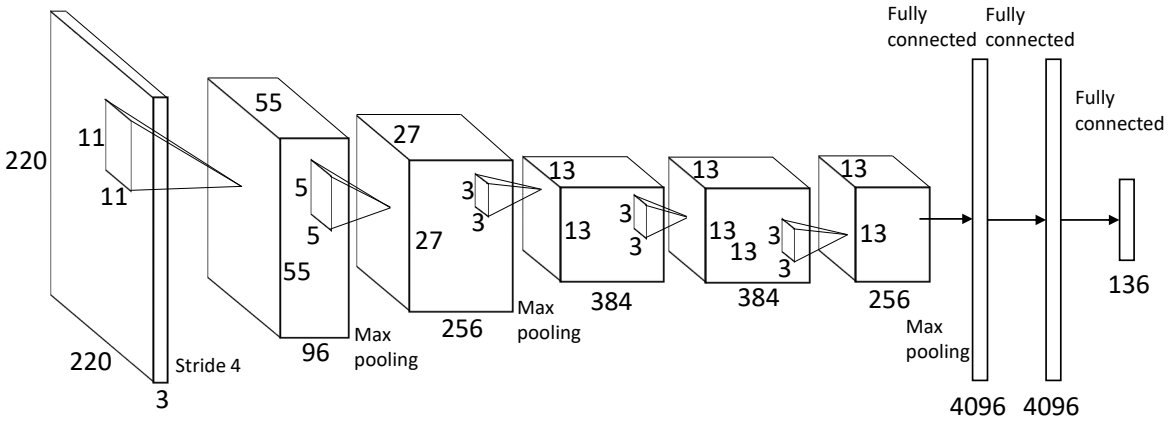


図 2.3　AlexNet

画像のみであり，首の pitch, yaw 回転を再現する事は出来ないものの，画像を回転させる事で擬似的に roll 回転のみ再現する．

shifting = ±10 pixel, rotating = ±0 degrees(無し), padding_inf = 1.5, padding_sup = 3.0 として iBUG FPA データセットを 1000epoch 分学習した際の損失を図 2.4 に示す．なお，損失は顔画像の高さ・幅のサイズを 1 とした場合の検出した顔特徴点座標と正解座標との平均二乗誤差と定義する．training loss は学習時の正解特徴点位置と予測された特徴点位置との誤差を表し，test loss は同時にテスト用データで検証した際の誤差である．

図 2.4 から，学習の初期段階 (epoch=30) において局所解に陥り，学習がうまく進んでいない．CNN の初期段階の層では，画像のエッジやコーナーなどに相当する低レベルなフィルタが形成される事が理想である．しかし，本学習条件では拡張幅が大きすぎるた

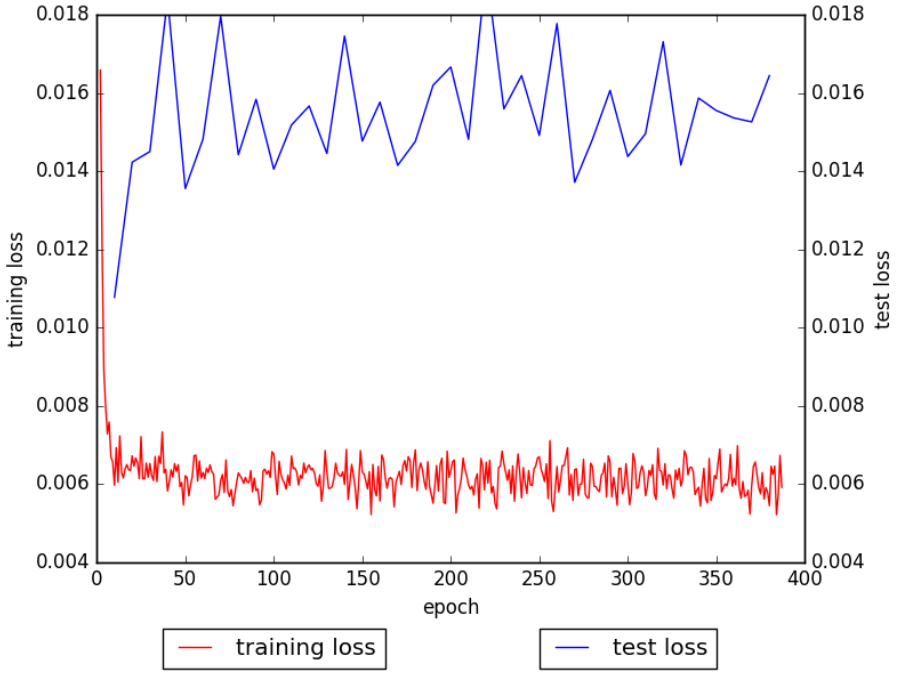


図 2.4　Loss curve1: shifting = ±10 pixel, rotating = ±0 degrees(無し), padding_inf = 1.5, padding_3.0

め，それらのフィルタが効果的に学習されなかった事が推察される．

一般に，学習対象が難しい時，条件を緩めた対象を学習し，後に難しい対象を学習するカリキュラムラーニングと呼ばれるテクニックがある．

カリキュラムラーニングの思想を取り入れ，shifting = ±5 pixel, rotating = ±0 degrees(無し), padding_inf = 1.5, padding_sup = 2.0 と条件を緩和して同データセットを 1000epoch まで学習した際の損失を図 2.5 に示す．

図 2.5 から分かるように，緩やかに損失が減少し，学習がうまく進んでいる事が伺える．しかしながら，epoch = 400 あたりから飽和が発生し，test loss の減少が頭打ちとなっている事が確認出来る．

続いて rotation を加え，shifting = ±5 pixel, rotating = ±20 degrees(無し), padding_inf = 1.5, padding_sup = 2.0 とし，図 2.5 における epoch=1000 の状態の重みを初期値として再学習を行った結果を図 2.6 に示す．図 2.6 から，rotating の data augmentation を加えた事により epoch 1010 付近の train loss が増加しているが，すぐに再学習し epoch 1050 付近で rotating 無しの時と同等まで減少した．なお，図 2.5 と図 2.6 を比較すると，training loss の値が上昇しているが，図 2.6 では rotation が加わったためより難しい条件で学習しているためである．

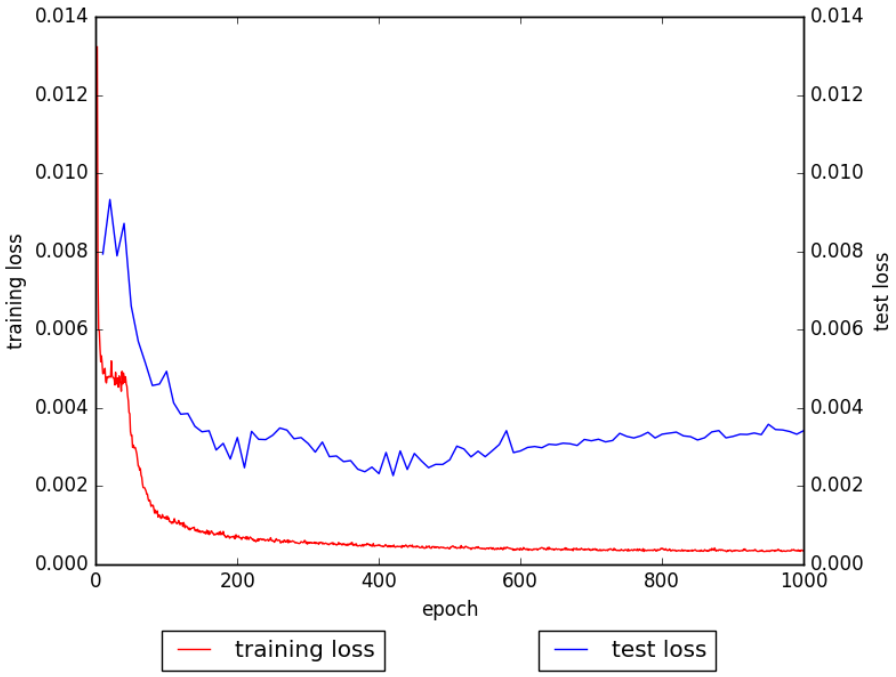


図 2.5　Loss curve2: shifting = ±5 pixel, rotating = ±0 degrees(無し), padding_inf = 1.5, padding_sup = 2.0

図 2.6　Loss curve3: shifting = ±5 pixel, rotating = ±20 degrees(無し), padding_inf = 1.5, padding_sup = 2.0

2.2　目形状検出器

2.1 節では，単一のネットワークによってすべての顔特徴点を推定した．しかしながら，入力層のサイズは 220x220 であり，得られる顔特徴点の推定精度もこのサイズに準ずる．また，7 層と小規模のネットワークであるため，表現力に限界があり，いくつかの特徴点位置，特に目形状において，我々の要求する精度を満たすものではなかった．これを解決するため，目形状のみを検出する新たなネットワークを作成し，2 階層目のカスケードとして 2.1 節のネットワークと直列に接続し，目形状の詳細な検出を目指す．この時，予測された顔の角度は既知であるため，顔の rotation 角度に応じて目領域矩形を回転させ，顔回転角をキャンセルするようにアフィン変換を施し切り抜いた目画像を 2 階層目の cascade の入力とする．

学習画像は，目形状を構成する 6 点を囲む矩形を，2.1 節と同様に拡張した．Data Augmentation についても同様であるが，目の形状は左右非対応であるため (目尻と目頭の形状は異なる)，切り抜いた目画像の中での flipping は行わない．その代わり，画像全体を flip し，全てのデータに対して，左側の目のみ学習することとした．つまり，flip していない画像については左目を，flip した画像については本来は右目であったものを左目と捉えて学習した．

shifting = $\pm$7 pixel, rotating = $\pm$0 degrees(無し), padding_inf = 1.5, padding_sup = 3.0 の条件で学習した過程を図 2.7 に示す．

2.1 節の顔特徴点検出ネットワークを 1st cascade，本節の目形状検出ネットワークを 2nd cascade と呼称し，処理の流れを図 2.8 に示す．

また，処理結果例を図 2.12 に示す．"Faces with predicted points"では，入力顔画像に 1st cascade の結果を赤点で，2nd cascade の結果を水色の点で示す．また，"Left eye"および"Right eye"には 2nd cascade に入力された左目，右目の目画像と検出結果を示す．図より，(a),(b) など顔向きが正面の時は 1st cascade と 2nd cascade の結果に大きな違いは無いが，(c),(e) など顔向きが正面以外の時は 2nd cascade の結果が良い事が認められる．(d) では眼鏡フレームの影響で右目の 1st cascade の検出結果が上にずれているが，2nd cascade の結果はより良い．(k) の右目に着目すると，1st cascade で得られた結果の大きな誤差のため，赤線で表す目矩形の中に実際の目の下瞼が収まっていない．しかし，2nd cascade の結果を見ると，正しく特徴点を捉えられている事が分かる．

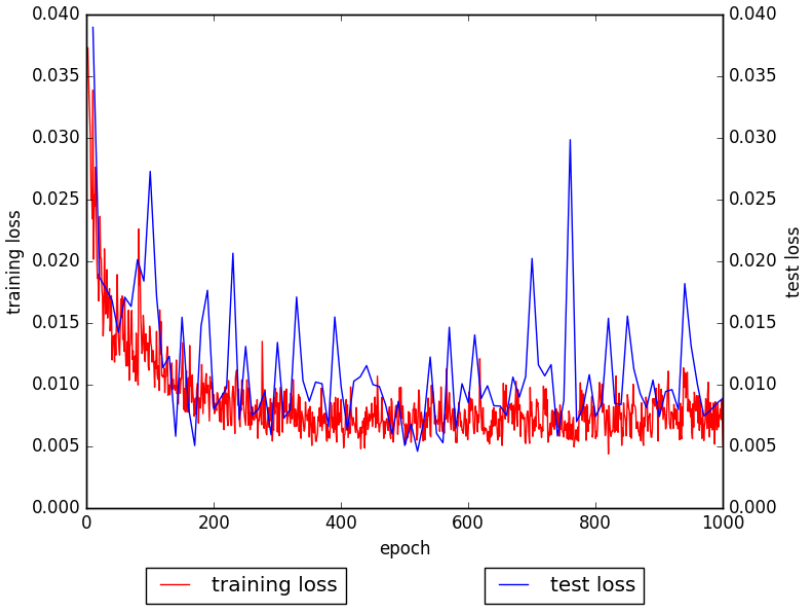
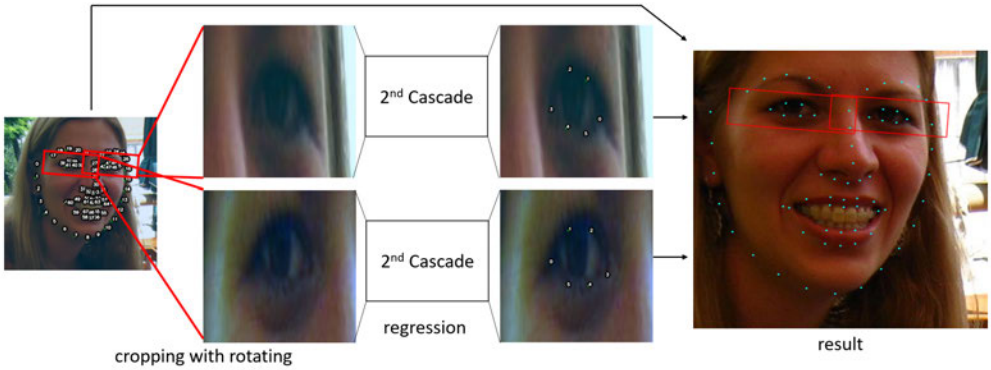


図 2.7　Loss curve4: shifting = ±7 pixel, rotating = ±0 degrees(無し), padding_inf = 1.5, padding_sup = 3.0 の時

