卒 業 論 文 概 要 書

Summary of Bachelor's Thesis

Date of submission: 1/31/2023

学科名 Department	情報通信	氏名 Name	杉山秀治	指 導 教 員 Advisor	渡辺 裕 ㊞
研究指導名 Research guidance	オーディオビジュアル 情報処理研究	学籍番号 Student ID number	1W192198-5 (CD)		
研究題目 Title	アピアランスベースの視線推定における精度改善 Accuracy Improvement in Appearance-Based Gaze Estimation				

1. まえがき

近年，エンゲージメント調査やメタバースなど人間の視線情報はあらゆるところで活用される．人間の視線に関する需要が高いことから，視線推定は Computer Vision の分野で注目されている．視線推定の手法のうち，アピアランスベースの手法は，顔や目の画像から直接視線方向を推定するアプローチをとる．この手法は，入力画像に対する高度な要求がないという利点があるものの，撮影環境における外的要素や被写体となる人物ごとの際などの考えうる変動要因が多いことから，正確な推定が難しいという問題がある．しかし，2015 年に Zhang ら[1]が, CNN を用いることで精度は十分でないものの，変動要因に起因する推定の困難さをかなり解決した．そのため，これ以降 CNN ベースのモデルがアピアランスベースの視線推定の手法の主流である．

本論文では，アピアランスベースかつ人物に依存しない視線推定手法の一つである L2CS-Net[2] の精度改善を目的とする．まず，L2CS-Net の問題点を挙げ，それに対する改善手法を提案する．そして，評価実験により提案手法の有効性を示す．

2. 従来手法とその改善点

2.1 L2CS-Net

L2CS-Net はロバストな視線推定を可能とする三次元視線推定のモデルである．顔画像を入力とし，バックボーンとして ImageNet で学習済みの ResNet-50 を用いる．さらに，得られた顔の特徴量から視線方向を推定する．L2CS-Net は，視線方向の推定に，一般的な推定値であるベクトルではなく，ピッチ角とヨー角という三次元空間におけるオイラー角を独立に推定しているという特徴がある．L2CS-Net のネットワークを図 1 に示す．

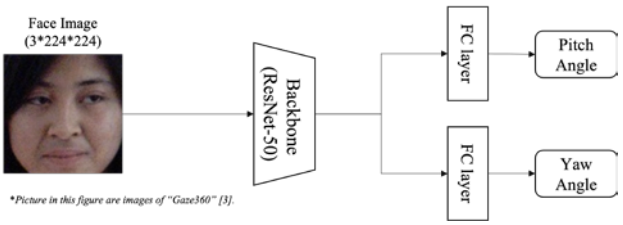

図 1 L2CS-Net のネットワーク

L2CS-Net は，ロバストな三次元視線推定を可能とするために作成されたデータセット Gaze360[3] において，CNN を用いたアピアランスベースの視線推定手法としては最良の手法となっている．

2.2 L2CS-Net の問題点

L2CS-Net は，Gaze360 において平均角度誤差が 10.41°という精度である．角度誤差とは推定視線方向と正解視線方向とのなす角であり，誤差を評価する指標である．この平均角度誤差より，L2CS-Net はおおよその視線は推定できているといえる．しかし，推定対象の人物の顔とその人物の目線の先の対象物との距離が$1m$の場合，10.41°という角度誤差により生じるずれは約$18cm$となる．正確な視線推定を求められる際，この性能は不十分であるため，L2CS-Net を改善する必要がある．

また L2CS-Net は, 顔画像のみを入力に用いるモデルである．そのため，視線方向の詳細な情報を含む両目のサイズが顔画像全体のサイズに対して小さいことと，顔画像には視線方向の推定に必要のない情報を含んでいることが L2CS-Net の問題点として挙げられる．これらの問題点より，L2CS-Net は視線方向の情報を含む目の分析が不十分であると考える．したがって，L2CS-Net と比較して，目の分析をさらに行うことで，精度の改善が期待される．

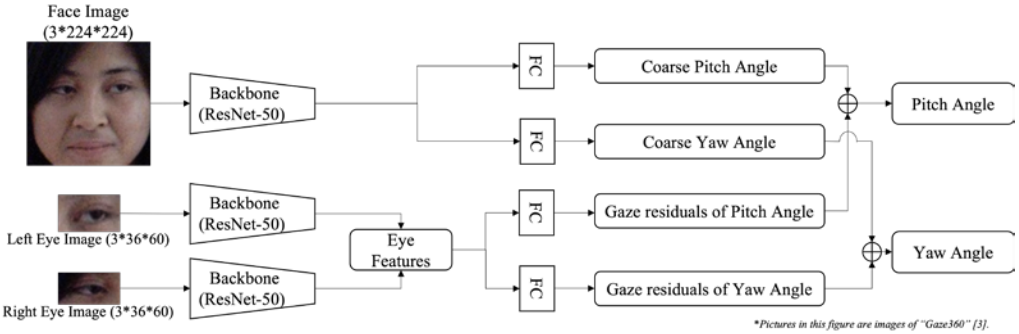

図 2 提案手法の視線推定ネットワーク

3. 提案手法

本論文では, 第 2.2 節で述べた問題点に対して, 目の詳細な情報を付加し，その情報を適切に利用するモデルを提案する．提案手法では，顔画像と対応する両目の画像を入力に使用する．提案手法の視線推定ネットワークを図 2 に示す.

顔画像と対応する両目の入力画像に対し, ImageNet で学習済みの ResNet-50 をバックボーンとして特徴量抽出を行う．提案手法では, ResNet-50 の一層目の畳み込み層以外をファインチューニングするため，画像サイズや構造の違いから顔と目のバックボーンを分離する．さらに，顔の特徴量から大体の視線方向（粗い視線），両目の特徴量から粗い視線を正しい視線に近づけるような残差（視線残差）を推定する．最終的に粗い視線方向と視線残差を足し合わせて視線方向を決定する．粗い視線，視線残差ともに，ピッチ角とヨー角を独立に推定する.

次に提案手法で用いる損失関数について述べる. 推定した視線方向の角度と正解の視線方向の角度の L1 ロスをとる．提案手法では，粗い視線と視線残差の二つの角度を推定していることから，視線残差に対する重み係数をγとして定義する. 提案手法で用いる損失関数を式(1)に示す.

$$Loss = l_{L1}(\theta, \theta') + \gamma l_{L1}(\theta, \theta'') \qquad (1)$$

ここで，正解視線方向の角度をθ，粗い視線方向の角度をθ'，粗い視線方向と視線残差を足し合わせた推定した視線方向の角度をθ''とする.