(a) 1st cascade layer



(b) 2nd cascade layer

図 2.8　顔特徴点検出の処理の流れ

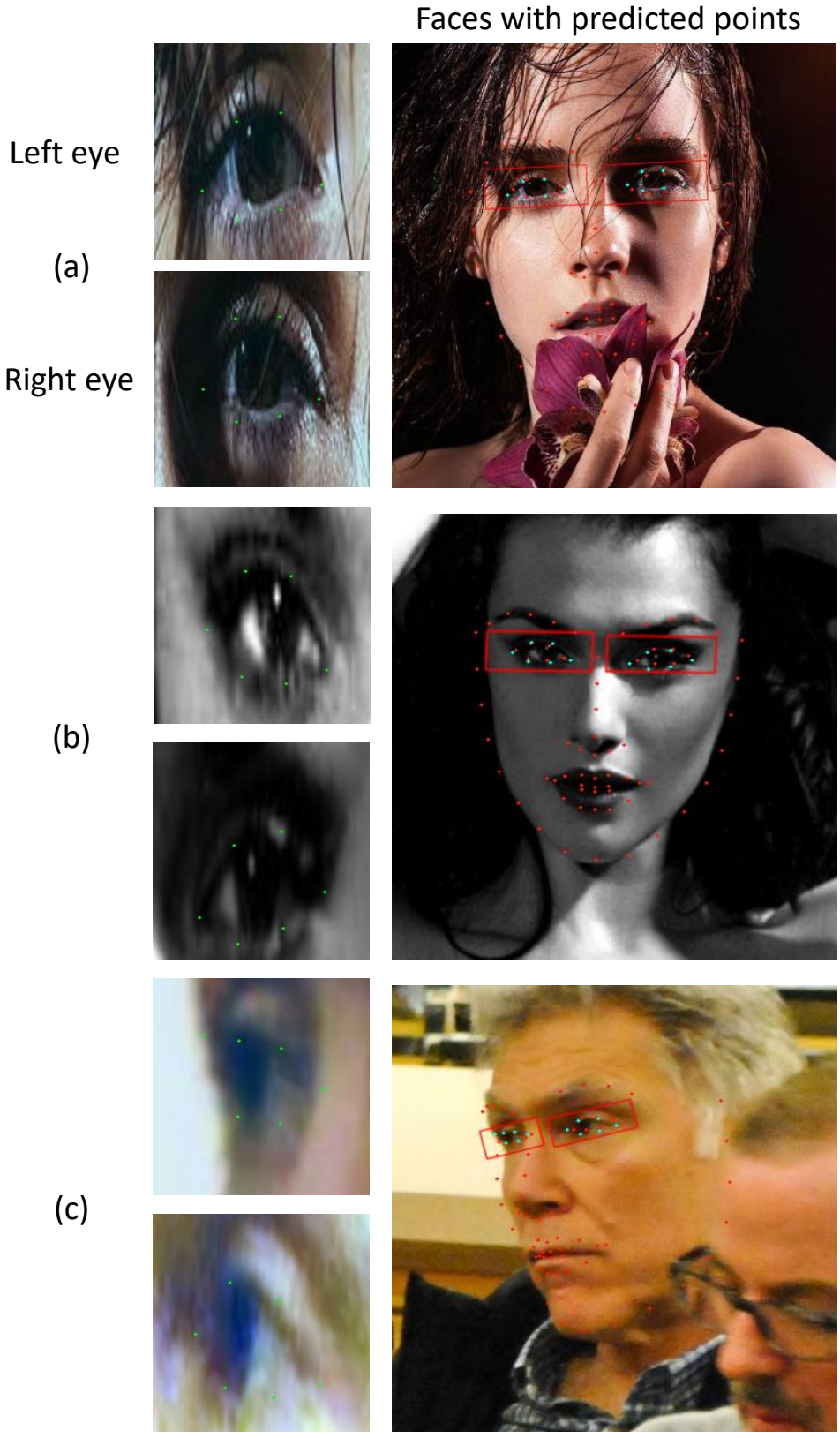


図 2.9　顔特徴点検出結果 1

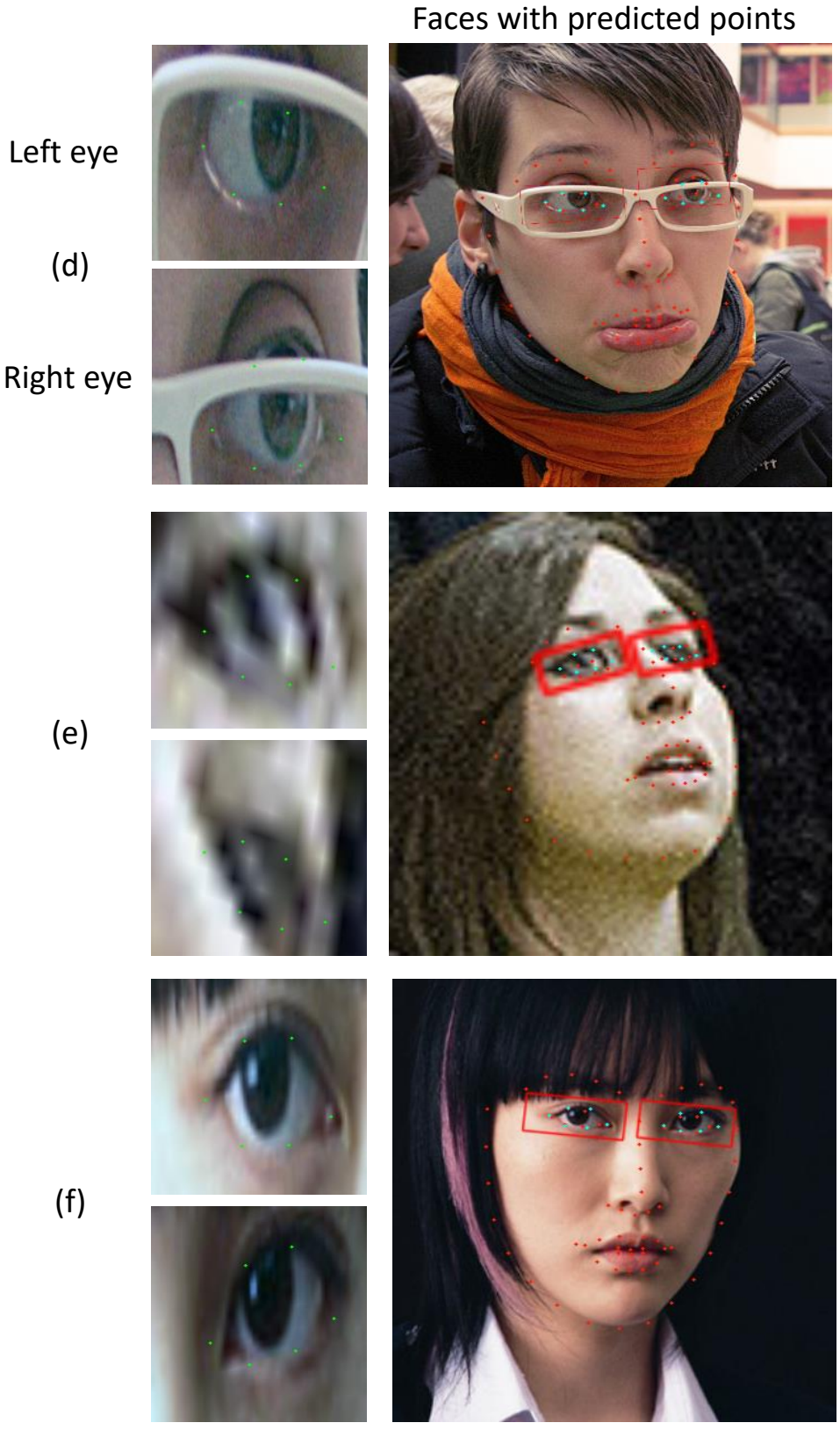


図 2.10　顔特徴点検出結果 2

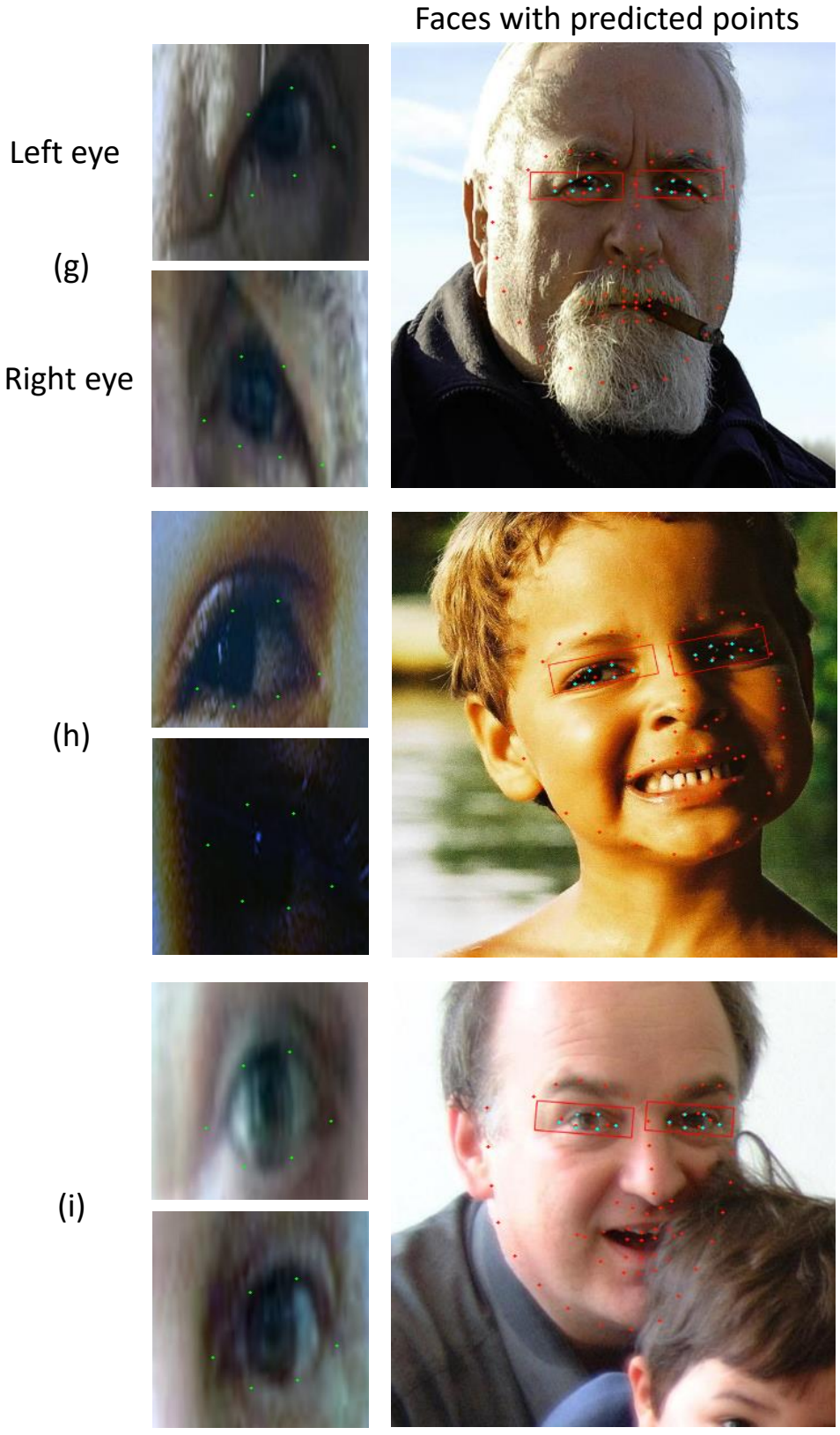


図 2.11　顔特徴点検出結果 3

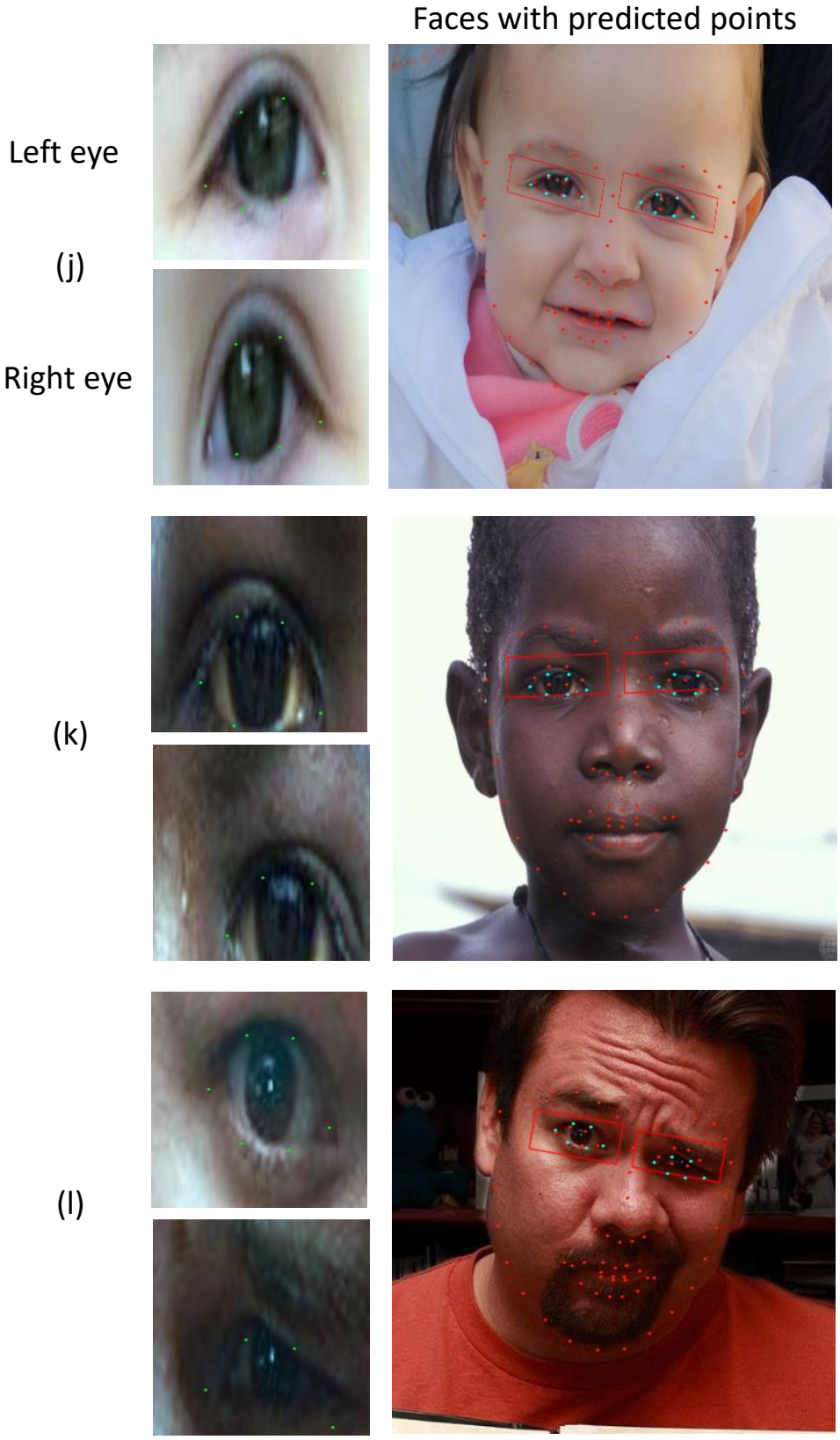


図 2.12　顔特徴点検出結果 4

2.3 従来手法との比較

2.3.1 定量的評価

提案手法の顔特徴点検出精度の有効性を検証するため，CLNF ベースの顔特徴点検出手法である OpenFace [63] との精度比較実験を行った．実験は iBUG FPA データセットのうち，学習に用いなかった 443 枚について，68 点の顔特徴点を OpenFace および提案手法で検出し，正解座標との誤差を root mean squre error (RMSE) で比較した．尚，データセットに含まれる顔の大きさは統一されていないため，顔矩形の高さ，幅の平均値が 220 になるよう正規化した．結果を表 2.1 に表す．

表 2.1 顔特徴点検出精度の比較

	OpenFace [63]	提案手法
RMSE[pixel]	9.2	7.3

なお，RMSE は回帰モデルの質を評価するために使用される統計量であり [64]，次の式によって算出される．

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y_j})^2} \tag{2.1}$$

ただし，モデルにより予測された値を y_j, 真値を $\hat{y_j}$ で表す．RMSE の値は真値からのばらつき具合を示し，68% のデータが真値から 1RMSE 以内に，95% のデータが真値から 2RMSE 以内に収まる．

2.3.2 定性的評価

図 2.13 に，OpenFace と提案手法それぞれについて成功例と失敗例をまとめた．ここでは，顔輪郭が大きく外れた場合を失敗とした．顔特徴点について，真値を緑色，OpenFace による結果を紫色，提案手法による 1 段目のカスケードの結果を赤色，2 段目のカスケードの結果を水色の点でそれぞれ表す．(a) 群は OpenFace，提案手法ともに顔検出に成功した例である．顔向きが正面に近く，また画像が鮮明なデータは両手法で成功する傾向にあった．(b) 群は OpenFace で成功したが，提案手法で失敗した例である．上段の例は，入力画像中に占める顔のスケールが大きく，今回の学習時に含まれない条件であったため失敗した．学習時の顔矩形切り出しスケールのバリエーションを増やすことで改善可能と思われる．下段の例では，画像左側の輪郭が外れていた．(c) 群は OpenFace で失敗し，提案

手法で成功した例である．顔向きが正面から大きく離れている場合，OpenFace は失敗する傾向にあった．対して提案手法では，顔向き変化に頑健であった．これはデータセットに大きな顔向きのバリエーションが含まれているためと考えられる．(d) 群はどちらの手法も失敗した例である．コントラストが非常に大く，画像が不鮮明など悪条件においては失敗する傾向にあった．テストデータ内で (b) 群の占める割合は 3.4% と小さく，一方 (c) 群は 14% と少なくない割合を占めている．提案手法により，従来では検出に失敗していた顔向き変化の大きいような条件において，提案手法では頑健に顔特徴点検出を実現し，視線推定が可能となった．

2.4　頭部姿勢推定

単眼カメラによって頭部姿勢変動に対応する視線推定のためには，頭部姿勢の推定が二つの理由から不可欠である．一つ目の理由は，カメラに対する人物の頭部位置は，求めたい視線の始点位置であるためである．頭部位置推定精度は最終的な注視点推定精度に影響するため，推定位置の精度検証を第 5 章にて扱う．二つ目の理由は，顔画像中の眼球中心位置推定のためである．前節で検出した目形状，つまりまぶたに対する眼球中心位置は，頭部姿勢により異なる．例えば，被験者が上を向いていれば，眼球中心位置はまぶたの位置に対して下側にシフトする．眼球中心位置推定に関しては 2.5 節にて説明する．

提案手法では，予め作成した 3 次元顔モデルを，検出された顔特徴点位置にフィッティングさせる事で頭部姿勢を求める．

2.4.1　3 次元顔モデル作成

予め撮影した 200 枚の顔画像を元に，Structure from Motion 法 [65] により 3 次元顔モデルを作成する．Structure from Motion 法は，ある対象物を様々な角度から撮影した画像群から，その対象物の 3 次元形状とカメラ姿勢を同時に復元する手法であり，その実践的な解法として Tomasi ら [66] の提案した因子分解法が使われる．

i 枚目の顔画像の j 番目の 2 次元顔特徴点の座標を $\boldsymbol{x}_j^i$ とし，i 番目の画像内のすべての顔特徴点の重心を $\bar{\boldsymbol{x}}^i$ とする．式 2.2 のように F フレーム分全てについて P 個の顔特徴点情報を $2F \times P$ 行列である計測行列 W に格納する．続いて，W を特異値分解し 2 つの行列に分解する．$2F \times 3$ 行列であるモーション行列 M が各画像の回転・並進を表す．また，$3 \times P$ 行列である形状行列 S が画像ごとに不変な共通因子であり，今回求めたい顔特徴点の 3 次元モデルである．図 2.14 に作成した 3 次元顔モデルを示す．

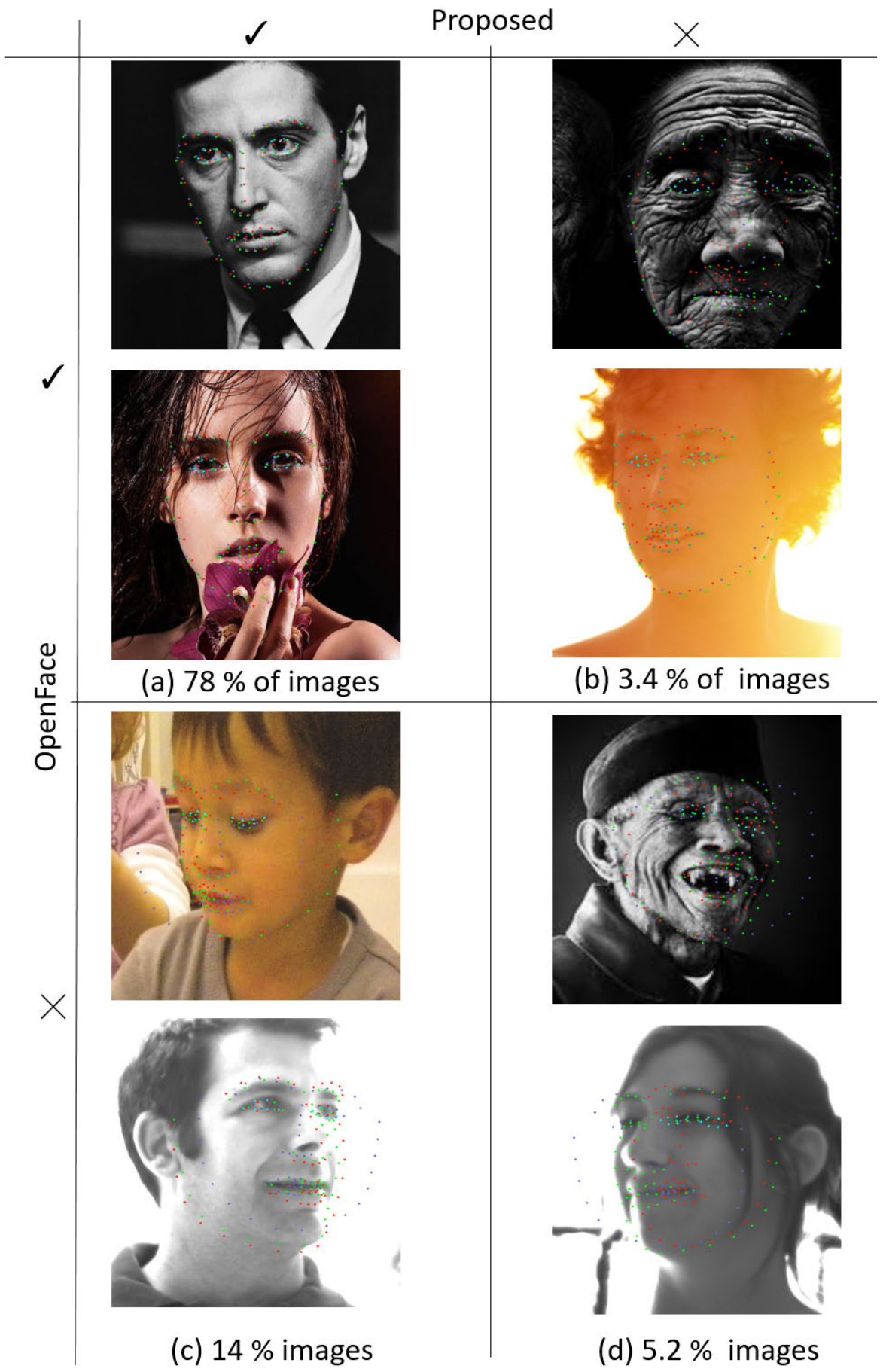


図 2.13　顔特徴点検出の従来手法と提案手法の定性的比較

2.4.2　ピンホールカメラモデルによる頭部姿勢推定

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^1 - \bar{\boldsymbol{x}}^1 & \cdots & \boldsymbol{x}_{\boldsymbol{P}}^1 - \bar{\boldsymbol{x}}^1 \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_1^{\boldsymbol{2F}} - \bar{\boldsymbol{x}}^{\boldsymbol{2F}} & \cdots & \boldsymbol{x}_{\boldsymbol{P}}^{\boldsymbol{2F}} - \bar{\boldsymbol{x}}^{\boldsymbol{2F}} \end{bmatrix} \tag{2.2}$$

$$\boldsymbol{W} = \boldsymbol{MS} \tag{2.3}$$

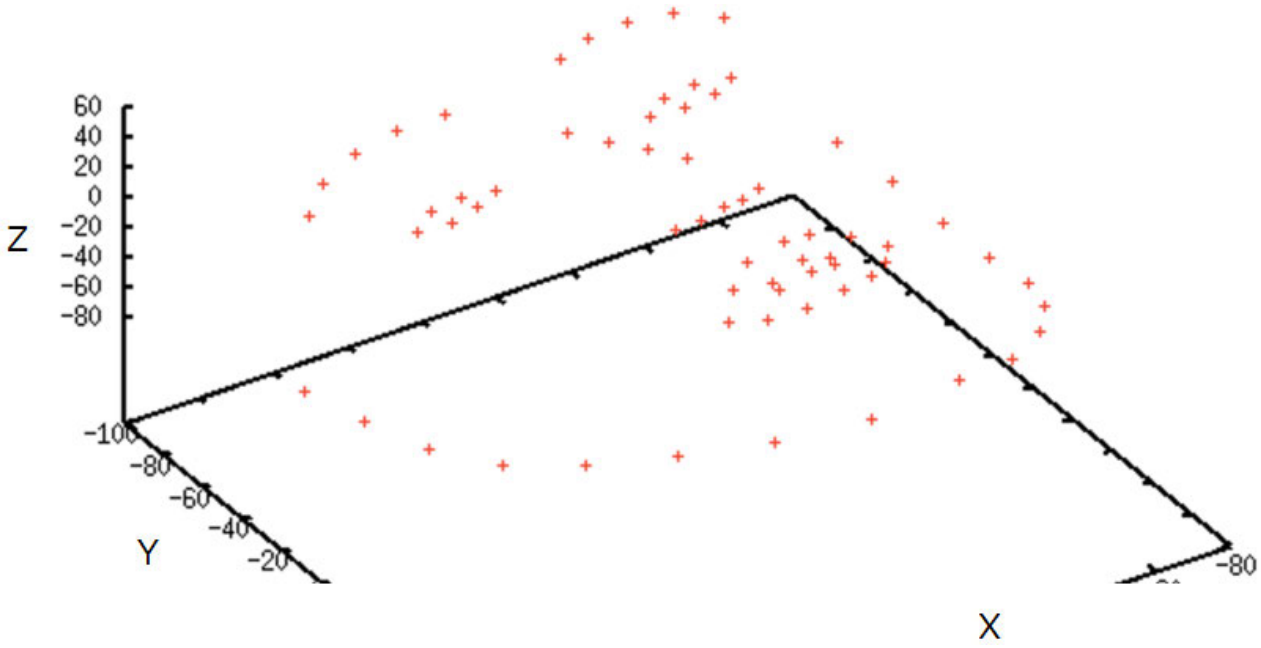


図 2.14　3 次元顔モデル

ここで作成した 3 次元顔モデルをもとに，現フレームに対して観測された顔特徴点座標から頭部の位置・方向を求める頭部姿勢推定を行う．頭部姿勢推定は，ピンホールカメラモデルを前提とする．ピンホールカメラモデルは式 2.4 で表され，3 次元物体が 2 次元画像上にどう写るかを定義するモデルである．3 次元顔モデルの座標 (X, Y, Z) を回転行列 R 及び並進行列 t で座標変換し，カメラ内部行列 A によって 2 次元画像上の点 (u, v) に写像する．ここで，f_x, f_y はそれぞれカメラの x 方向及び y 方向の焦点距離を表し，(c_x, c_y) はカメラの中心座標を表す．f_x, f_y, c_x, c_y はカメラに固有であり，予めチェスボードを用いたカメラキャリブレーションにより実測した．

ピンホールカメラモデルによって，作成された 3 次元顔モデル (X, Y, Z) から，そのフレームでの姿勢 (回転・並進) を表す $[R|t]$ とカメラ内部行列 A によって，画像上の 2 次元顔特徴点座標 (u, v) への関係が記述される．

ここで未知なのは姿勢 $[R|t]$ の 12 変数であるので，68 点の顔特徴点から行列変換により解ける．$\boldsymbol{t}$ はそのまま頭部位置ベクトルを表す．得られた回転行列 R をオイラー角に変換した後，頭部中心から顔の前方正面方向に向かうベクトルとして頭部方向 $\boldsymbol{r}$ を求める．

$$s \quad = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.4}$$

$$s \quad = \quad A\,[R|t]\,M \tag{2.5}$$

以上により，人物の頭部位置 $\boldsymbol{t}$ および方向 $\boldsymbol{r}$ が得られた．

2.5　眼球中心位置推定

モデルベース視線推定では，視線は眼球中心位置と虹彩中心位置を結ぶベクトルと定義される．このうち眼球中心位置は画像中から直接観測できないため，顔特徴点位置および頭部姿勢から推定しなければならない．

左右それぞれの目について，3 次元世界座標系において，図 2.15 のように，目尻と目頭の中点から頭部方向 $\boldsymbol{r}$ と逆方向に（顔の内部方向に）12　　mm 移動した点が世界座標系の眼球中心位置であり，それを画像内に投影する事で画像中の眼球中心位置 $\boldsymbol{e}_l$,$\boldsymbol{e}_r$ を得る．

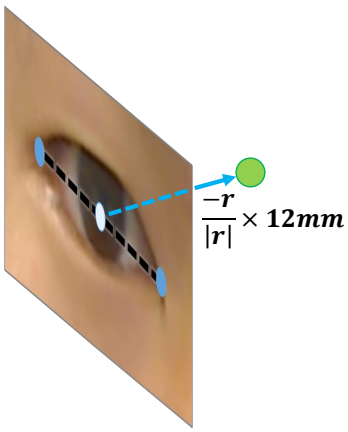


図 2.15　眼球中心位置推定

2.6　本章のまとめ

本章では，視線推定のための前処理として画像から顔特徴点検出および頭部姿勢を推定する手法を説明した．顔特徴点検出では，CNN により学習した回帰モデルによって，照

度変化や遮蔽の激しい環境に頑健な検出手法を提案した．視線推定という本論文の目的のため，目の形状について更に詳細な改善を行うため目領域のみを学習したネットワークを接続し，２階層のカスケード型検出器を作成した．データセットによる検証を通して従来の最先端の手法を超える精度を確認した．さらに，ピンホールカメラモデルを前提に，単眼カメラから頭部姿勢 (位置 $\boldsymbol{t}$, 方向 $\boldsymbol{r}$) を推定した．最後に，頭部回転角を元に画像中の眼球中心位置 $\boldsymbol{e}_l, \boldsymbol{e}_r$ を推定した．これらの結果は第３章，第４章の入力として使われる．

第 3 章

虹彩追跡手法

本章では，第 2 章で得られた眼球中心位置 $\boldsymbol{e}_l$, $\boldsymbol{e}_r$ を元に画像中から高精度に虹彩位置を追跡する手法について説明する．提案手法ではモデルベースの手法を取り，虹彩位置を 3.1 節で記述する 3 次元眼球モデルを元に探索する．従来手法においては低解像度環境における追跡精度が課題であったが，提案手法では，虹彩エッジの強度および勾配方向に着目した尤度評価と密なサンプリングにより低解像度でも頑健な虹彩追跡を実現する．虹彩追跡は，3.2 節で説明する計算量の削減のためのテンプレートマッチングによる初期探索と，3.3 節で説明する Particle Filter による高精度探索からなる 2 ステップで構成される．また，第 2 章で得られた瞼形状から，瞼エッジを除去するフィルタについて説明する．

3.1　3 次元眼球モデル

本節では，探索に用いられる 3 次元眼球モデルについて説明する．提案手法では，低い解像度においても精度を担保するため，虹彩の持つエッジを出来る限り捉え，それらに最も適合する虹彩楕円を探索する事で，サブサンプリングの精度を目指す．言い換えれば，画像中にアピアランスとして現れる特定の顕著な特徴を捉えるのではなく，虹彩エッジ全体を使い虹彩中心点を推定するため，データの持つ解像度以上の粒度で虹彩位置を推定する．そのために，提案手法ではアピアランスベースではなくモデルベースの手法を採用する．

我々が作成した 3 次元眼球モデルを図 3.1 に表す．

この眼球モデルは，人間の眼球を球体と仮定し，虹彩を円盤に見立て，眼球回転角に従っ

て虹彩が半径 R の眼球の上をスライドする．モデルは隠れ変数に眼球回転角 $yaw, pitch$ および虹彩半径 d を持つ．R は顔のスケールから決定される定数である．ここで，虹彩半径 d は本来，解剖学的には個人差が少ないため定数とする項である．しかし，この眼球モデルのスケールは顔の大きさによって決定されるため，例えば顔が大きい人は本来の目よりも大きく，逆に顔が小さい人は本来の目よりも小さく眼球モデルが正規化される．もし，d の値が実際の虹彩半径と大きく異なっていた場合，探索の過程で複数の局所解が発生し，結果が安定しない．よって提案手法では，顔の大きさの個人差を吸収する意味で d を変数とし，実際の画像への安定的な適合を優先させる．尚，R については，1 枚の入力画像のみでは最適化が不可能のため定数とする．R のずれに起因する誤差は第 4 章の注視点推定時に吸収する．

虹彩探索のため，この眼球モデルを第 2 章で得た画像中の眼球中心位置 $\boldsymbol{e}_l$，　$\boldsymbol{e}_r$ に配置する．以後簡便のため，左目 $\boldsymbol{e}_l$ のみについて説明する．

画像上の虹彩探索は，目尻-目頭間の距離を a とすると，縦 a 横 $2a$ の大きさで $\boldsymbol{e}_l$ が中心となるよう切り抜いた画像内で実行される．これを目領域画像と呼ぶ．目領域画像は 100×200 にリサイズされ，座標軸を (u, t) と定義する．目領域画像内の眼球中心位置 (e_{lu}, e_{lt}) は常に $(100, 50)$ で一定となる．

眼球回転角が $yaw, pitch$, 虹彩半径が d のとき，目領域画像での虹彩楕円形状は式 3.1 に従って決定される．尚，カメラに対して目の大きさは十分小さいと仮定し，空間を弱透視投影モデルとして捉える．そのため，眼球から目領域画像への変換は正射影として計算する．

$$
\begin{aligned}
c_u &= -R \times \sin(yaw) \times \cos(pitch) + e_{lu} \\
c_t &= -R \times \sin(pitch) + e_{lt} \\
h &= d \\
w &= |d \times \cos(yaw) \times \cos(pitch)| \\
\theta &= \arctan\left(\frac{\sin(pitch)}{\cos(pitch) \times \sin(yaw)}\right)
\end{aligned}
\tag{3.1}
$$

ここで，(c_u, c_t) は虹彩楕円の中心，h, w は楕円の長軸，短軸の半径，θ は楕円の反時計回りの回転角を表す．

3.2　初期探索

眼球モデルを元に $yaw, pitch$ の取りうる全範囲において虹彩探索をする事は，計算コストの観点から冗長である．計算コストの削減のため，まず目領域画像内でテンプレート