4. 提案手法による評価実験

提案手法の評価実験で用いるデータセットは, Gaze360 である．評価指標には，平均角度誤差, Mean Angular Error(MAE)を用いる．角度誤差は式(2)で算出される.

$$Angular\ Error = \arccos\left(\frac{\boldsymbol{g}\cdot\hat{\boldsymbol{g}}}{\|\boldsymbol{g}\|\|\hat{\boldsymbol{g}}\|}\right) \qquad (2)$$

ここで，$\boldsymbol{g}$は正解の視線方向のベクトル，$\hat{\boldsymbol{g}}$は推定した視線方向のベクトルである.

実験条件としては，L2CS-Net との比較のためにエポック数を 50，バッチサイズを 16，最適化手法を Adam，学習率を 0.00001 とした.

視線残差に関するロスの重みの値が$\gamma = 2$のときの提案手法の実験結果を表 1 に示す.

表 1 提案手法の実験結果

	L2CS-Net	提案手法 ($\gamma = 2$)
MAE(°)	10.41	**10.16**

表 1 の実験結果より，提案手法は従来手法である L2CS-Net より，平均角度誤差を0.25°改善できることがわかった.

5. 結論

本論文では, 従来手法である L2CS-Net の問題点として，目の分析が不十分であることを挙げた. そこで，顔画像のみでなく両目の画像を入力することで，目の情報を追加するモデルを提案した. さらに，顔の特徴量から粗い視線，目の特徴量から視線残差を推定することで，予備実験で明らかになった目の特徴量を十分に活用できないという課題を解決した.

参考文献

[1] Xucong Zhang *et al.*, “Appearance-Based Gaze Estimation in the Wild”, CVPR, pp.4511-4520, Jun. 2015.

[2] Ahmed A. Abdelrahman *et al.*, “L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments”, arXiv preprint arXiv:2203.03339(2022).

[3] Petr Kellnhofer *et al.*, “Gaze360: Physically Unconstrained Gaze Estimation in the Wild”, ICCV, pp.6912-6921, Oct. 2019.

2022 年度　卒業論文

アピアランスベースの視線推定における精度改善

Accuracy Improvement in Appearance-Based Gaze Estimation

提出日　2023 年 1 月 31 日

指導教員　渡辺 裕 教授

早稲田大学基幹理工学部 情報通信学科

1W192198-5

杉山　秀治

目次

第１章　序論

１．１　研究の背景

視線推定とは画像から人の視線を推定するタスクのことを指す．近年では，人間の視線はマーケティングにおけるエンゲージメント調査や AR，VR などの様々な場所で活用されている．また，最近ではメタバースといったワードが多く使われるように，仮想空間への拡張の指向も高まっていることから，人間の視線の需要があることが予想される．このように人間の視線情報は有益であるため，視線推定は Computer Vision の分野で注目されている．

視線推定は，大きく二つの軸においてそれぞれ二種類ずつ，計四つのクラスに分類される．一つ目の軸は，視線方向を推定するためのアプローチの違いであり，モデルベースの手法とアピアランスベースの手法に分類される．二つ目の軸は，視線方向を推定するための学習方法の違いであり，人物に依存しない手法と個人に特化した手法に分類される．四つの手法について以下で詳細に述べる．

まず，一つ目の軸に関する手法について述べる．モデルベースの手法は，画像に映る角膜の反射や虹彩の様子などの目の情報を幾何学的にモデル化し，そのモデルから視線方向を推定するアプローチのことである．この手法の場合，高画質な画像でないと十分に目の情報を取得できないため，カメラの性能に依存する面が大きく，高価なハードウェアを要求する．一方で，アピアランスベースの手法は，顔や目の画像から直接視線方向を推定するアプローチのことである．この手法はモデルベースの手法と異なり，高度な入力を要求しないので，導入が容易であるという利点がある．しかし，アピアランスベースの手法を実生活で使用する場合には，想定されるさまざまな変動要因を考慮する必要があるため，汎化性能のあるモデルの作成が難しいという点が挙げられる．この問題に対して， 2015 年に Zhang[1]らが Convolutional Neural Network(CNN)を用いたアピアランスベースの視線推定のモデルを発表した．これは変動要因に起因するアピアランスベースの視線推定が困難であることに対して，CNN を用いることで視線方向の推定が可能になるというものであった．したがってこれ以降，アピアランスベースの視線推定において CNN ベースのモデルが主流である．

次に，二つ目の軸に関する手法について述べる．人物に依存しない手法は，複数人物の画像を用いて学習した汎用的な手法のことである．一方，個人に特化した手法は，汎用的な視線推定モデルに対して特定の人物の画像を用いて再学習させる手法のことである．現実世界での運用を考えると人物に依存しないロバストなモデルの需要がある．

しかし，ある個人に特化したモデルを作成することで，より正確な視線推定を行うことが可能となるとなることから，用途に合わせて使い分けられる．

以上のように，視線推定はモデルベースの手法とアピアランスベースの手法の分類と，人物に依存しない手法と個人に特化した手法の分類と四つのクラスが存在するが，日常生活における導入を考えた際に，リアルタイム性やコスト面からアピアランスベースかつ人物に依存しない手法の方の需要が高いことがわかる．

１．２　関連研究と問題点，および研究目的

第 1.1 節で述べたようにアピアランスベースの視線推定技術は，CNN を用いることで大まかな視線推定を可能としている．しかしながら，視線推定というタスク自体が難しいこともあり，完全に視線方向を推定できないのが現状である．本節では，本論文で比較対象とする視線推定モデルを紹介し，それの問題点を述べる．さらに，研究目的について述べる．

本論文の比較対象となる手法で，アピアランスベースかつ人物に依存しない視線推定手法の一つである L2CS-Net[2]は，Gaze360[3]というデータセットにおいて，CNN ベースの手法の中では最も精度の高いモデルである．

本論文でも使用するデータセット Gaze360 は，ロバストな三次元視線推定を可能とするために作成された，視線推定のためのデータセットの中では大規模なデータセットである．具体的には，Gaze360 は環境や人物に左右されない視線推定を行うために，238 人の被験者に対して，屋内外の環境で撮影および視線方向が記録されたデータセットである．すなわち，L2CS-Net は，人物の顔画像を用いて学習するアピアランスベースの手法であり，Gaze360 をデータセットに用いるため人物に依存しないロバストなモデルである．具体的なモデルの構造については第 2 章で述べる．

L2CS-Net の精度は，Gaze360 において平均角度誤差が10.41°という性能を示している．ここで角度誤差とは，推定した視線方向のベクトルと正解である視線方向のベクトルのなす角であり，推定した視線方向の誤差を示す角度のことである．この10.41°いう角度についておおまかな視線方向は推定できていると言える．しかし，推定対象となる人物の顔とその目線の先の物体の距離が 1m であった場合，10.41° の角度誤差があると，約 18cm のずれが生じる．このずれは視線推定の目的に応じて意味が異なるが，第 1.1 節で述べたエンゲージメント調査においては消費者が見ている商品が全く異なる可能性が生じるため問題となる．このように大まかな視線方向の推定では，現実世界での運用の際に支障をきたす場合がある．したがって，L2CS-Net は改善の必要がある．ゆえに本論文では L2CS-Net を比較対象として，視線推定の精度改善を目的とする．以下ではそのための提案手法を示し，その手法の有効性を実験により示す．

１．３　本論文の構成

以下に本論文の構成を示す.

第１章　研究の背景及び目的について述べる．まず，研究の背景について述べたのち，従来手法である L2CS-Net の問題点を示し，研究目的を明らかにする．

第２章　CNN ベースの視線推定手法について述べる．まず，CNN ベースの視線推定手法の概要について述べる．続いて，第 1 章で述べた L2CS-Net について詳細に説明する．さらに，他の従来手法である CA-Net について説明する．

第３章　まず， L2CS-Net の構造的な問題点を述べる．さらに，その問題点に対する提案手法を示し，予備実験を行う．

第４章　予備実験で明らかになった課題に対して，本論文における提案手法について述べる．

第５章　提案手法の有効性を評価実験により確認する．

第６章　本研究の結論と今後の課題を述べる．

第２章　CNN ベースの視線推定手法

２．１　まえがき

本章では，本論文で研究対象とする CNN ベースの視線推定手法について述べる．まず，CNN ベースの視線推定手法の概要を述べる．さらに，従来の CNN ベースの視線推定手法について説明する．最初に本論文での比較対象となる，L2CS-Net について説明する．次に，他の従来手法である CA-Net[4]について説明する．

２．２　従来の CNN ベースの視線推定手法

CNN ベースの視線推定手法とは，対象人物の外見の画像を入力とし，入力画像に対して CNN で特徴量抽出を行ったのち，その特徴量から視線方向を回帰により推定する手法のことである. 外見の画像として顔画像や目の画像を入力に用いる手法が現在主流である．第 1 章で述べたように，アピアランスベースの視線推定において CNN を用いて推定を行う技術は Zhang らによって最初に考案された.

２.２.１　L2CS-Net

L2CS-Net は, Abdelrahman らによって提案された三次元視線推定の手法である. L2CS-Net は，Gaze360 データセットにおいて，CNN ベースの視線推定のモデルの中で最も精度の良いモデルである．L2CS-Net のネットワークを図 ２.１に示す.

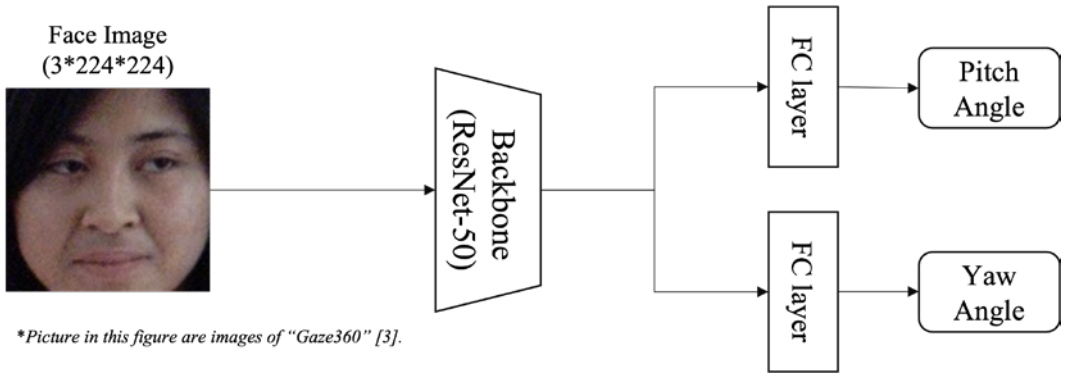

図 ２.１　L2CS-Net のネットワーク

図　2.1 の通り，L2CS-Net は顔画像を入力に用いて視線方向を推定する．顔画像の特徴量抽出には，ImageNet で学習済みの ResNet-50 をバックボーンとして用いる．さらに，得られた特徴量を全結合層に入力し，視線方向を推定する．また図　2.1 が示すように，L2CS-Net は視線方向の推定に必要となるピッチ角とヨー角を独立に推定している．　三次元視線推定のモデルの多くは，視線方向の推定にベクトルを用いるのに対して，L2CS-Net は二つの角度を用いている点が L2CS-Net の特徴である．