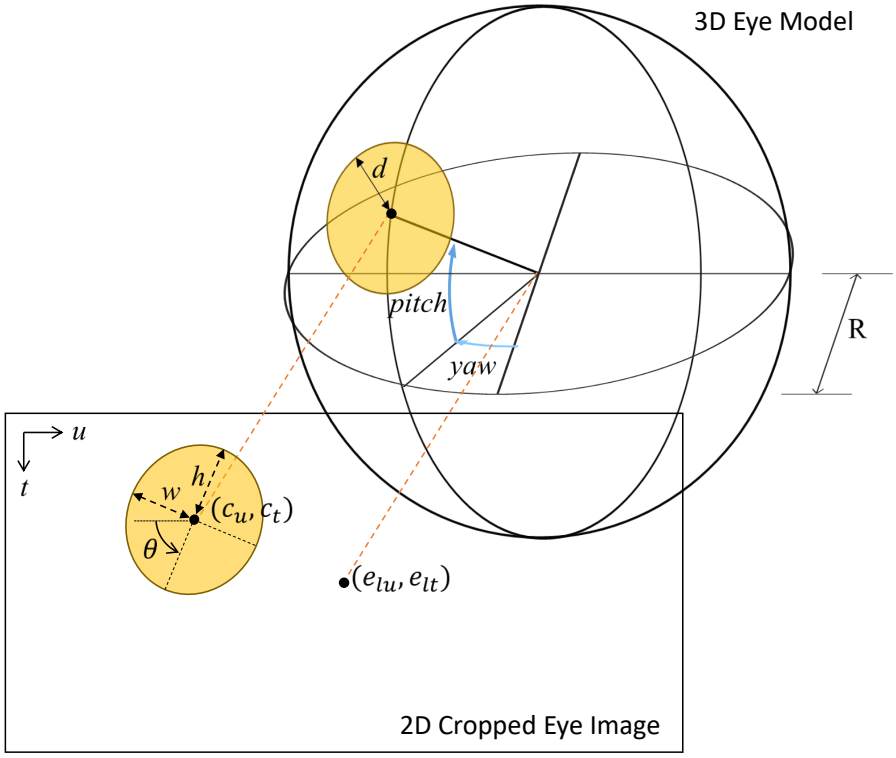


図 3.1　3 次元眼球モデル

マッチングにより高速かつラフな虹彩位置の絞り込みをする．目領域画像をグレイスケールに変換し (図 3.2)，その中で虹彩を模したテンプレート画像とマッチングを行う．テンプレート画像は，目領域画像の高さ a の 0.55 倍を直径とする黒円盤画像とする．なお，0.55 という数値は実験的に定めた．テンプレート画像を図 3.3 に示す．

室内環境においても逆光や外光の影響により輝度値は安定しない．画像の明るさ変動を吸収するため，式 3.2 で表す normalized cross correlation method を指標としテンプレートマッチングを行う．ピーク位置を目領域画像上の大まかな虹彩位置 (u_{base}, t_{base}) とする．

$$S_{NCC}(d_x, d_y) = \frac{\sum\sum g(d_x + i, d_y + j) f(i, j)}{\sqrt{\sum\sum (g(d_x + i, d_y + j))^2} \sqrt{\sum\sum f(i, j)^2}} \tag{3.2}$$

3.3　高精度探索

続いて，テンプレートマッチングで得られた大まかな虹彩位置 (u_{base}, t_{base}) を元に，Particle Filter を使用した高精度な探索を行う手法について説明する．まず，Particle Filter の概要を示し，続いて虹彩追跡の詳細な流れを説明する．

図 3.2　目領域グレイスケール画像

図 3.3　黒円盤テンプレート画像

3.3.1　Particle Filter

Particle Filter [67] は事後確率分布をランダムサンプリングによるモンテカルロ近似によって推定することで，高次元の状態空間に対して効率よく状態推定を行う手法であり，パラメトリックな状態ベクトルで表現可能な任意のモデルに適応可能である．観測モデルの尤度計算方法も自由に設計可能なため様々な分野に応用されている．以下では，Particle Filter の基本的な考え方を説明した後に，それを応用した虹彩追跡手法について説明する．

Particle Filter の基本的な考え方は，事前確率分布や事後確率分布を図 3.4 のように，これらの分布に従って生成した多数のサンプル（Particle）の密度によって近似するというものである．アルゴリズムのポイントは 2 種類のサンプル集合を事前分布と事後分布から生成することである．処理手順は図 3.5 のようになっている．

3.3.2　エッジベース虹彩追跡

$$pitch_{base} = \arcsin\left(\frac{-(t_{base} - c_{lt})}{R}\right), \tag{3.3}$$

$$yaw_{base} = \arcsin\left(\frac{-(u_{base} - c_{lu})}{R\cos(\theta)}\right). \tag{3.4}$$

Particle Filter を虹彩検出に応用する．3.2 節で得られた大まかな虹彩位置 (u_{base}, t_{base}) から，式 3.1 を変形した式 3.3,3.4 により大まかな眼球回転角 $(yaw_{base}, pitch_{base})$ を

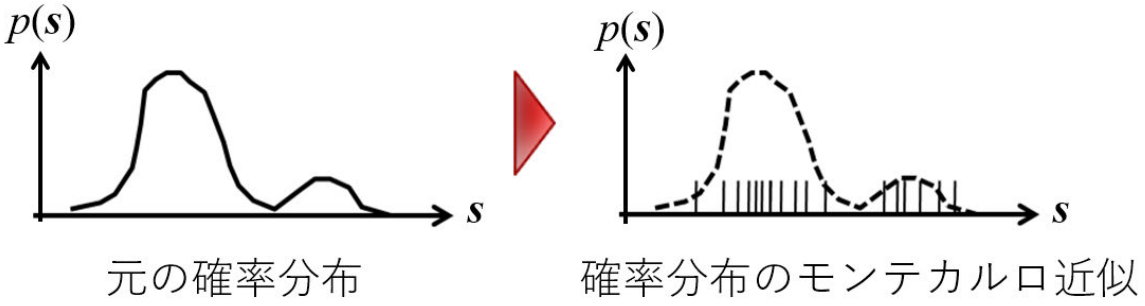


図 3.4　サンプル集合による分布の近似

算出する．また，d も既定値として 25 を定める．Particle Filter で推定する状態は，$(yaw_{base}, pitch_{base})$ を基準とした実際の眼球回転角との差 $(yaw_{diff}, pitch_{diff})$ と虹彩半径の既定値との差 d_{diff} とする．

システムモデル

システムモデル (状態遷移関数) には，対象の動きが複雑な動きに用いるランダムウォークと対象の動きが速い場合に用いる等速直線運動があるが，眼球はサッカードと呼ばれる非常に早い動きや，眼振と呼ばれる不随意な振動など複雑かつ規則性が無い．これらの動きは一般的な単眼カメラのフレームレートである 30fps で捉える事ができない．そのため，各フレームの動きに連続性は期待できない．そのため，ここではシステムモデルはランダムウォークとする．

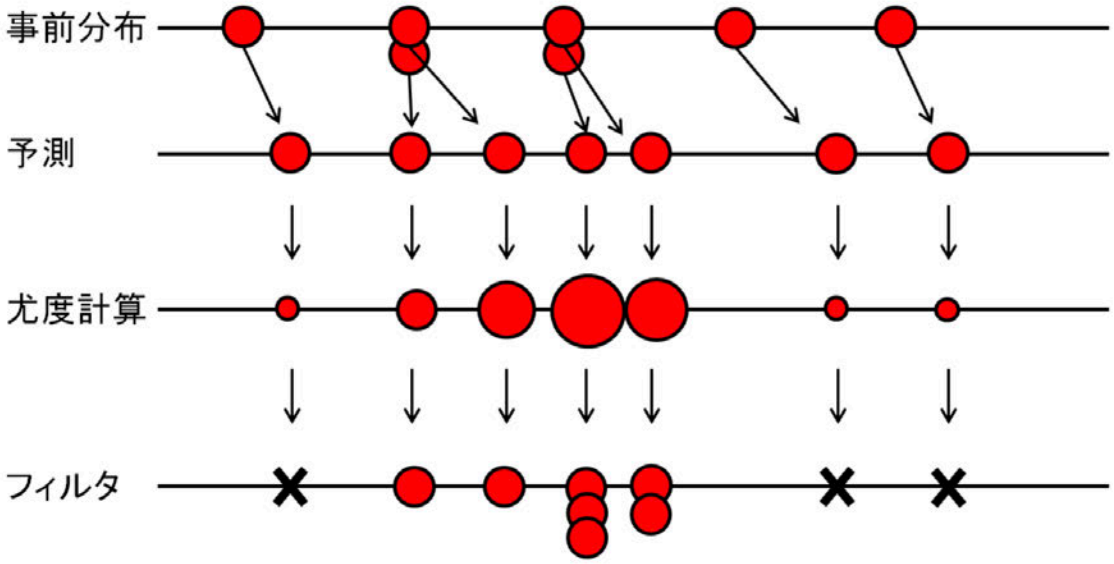
$$\mathbf{s}_t = (yaw_t, pitch_t, d_t)^{\mathbf{T}} \tag{3.5}$$

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.6}$$

$$\mathbf{s}_t^{(i)} = \mathbf{F}\mathbf{s}_{t-1} + n_t^{(i)} \tag{3.7}$$

ここで，$\mathbf{s}_t$ は状態ベクトル，yaw_t は眼球の yaw 回転，$pitch_t$ は眼球の pitch 回転，d_t は虹彩の半径，$n_t^{(i)}$ は正規分布に従ってランダムなノイズを加えることを意味している．実験的に，付加するランダムノイズの範囲を，$n_t^{(i)} = (\pm 5, \pm 5, \pm 2)$ と決定した．上式より，テンプレートマッチングで得られた $(yaw_{base}, pitch_{base})$ の近傍に Particle を散布する事で予測サンプル $\mathbf{s}_t^{(i)}$ を生成することができる．Particle 数 I は 200 とした．

また，システム行列 $\mathbf{F}$ を変えることで様々な線形運動モデルや等加速度運動を仮定したモデルに拡張することができる．例えば 200fps や 500fps などフレームレートの非常に早いカメラを使用する際は，ランダムウォークではなく等速直線運動をシステムモデルと



Step0　初期化

・あらかじめ与えられた初期分部に従って Particle を生成

Step1　予測

・事前分布から自分で設定したシステムモデルに従って Particle を移動

・さらにシステムノイズによるばらつきを与えて最終的な予測サンプルを生成

・これにより対象の不規則な動きにも対応できる

Step2　尤度計算

・各 Particle の重みを観測モデルに従って推定

・この重みが大きい程、その状態である確率が高いことを示す

Ste3　フィルタ

・重みに比例する割合となるよう Particle を復元抽出

・このとき、Particle の数が一定になるようにする

・全ての Particle の重み付き平均によって、期待値として現在の状態の推定値を得る

・この事後分布を事前分布として Step1 の処理へ戻る

図 3.5　Particle Filter の処理の流れ

した方がよい．そのときのシステムモデルは式 3.8,3.9,3.10 になる．

$$\mathbf{s}_t = (yaw_t, pitch_t, \dot{yaw}_t, \dot{pitch}_t, d_t)^{\mathbf{T}} \tag{3.8}$$

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.9}$$

$$\mathbf{s}_t^{(i)} = \mathbf{F}\mathbf{x}_{t-1} + n_t^{(i)} \tag{3.10}$$

尤度評価

尤度計算では，各予測サンプル $\mathbf{s}_t^{(i)}$ を尤度によって重み付ける．各 Particle $\mathbf{s}_t^{(i)} = (yaw_t^{(i)}, pitch_t^{(i)}, d_t^{(i)})^{\mathbf{T}}$ は眼球回転角の隠れ状態であるが，式 3.1 により隠れ状態 $\mathbf{s}_t^{(i)}$ を観測状態 $\mathbf{o}_t^{(i)}$ に変換する．$\mathbf{o}_t^{(i)} = (c_u^{(i)}, c_t^{(i)}, h^{(i)}, w^{(i)}, \theta^{(i)})_t$ は目領域画像中の楕円である．ここから，簡便のためフレーム番号 $_t$ を省略する．各 $\mathbf{o}^{(i)}$ について，尤度

各 $\boldsymbol{o}^i$ について，尤度 $L(\boldsymbol{o}^i)$ を式 3.11 により計算する．

$$\begin{aligned}
L(\boldsymbol{o}^i) &= \frac{1}{|\boldsymbol{o}^i|_0} \sum_{\boldsymbol{p}^j \in \boldsymbol{O}^i} m(\boldsymbol{p}^j) \times g(\boldsymbol{p}^j) \times w(\boldsymbol{p}^j) \\
m(\boldsymbol{p}^j) &= \sqrt{dx(\boldsymbol{p}^j)^2 + dy(\boldsymbol{p}^j)^2} \\
g(\boldsymbol{p}^j) &= \frac{1}{(grad_{im}(\boldsymbol{p}^j) - grad_{md}(\boldsymbol{p}^j))^2 + 1} \\
grad_{im}(\boldsymbol{p}^j) &= \arctan\left(\frac{dt(\boldsymbol{p}^j)}{du(\boldsymbol{p}^j)}\right) \\
grad_{md}(\boldsymbol{p}^j) &= \arctan\left(\frac{t(\boldsymbol{p}^{j+1}) - t(\boldsymbol{p}^j)}{u(\boldsymbol{p}^{j+1}) - u(\boldsymbol{p}^j)}\right) \\
q(\boldsymbol{p}^j) &= \begin{cases} 1 & \left(\begin{array}{l} |grad_{im}(\boldsymbol{p}^j)| \leq 45^\circ \text{ or} \\ |grad_{im}(\boldsymbol{p}^j) - 180| \leq 45^\circ \end{array} \right) \\ 0.01 & (otherwise) \end{cases}
\end{aligned} \tag{3.11}$$

ここで, $\boldsymbol{p}^j$ は楕円弧上に均等に散布した 120 点のうち j 番目のエッジ点を表す．$du(\boldsymbol{p}^j)$ および $dt(\boldsymbol{p}^j)$ はそれぞれ点 $\boldsymbol{p}^j$ における u 方向および t 方向の画像輝度の微分値である．$u(\boldsymbol{p}^j)$ と $t(\boldsymbol{p}^j)$ は点 $\boldsymbol{p}^j$ の u,t 座標を表す．

$m(\boldsymbol{p}^j)$ は $\boldsymbol{p}^j$ でのエッジ強度であり，$g(\boldsymbol{p}^j)$ は画像と楕円サンプル形状のエッジ勾配類似度である．$q(\boldsymbol{p}^j)$ によって, もしエッジ勾配方向が水平から 45° 以内なら 1 をかけ, ほかの場合は 0.01 をかける．なぜならエッジの勾配方向が 45° から 135° のもの，および

225° から 315° のものは虹彩エッジに属さないと考えられるため，$q(\boldsymbol{p}^j)$ によってこれらのエッジを除外する. つまり, 画像中の楕円上に多くの強いエッジを持ち，かつモデルと近い勾配方向をもつとき，$\boldsymbol{o}^i$ の尤度は大きくなる．

最後に，最も尤度が高いと推定される状態 $\boldsymbol{s}^* = (yaw^*, pitch^*, d^*)$ が，全ての Particle について重み付き平均を取ることで得られる. 続いて視線ベクトル $\boldsymbol{g}$ は式 3.12 のように得られる．以上のように，各目の視線ベクトル $\boldsymbol{g}_l$,$\boldsymbol{g}_r$ が得られた.

一連の処理の流れを図 3.6 にまとめた．なお，図 3.7 のように，モデルの勾配方向を hue で，尤度を明度で表現している．

$\boldsymbol{p}^j$ はエッジ上の点として表現されており, 色（hue）は エッジ勾配方向をあらわし明度は $m(\boldsymbol{p}^j) \times g(\boldsymbol{p}^j) \times q(\boldsymbol{p}^j)$ の値を表す.

$$\boldsymbol{g} = (g_x, g_y, g_z) = (1\times\cos(pitch^*)\times\sin(yaw^*), 1\times\sin(pitch^*), -1\times\cos(pitch^*)\times\cos(yaw^*) \tag{3.12}$$

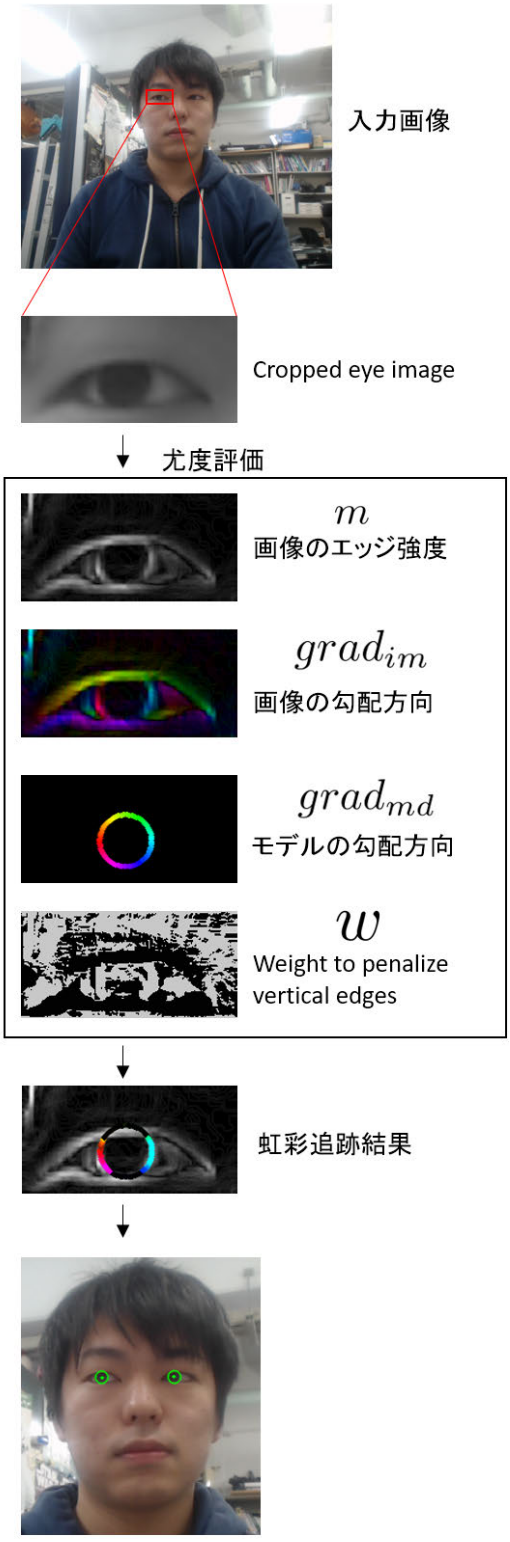


図 3.6　虹彩追跡の処理の流れ

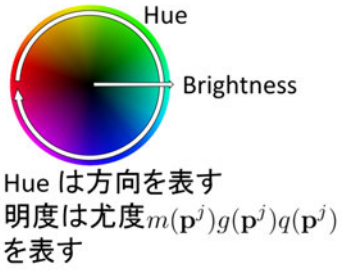


図 3.7　HSV 色空間による勾配および尤度の表現

3.4　瞼形状フィルタ

本節では，第 2 章で推定した目形状を元に，瞼エッジを除去し，虹彩追跡の精度を向上させる試みについて説明する．まず，第 2 章で得られた目形状は，片目あたり 6 点のみである．瞼の曲線を再現するため，B-spline 曲線に近似によりなめらかな瞼形状を得る．尚，目尻，目頭点は滑らかである必要はないため，曲線近似はしない．そのため，上瞼，下瞼に分離し，それぞれで近似曲線を得たあと統合し，瞼形状とする．

続いて，瞼形状の内部には 1.0, 外部には 0.3 を瞼形状フィルタとして 3.3 節の尤度評価の項にかける．処理結果を可視化したものが，図 3.8 である．上段のグレイスケール画像にて，赤線で瞼近似曲線，緑線で検出された虹彩を表す．下段のエッジ画像では，第 2 章で得られた 6 点の目形状点を赤点で表している．また，瞼形状の外部のエッジ強度は 0.3 倍されているので暗く見える．虹彩エッジ上の点群について，図 3.7 と同じように，その色はエッジ勾配方向を，明度はエッジ強度を表す．(b),(c),(e) など多くの例において，虹彩追跡において外乱となる瞼エッジが除去されている事が確認できるが，(a),(j) など検出された瞼形状が実際よりも小さいため，本来は虹彩にあるエッジが除去されてしまった例もある．また，(l) の目頭部のように実際の瞼エッジが検出された瞼形状内部に残っており，虹彩エッジと混同し強い尤度を持ってしまった例もある．瞼形状フィルタの視線推定精度への影響については第 5 章で議論する．

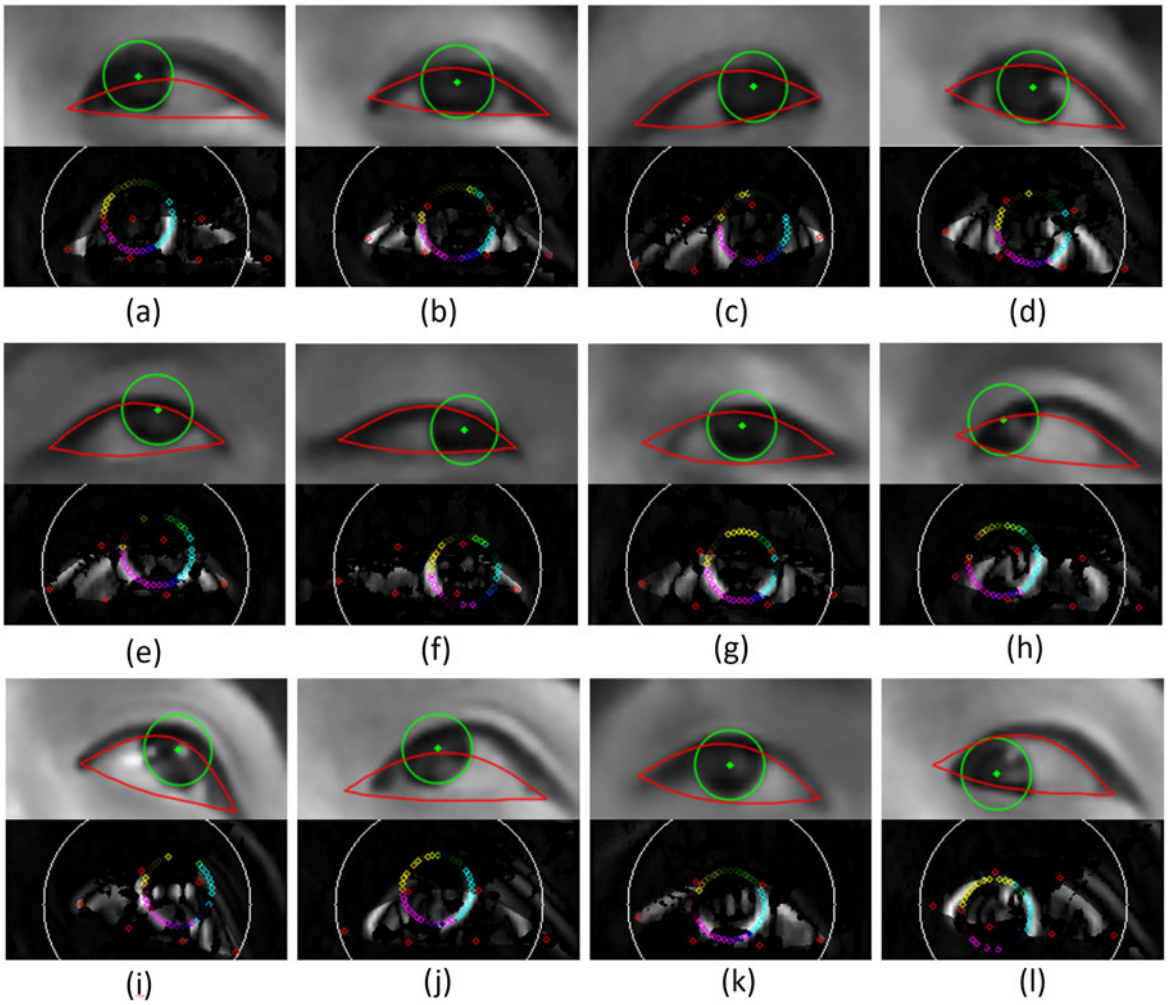


図 3.8　目のエッジ画像とモデルの例

3.5　本章のまとめ

本章では，眼球中心位置 $\boldsymbol{e}_l$,$\boldsymbol{e}_r$ を元に高精度に虹彩位置を追跡する手法について説明した．提案手法では，３次元眼球モデルを元にテンプレートマッチングによる初期探索と，Particle Filter による高精度探索からなる２ステップで構成される探索手法である．虹彩エッジの強度および勾配方向について眼球モデルと比較し，類似度が高いものほど高い尤度を持つように設計し，また勾配方向が水平に近い場合は尤度を限りなく小さくするようにした．この処理によりまぶたのエッジを除去しつつ適切に虹彩のエッジを捉える．そして，これらの Particle の密なサンプリングを行うことで，低解像度でも頑健な虹彩追跡を

実現した．また，第 2 章で得られた瞼形状から瞼エッジを除去するフィルタについて説明した．

第 4 章

注視点推定

本章では，広範囲な頭部位置における注視点推定手法について説明する．個人キャリブレーションフリーかつ頭部位置非拘束を実現するため，そして検証のため，我々は独自のデータセットを作成した．そのデータセットの特性と追跡のための難しさについて触れ，学習による注視点推定手法について解説する．

4.1 Room Scale Gaze Dataset (RSGD)

4.1.1 データセット作成のねらい

我々は，キャリブレーションフリーかつ広い空間内での非拘束視線推定を実現するため，また検証のため Room Scale Gaze Dataset を作成した．

モデルベース視線推定では，ユーザの視線ベクトルや注視点は幾何的に求められる．よってデータセットによる学習等をすることなく高い精度を実現可能であることがモデルベース視線推定手法の強みである．しかしながら，モデルフィッティングで前提とされる眼球中心位置や顔のサイズ，頭部姿勢といったパラメータ，顔特徴点のアラインメントにはそれぞれ誤差を含んでいるため，単純に視線を幾何的に求めるのみでは，顕著な誤差を持つ．これは 4.3 節で検証する．つまり，モデルベース手法は耐環境性が弱みである．

一方アピアランスベース視線推定では，データセットをもとに注視点を推定する．もしデータセットに多様な頭部位置が含まれ，かつ偏りなく膨大なデータ数を取得できれば，環境変化や個人差に頑健な視線推定を実現できるだろう．関連研究の節でも触れた通り，アピアランスベース手法の強みは耐環境性である．

ただし，このように膨大な数のデータを取得することは容易ではない．顔特徴点追跡の

分野においては，研究が非常に盛んであり，研究者や研究機関の母数が多いため，様々な良質なデータセットが多数公開されている．最近のトレンドは in-the-wild つまり環境光変化の激しい条件における頑健性である．本論文でもそれを利用し第 3 章でアピアランスベースの顔特徴点追跡手法を扱い，耐環境性の高い顔特徴点追跡手法を実現した．しかしながら視線推定用に公開されているデータセットは現状では限られている．そして，それらの既存のデータセットは，頭部位置を固定，もしくは，非拘束であってもカメラの近傍で取得されたものであり，遠く離れた位置における検証が十分であるとは言えない．

公開されている視線データセット

例えば，Mora ら [68] の提案した EYEDIAP データセットでは, 図 4.1 のようにデスクユーザを対象に取得された．注視の位置は 3 次元空間内に浮かぶ球であり，注視位置は連続値であるものの，頭部位置はカメラに対して固定されている．Arantxa ら [69] の手法で使われたデータセットも頭部位置が固定であり，また視線ターゲット数が 12 に限られている．

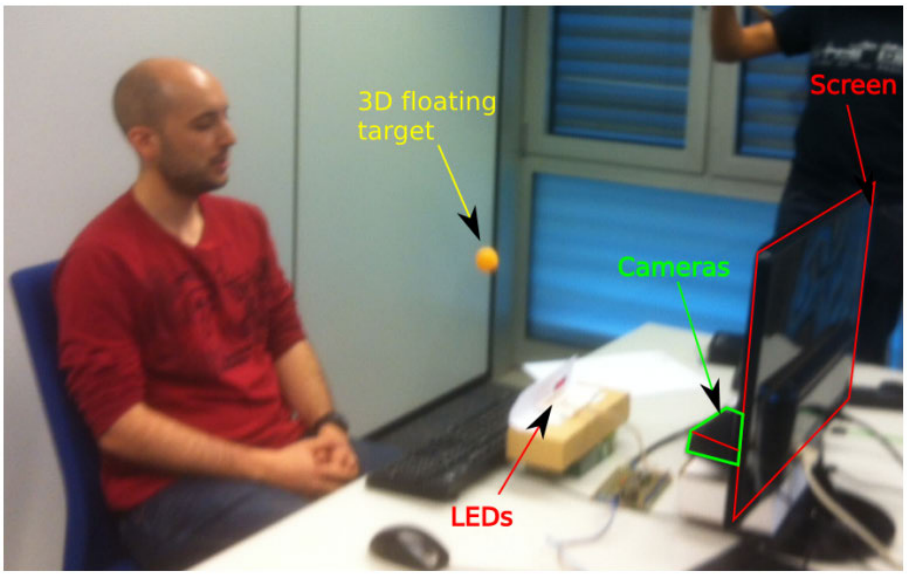


図 4.1　EYEDIAP データセット取得風景 [68]

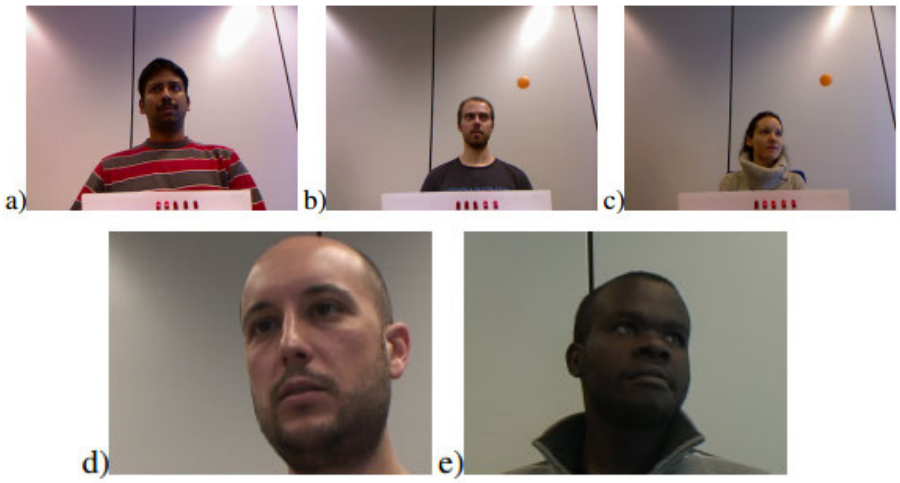


図 4.2 EYEDIAP データセット [68] 論文より引用
a-c は RGB-D カメラで，d-e は RGB カメラで取得された画像である

また，Zhang ら [20] の公開している MPII gaze データセットは，ノート PC を使用するユーザの顔写真を長期間に渡って取得し，様々な環境の中で取得された．図 4.4 にその例を示す．このデータセットは wild 環境での検証を目的としている．しかし，ユーザの顔位置はカメラの近傍に限られている．

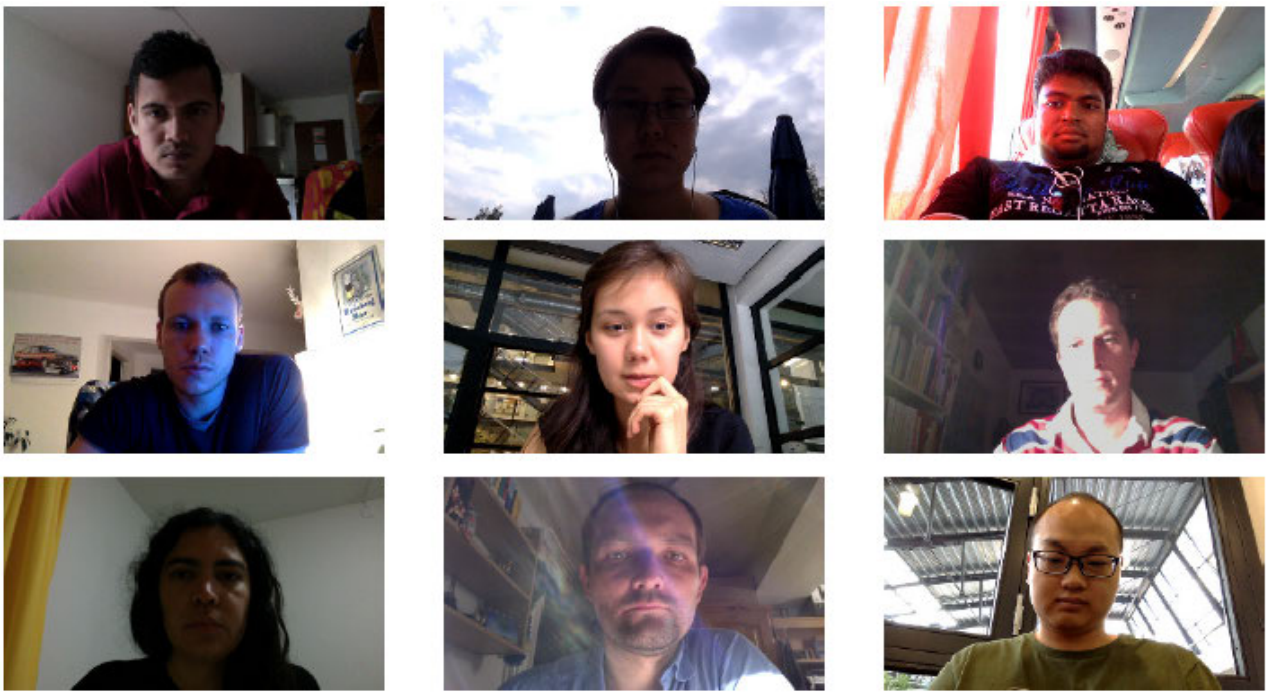


図 4.3 MPII gaze dataset 1 [20] 論文より引用

図 4.4　MPII gaze dataset 2 [20] 論文より引用

図 4.5　MPII gaze dataset に含まれる実際の画像 [20]

また，Smith らのデータセットは図 4.6 のように頭部を器具で固定している．

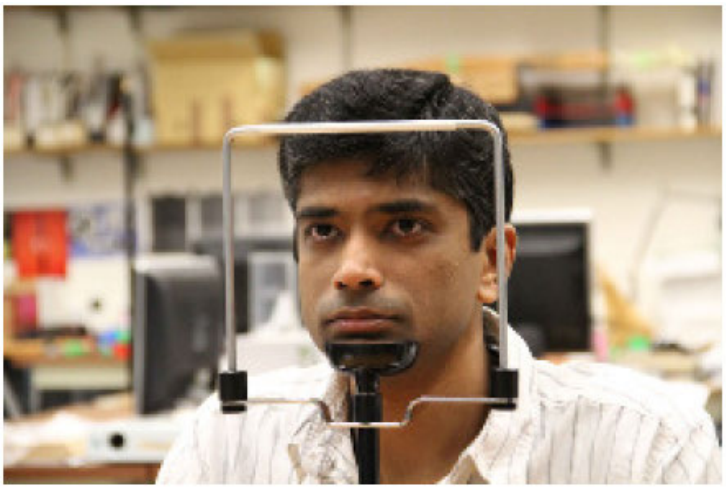


図 4.6 Smith ら [70] のデータ取得風景 論文より引用

以上で見られるように，これまで様々な目的の元視線推定用のデータセットが作成されてきたが，従来のデータセットは頭部位置のバリエーションが限られている．しかし，社会における視線推定技術の応用を考えた場合，より広範囲で使用可能である事が望まれる．

4.1.2 データセット作成

我々は，TV 視聴者やデジタルサイネージの利用者を想定し，カメラに対して広いアングル，頭部姿勢変化，位置変化，表情変化を含む被験者がディスプレイ上のマーカを注視した際の画像を集めた．このデータセットを Room Scale Gaze Dataset (RSGD) と呼称する．RSGD を用いて，注視点推定の学習と評価を行う．

我々はデータ取得のため，図 4.7 に示した縦 1.2m ×横 0.8m の 50 インチディスプレイと，Kinect v2 を用いた． Kinect v2 は図 4.8 に示すように，ディスプレイ下部中央に設置し，ディスプレイを注視する被験者の顔を撮影する．Kinect v2 は解像度 1920x1080 の RGB カメラと Depth カメラから構成される．今回 Depth 値は，RGB 画像からの 3 次元顔位置推定の妥当性検証のためだけに用い，視線推定学習用のデータには含まない．

データ取得中はディスプレイ上に，半径 3.0 センチの円形マーカを表示する．このマーカは 1 秒間の移動と 3 秒間の静止を繰り返し， 3 秒間の静止中，マーカに縮小・拡大のアニメーションを施し，被験者に注視すべきタイミングを伝える．このアニメーション中に被験者はマーカ中心部を注視し，数枚の画像を保存する．マーカ移動中は写真撮影を行わない．マーカ出現位置はディスプレイ上で一様分布に従いランダムに決定されるが，被験

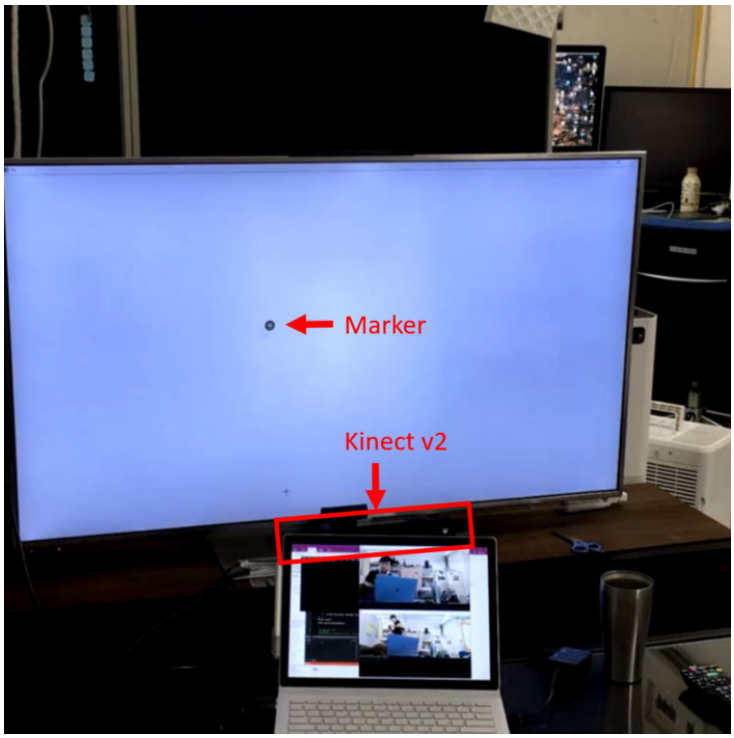


図 4.7　マーカ表示に用いるディスプレイ

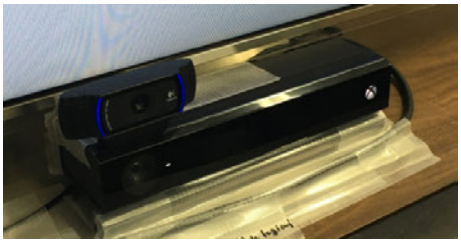


図 4.8　ディスプレイの下部中央に設置された Kinect v2

者の目の負担を考慮し，マーカ移動距離が大きくなりすぎないように，ディスプレイを 6 領域に分割し，左上→中央上→右上→左下→中央下→右下の順番で出現するようにした．

被験者はテレビの前の好きな位置に座り，データ取得をスタートした．手元のコントローラで自由に一時停止が可能であり，10 分に一回の小休止をはさみつつ一時間のデータ取得を行い，被験者一人あたり約 3000 枚の画像を取得した．また，60 枚に 1 回自動的に一時停止し，被験者の座る位置や顔姿勢を変えてもらい，データセット全体で幅広い顔位置，顔姿勢が含まれるようにした．実際の被験者画像群を図 4.9 に示す．

4.1.3　データセットの説明

このセクションでは，RSGD の特性について説明する．

図 4.9 RSGD の画像例

広範囲な頭部位置

従来のデータセットでは，ノート PC やタブレットを対象としたものであり，顔位置変化は限られていた．注視点推定において，顔位置変化は精度に影響を与える要因であるため，幅広い位置において検証が必要である．本データセットでは，カメラに対する顔位置のダイナミクスは図 4.10 にしめすように，奥行方向では TV の目の前から 2.5 メートルの距離まで，左右方向は 1 mから 1.5m の幅に分布している．この分布は今回使用したカメラで捉えられる画角全体に相当しており，一般的な TV 視聴者の利用者の範囲をカバーしている．

また，顔向きの分布は図 4.10 に示ように，水平方向で 40 度，垂直方向で 30 度となっている．

目領域解像度

一般に，目領域解像度は精度に大きな影響を与える．特にモデルベース視線推定手法では，高い目領域解像度を要求するものが多いと言われており，低解像度での精度評価は重要である．

RSGD に含まれる頭部位置は従来のデータセットよりカメラから遠く，広い範囲に分布

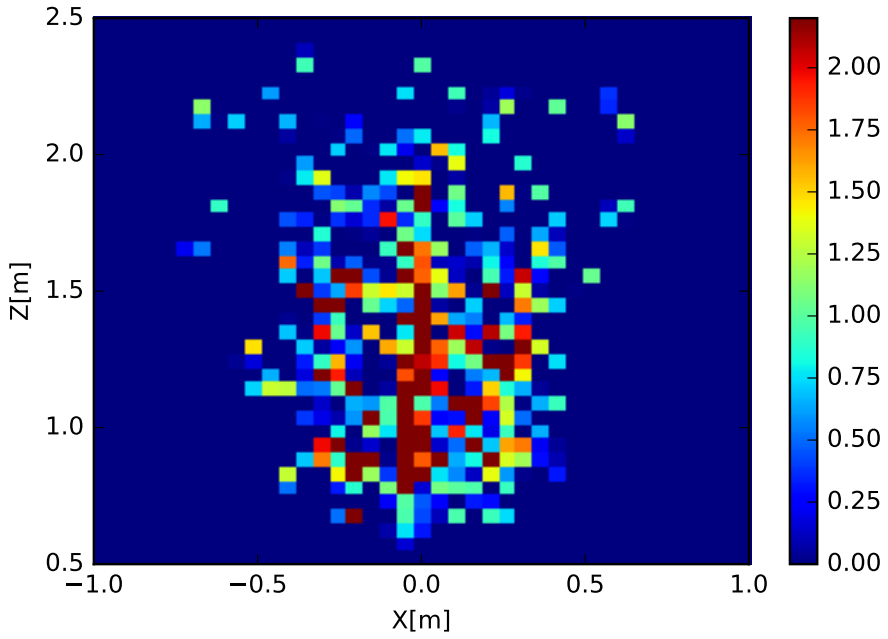


図 4.10　頭部位置分布

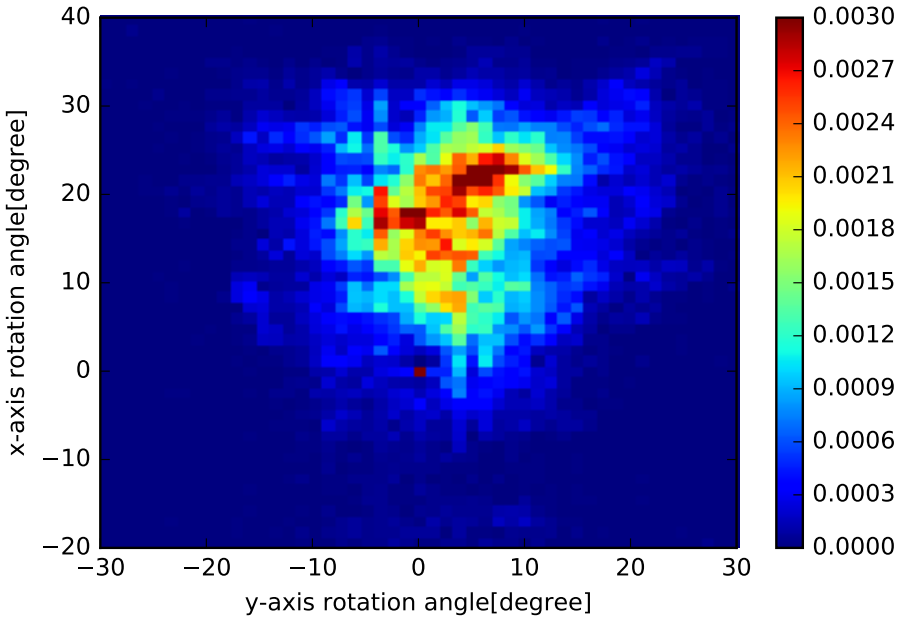


図 4.11　頭部方向分布

するが，目領域の解像度とカメラからの距離は別の問題である．使用しているカメラの解像度が同じ場合，カメラからの距離に伴い目領域解像度は低下するが，カメラの解像度が高ければ遠くの人物の被験者の目領域解像度も高い．RSGD で用いた Kinect v2 の RGB センサは HD(1920x1080) 解像度である．一方，従来のデータセットにおけるカメラ解像度は VGA(640x480) もしくは HD である．近年のスマートフォンの外向きのカメラには，

4K 解像度を超える素子が採用されているものの，ユーザ自身を撮影するための内向きのカメラとしては HD 解像度は高い部類に入る．そこで，RSGD と頭部姿勢変動を許容する従来のデータセット [20] の目領域解像度を比較した．各データセットにおける目領域解像度 (Eye resolution) の分布を図 4.12 に記す．ただし，ここでは目領域解像度を目尻と目頭の距離 [pixel] と定義する．

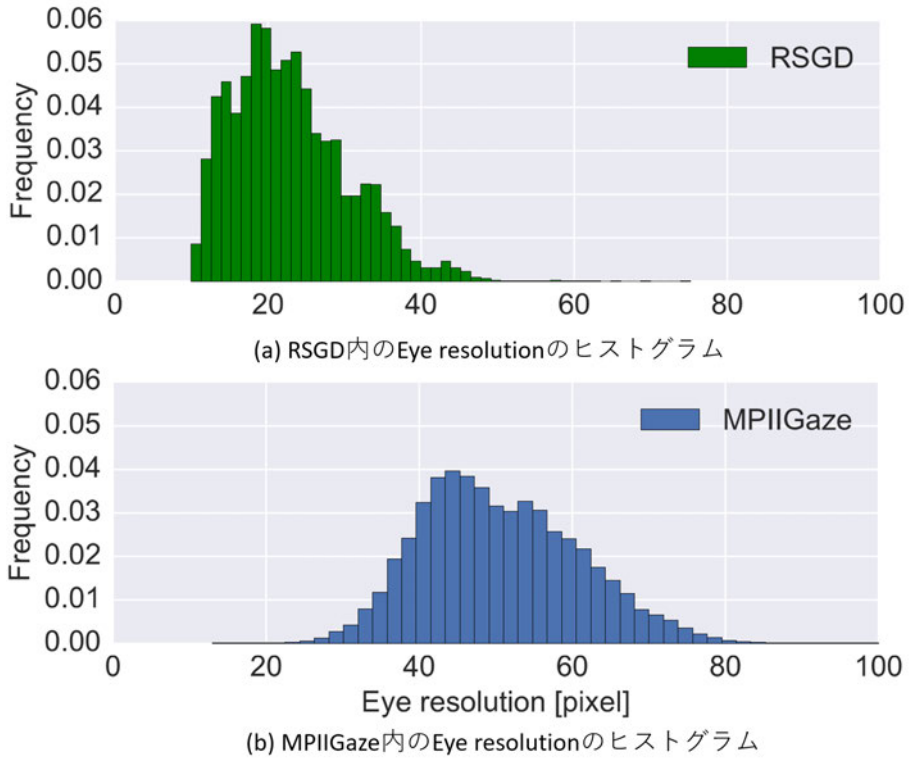


図 4.12　目領域解像度のヒストグラム

グラフから，RSGD は従来のデータセットよりも目領域解像度の低いデータが多くある事が読み取れる．これは，RSGD ではカメラの解像度こそ高いものの，人物の距離がカメラから離れている事や Kinect　v2 センサ内蔵の RGB カメラが広角であるためである．参考までに，図 4.14 に Kinect v2 と市販のウェブカメラである logicool C920R で同じシーンを撮影した例を記す．図より Kinect v2 の画角が広い事が分かる．

4.2　注視点回帰モデル

第 2 章にて頭部姿勢 $\boldsymbol{t}, \boldsymbol{r}$ が，第 3 章にて視線ベクトル $\boldsymbol{g}_l, \boldsymbol{g}_r$ が得られたため，単純にこれらを幾何的に加算すれば幾何的に視線ベクトルは求まる．しかし，4.3 節で示す実験によって，その視線ベクトルは大きな誤差を持つことが分かった．その理由は，眼球中心位置推定値の不正確性である．画像から直接観測できる黒目とは異なり，眼球中心位置は顔

図 4.13　RSGD の例

図 4.14　図 4.1.3 を logicool C920R で撮影したもの

の形状，向きから推定しなければならない．しかし，個人ごとの顔の形状差のために，顔向きが正面から離れるに従って，実際の眼球中心位置との系統誤差が大きくなる．そしてその誤差は，顔位置や顔向きに相関がある．Lu ら [25] はアピアランスベースの視線推定手法において，頭部姿勢と視線誤差の相関に着目し，頭部姿勢変動に起因する誤差を補正する手法を提案した．

本論文ではモデルベースで求めた視線出力結果を頭部姿勢情報から補正する．頭部姿勢 $\boldsymbol{t}$,$\boldsymbol{r}$ と視線ベクトルを説明変数に，注視点位置を目的変数とする回帰器を作成し，RSGD を用いて学習する．学習には，Gradient Boosting Regression Tree を用いる．

4.2.1　Gradient Boosting Regression Tree

本節では Gradient Boosting Regression Tree(GBRT) のアルゴリズムがなぜ本問題に適しているかを述べる．

GBRT は，回帰木を弱学習器としたアンサンブル学習の一種である．

回帰木

回帰木は，データ集合を説明変数の値のしきい値の大小で分割する 2 分木を階層的に組み合わせる事で，複雑な識別境界を得る方法である．回帰木の例を図 4.15 に示す．この例では，頭部位置 x 座標 t_x が 1.2 以上で，視線ベクトルの y 方向 g_y が 0.5 未満の場合，注視点の x 座標 p_x は 0.2 と予測される．回帰木で作成される識別境界は，入力変数が d 次元の場合 $d-1$ 次元の超平面であり，非線形問題を扱う能力を持つ．入力データに対して，損失が最小となるよう各ノードでの特徴軸の選択としきい値の決定が行われる．

提案手法での説明変数 $\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{g}_l, \boldsymbol{g}_r$ は位置や角度といった混合型のデータである．回帰木の各ノードでの特徴軸は独立であるため，異なるタイプ・スケールのデータを扱いに適している．