ここで，視線方向の決定に必要な角度について補足する．

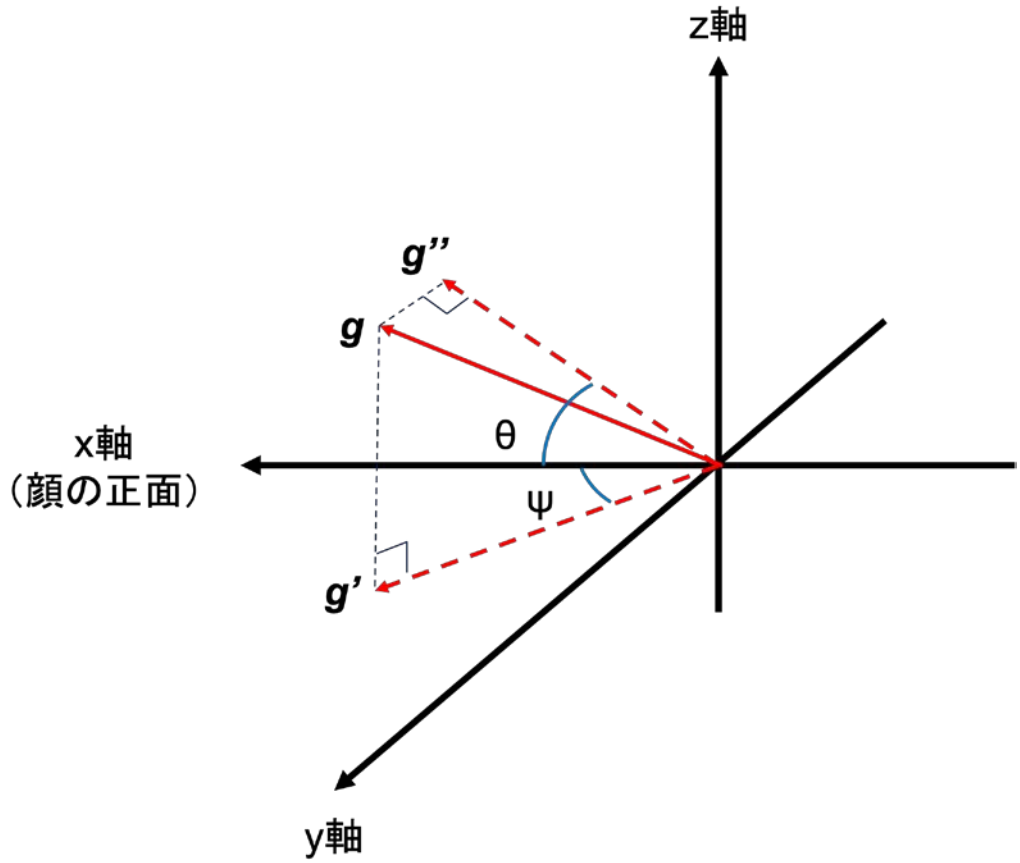

図　2.2　視線方向の決定の例

図　2.2 は，顔の正面をx軸とする，顔を固定座標系とした三次元空間である．そして，視線方向が$\boldsymbol{g}$であったとき，xy平面に射影したベクトルは$\boldsymbol{g}'$，xz平面に射影したベクトルは$\boldsymbol{g}''$である．また，三次元空間におけるオイラー角では，z軸回りの回転角をヨー角 ϕ，y軸回りの回転角をピッチ角 θ と呼ぶ．したがって，ヨー角とピッチ角を求めることで，$\boldsymbol{g}'$と$\boldsymbol{g}''$が求められるので，視線方向 $\boldsymbol{g}$を決定することが可能となる．さらに，　L2CS-Net では，ヨー角とピッチ角を独立に求めることで，従来の CNN ベースの視線推定技術より精度の良い推定を可能としている．

次に，L2CS-Net のロスについて述べる．L2CS-Net では，視線方向を一定の角度で離散化し，分類問題に近似した場合の Cross Entropy 誤差と，推定した角度と正解角度の MSE ロスを足し合わせたロスを用いている．式(2.1)が L2CS-Net が用いるロスである．

$$Loss = l_{Cross\ Entropy} + l_{MSE} \tag{2.1}$$

２.２.２　CA-Net

CA-Net は，Yihua らによって提案された三次元視線推定の手法である．CA-Net の大まかなネットワークを図 ２.３に示す.

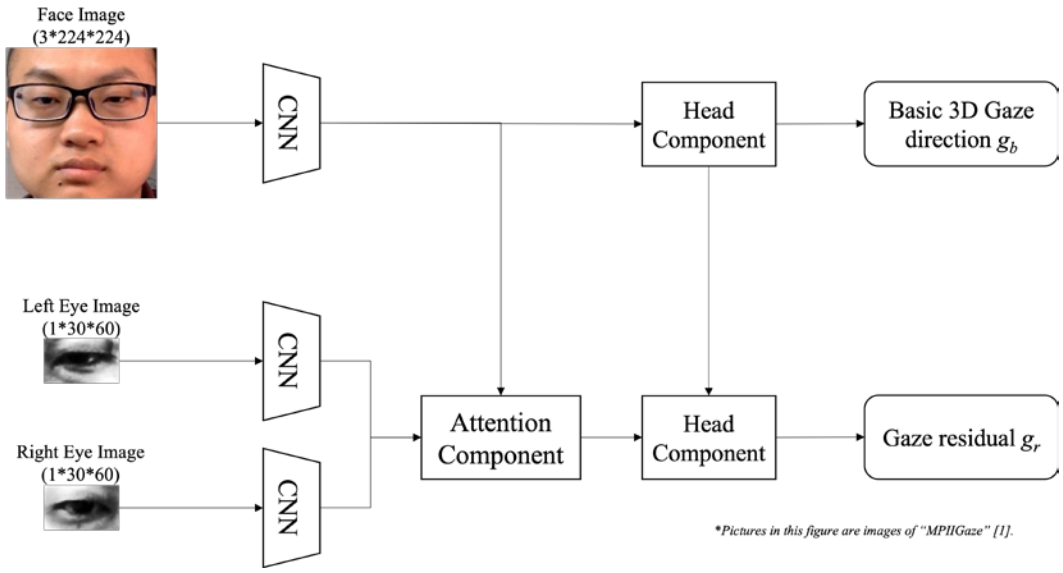

図 ２.３　CA-Net のネットワーク

CA-Net は，顔画像から基本的な視線方向（粗い視線），その顔画像における眼球画像から残差となるより細かな視線方向（視線残差）を推定し，それらを足し合わせることで視線方向を推定するネットワークである．さらに，顔の画像と両目の画像との関係性を考慮するために，Attention Component と Gated Recurrent Unit(GRU)を用いた Head Component を使用している．

ここで Attention[5]とは，自然言語処理(NLP)の分野を中心に発展した技術の一つで，入力されたデータの重要度を算出する仕組みことである．近年では，Attention は自然言語に限らず画像認識に用いられる技術となっており，入力画像のうちの重要な領域のみを認識可能となる．CA-Net では，目の細粒な特徴を捉えるために Attention Component を導入している．

また，GRU[6]も自然言語処理の分野を中心に用いられる技術の一つであり，Long Short Term Memory(LSTM)を簡略化したモデルである．GRU は更新ゲートとリセットゲートにより，LSTM 同様長期的特徴の学習が可能である．CA-Net では，顔の特徴朝から粗い視線，目の特徴量から視線残差の推定に適切な特徴量のみを取り出すために GRU を導入している．

CA-Net では図 2.3 に示すように，顔画像，両目の画像それぞれを CNN に入力し，特徴量を抽出した上で，Attention Component に入力する．すなわち，顔の特徴量f_f，左目の特徴量f_l，右目の特徴量f_rを Attention Component に入力する．ここで，Attention Component のネットワークを図 2.4 に示す．