これに加えて，顔向きや顔位置に起因する系統誤差を，異なる枝として分岐し吸収する事が期待されるため，提案手法で解決したい問題に適している．

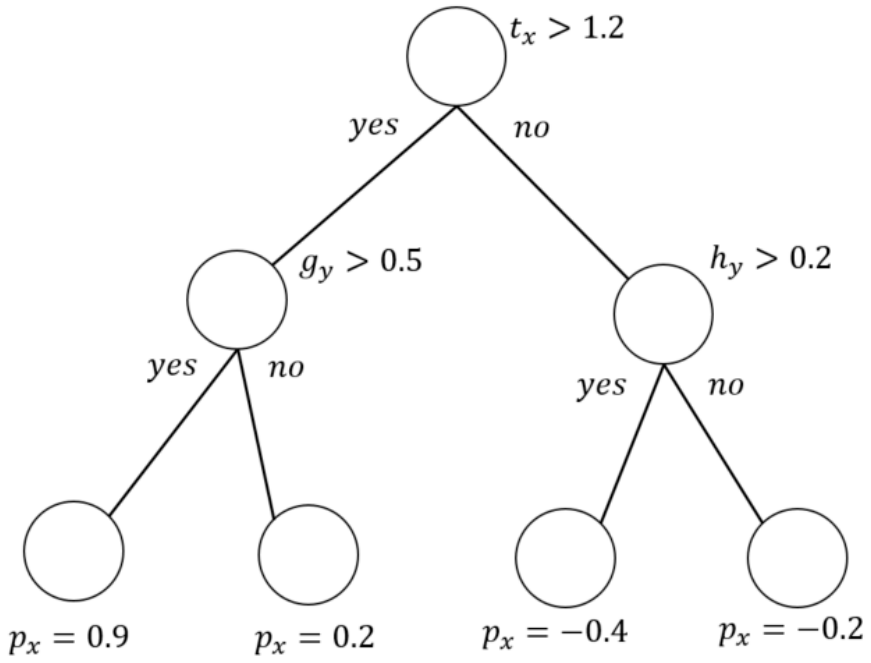


図 4.15　回帰木の例

アンサンブル学習

アンサンブル学習は，複数の弱学習器を組み合わせ，強力な学習器を構成する手法である．弱学習器の作成には，バギングとブースティングの二つの手法がある．

バギングとは，学習データから重複を許してサブデータセットを作成し，サブデータセット内で学習した弱学習器を組み合わせるという手法である．バギングでは個々の弱学習器の学習を並列化できるというメリットがあるが，データセットの持つ偏りが各学習器にも反映されるため，性能向上しづらい．一方，ブースティングでは弱学習器を直列的に学習する．この手法では前の弱学習器の学習結果から適切に識別できなかったデータに対して重みを大きくし学習していくため，後の弱学習器ほど，誤差の大きいデータを重点的に扱うようになる．逐次的に学習するため並列化出来ないが，データセットの偏りに依らない学習を実現する．GBRT ではその名にあるように，ブースティングにより弱学習器を更新する．

GBRT は式 4.1 で表現されるような加法モデルと捉えられる．ここで，$h_m(x)$ は m 番目の弱学習器に相当し，単一の回帰木である．γ_m は重みであり，学習の過程で，正解データとの誤差が小さいほど大きな重みを持つように更新される．

m 番目の弱学習器は式 4.2 のように 1 番目から $m-1$ 番目までの弱学習器の全ての回帰木の線形結合の結果を用いて，作成される．作成は式 4.3 で表したように，損失関数 L の最小化である．GBRT では勾配降下法でパラメータを最適化する．

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{4.1}$$

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{4.2}$$

$$F_m(x) = F_{m-1}(x) + \arg\min_{h} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x)) \tag{4.3}$$

図 4.16 に GBRT による学習の流れを示す．m 番目の弱学習器への入力は，$m-1$ 番目までの線形結合による回帰で誤差の大きかったデータを重点的に学習するよう w^m にて重みを付けられ，損失が最小となる回帰木が形成される．m 番目の回帰木 h_m で予測された結果が，ランダム回帰より良ければ $\gamma_m > 1$ の重みがつけられ，誤りが小さいほど大きな重みとなる．

よって，同じ学習データ群を入力としても，重み w により弱学習木 h_m ごとに異なる空間に写像され，独立した学習木郡が形成される．頭部位置や方向毎に異なる弱識別器集合を通して，RSGD のようにバリエーションの大きいデータセットを適切に学習する．

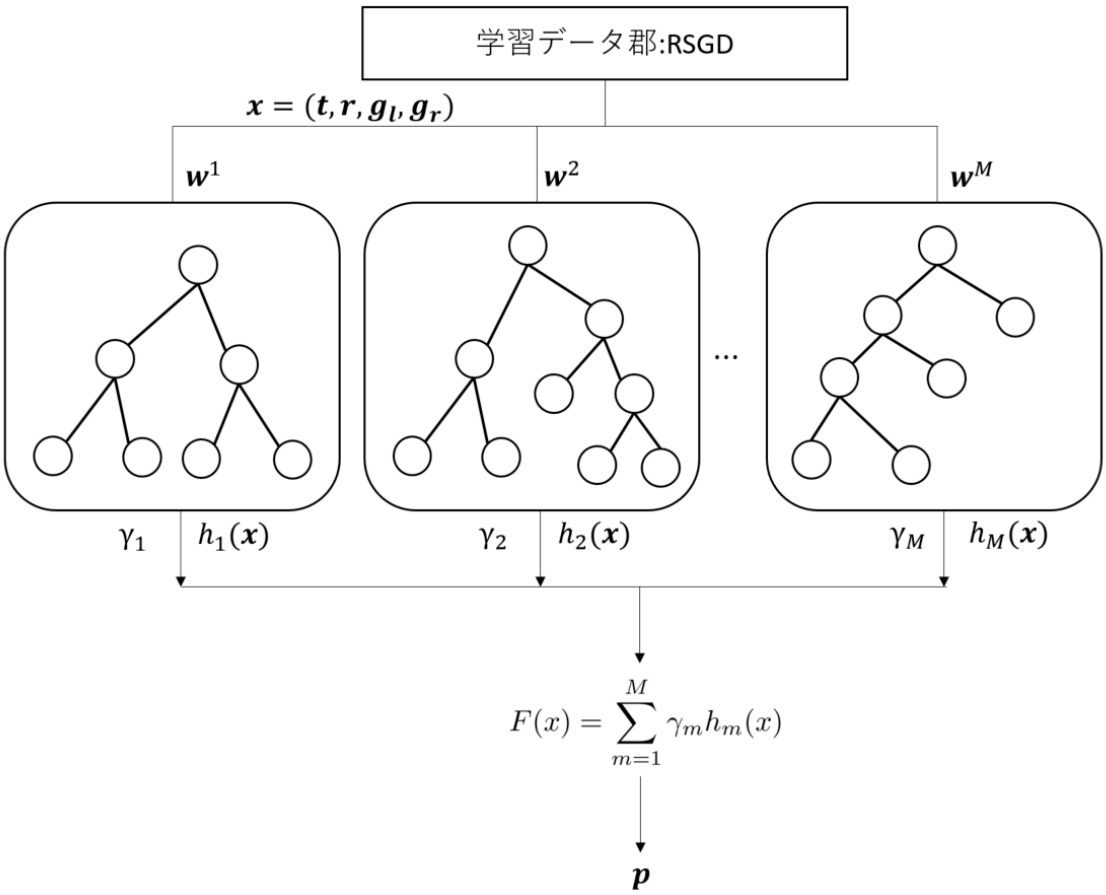


図 4.16　GBRT の学習の流れ

4.2.2　GBRT による RSGD の学習

本節は GBRT による RSGD の注視点推定について説明し，結果とともに考察する．

説明変数 $\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{g}_l, \boldsymbol{g}_r$ はそれぞれ，3 次元頭部位置，3 次元頭部方向ベクトル，3 次元左目視線ベクトル，3 次元右目視線ベクトルであり，計 12 次元の特徴ベクトルである．目的変数はディスプレイ上の注視点の x 座標 $\boldsymbol{p}_x$，および y 座標 $\boldsymbol{p}_y$ である．

クロスバリデーション

提案手法では，汎化性能の評価のため，被験者 11 人のうち 1 人をテスト用として取り出し，残り 10 人分のデータを学習する，leave-one-person-out 法にて精度評価を行う．

ブースティング試行回数

ある被験者におけるブースティングの試行回数毎の損失関数を図 4.2.2 に示す．x 座標については早期に学習が飽和しているが，y 座標では試行回数 100 を超えても徐々に最適化が進んでいる事が確認できる．提案手法では，計算時間も考慮し試行回数を 200 と設定した．

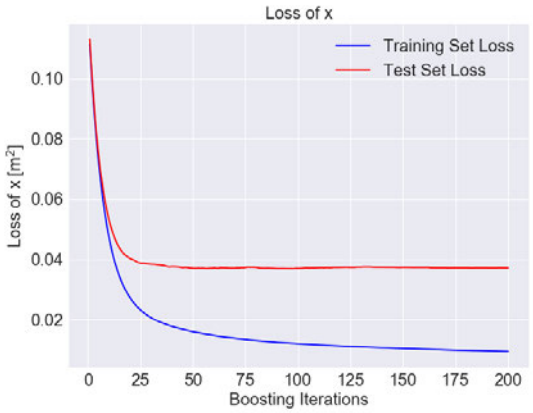


図 4.17　ブースティング試行回数毎の x 座標の損失関数

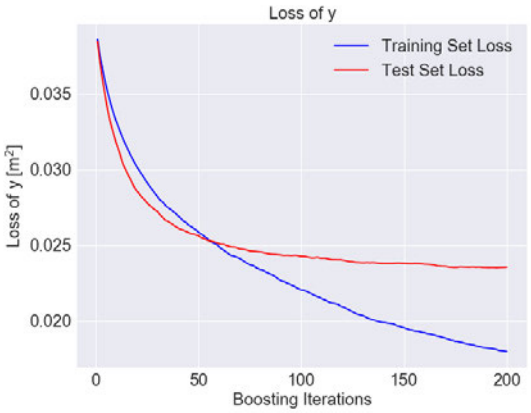


図 4.18　ブースティング試行回数毎の y 座標の損失関数

図 4.19 に，試行回数 200 回後の注視点予測結果の分布を記す．この図の座標は実際の縦 0.68 m, 横 1.2 m のディスプレイの座標 [m] に相当し，黄色い点は表示されたマーカ位置を，緑の点はそれに対して予測された注視点の位置を表す．本来のマーカ表示位置よりも狭い範囲であるものの，全域に渡って分布していることが確認できる．

同じデータについて正解位置と予測位置との相関関係を散布図で図 4.20 に示した．ここで，ディスプレイ左上を原点とし，赤い点はディスプレイ上の水平方向について，青い点は垂直方向について，グラフの横軸を正解座標，縦軸を予測された座標としてそれぞれプロットした．相関係数が 1 なら，$y = x$ の直線となる．

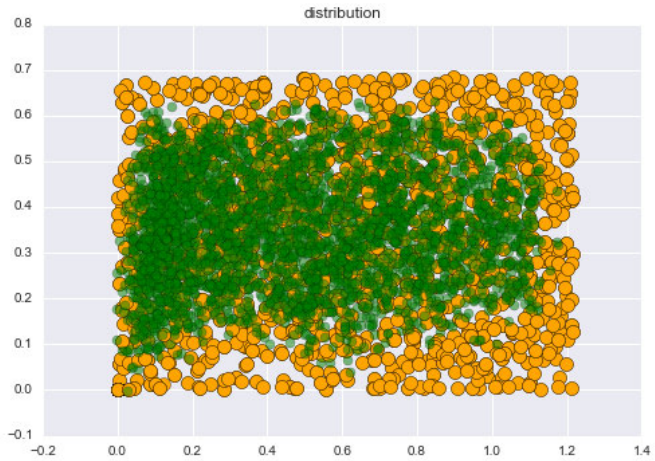


図 4.19　注視点予測結果の分布

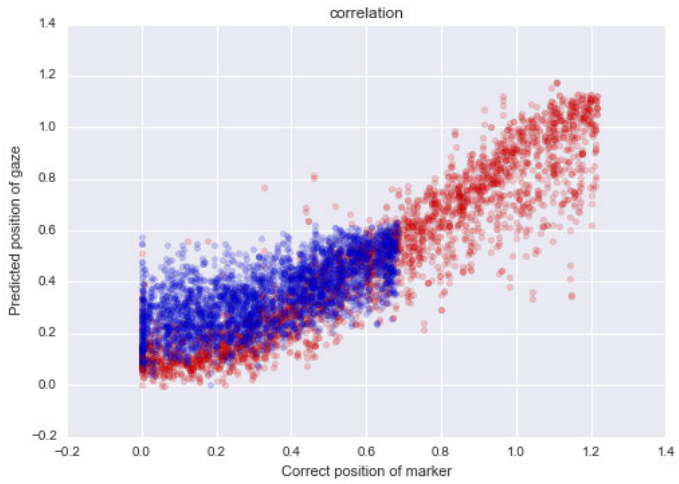


図 4.20　正解座標と予測座標の相関

過学習の制御

GBRT では回帰木の深さ max depth=h によって変数間の相互作用のレベルを定義する．そのような木は最大で 2^h 個の末端ノードと $2^h - 1$ 個の分割ノードを持つ．深さが足りないと十分な表現力を持たないが，深すぎても過学習を招く．よって過学習を制御するため，木の深さの最大値を決めなければならない．

図 4.21，4.22 および表 4.1 は被験者 A における木の深さと推定誤差 (RMSE) をまとめたものである．また，他の被験者 B についても表 4.2 に示す．

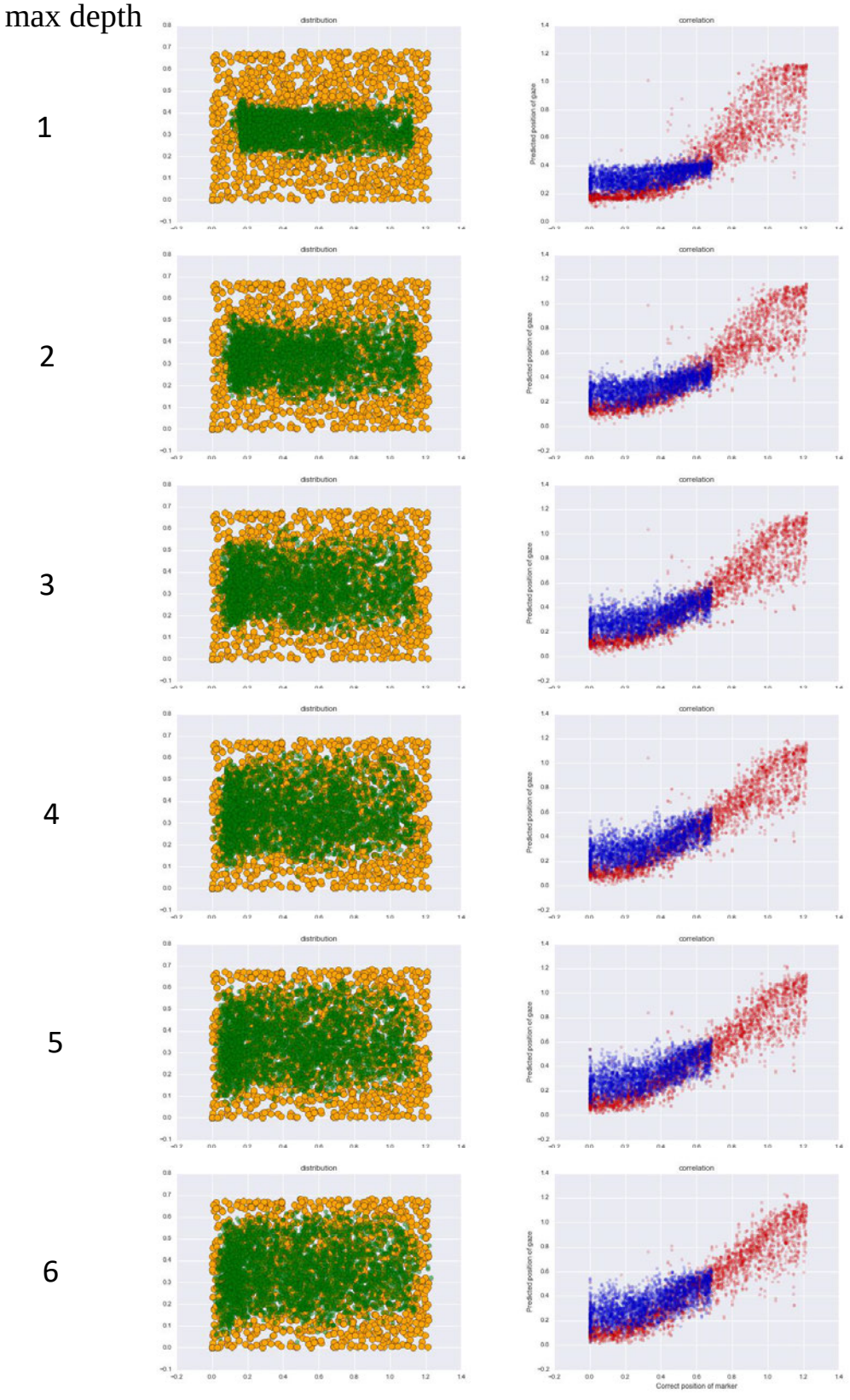


図 4.21　max depth における予測点分布と相関図のまとめ (max depth1:6)

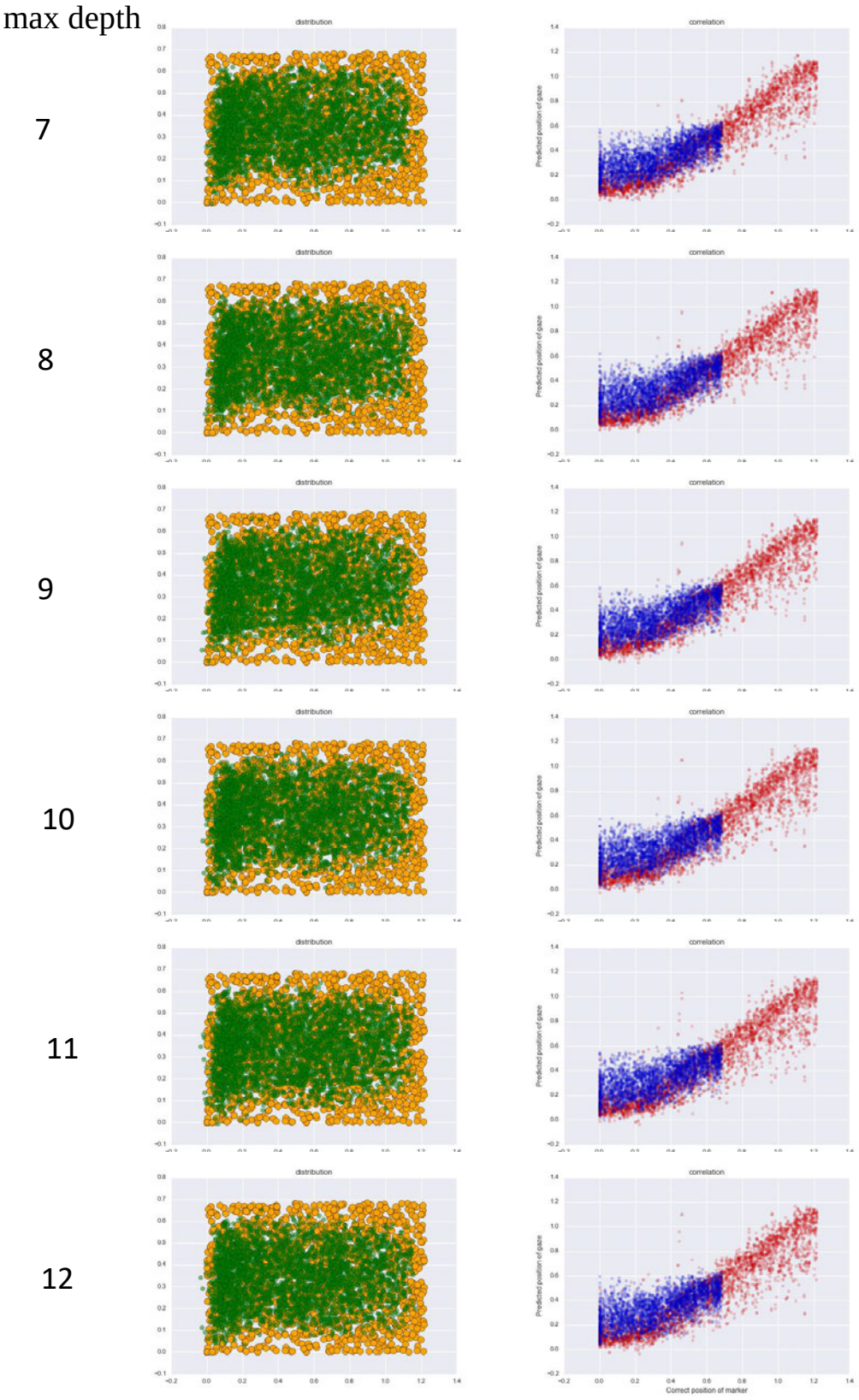


図 4.22　max depth における予測点分布と相関図のまとめ (max depth7:12)

表 4.1　被験者 A における max depth と推定誤差の比較

h	RMSE of x	RMSE of y
1	0.1963641	0.17545651
2	0.18483933	0.16461499
3	0.17967503	0.15621939
4	0.17424534	0.15108104
5	0.17184691	0.14877546
6	0.17029894	0.14856163
7	0.17218341	0.1481213
8	0.17271508	0.14956491
9	0.17450912	0.14866678
10	0.17703254	0.1524827
11	0.18180107	0.15422213
12	0.18538072	0.15325820
13	0.18630203	0.15563586

表 4.2　被験者 B における max depth と推定誤差の比較

h	RMSE of x	RMSE of y
1	0.14372301	0.18395458
2	0.13268566	0.17642487
3	0.12461003	0.17339411
4	0.12336224	0.17022172
5	0.12584975	0.16773166
6	0.12688238	0.16771391
7	0.12465736	0.16975756
8	0.12752226	0.17029232
9	0.13103416	0.17402231
10	0.13231125	0.17430013
11	0.1310731	0.17505721
12	0.13123266	0.17747127
13	0.13245193	0.17904151

図 4.21, 4.22 を見てみると，max depth が 1 から 3 では，予測位置分布がディスプレイの中央付近に集中している．これは，木の深さが浅いため表現力が十分では無い事を意味する．特にディスプレイの y 軸方向に関しては，学習データセット内の平均値に近い値を出力しており，相関が低い事が分かる．x 軸方向に関しては，0.6 以上，つまりディスプレイの右側の領域の予測で分散が大きくなっている．max depth が 6,7 あたりでは，予測点がディスプレイ全体に分布しており，相関図の分散も小さい．つまり，相関係数が 1 に近づいており，精度が高くなっている．max depth 以上では，分布図には大きな変化は起きていないように見えるが，相関図の分散がやや増加し始めており，過学習が発生している．

表 4.1 からは，x 軸では max depth が 6, y 軸では max depth が 7 のときに最も誤差が小さい．別の被験者 B について見れば，表 4.2 から，x 軸では max depth が 4, y 軸では max depth が 6 の時に最も誤差が小さい．

以上の結果のように最適な max depth は被験者ごとに異なるが，安定的に低い誤差を出した 6 と決定する．

4.3　幾何的注視点推定の問題点

ここでは，幾何的に注視点を求める手法の問題点について実験を通して議論する．

4.3.1　実験

実験は RSGD を用いて行った．ディスプレイ上に実際に表示されたマーカ位置と，予測された位置との距離を誤差とし，メートルで表記する．

CLNF 特徴量を用いたモデルベース手法 [55] では，モデルフィッティングの結果からユーザの頭部姿勢 $\boldsymbol{t}, \boldsymbol{r}$ と眼球回転角 $\boldsymbol{g}_l$,$\boldsymbol{g}_r$ が得られる．これらを統合すれば，視線ベクトル $\boldsymbol{p} = \boldsymbol{h} + l\boldsymbol{g}$（$l$ は媒介変数）が得られる．カメラを原点とする世界座標系にて，モニターの位置が与えられれば，視線ベクトルとモニター平面の交点を計算することで，画面上の注視点を推定する．これを [CLNF+Geometric] とする．

一方，同じ $\boldsymbol{t}$,$\boldsymbol{r}$,$\boldsymbol{g}_l$,$\boldsymbol{g}_r$ を説明変数として 4.2.2 項で提案した回帰モデルにより注視点を推定する手法を [CLNF+Training] とする．これら 2 つの手法を比較し，学習による効果を確認する．