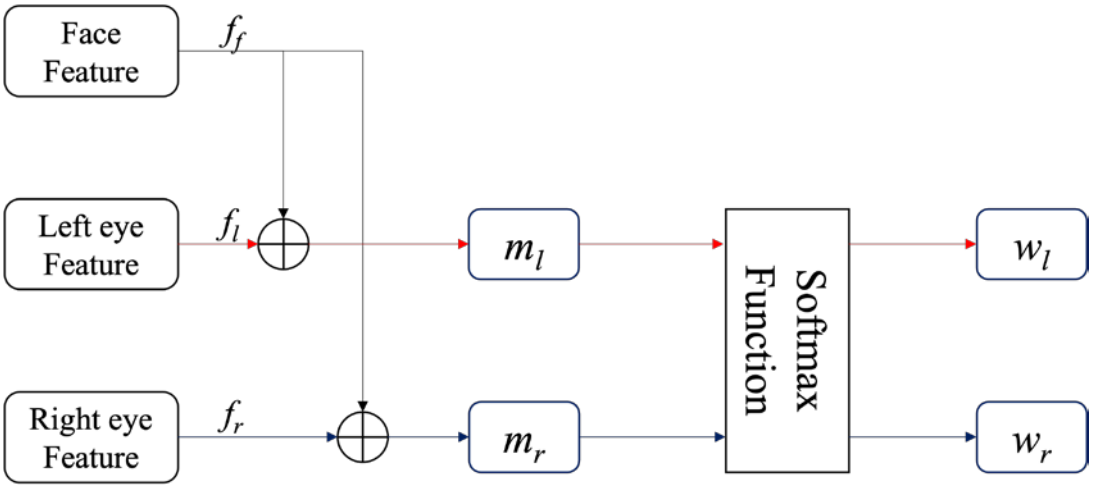

図 2.4 Attention Component のネットワーク

図 2.4 に示すように，Attention Component は顔の特徴量と左目の特徴量，顔の特徴量と右目の特徴量の Attention をとり，Attention Weight m_l, m_rを算出する．さらに，左右の目の Attention Weight を Softmax 関数により正規化し，左右の目の重みw_l, w_rを得る．Attention Weight および左右の目の重みの算出式は，以下の式(2.2)，式(2.3)，式(2.4)に示す．

$$m_l = v^T \tanh\left(W_1^T f_f + W_2^T f_l\right) \tag{2.2}$$

$$m_r = v^T \tanh\left(W_1^T f_f + W_2^T f_r\right) \tag{2.3}$$

$$[w_l,\ w_r] = softmax([m_l, m_r]) \tag{2.4}$$

そして，上記で得られた左右の目の重みをそれぞれの特徴量にかけて足し合わせ，式(2.5)に示すように目の特徴量f_eを生成する．

$$f_e = w_l * f_l + w_r * f_r \tag{2.5}$$

このように，Attention Component により顔画像と左右の目の関係性を考慮した目の特徴量を生成可能である．

次に，Head Component について述べる．Head Component のネットワークを図 2.5 に示す．

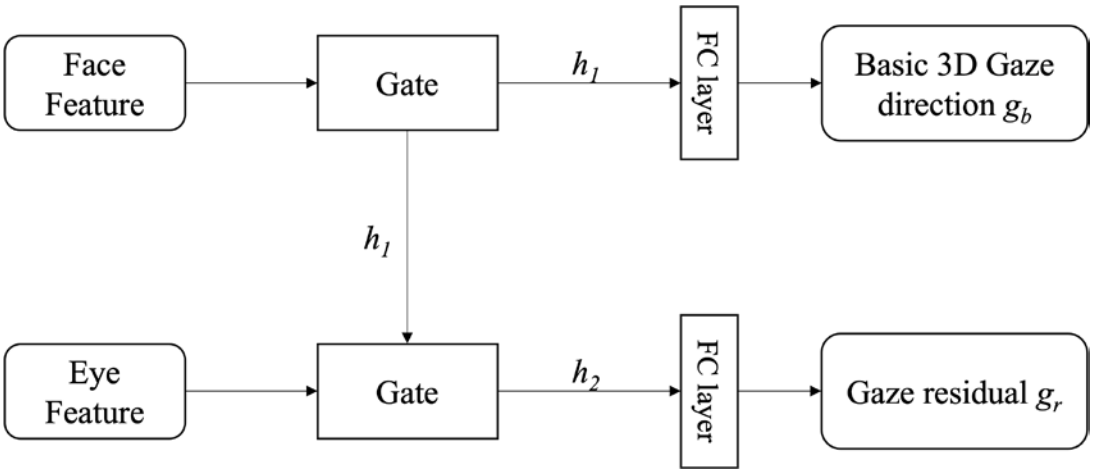

図 2.5 Head Component のネットワーク

前述した Attention Component で得られた顔の特徴量f_fと目の特徴量f_eを GRU に基づくゲートに通すことで，視線残差に適切な特徴量のみを抽出可能となる．各ゲートの定義式を式(2.6)~式(2.9)に示す．式(2.6)は更新ゲート，式(2.7)はリセットゲートの定義式である．

$$z_i = \sigma(W_z \cdot [h_i, f]) \tag{2.6}$$
$$r_i = \sigma(W_r \cdot [h_i, f]) \tag{2.7}$$
$$\widetilde{h_{i+1}} = ReLU(W_h \cdot [r_i * h_i, f]) \tag{2.8}$$
$$h_{i+1} = (1 - z_i) * h_i + z_i * \widetilde{h_{i+1}} \tag{2.9}$$

図 2.5 に示すように，顔の特徴量f_fは一つ目のゲートを通過して顔の適切な特徴量h_1，目の特徴量は二つの目のゲートで顔の適切な特徴量と共に入力されることで，顔との関係性を考慮した視線残差の推定に適した特徴量h_2となる．そして，顔と目の適切な特徴量をそれぞれ全結合層に入力し，粗い視線$\boldsymbol{g_b}$と視線残差$\boldsymbol{g_r}$を推定する．

最後に，CA-Net が用いるロスについて述べる．まず角度誤差とは，正解方向のベクトル$\boldsymbol{g}$と推定方向ベクトル$\boldsymbol{g'}$のなす角のことであり，推定結果の誤差を示す指標であ

る．角度誤差の算出式を式(2.10)に示す．

$$L_{angular}(\boldsymbol{g}, \boldsymbol{g}') = \arccos\left(\frac{\boldsymbol{g} \cdot \boldsymbol{g}'}{\|\boldsymbol{g}\|\|\boldsymbol{g}'\|}\right) \tag{2.10}$$

CA-Net が用いるロスは，視線方向の算出が$\hat{\boldsymbol{g}} = \boldsymbol{g_b} + \boldsymbol{g_r}$であることを踏まえて以下の通りである．

$$Loss = \alpha L_{angular}(\boldsymbol{g}, \boldsymbol{g_b}) + \beta L_{angular}(\boldsymbol{g}, \hat{\boldsymbol{g}}) \tag{2.11}$$

ここでα, βは粗い視線，視線残差に関するロスの重みであり，CA-Net では$\alpha = 1, \beta = 2$が最良であったとしている．

２．３　むすび

本章では，本研究でテーマとする CNN ベースの視線推定手法の導入を述べた．さらに本研究の比較対象となる L2CS-Net の概要およびネットワークを示した．また，他の CNN ベースの視線推定手法である CA-Net の概要およびネットワークも示した．

第３章　予備実験

３．１　まえがき

本章では，従来手法である L2CS-Net の問題点を述べ，それに対する改善手法を検討するための予備実験を行う．

３．２　L2CS-Net の問題点

第 2 章で述べた通り，L2CS-Net は顔画像のみをネットワークに入力とするモデルである．顔画像のみをネットワークに入力する場合，視線方向の情報を持つ両目のサイズは顔全体のサイズに対して小さいことと，顔画像は視線方向の推定に必要のない情報も含んでいることが L2CS-Net の問題点として挙げられる．これらの問題点より，L2CS-Net は目の分析が不十分であると考える．したがって，より正確な視線方向を推定するためには粒度の高い目の特徴を捉えることが可能となるモデルが必要であると考えられる．

３．３　予備実験

３．３．１　予備実験における提案手法

第 3.2 節で述べた，従来手法 L2CS-Net の問題点に対して，顔画像と共に両目の画像を入力に用いることで，詳細な視線方向の推定に必要な目の情報を付加できると期待する．そのため，予備実験によりこの改善手法を検討する．

予備実験で検討する手法を提案手法 A とし，提案手法 A について述べる．前述のとおり，提案手法 A は，顔画像とそれに対応する両目の画像を入力に使用するモデルである．提案手法 A の視線推定ネットワークを図 3.1 に示す．

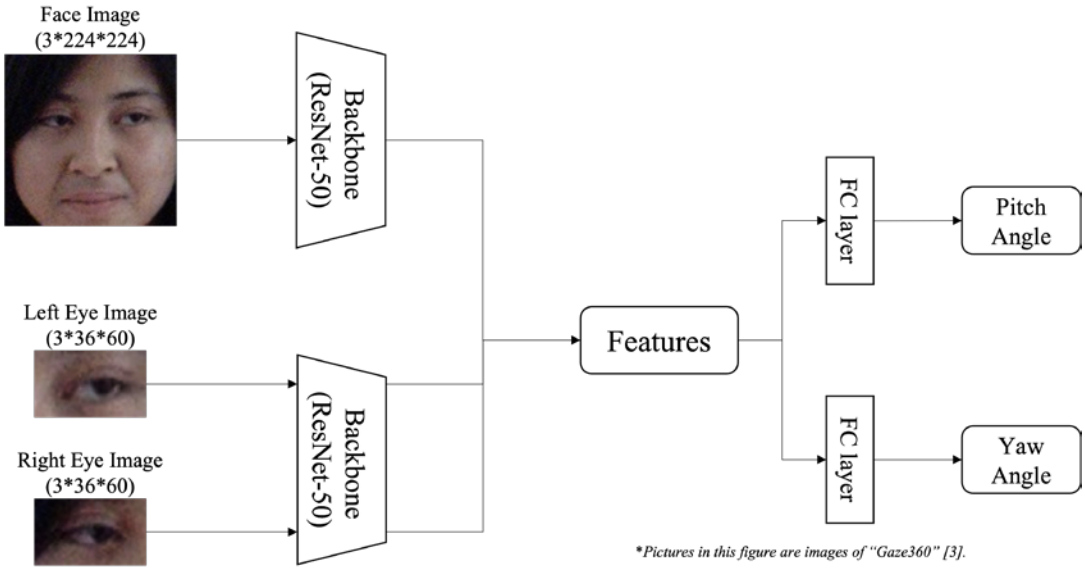

図 3.1 提案手法 A の視線推定ネットワーク（2023 年電子情報通信学会総合大会発表予定）

提案手法 A は，入力である顔画像と両目の画像に対して，それぞれ ImageNet で学習済みの ResNet-50 をバックボーンとして特徴量抽出を行う．そして，これらの特徴量を結合して一つの特徴量とする．結合した特徴量を全結合層に入力し，ピッチ角とヨー角を独立に推定する．

また，提案手法 A では，バックボーンに使用している ResNet-50 の一層目の畳み込み層以外をファインチューニングしている．顔と両目の画像は画像サイズや構造的差異が大きいため，バックボーンを分離している．一方で，提案手法 A は，左右の目に関してはバックボーンの分離を行なっていない．そこで，左右の目の構造の差異が推定精度にどの程度影響を及ぼすかを確認するために，左右の目のバックボーンを分離した場合の手法も検討する．この手法を提案手法 B とし，提案手法 B の視線推定ネットワークを図 3.2 に示す．