4.3.2　結果

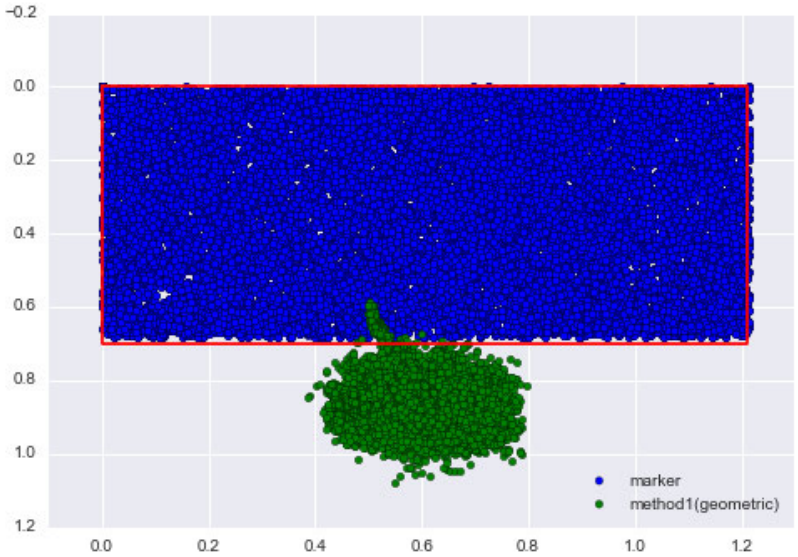


図 4.23　CLNF+Geometric で予測された注視点の分布

ディスプレイ上に表示したマーカの位置を青い点で，ディスプレイの大きさを赤

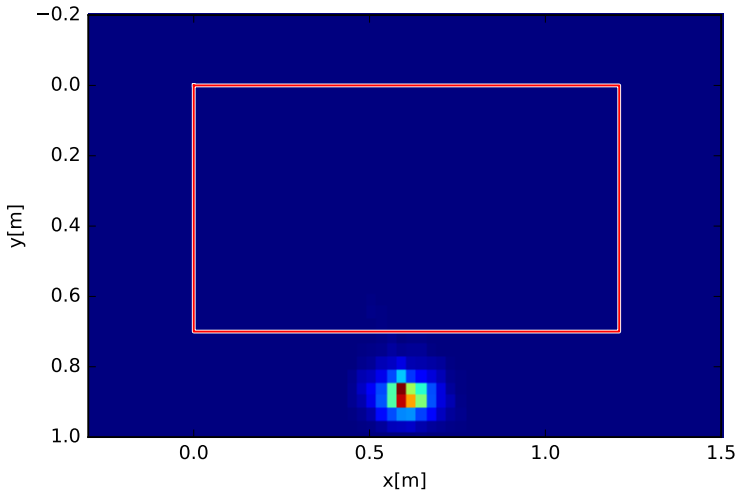


図 4.24　CLNF+Geometric で予測された注視点のヒストグラム

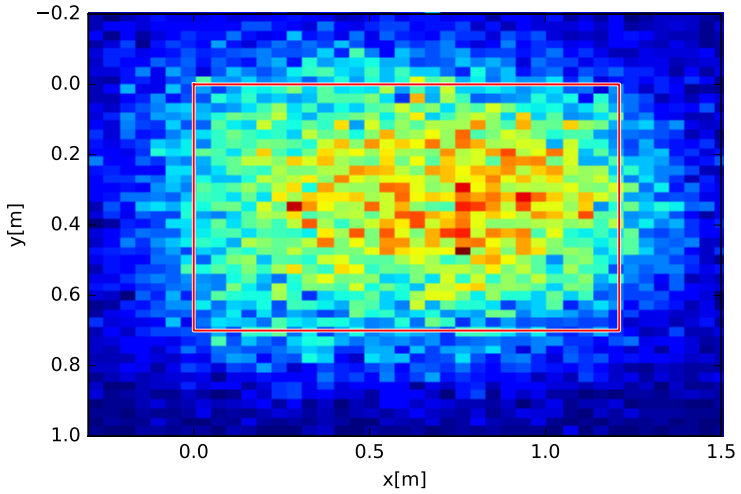


図 4.25　CLNF+Training で予測された注視点のヒストグラム

い枠で図 4.23 に示す．[CLNF+Geometric] で推定した注視点を緑色の点で示す．図の水平，垂直方向は実際のディスプレイの方向に対応している．緑色の点の分布から [CLNF+Geometric] で推定された注視点はディスプレイの中央下エリアに集中している事が分かる．図 4.24 はそれをヒストグラムとしてヒートマップで表示したもので，予測注視点が非常に狭い範囲に集中していることが分かる．この範囲の中心点は $(x, y) = (0.60, 0.87)$ で，カメラの設置箇所に相当する．このことから，推定された全ての視線ベクトルがカメラ中心に集中している．この様子を図示すると図 4.26 のようになり，推定された眼球位置が実際の眼球位置よりもカメラ光軸に対して外側にずれている事

が考えられる．

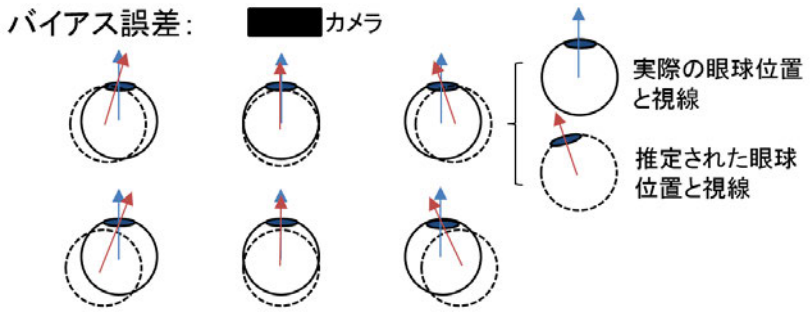


図 4.26　推定された眼球中心位置のずれ

眼球位置のずれは，目尻・目頭の顔特徴点検出の誤差，画像上の眼球半径値の誤差，および頭部姿勢推定の誤差に起因する．このうち目尻・目頭の検出精度については第 2 章にて従来手法からの改善を行っており，またバイアス誤差は認められなかった．頭部姿勢推定には誤差が生じる．本論文は，デプスセンサやステレオカメラではなく単眼カメラを用い，さらに個人キャリブレーションレスを前提としている．2.4.1 項で作成した一般顔モデルに基づき頭部姿勢を推定するため，誤差は避けられない．5.1.2 項で検証するが，頭部推定位置にはバイアス誤差が生じていた．画像上の眼球半径値についても誤差が生じる可能性がある．実際の眼球半径は 12mm の近似で十分であると報告されている [40] が，その 12mm が画像上で何 pixel に相当するかを計算するためには正確な頭部姿勢（位置および方向）が必要である．頭部姿勢は上記の理由で正確な推定が出来ないため，眼球半径についても誤差が生じた．個人キャリブレーションによって精度を向上させる手法は多く提案されているが，本論文の前提とするキャリブレーションレス条件下では良い精度を得ることは難しい．これらのバイアス誤差を機械学習により補正する事が本論文の狙いである．

[CLNF+Geometric] 手法と同じ入力変数を用い，学習から補正した結果である [CLNF+Training] による予測注視点のヒストグラムを図 4.25 に示す．図の通り，実際のディスプレイの範囲にマッピングされており，改善が確認できる．RMSE は [CLNF+Geometric] の $(x, y) = (0.38, 0.56)$ から，[CLNF+Training] では $(x, y) = (0.30, 0.18)$ に改善した．決定木ベースのアンサンブル学習であり，頭の位置や方向別に異なる回帰係数を得るという特性が今回の目的に適している．

4.4　本章のまとめ

本章では，広範囲空間内の頭部位置に頑健な視線推定を実現するため，既存の視線データセットよりもカメラに対して距離・角度ともに広い頭部位置分布を持つ RSGD の作成について説明した．このデータセットは被験者にディスプレイ上に表示されたマーカを注視してもらい，その時の画像および頭部位置，正解マーカ座標を記録したものである．１１名の被験者に対しそれぞれ平均 3000 枚の画像を取得した．そして，RSGD の目領域解像度は従来のデータセットそれよりも低く，難しい課題であることを明らかにした．

続いて，Gradient Boosting Regression Tree による個人キャリブレーションを行わない注視点推定の学習について，その特性とともに記した．

また，個人別の推定結果から，最適な学習パラメータについて議論した．

第 5 章

提案手法の有効性の検証

本章では，これまでに提案した視線推定手法の有効性を示すために行った 4 つの実験について説明する．実験は全て RSGD を使用し，頭部姿勢推定，虹彩追跡のそれぞれの要素技術について，従来手法と提案手法について比較し，その特性について議論する．これに加えて，提案した虹彩追跡手法のカメラからの距離に対する頑健性を示す．

5.1 実験

本章では，提案手法の有効性を示すために 4 つの実験を行った．これらの実験は全て第 4 章で述べた RSGD を用いた．ここでは，誤差はディスプレイ上に実際に表示されたマーカ位置と，顔画像から予測された注視点位置との距離と定義し，単位は全てメートルである．被験者ごとに平均二乗誤差を算出し比較する．実験で用いた手法についてまとめたものを表 5.1 に示す．

5.1.1 虹彩追跡に関する比較

虹彩追跡そのものの精度を比較するため，頭部姿勢推定手法を Open Face に，注視点推定を GBRT に固定し，虹彩追跡について CLNF 手法 [55]，x-sobel エッジに着目したパーティクルフィルタ手法 [71]，提案手法の 3 つの手法を比較する．以後 [CLNF+Training], [PF+Training], [Proposed-1] と呼称する．

CLNF とは，Constrained Local Neural Field の略であり，虹彩特徴の patch exparts を 3 層のニューラルネットワークで学習し，視線を推定する手法である．この手法は MPII Gaze dataset において state-of-the-art となる精度を達成し，CNN を用いたアピアランス

表 5.1　実験で用いた各手法

Method	Head Pose Est.	Eyelid Filter	Iris Tracking	PoG Estimation
CLNF+Training	openface [63]	-	CLNF [63] [55]	GBRT
PF+Training	openface	-	PF(x-Sobel) [71]	GBRT
Proposed-1	openface	-	PF(edge-gradient)	GBRT
Proposed+Depth	Kinect	-	PF(edge-gradient)	GBRT
Proposed-2	CNN2.1 節	✓3.4 節	PF(edge-gradient)	GBRT

ベース手法を超えたと主張している．よって，本論文での比較対象として選んだ．ただし，Baltrusaitis らの手法により注視点を推定した結果は，4.3 節で検証したとおり，実際のディスプレイ位置と異なり，全てカメラ設置位置近傍に集中していた．しかしながら，系統誤差はあるものの視線方向には有効な情報があると考え，4.2 節で述べた回帰によって，提案手法と同じ枠組みで学習した．

5.1.2　単眼カメラと RGB-D カメラでの比較

本論文が想定する環境のようにカメラに対して被験者の頭の位置の変化が大きい場合，被験者の頭部姿勢推定は必須である．単眼カメラから頭部姿勢推定は，カメラの焦点距離，画像中の頭部位置，顔のスケールに基づく．しかし，顔の形状や大きさの個人差，設置環境により単眼カメラのみからの頭部姿勢推定には誤差が発生する．そこで，頭部姿勢について，単眼カメラのみから推定された情報を用いた場合 ([Proposed-1] と，Li ら [72] の手法のように Kinect v2 の深度センサにより得られた正解値を用いた場合 [Proposed+Depth] について比較し，単眼カメラのみを用いた場合の妥当性について検証する．

5.1.3　瞼形状フィルタの比較

瞼やまつげのエッジは虹彩追跡において外乱となりうる．そこで，それらのエッジを除去しない場合 [Proposed-1] と 3.4 節で説明した，瞼の内側のみのエッジに重点を置くフィルタにより除去した場合 [Proposed-2] で比較を行う．

5.2　結果

5.2.1　虹彩追跡に関する比較

各手法による予測誤差を図 5.1 に示す．水平方向の誤差を赤色，垂直方向を緑色，それらを統合した誤差を青色のバーで表す．また，[CLNF+Training] および [Proposed-1] による虹彩追跡結果画像を図 5.2 および図 5.3 に示す．

各手法の RMSE は，[CLNF+Training] が 0.356 m なのに対し，[PF+Training] で 0.256 m，[Proposed-1] で 0.239 m となっており，提案手法にて精度向上が確認された．頭部姿勢推定及び注視点推定は全て同じプロセス，同じパラメータ数で行われているため，この精度の差は虹彩追跡そのものの性能の差に起因する．CLNF 特徴量に基づく手法は，視線追跡の精度評価では MPII Gaze dataset において state of the art のスコアを実現していた．しかし，今回のデータセットについては，図 5.2(a)(c)(d) 等の [CLNF+Training] Result 行の画像から確認できるように，目領域の解像度の低さやノイズ故に，黒目の特徴を捉えきれず誤追跡が多く見受けられた．一方，本論文で提案する勾配 + パーティクルフィルタのエッジベースの手法は，[Proposed-1] 行から確認できるように，モデルの勾配に着目する事で影やまぶたなど外乱エッジを無視し，黒目のエッジのみを効果的に抽出するために，このような悪条件画像でも破綻する事無く追跡に成功した．

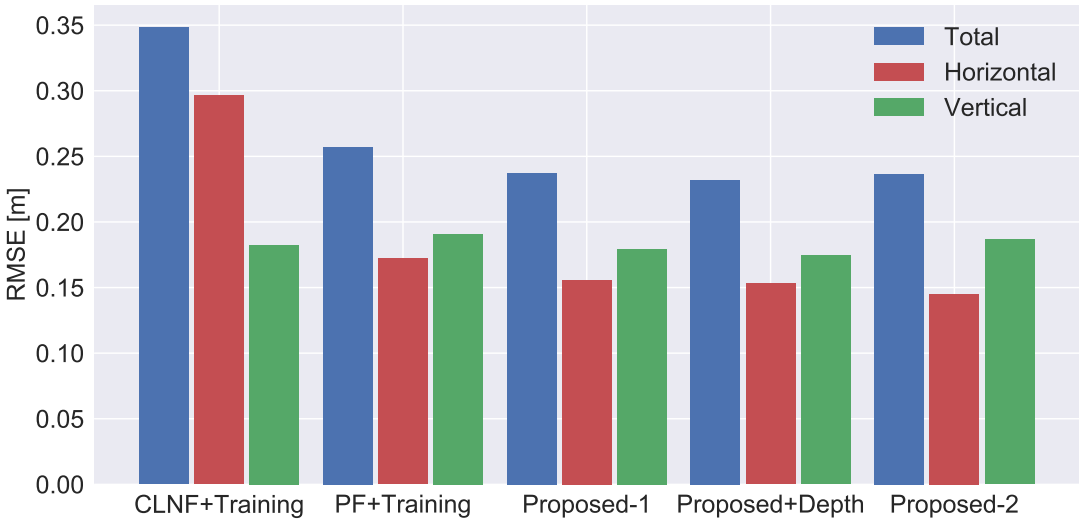


図 5.1　各手法での RMSE

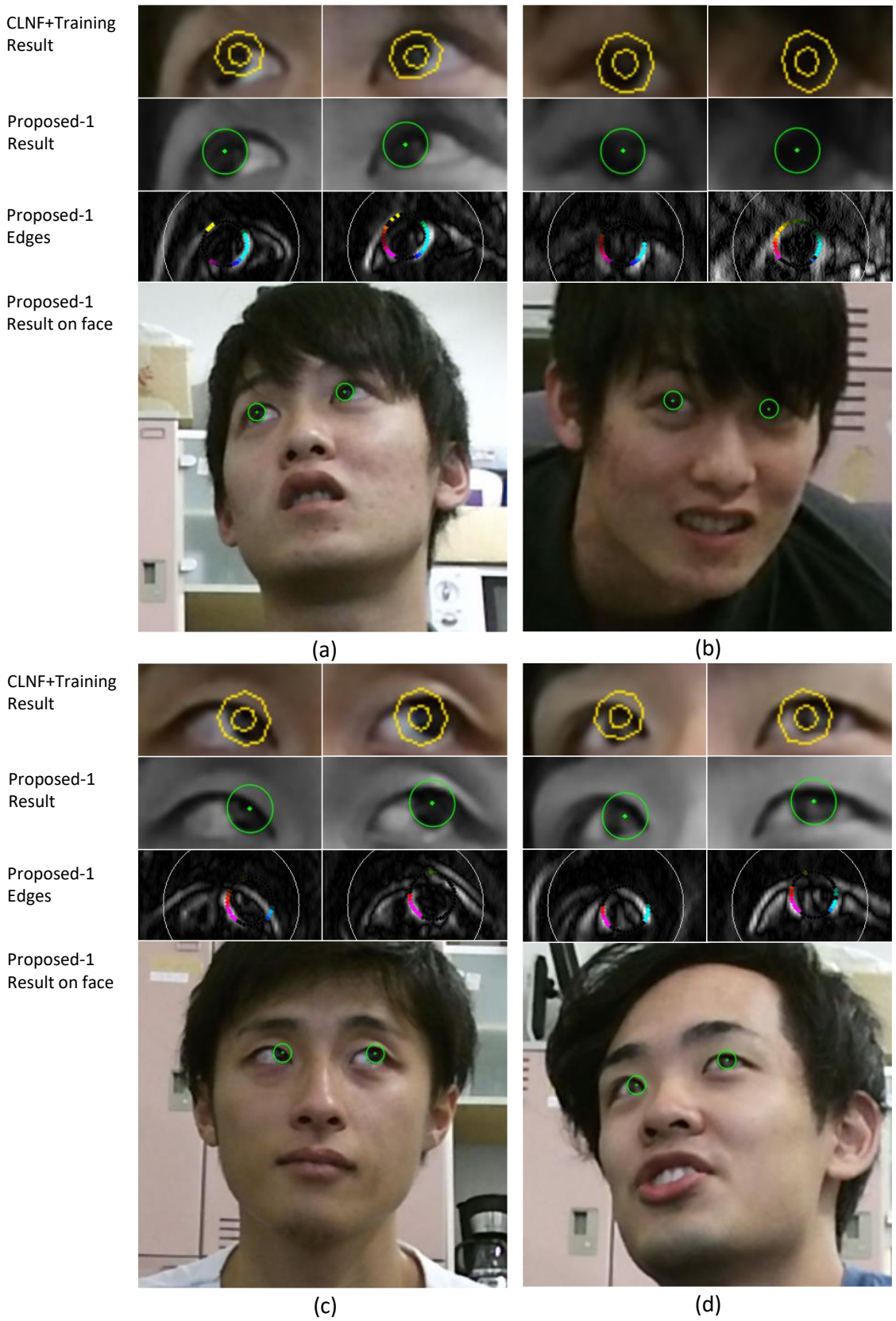


図 5.2　[CLNF+Training] および [Proposed-1] による虹彩追跡結果の比較 1/2

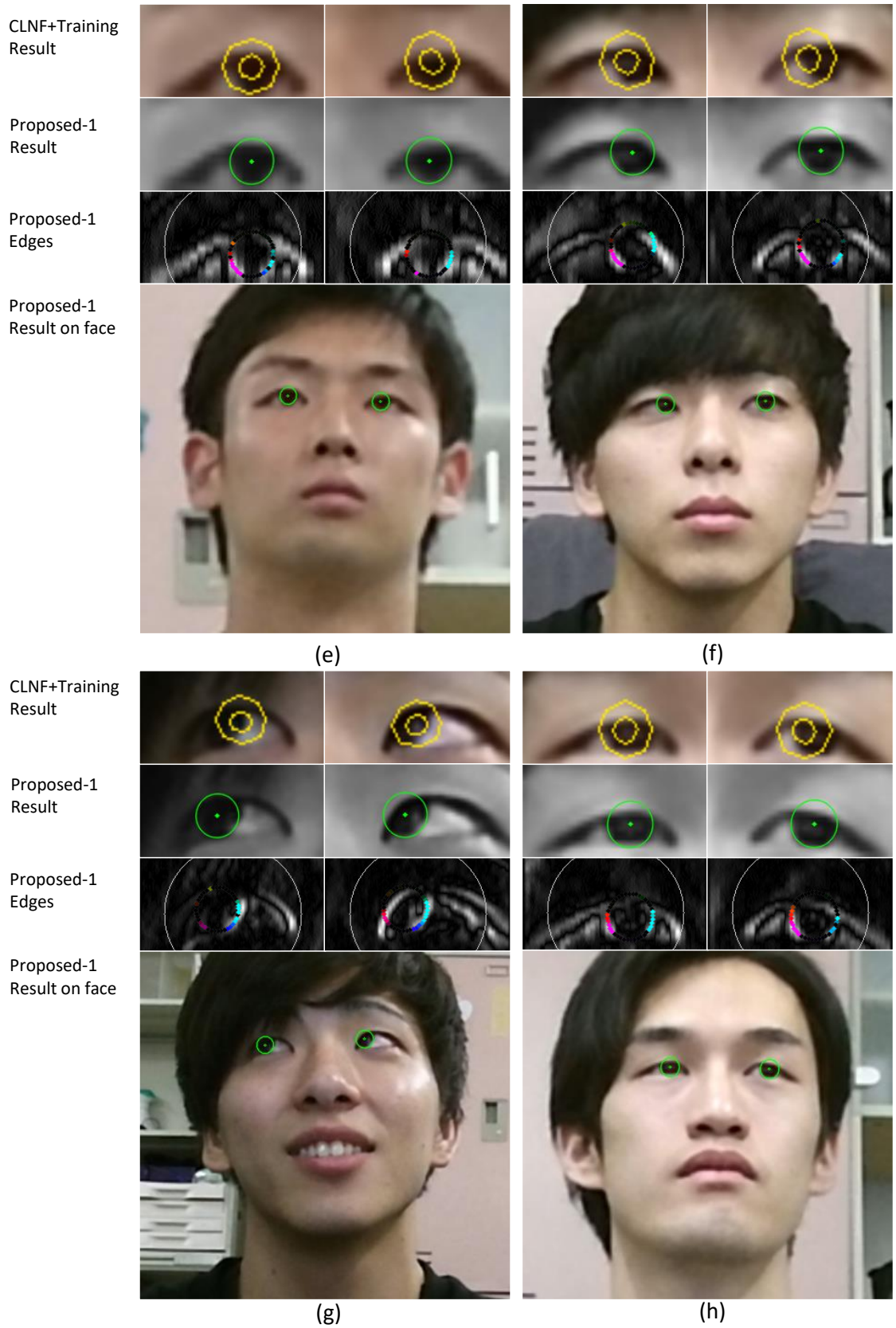


図 5.3 [CLNF+Training] および [Proposed-1] による虹彩追跡結果の比較 2/2

図 5.1 の垂直および水平誤差に着目すると，水平方向では特に従来手法 [63] に対して大きく改善が確認できる．目は多くのケースにおいて，黒目輪郭の上下がまぶたにより遮蔽が発生し，左右の黒目輪郭しか見えない．提案手法では黒目輪郭の周回 120 点のサンプリングを行い，モデルと整合性の高い点のみを抽出し，モデルパラメータを推定する．それにより，提案手法ではこの限られた左右のエッジから眼球回転角の yaw 方向を高精度に推定する．一方，[CLNF+Training] では，黒目輪郭に対し 8 点の patch しか散布しないため，まぶたによる遮蔽の影響を受けやすい事が理由として考えられる．

図 5.4 は [CLNF+Training], [Proposed-1] の各被験者での RMSE の値を示す．また，図 5.5，図 5.6，図 5.7, 図 5.8 では，[CLNF+Training], [Proposed-1] について，各被験者の正解値と予測値との相関図を x 方向と y 方向それぞれについて示す．x 方向については全ての被験者で安定して精度が向上した事が確認できる．一方，y 方向については大きな改善は見られなかった．

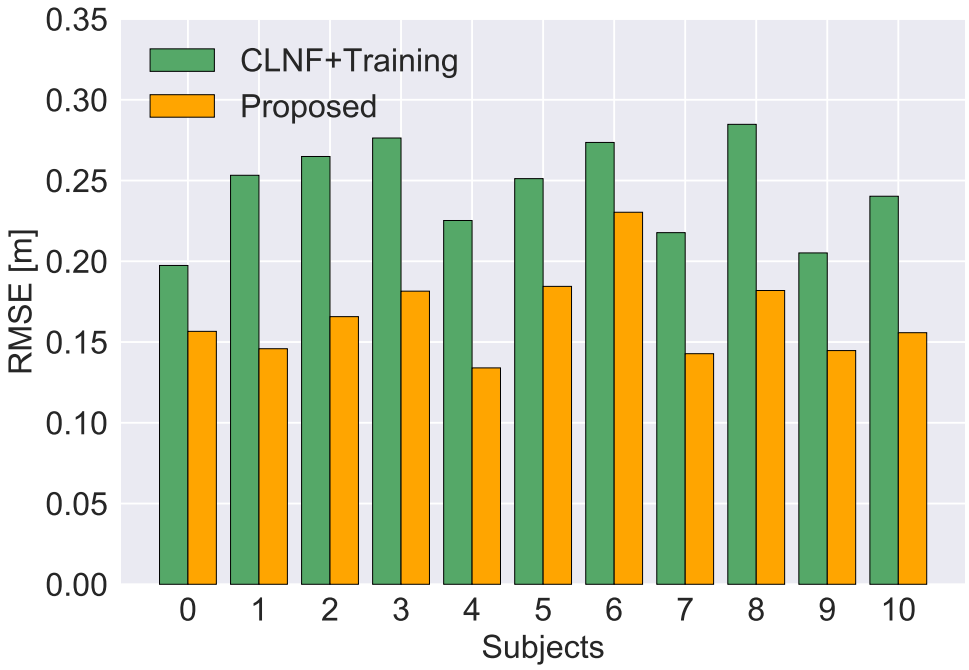


図 5.4　各被験者における RMSE

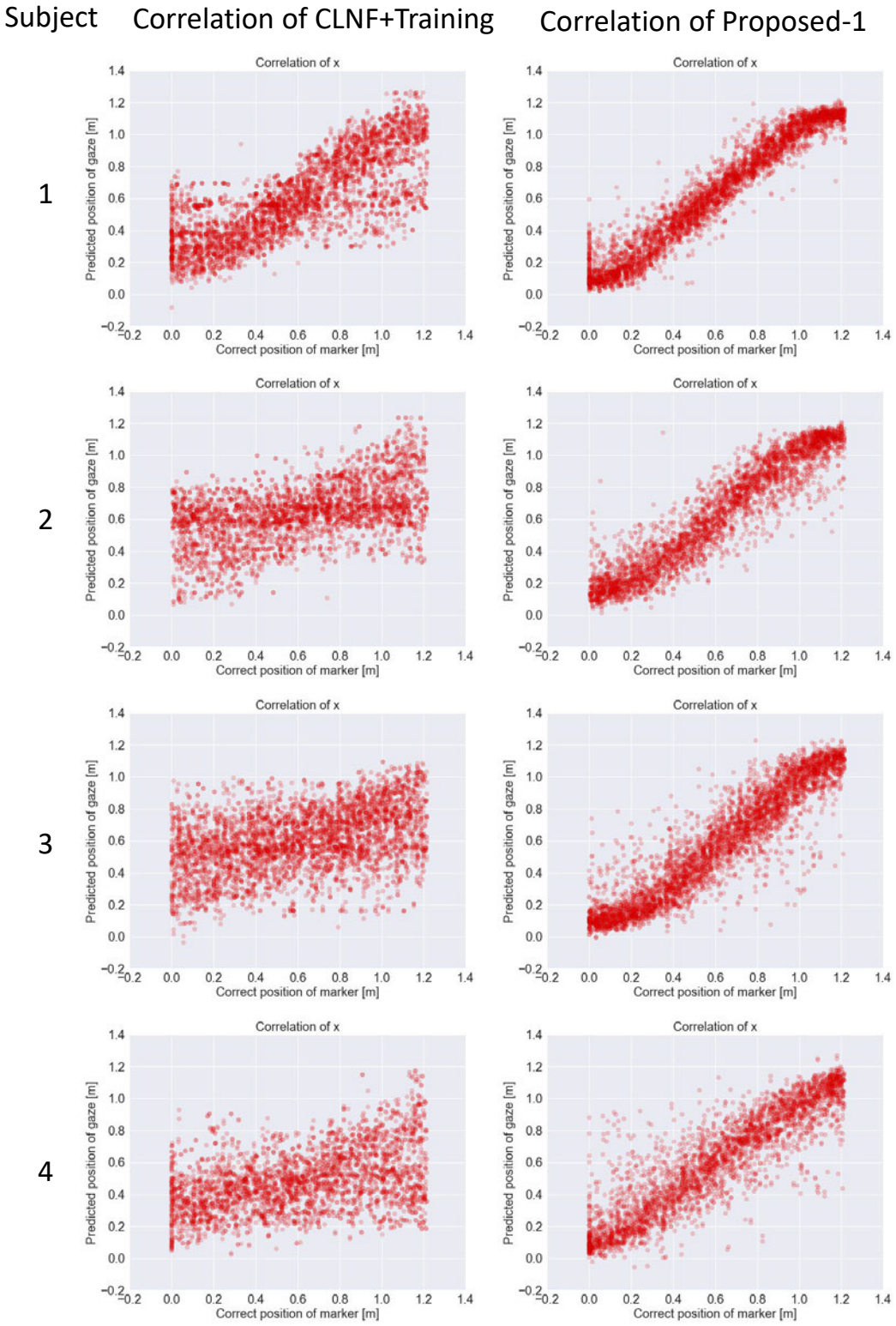


図 5.5　各被験者における x 方向の正解値と予測値の相関図 1

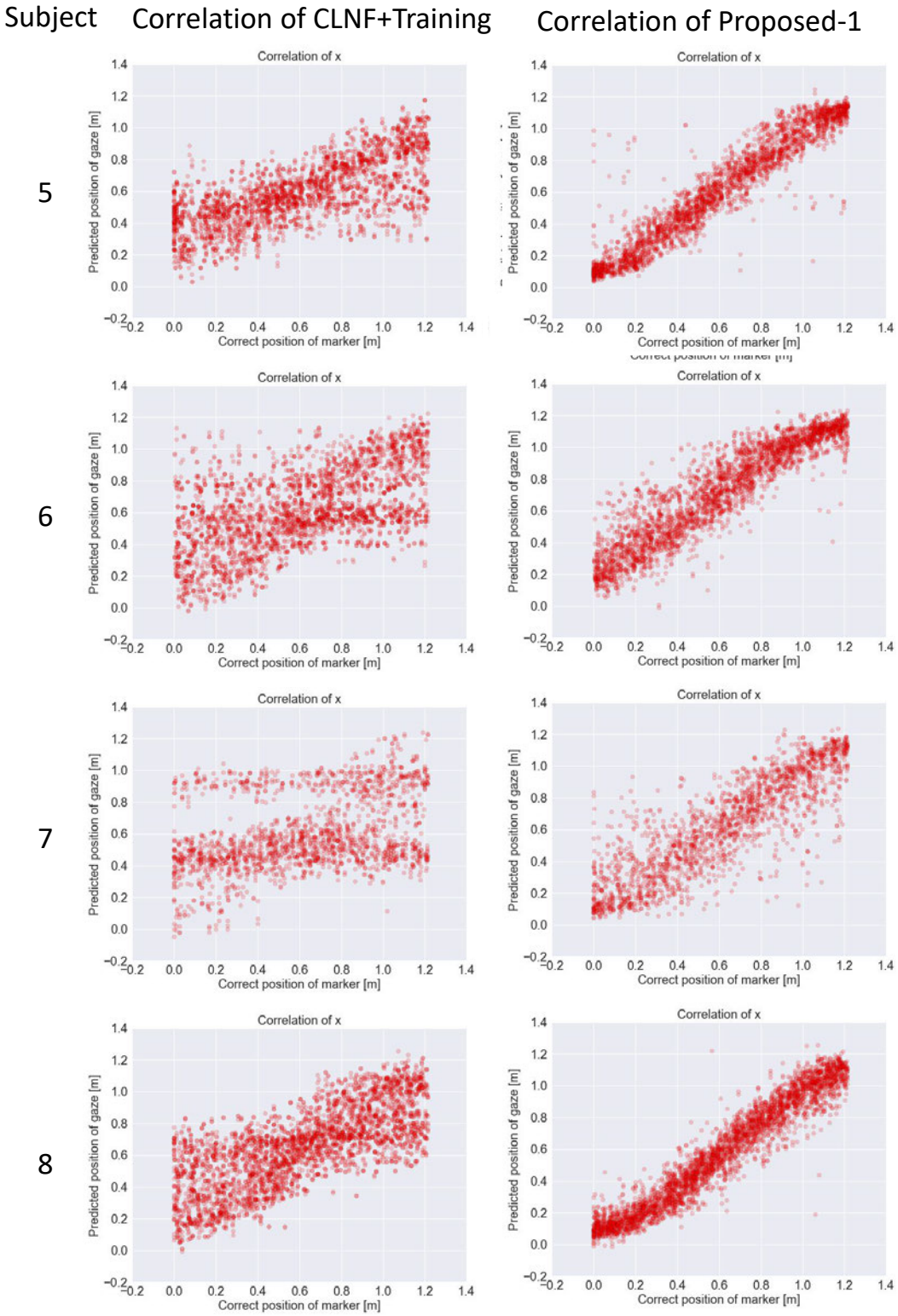


図 5.6　各被験者における x 方向の正解値と予測値の相関図 2

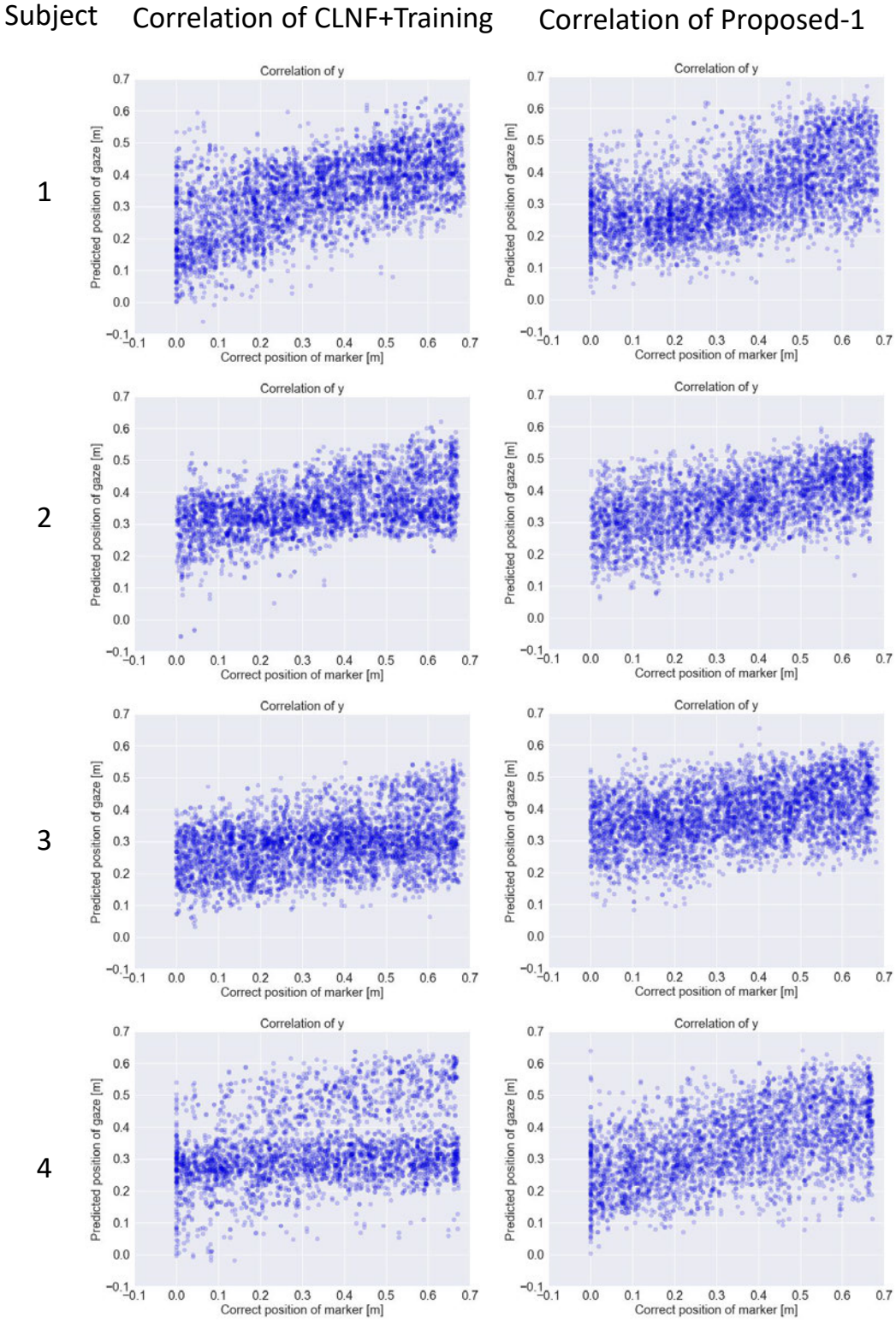


図 5.7　各被験者における y 方向の正解値と予測値の相関図 1

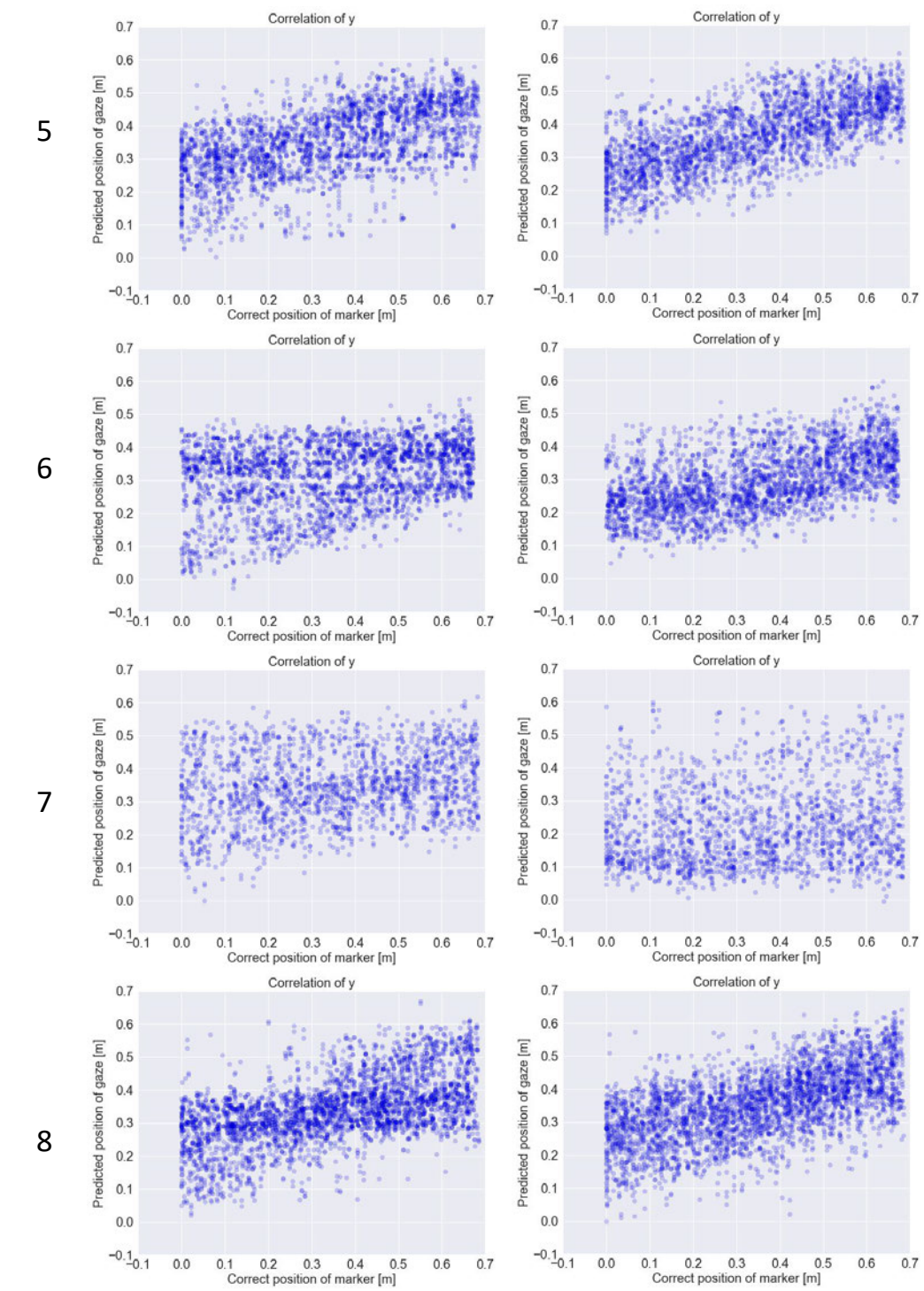


図 5.8　各被験者における y 方向の正解値と予測値の相関図 2

5.2.2 単眼カメラと RGB-D カメラでの比較

[Proposed-1] では，画像から推定された頭部の位置 $\boldsymbol{t}$ を使用した．[Proposed+Depth] では RGB-D センサで得られた正確な頭部位置を使用した．$\boldsymbol{t}$ の推定誤差は x,y,z 方向それぞれ $(t_{\epsilon x}, t_{\epsilon y}, t_{\epsilon z}) = (0.032 \pm 0.034\mathrm{m}, -0.30 \pm 0.26\mathrm{m}, -0.14 \pm 0.068\mathrm{m})$ となった．それぞれ RMSE で $0.046 \quad \mathrm{m}, \quad 0.40 \quad \mathrm{m}, \quad 0.15 \quad \mathrm{m}$ である．個人間で顔のスケールが異なるため，誤差の発生は避けられない．そのため，結果としては確かに注視点推定精度は [Proposed+Depth] が上回っているが，その影響は非常に限定的である．$\boldsymbol{t}$ の推定誤差のうち，バイアス誤差については学習によって吸収されると考えられるため，問題となるのは偶然誤差だ．x 方向に関しては，そもそも $t_{\epsilon x}$ の推定誤差が非常に小さいため，注視点推定の水平方向誤差は [Proposed-1] と [Proposed+Depth] でほとんど差が見られない．一方で，y 方向に関しては，$t_{\epsilon y}$ の偶然誤差が x 方向や z 方向と比較して大きいため，注視点推定の垂直方向誤差に若干の差がある．しかしその差も大きいとはいえない．以上のことから，RGB-D センサを利用した時と比較して，単眼カメラのみでも遜色ない推定精度が得られる事が分かった．

5.2.3 瞼形状フィルタの比較

瞼形状フィルタを用いた結果，ディスプレイの水平方向について RMSE が [Proposed-1] で 0.155，[Proposed-2] で 0.145 となり，精度向上が確認された．外乱となる瞼エッジや眼鏡等の影響を排除した効果である．

しかしながら，垂直方向については，RMSE が [Proposed-1] で 0.179，[Proposed-2] で 0.186 へと誤差増大が発生した．上記，虹彩追跡に関する考察で述べた通り，虹彩は瞼で囲まれた狭い領域内のエッジをサンプルする．特に垂直方向については，虹彩エッジ曲線への僅かな向きの差が重要な手がかりとなる．

瞼形状フィルタにより，図 3.8 にあるように，瞼近傍の虹彩のエッジも除去してしまった例があり，この結果垂直方向の視線推定誤差が増大したと考えられる．方向別で異なる処理とする，もしくは，瞼形状の表現力を高める等，改善する必要がある．

5.2.4 カメラからの距離に対する頑健性

カメラから被験者までの距離が離れている時 (1 m 以上)，目画像の解像度が検出に十分で無いためにモデルベース視線推定手法は失敗する傾向になる．このような目画像の解像度が低い条件下において視線を推定するために，Cazzato ら [45] は頭部姿勢のみから視線を推定する手法を提案した．大規模な実験の結果，他の state-of-the-art な手法と遜色ない

精度を持つ事が報告されている．

本論文で提案する手法も同様にカメラから離れた人物 (< 2.5 m) の視線を推定することを試みており，頭部姿勢情報も視線推定のために用いている．そのため，視線が頭部姿勢のみから推定され，虹彩追跡がもはや意味を成していないという懸念が生じる．このような遠く離れた距離において提案した虹彩追跡手法が有効か確かめるために，提案手法 [Proposed-1] と Cazzato らのように頭部姿勢情報のみを用いて推定する手法 ([Head] と呼称する) を比較した．これに加え，[CLNF+Training] 手法も比較した．

図 5.9 は 3 つの手法の RMSE について，カメラからの距離が 0.5 m から 2.6 m までの間の 0.3 m 毎にプロットしたものである．図 5.2.4(a) は水平方向について，図 5.2.4(b) は垂直方向についての結果を表す．図の x 軸はカメラからの距離を表し，y 軸は RMSE を表す．

垂直方向に関しては，1.5 m 以上の距離では各手法の RMSE はほぼ同等となった．このことは，このような距離においては虹彩追跡が有効でないことを意味する．

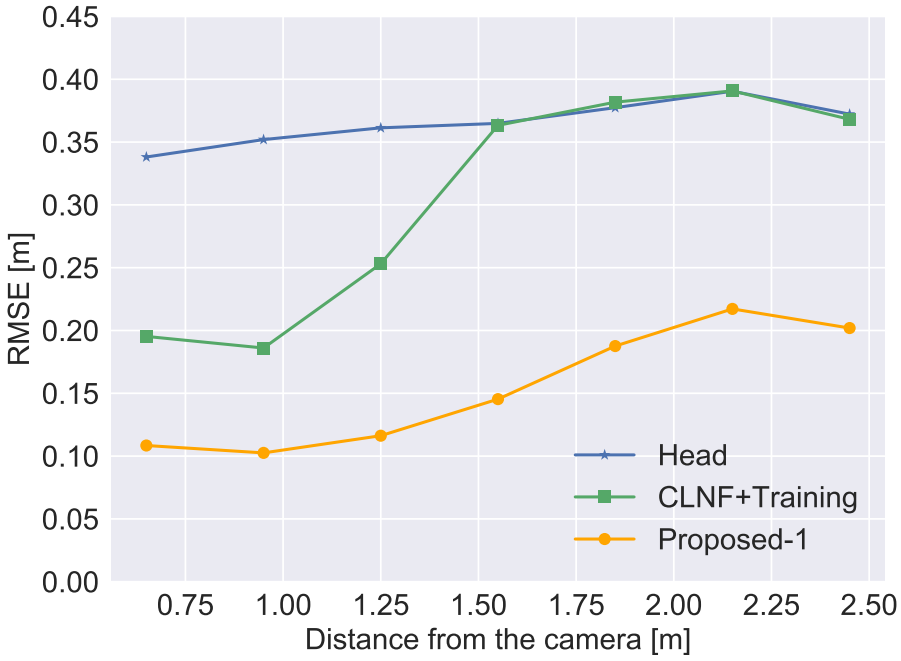
しかしながら，水平方向に関しては，[Proposed-1] はカメラ近傍のみでなく遠い距離においても誤差が低い事が分かる．[CLNF+Training] 手法では 1.5 m より遠い距離では虹彩追跡が破綻し，視線推定に寄与していない．一方提案手法では 2.5 m の距離まで虹彩追跡が有効である．

具体的には，1.5 m 地点において従来手法の水平方向 RMSE は 0.35 m であったが，提案手法では 0.15 m まで誤差を低減した．これは，従来手法では半径 0.35 m の目標内を 68% の確率で注視しているといえる事に対し，提案手法では半径 0.15 m の範囲内を注視しているといえる事を意味する．例えば，デジタルサイネージへの応用では，従来手法ではディスプレイの右側か左側かなど大域的にどこを見ているかしか推定できなかったのに対し，提案手法ではより詳細な注視領域（例えばディスプレイ内の人物や商品の位置など）を推定できる．一方で次世代型自動販売機に表示されている飲料の位置までは特定できない．

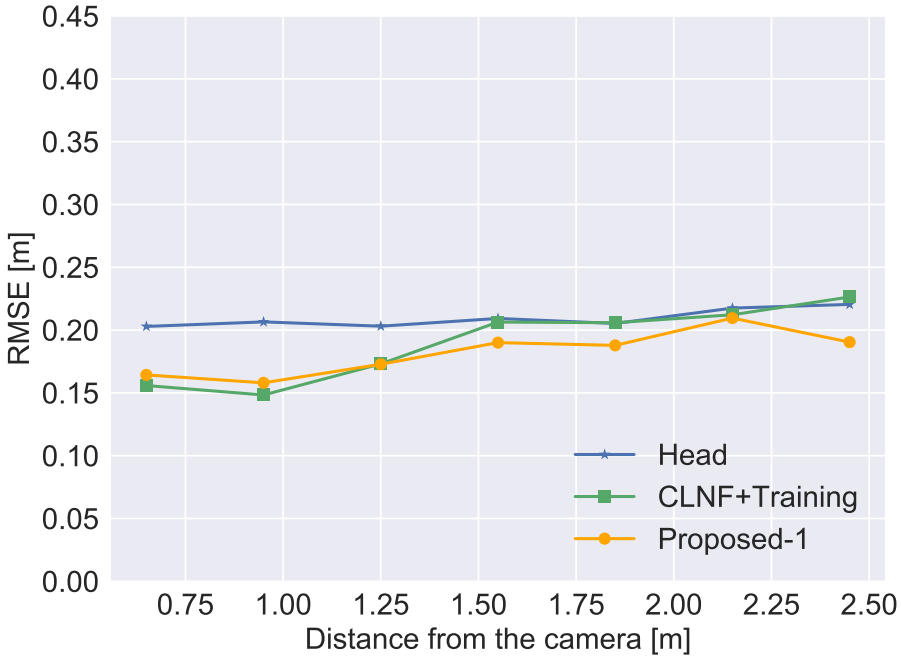
RSGD では虹彩追跡が有効な最大距離を確認できなかった．言い換えれば，[Head] と [Proposed-1] の線が交差する地点を確認することはできなかった．しかし，図 5.2.4 は虹彩追跡が 2.5 m より遠い地点でも有効である可能性を示しており，興味深い結果と言える．

5.2.5　他の最先端手法との同一基準による比較

本項では，他の代表的な最先端手法で報告されている精度と提案手法との比較を行う．表 5.2 は頭部姿勢変動に対応した手法を頭部位置，方向，手法のカテゴリ，誤差，被験者の数とカメラの数とともにまとめたものである．他の手法は誤差を度数で報告しているため, 提案手法の誤差を [m] から [degree] に変換し, 平均絶対誤差と標準偏差について記し



(a) 水平方向



(b) 垂直方向

図 5.9　[Head], [CLNF+Training], [Proposed-1] の 3 手法の RMSE

表 5.2　他の最先端手法との比較

Method	Head Positions [m]		Head Rotations	Category	Reported Error [°]	Subjects	Camera
Proposed	continuous (wide)	X: -0.9∼0.9 Y: -0.2∼0.5 Z: 0.5∼2.5	continuous	model	Hor.: 4.02±2.57 Ver.: 5.94±2.79 Total: 7.58±4.48	16	RGB
CLNF+Training	continuous (wide)	X: -0.9∼0.9 Y: -0.2∼0.5 Z: 0.5∼2.5	continuous	model	Hor.: 8.80±4.06 Ver.: 6.04±2.65 Total: 11.25±6.15	16	RGB
Yamazoe *et al.* [73]	1	X: 0 Y: 0 Z: 2.2	continuous	model	Hor.: 5.3 Ver.: 7.7	5	RGB
Cazzato *et al.* [45]	discrete (wide)	X: unknown Y: unknown Z: 0.7, 1.5, 2.5	continuous	model	Hor.: 4∼12 Ver.: 4.5∼8	6	RGB-D
Sugano *et al.* [74]	continuous (limited)	range of X: 0.22 range of Y: 0.05 range of Z: 0.20	continuous	appearance	Total: 4∼5	3	RGB
Lu *et al.* [25]	continuous (limited)	X: -0.1∼0.09 Y: 0∼0.07 Z: 0.54∼0.67	continuous	appearance	Total: 2∼3	7	RGB
Zhang *et al.* [20]	continuous (limited)	X: -0.1∼0.1 Y: -0.1∼0.1 Z: 0.3∼0.8	continuous	appearance	Total: 6.3	15	RGB

た．　変換は世界座標系内で余弦定理により行った．つまり，提案手法によって予測された視線ベクトルと，頭部位置からディスプレイ上のマーカへ向かう正解視線ベクトルのなす角を計算した．なお，頭部位置について，ここでは Kinect® の Face Tracking 機能から得られた値を使用した．なぜなら，深度情報に基づくためより高精度と思われるためである．

アピアランスベース手法 [20, 25, 74] は提案手法より高い精度を示しているが, 全ての検証はカメラに非常に近い位置 (1 m 以内) において行われている．さらに, 頭部位置の x 軸方向の並進幅が 0.2 m 以下に限られている．よって, これらの手法は非常に狭い範囲でのみ検証されている．

Cazzato ら [45] は視線推定精度の検証をカメラから 0.7 m, 1.5 m, 2.5 m の地点で行った. 彼らは包括的な調査のために被験者を 3 つのグループに分けた [75]. 第 1 グループの被験者はシステムがどのように動作するかを知らされており，システムを経験済みである. 第 2 のグループは，システムの仕組みについては知らされていたが、経験は無い． 第 3 群の被験者はそれら両方について未知である．本研究の実験は、この分類に従えば 5 人の被験者が第 2 グループにあたり、11 人の被験者が第 3 グループにあたる. RSGD では 2 つのグループ間で大きな差が見られなかったものの, 同じ条件で比較するために [45] から第 2 グループと第 3 グループの結果を選択し，　表内に記した. その方法は RGB-D セ

ンサを必要とするが、提案された方法の精度は、垂直方向において同様であり、水平方向においてより高い。 この結果は、 5.2.4 項の結果と矛盾しない.