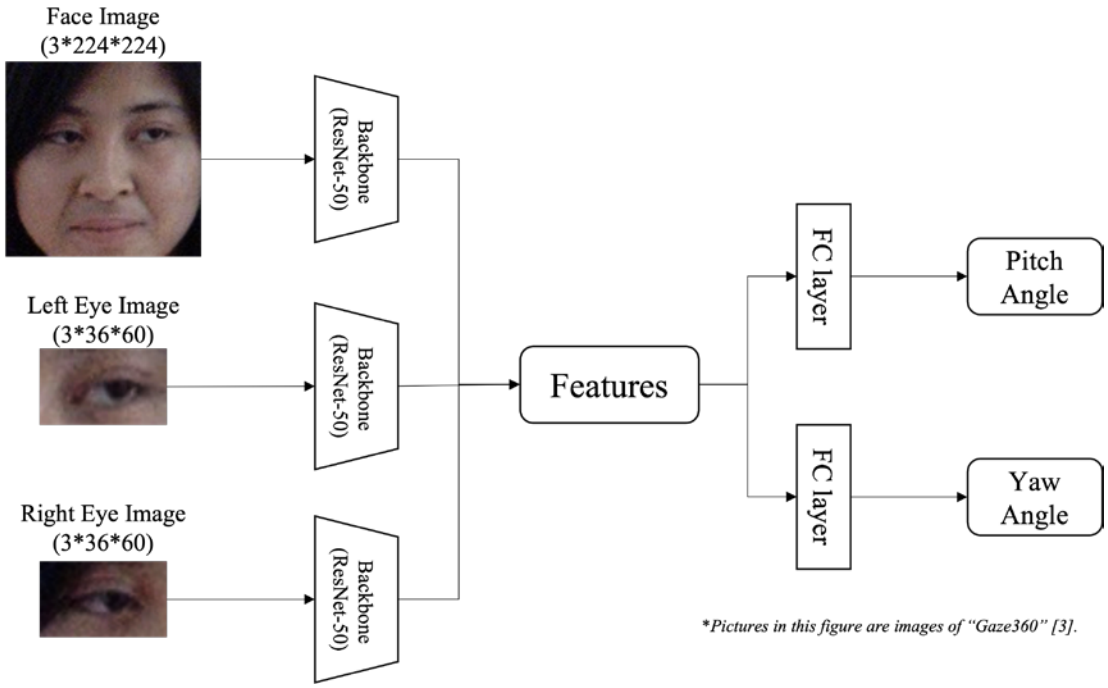

図 3.2 提案手法 B の視線推定ネットワーク

なお，提案手法 A，B のロスには，式(2.1)で示した L2CS-Net と同様のロス関数を用いる．

3.3.2 実験方法

予備実験の方法について述べる．提案手法 A，B の学習には，第 1.2 節の関連研究で述べたロバストな視線推定のためのデータセットである，Gaze360 の学習用画像を用いて行う．L2CS-Net と同様に Phi-ai Lab[7]が公開している Gaze360 の事前処理を施す．事前処理では，二つの処理を行う．一つ目の処理は，アピアランスベースの視線推定を行う際に正確な推定を行うことのできない, 被験者が後ろ向きである場合の画像の除去である．さらに，Gaze360 の画像は顔のバウンディングボックスと左右の目のバウンディングボックスがアノテーションされている．二つ目の処理では，この情報をもとに顔画像，左目の画像，右目の画像の三種類の画像を元の画像から切り抜く処理である．このとき，顔画像のサイズは 224pixel*224pixel，両目の画像は 36pixel*60pixel で統一されている．さらに, 学習条件について述べる．L2CS-Net との比較のためにエポック数は 50, バッチサイズは 16 で統一する．最適化手法には Adam を使用し，学習率は 0.00001 とする．ファインチューニングについては，ImageNet で学習済みの ResNet-50 の一層目の畳み込み層のみパラメータを固定して，他の層を学習させる．また，評価方法について述べる．提案手法 A，B の評価には， Gaze360 のデータセットの評価用画像を用いる．

まず，角度誤差は第 2.2.2 節で述べた通り，式(2.10)で算出される，正解方向のベクトルと推定した方向のベクトルの誤差を示す指標である. 評価指標としては, 平均角度誤差, Mean Angular Error(MAE)を用いるものとし，式(3.1)で算出する．なお，$\boldsymbol{g_i}$は正解視線方向のベクトル，$\boldsymbol{g_i}'$は推定した視線方向のベクトルである.

$$Mean\ Angular\ Error = \frac{1}{N}\sum_{i=1}^{N} L_{angular}(\boldsymbol{g_i}, \boldsymbol{g_i}') \tag{3.1}$$

予備実験では，提案手法 A，B と従来手法である L2CS-Net と精度の比較を式(3.1)で得られた平均角度誤差の値により行う.

３．３．３　実験結果

提案手法 A および提案手法 B の実験結果を表 ３.１に示す.

表 ３.１　提案手法 A および提案手法 B の実験結果

	L2CS-Net	提案手法 A	提案手法 B
MAE(°)	10.41	**10.30**	10.32

表 ３.１より，平均角度誤差が提案手法 A では10.30°，提案手法 B では10.32°であることから改善できることがわかった.

３．３．４　結果のまとめと考察

予備実験の結果より，提案手法 A，B ともに従来手法 L2CS-Net より，視線推定の精度を改善できることがわかった．したがって，顔画像だけでなく，両目の画像も入力に用いることで，詳細な目の情報を追加可能となり，アピアランスベースの視線推定の精度改善が図れることを示した．また，提案手法 A と提案手法 B の比較により，アピアランベースの視線推定において, 左右の目のバックボーンの分離は有効でない可能性を示した.

一方で，予備実験において詳細な目の情報を付与したにもかかわらず，平均角度誤差は最大で0.11°の改善となっている．この結果より，詳細な目の情報を視線推定に最大限活用できていない可能性が考えられる．そのため，予備実験より，目の特徴量をより視線推定の精度改善に活用できるモデルの作成が課題となる.

３．４　むすび

本章では，従来手法である L2CS-Net の問題点より考えられる手法の検討を予備実験により行なった．さらに，実験により手法の有効性を示した上で，更なる課題を明らかにした．

第４章　　提案手法

４．１　　まえがき

本章では，従来手法である L2CS-Net の問題点および，第 3 章の予備実験により明らかになった課題に対する手法を提案する．

４．２　　提案手法

第 3 章の予備実験により，アピアランスベースの視線推定において，顔画像のみでなく，それに対応する両目の画像も入力に用いることで，視線推定の精度を改善可能なことが示されている．したがって，提案手法においても顔画像と両目の画像を入力に用いる．

ここで，人間の視線方向について考える．人間のおおよその視線方向は，顔の向きを含む顔全体の情報から判断可能である．これは，従来手法の L2CS-Net が顔画像のみの入力で従来の CNN ベースの視線推定モデルの中で最良であったことや, CA-Net が顔画像からおおよその視線方向を推定可能であったという結果から裏付けられる．さらに，人間は顔の向きに加え，目を動かすことで対象物を捉える．したがって，この人間の視線方向の仕組みを組み込んだモデルを提案する．提案手法では，顔の特徴量からおおよその視線方向（粗い視線）と両目の特徴量から粗い視線を正しい視線に補正する詳細な視線方向（視線残差）を推定する．この手法により，第 3 章の予備実験で明らかにした，目の特徴量が有効活用できていないという課題を解消し，視線推定の精度改善が期待される．