Yamazoe ら [73] は単眼カメラのみを用いて 2.2 m 離れた被験者の視線を推定した．目の画像のサイズは 30 × 15 ピクセルと報告されているため，RSGD の条件に近い．しかし，被験者は実験の固定された椅子に座っていたため，頭部位置にばらつきはなかった．一方提案手法では，被験者は広い空間の任意の位置に移動することができる．また，提案手法の精度は水平方向と垂直方向の両方で [73] よりも高かった.

上述したように，他の代表的な視線推定方法は，被験者の位置をカメラの近傍に制限するか，または追加のハードウェアを必要とする．これに対して提案手法は，単眼カメラのみを用いて広い空間の任意の頭部位置に適用可能である．さらに，精度は従来の方法と同等かそれ以上である．提案方法はデジタルサイネージやテレビ視聴者の視線推定アプリケーションのための実用的な方法であるといえる.

処理速度に関しては、C++ 環境のシングルスレッドで実装されており、フレームレートは 640 ×480 で 20 fps，Intel Core i7 3.4 GHz CPU で 1920 × 1080 の解像度で 16 fps だった．マルチスレッドで実装すると，両方の画像解像度でフレームレートが 30 fps となった. 報告された処理速度は [45] で 30fps，[73] で 10 fps であった．入力データとマシン仕様は異なるものの，提案手法は効率的にリアルタイムで動作するといえる.

5.3 本章のまとめ

本章では，提案手法の有効性を示すための実験について扱った．第 4 章で作成したデータセットを通して，従来のデータセットより低い目領域解像度条件での比較を行った．その結果，提案手法の虹彩追跡は，影や髪の毛などの外乱の影響を排除し，従来手法よりも高い精度を持つことを確認した.

第 2 章で得た瞼形状によるフィルタにより，水平方向について精度が更に改善される事を確認した．しかし，垂直方向については精度低下が確認されたため，今後の課題としたい.

カメラからの距離に対する誤差の関係に着目した所，提案した虹彩追跡手法はカメラからの距離が離れていても誤差増大が抑えられており，解像度低下に頑健であることを明らかにした.

また，提案手法は単眼カメラのみを用いるため，頭部位置の推定誤差の発生は避けられない．この誤差の注視点推定への影響を検証した．距離センサによる正確な頭部位置情報を使った場合と比較すると，確かにわずかに精度が劣るものの，その差は非常に限定的であり，単眼カメラのみを使用した場合でも RGB-D カメラと遜色無い精度であることを確認した.

第 6 章

結論

本章では，本論文を総括し，今後の課題と将来展望を示す．

6.1 総括

本論文は，広く普及しているデバイス内蔵型単眼カメラの応用に着目し，ユーザの意図推定やインタラクションへの活用といった実社会での利用を想定した，広範囲空間の人物の視線をキャリブレーションフリーで推定することを目標としている．この目標のためには遠い位置にいる人物の低解像度目画像への頑健性や，未知のユーザの広範囲空間内の自由な移動に対応する技術が必要となるが，これまで提案されてきた視線推定技術では実現が困難である事を示した．最も広く使われる視線推定手法である角膜反射法は，投光する近赤外光の到達範囲に限度があり，同時に高解像度目画像を要求するため，使用可能範囲が限られていた．可視光を用いた手法も多く提案されているが，使用前の個人キャリブレーションの必要性や頭部位置・姿勢の制約がある事，目画像が低解像度となる時十分な精度を得られない事を示した．

これらの課題に対し本研究では，モデルベース手法のアプローチを取りつつ，顔特徴点検出および虹彩検出について従来手法よりも精度を上げ，さらに独自に作成したデータセットによる回帰モデルの作成により，カメラに対して広い空間内にいるユーザの視線を個人キャリブレーション無しで定する事を可能とした．提案手法は頭部姿勢推定，虹彩追跡，注視点推定で構成される．頭部姿勢推定においては，まず，カメラから取得した画像内から手法 [62] により検出された顔矩形を入力情報とし，顔特徴点を検出する．顔特徴点検出器は，自然環境光下で撮影された様々な人種・顔向き・表情を含む iBUG FPA データセットを用いて，畳み込みニューラルネットワークを学習することで作成した．実験に

よって，得られた顔特徴点検出器が従来手法より精度が高く，特に顔向き変化が大きいシーンにも頑健であることを示した．続いて，予め再構成した 3 次元顔モデルを検出された顔特徴点位置に当てはめ，頭部姿勢および画像中の眼球中心位置を推定する．虹彩追跡においては，テンプレートマッチングにより虹彩位置を高速に絞り込んだ後，3 次元眼球モデルに基づき，パーティクルフィルタにより高精度に虹彩位置を検出する．密なパーティクル散布と尤度評価の改善により，低解像度目画像においても高精度な虹彩追跡手法を実現した．実験から，従来手法ではカメラからの距離が 1.5 m を超えると虹彩追跡に失敗し水平方向 RMSE が 0.35 m であったが，提案手法では水平方向 RMSE が 0.15 m に改善された事を示した．また，2.5 m の距離においても水平方向 RMSE は 0.22 m に抑えられており，従来より詳細な注視領域を推定可能であることを示した．注視点推定においては，ユーザの頭部姿勢を拘束せず，かつキャリブレーションフリーを実現するため，アピアランスベース手法である [20] に倣い，視線データセット RSGD を作成した．RSGD を用いて回帰器を学習し，広範囲空間内で頭部姿勢を制約しない視線推定を実現した．RSGD には従来のデータセットより遥かに広い範囲の頭部位置を含んでいるが，提案手法は広い範囲においても従来手法がカメラ近傍で実現していた精度と遜色ない視線推定精度を持つ事を確認した．以上のことから，実社会での利用に即した視線推定手法として，本研究は目標に繋がる技術であると考えられる．

6.2 課題

6.2.1 設置環境

提案手法では，ディスプレイのサイズ・位置やカメラの設置位置を固定している．そのため，異なる設置環境にそのまま適用する事はできない．しかしながら，提案手法の出力値を，注視点座標ではなく視線ベクトルとすれば，異なる設置環境への対応が可能である．その際，4.3 節で示した幾何的手法の問題点とは異なり，眼球中心位置の推定誤差に起因する視線ベクトルの誤差は低減可能であると考えられる．

6.2.2 対環境性

実社会での応用を考える時，照明環境や人物の年齢，人種による影響も考慮しなければならない．本研究の顔特徴点追跡手法では，照明環境，カメラ向き，人種，遮蔽の有無，サングラスの有無など様々な環境下で作成されたデータセットで学習し，検証も同じデータセット内で行った．そのため，顔特徴点追跡や頭部姿勢推定には耐環境性があると言える．

しかしながら，虹彩追跡および視線推定に関しては，室内環境で撮影された 20 代前半

の被験者による視線データセットにて学習および評価を行った．そのため，屋外や逆光，サングラスなど，悪条件環境における頑健性の検証は十分ではない．また，年齢による推定精度への影響は特に検証しなければならない．一般に高齢になるほど瞼が垂れ下がり，虹彩追跡を難しくするためである．本論文では，瞼形状に基づくフィルタの効果についても検証した．その結果，本データセットにおいては，水平方向は改善されたものの，垂直方向で精度低下が見られた．今後は年齢，性別，人種，環境といった側面からデータセットを多様化し，これらの条件による影響を検証したい．

6.2.3 垂直方向精度

第 5 章でも述べたように，視線推定の垂直方向の精度の本研究による改善は限定的であった．これは，瞼によって虹彩と強膜 (白目) 間のエッジのうち上側と下側が遮蔽されているため，視線の上下移動に伴う画像中の虹彩エッジの見えの変化が小さいためである．一方で，視線の上下移動に伴い瞼形状は変化する．そのため，瞼形状の追跡結果も合わせて説明変数とすることで，垂直方向の精度が向上する可能性がある．

6.3 展望

本研究では，ディスプレイ上の注視点を推定対象としたが，より汎用性の高いシステムのため，被験者の視線方向推定へと拡張したい．さらに，今回作成した視線データセットの人数および頭部位置の範囲を拡張し，例えば 5m，10m 先などより遠い位置の人物の視線を推定するシステムを実現したい．

また，将来的に実製品への応用も実現したい．これまで学会や展示会で，本研究の視線推定技術を搭載した，離れた位置から視線で操作可能なデジタルサイネージ (図 6.1) や，利用者の注意に応じて行動計画を決定するロボット (図 6.2) を展示してきた．本研究は矢崎総業株式会社との共同研究であり，ドライバーの視線検出装置として，本技術が実社会へ展開される事を最終的な目標としたい．

視線方向によりボタンの選択やページのスクロールが可能なデジタルサイネージ

図 6.1　デジタルサイネージ展示例

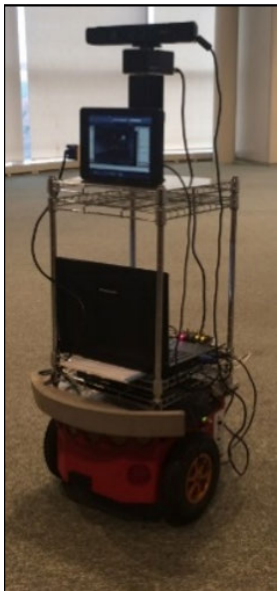


図 6.2　人の注意方向に応じて動く移動ロボット

謝辞

本論文は，筆者が慶應義塾大学 青木義満研究室に在籍中に行った研究成果をまとめたものである．指導教員の青木義満教授には，筆者が学部 4 年生次から現在に至るまでの 6 年間に数多くのご指導，ご指摘を頂いた．その中で，青木教授の新しいことに挑戦する姿勢やどんなことも受け止める寛大な心や柔軟な思考など，研究者としてあるべき姿をご教示頂いた．また，本研究を行う機会を与えて頂くだけでなく，学会発表や他研究室との交流，海外留学など研究者として大きく成長できる環境を与えて頂いた．ここに深謝の意を表する．

そして，慶應義塾大学の池原雅章教授，岡田英史教授，田中敏幸教授，満倉靖恵准教授には副査として本研究を審査していただいた．池原雅章教授には，技術的な指導のみならず，進路や学生生活まで相談をさせて頂いた．岡田英史教授には，学部 4 年時から現在に至るまで，多くの的確なアドバイスを頂いた．田中敏幸教授には，研究室交流を通じ，学科の枠を超えた貴重な意見を多く頂いた．満倉靖恵准教授には，学内の交流会に加えて国際学会や研究会で数多くのご指導を頂いた．ここに感謝の意を表する．

最後に，常に温かい目で見守ってくれた両親や家族，多くの友人に心から感謝する．

参考文献

[1] “acure 自動販売機,” http://www.acure-fun.net/acure/, アクセス日 2017 年 1 月 7 日.

[2] 杵渕哲也, 新井啓之, 宮川勲, 安藤慎吾, 片岡香織, and 小池秀樹, “画像処理による広告効果測定技術,” *NTT* 技術ジャーナル, vol. 7, pp. 16–19, 2009.

[3] T. Ishikawa, “Passive driver gaze tracking with active appearance models,” 2004, technical Report.

[4] Q. Ji and X. Yang, “Real-time eye, gaze, and face pose tracking for monitoring driver vigilance,” *Real-Time Imaging*, vol. 8, no. 5, pp. 357–377, 2002.

[5] W.-B. Horng, C.-Y. Chen, Y. Chang, and C.-H. Fan, “Driver fatigue detection based on eye tracking and dynamk, template matching,” in *IEEE International Conference on Networking, Sensing and Control, 2004*, vol. 1, March 2004, pp. 7–12.

[6] X. Liu, F. Xu, and K. Fujimura, “Real-time eye detection and tracking for driver observation under various light conditions,” in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 2. IEEE, 2002, pp. 344–351.

[7] C. Hennessey, B. Noureddin, and P. Lawrence, “A single camera eye-gaze tracking system with free head motion,” in *Proceedings of the 2006 symposium on Eye tracking research & applications*. ACM, 2006, pp. 87–94.

[8] D. W. Hansen and Q. Ji, “In the eye of the beholder: A survey of models for eyes and gaze,” *IEEE Trans. on Pattern Analysis Machine Intelligence*, vol. 32, no. 3, pp. 478–500, March 2010. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2009.30

[9] T. Ohno, N. Mukawa, and A. Yoshikawa, “Freegaze: a gaze tracking system for everyday gaze interaction,” in *Proceedings of the 2002 symposium on Eye tracking research & applications*. ACM, 2002, pp. 125–132.

[10] J. W. Lee, C. W. Cho, K. Y. Shin, E. C. Lee, and K. R. Park, “3d gaze tracking method using purkinje images on eye optical model and pupil,” *Optics and Lasers in Engineering*, vol. 50, no. 5, pp. 736–751, 2012.

[11] S.-W. Shih, Y.-T. Wu, and J. Liu, “A calibration-free gaze tracking technique,” in *Pro-*

ceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 4, 2000, pp. 201–204 vol.4.

[12] A. Villanueva and R. Cabeza, "A novel gaze estimation system with one calibration point," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, no. 4, pp. 1123–1138, 2008.

[13] S. Milekic, "Gaze-tracking and museums: Current research and implications," in *Museums and the Web*, 2010, pp. 61–70.

[14] Z. Zhu and Q. Ji, "Eye gaze tracking under natural head movements," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 918–923.

[15] Z. Zhu, Q. Ji, and K. P. Bennett, "Nonlinear eye gaze mapping function estimation via support vector regression," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 1132–1135.

[16] T. Ohno and N. Mukawa, "A free-head, simple calibration, gaze tracking system that enables gaze-based interaction," in *Proceedings of the 2004 symposium on Eye tracking research & applications*. ACM, 2004, pp. 115–122.

[17] F. Lu, Y. Sugano, T. Okabe, and Y. Sato, "Adaptive linear regression for appearance-based gaze estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 10, pp. 2033–2046, 2014.

[18] B. Noris, K. Benmachiche, and A. Billard, "Calibration-free eye gaze direction detection with gaussian processes," in *In Proceedings of the International Conference on Computer Vision Theory and Applications*, no. LASA-CONF-2007-018, 2008.

[19] C. L. L. Jerry and M. Eizenman, "Convolutional neural networks for eye detection in remote gaze estimation systems," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1. Citeseer, 2008.

[20] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," *CoRR*, vol. abs/1504.02863, 2015. [Online]. Available: http://arxiv.org/abs/1504.02863

[21] D. Pomerleau and S. Baluja, "Non-intrusive gaze tracking using artificial neural networks," in *AAAI Fall Symposium on Machine Learning in Computer Vision, Raleigh, NC*, 1993, pp. 153–156.

[22] K. Liang, Y. Chahir, M. Molina, C. Tijus, and F. Jouen, "Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression," in *Proceedings of the 2013 Conference on Eye Tracking South Africa*. ACM, 2013, pp. 17–23.

[23] O. Williams, A. Blake, and R. Cipolla, "Sparse and semi-supervised visual mapping

with the sˆ 3gp," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 230–237.

[24] W. Sewell and O. Komogortsev, "Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2010, pp. 3739–3744.

[25] F. Lu, T. Okabe, Y. Sugano, and Y. Sato, "Learning gaze biases with head motion for head pose-free gaze estimation," *Image and Vision Computing*, vol. 32, no. 3, pp. 169–179, 2014.

[26] K. A. F. Mora and J.-M. Odobez, "Gaze estimation from multimodal kinect data," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 25–30.

[27] J. Choi, B. Ahn, J. Park, and I.-S. Kweon, "Appearance-based gaze estimation using kinect." in *URAI*, 2013, pp. 260–261.

[28] Y. Sugano, Y. Matsushita, and Y. Sato, "Unconstrained gaze estimation with learning from mouse operations," *MIRU2009*, pp. 266–273, 2009.

[29] ——, "Appearance-based gaze estimation using visual saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 329–341, Feb 2013.

[30] E. Wood and A. Bulling, "Eyetab: Model-based gaze estimation on unmodified tablet computers," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '14. New York, NY, USA: ACM, 2014, pp. 207–210. [Online]. Available: http://doi.acm.org/10.1145/2578153.2578185

[31] H. Wu, Q. Chen, and T. Wada, "Conic-based algorithm for visual line estimation from one image," in *Proc. the Sixth IEEE Int. Conf. on Automatic Face and Gesture Recognition*, ser. FGR' 04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 260–265. [Online]. Available: http://dl.acm.org/citation.cfm?id=1949767.1949816

[32] W. Zhang, T.-N. Zhang, and S.-J. Chang, "Eye gaze estimation from the elliptical features of one iris," *Optical Engineering*, vol. 50, no. 4, pp. 047 003–047 003, 2011.

[33] Y. Matsumoto and A. Zelinsky, "An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement," in *Proc. the Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition 2000*, ser. FG '00. Washington, DC, USA: IEEE Computer Society, 2000, p. 499. [Online]. Available: http://dl.acm.org/citation.cfm?id=795661.796234

[34] Y. Kitagawa, H. Wu, T. Wada, and T. Kato, "On eye-model personalization for automatic visual line estimation," *PRMU2007*, vol. 106, no. 469, pp. 55–60, 2007.

[35] J. Chen and Q. Ji, "3d gaze estimation with a single camera without ir illumination," in

Proc. 2008 19th Int. Conf. on Pattern Recognition, Dec 2008, pp. 1–4.

[36] W.-z. Zhang, Z.-c. Wang, J.-k. Xu, and X.-y. Cong, "A method of gaze direction estimation considering head posture," *Int. Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 2, pp. 103–112, 2013.

[37] J. Orozco, O. Rudovic, J. Gonzàlez, and M. Pantic, "Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises," *Image and vision computing*, vol. 31, no. 4, pp. 322–340, 2013.

[38] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 802–815, 2012.

[39] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Ave, "Automatic calibration of 3d eye model for single-camera based gaze estimation," *Trans. IEICE*, pp. 998–1006, 2011.

[40] D. Cazzato, A. Evangelista, M. Leo, P. Carcagnì, and C. Distante, "A low-cost and calibration-free gaze estimator for soft biometrics: An explorative study," *Pattern Recognition Letters*, 2015.

[41] N. Robertson, I. Reid, and J. Brady, "What are you looking at? gaze estimation in medium-scale images," in *Proc. the HAREM Workshop (in assoc. with BMVC), Oxford, UK*, vol. 9, 2005.

[42] S. O. Ba and J.-M. Odobez, "Recognizing visual focus of attention from head pose in natural meetings," *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 16–33, 2009.

[43] M. J. Reale, S. Canavan, L. Yin, K. Hu, and T. Hung, "A multi-gesture interaction system using a 3-d iris disk model for gaze estimation and an active appearance model for 3-d hand pointing," *IEEE Trans. on Multimedia*, vol. 13, no. 3, pp. 474–486, 2011.

[44] K. Alberto Funes Mora and J.-M. Odobez, "Geometric generative gaze estimation (g3e) for remote rgb-d cameras," in *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1773–1780.

[45] D. Cazzato, M. Leo, and C. Distante, "An investigation on the feasibility of uncalibrated and unconstrained gaze tracking for human assistive applications by using head pose estimation," *Sensors*, vol. 14, no. 5, pp. 8363–8379, 2014.

[46] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[47] B. Czupryński and A. Strupczewski, *High Accuracy Head Pose Tracking Survey*. Cham: Springer International Publishing, 2014, pp. 407–420. [Online]. Available:

http://dx.doi.org/10.1007/978-3-319-09912-5_34

[48] F. De la Torre and J. F. Cohn, "Facial expression analysis," in *Visual analysis of humans*. Springer, 2011, pp. 377–409.

[49] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[50] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer vision and image understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[51] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor, "View-based active appearance models," *Image and vision computing*, vol. 20, no. 9, pp. 657–664, 2002.

[52] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2d+ 3d active appearance models," in *CVPR (2)*, 2004, pp. 535–542.

[53] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[54] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476–3483.

[55] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*. Sydney, Australia: IEEE, Dec. 2013, pp. 354–361.

[56] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. the IEEE Int. Conf. on Computer Vision Workshops*, 2013, pp. 397–403.

[57] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[58] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 896–903.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke,

and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[61] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385*, 2015.

[62] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[63] T. Baltrušaitis, P. Robinson, and L. P. Morency, “Openface: An open source facial behavior analysis toolkit,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–10.

[64] “Oracle data mining 概要,” 2 2017, https://docs.oracle.com/cd/E16338_01/datamine.112/e48231/regress.htm.

[65] S. Ullman, *The interpretation of visual motion.* Massachusetts Inst of Technology Pr, 1979.

[66] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[67] M. Isard and A. Blake, “Condensation-conditional density propagation for visual tracking,” *International journal of computer vision*, vol. 29, no. 1, pp. 5–28, 1998.

[68] K. A. Funes Mora, F. Monay, and J.-M. Odobez, “Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras,” in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA ’14. New York, NY, USA: ACM, 2014, pp. 255–258. [Online]. Available: http://doi.acm.org/10.1145/2578153.2578190

[69] A. Villanueva, V. Ponz, L. Sesma-Sanchez, M. Ariz, S. Porta, and R. Cabeza, “Hybrid method based on topography for robust detection of iris center and eye corners,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 9, no. 4, pp. 25:1–25:20, Aug. 2013. [Online]. Available: http://doi.acm.org/10.1145/2501643.2501647

[70] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, “Gaze locking: passive eye contact detection for human-object interaction,” in *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 2013, pp. 271–280.

[71] K. Tamura and Y. Aoki, “Eyelid and iris tracking method with novel eye models,” in *System Integration (SII), 2013 IEEE/SICE Int. Symp. on*. IEEE, 2013, pp. 449–453.

[72] L. Jianfeng and L. Shigang, “Eye-model-based gaze estimation by rgb-d camera,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

Workshops, 2014, pp. 592–596.

[73] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe, "Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions," in *Proceedings of the 2008 symposium on Eye tracking research & applications*. ACM, 2008, pp. 245–250.

[74] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, "An incremental learning method for unconstrained gaze estimation," in *Proc. European Conf. on Computer Vision*. Springer, 2008, pp. 656–667.

[75] F. Lu, T. Okabe, Y. Sugano, and Y. Sato, "A head pose-free approach for appearance-based gaze estimation." in *BMVC*, 2011, pp. 1–11.