以上を踏まえた提案手法 C の視線推定ネットワークを図　４．１に示す．

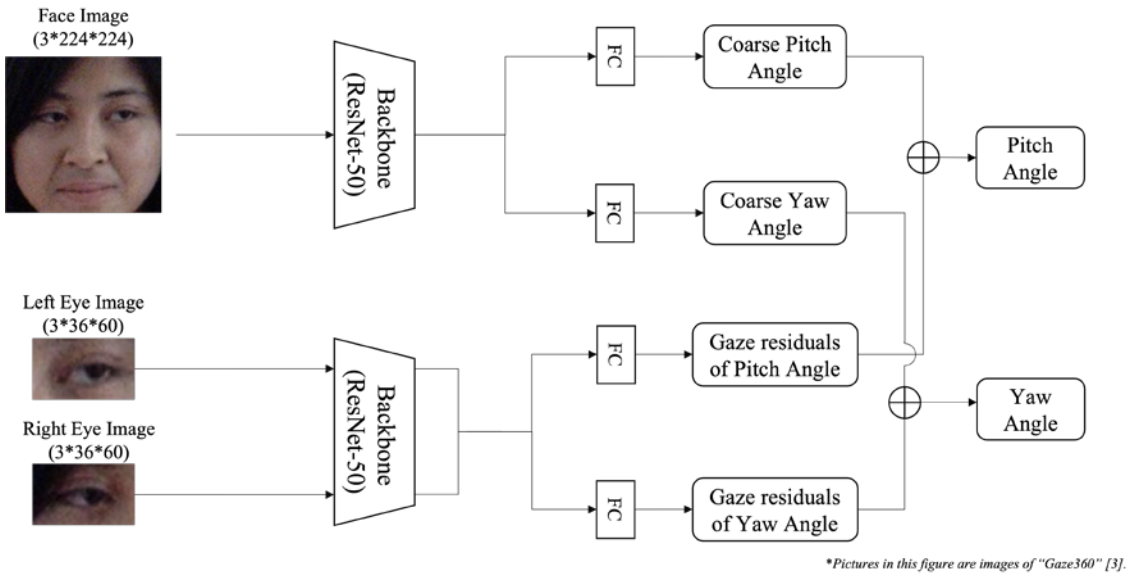

図 4.1　提案手法 C の視線推定ネットワーク

提案手法 C では，顔と両目の画像に対して，それぞれ ImageNet で学習済みの ResNet-50 をバックボーンとして特徴量抽出を行う．さらに，顔の特徴量から粗い視線のピッチ角とヨー角を独立に推定する．また，左右の目の特徴量は，結合して一つの特徴量としたのち，結合した特徴量から視線残差のピッチ角とヨー角を独立に推定する．最後に，粗い視線と視線残差を足し合わせることで，視線方向が推定できる．

また，予備実験と同様に左右の目のバックボーンを分離した場合の手法も提案する．この手法を提案手法 D とし，提案手法 D の視線推定ネットワークを図 4.2 に示す．

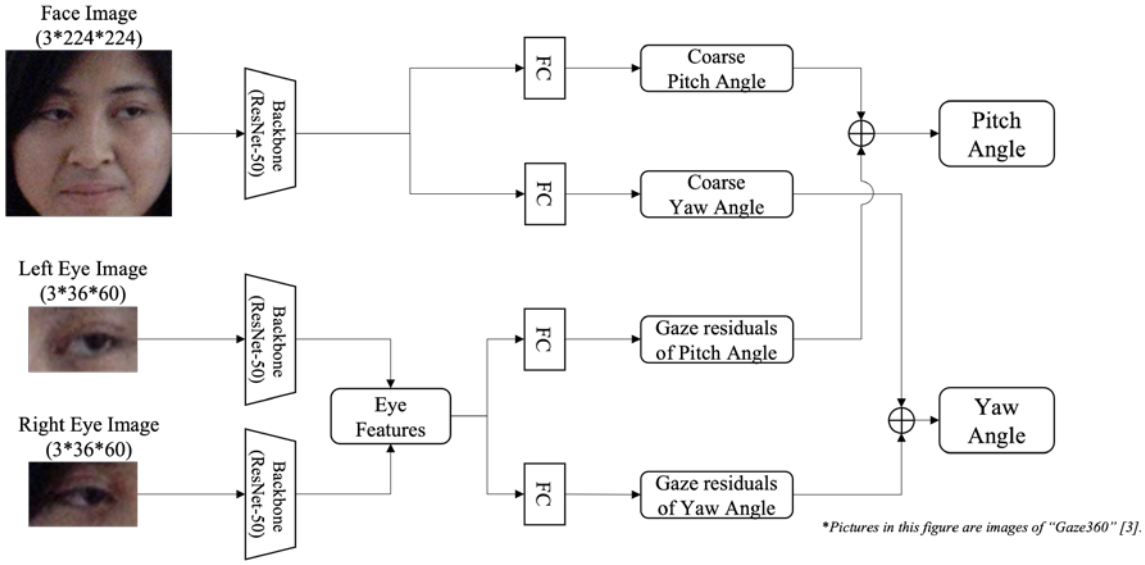

図 4.2　提案手法 D の視線推定ネットワーク

最後に提案手法 C および提案手法 D で使用するロスについて述べる．ロスは，正解角度と推定した角度の L1 ロスとする．提案手法 C，D では，粗い視線と視線残差を推定していることから，それぞれのロスを考慮する必要がある．以上を踏まえ，用いるロスを式(3.1)に示す．なお，推定した粗い視線方向の角度を θ''，視線残差を足し合わせた視線方向の角度をθ'，正解の視線方向の角度をθとする．なお，γは視線残差に関するロスの重み係数である．

$$Loss = l_{L1}(\theta'', \theta) + \gamma l_{L1}(\theta', \theta) \tag{3.1}$$

4.3　むすび

本章では，提案手法の概要およびネットワークの構造を図とともに示した．提案手法は，従来手法 L2CS-Net の問題点と予備実験の結果から考察される課題に対するアプローチとなっている．

第５章　　提案手法の評価実験

５．１　　まえがき

本章では，提案手法の評価実験を行い，結果を述べる．まず，評価実験の条件について述べる．続いて，評価実験による提案手法の実験結果を述べる．最後に提案手法の実験結果のまとめと考察を述べる．ただし L2CS-Net は， CA-Net より Gaze360 において平均角度誤差を1.85°削減しており，アピアランスベースの視線推定手法として優れている．そのため，本論文では提案手法との比較対象を L2CS-Net とする．

５．２　　提案手法の実験条件

提案手法は, 予備実験と同様に事前処理を施した Gaze360 の学習用画像を用いて学習する．学習条件は，L2CS-Net との比較のためにエポック数は 50，バッチサイズは 16 で統一する．最適化手法には Adam を使用し，学習率は 0.00001 とする．ファインチューニングについては, ImageNet で学習済みの ResNet-50 の一層目の畳み込み層のみパラメータを固定し，他の層を学習させる．続いて，提案手法の評価には，予備実験と同様に事前処理を施した Gaze360 の評価用画像を用いる．さらに評価指標には，式(3.1)で示した予備実験と同様の平均角度誤差を用いる．

５．３　　実験結果

第 5.2 節で述べた実験条件による提案手法 C および提案手法 D の実験結果を述べる．

まず，視線残差のロスの重みγに関して値の比較のために，$\gamma = 1, 2, 3, 4$とした場合の提案手法 C の実験結果を表 ５．１に示す.

表 ５．１　提案手法 C の実験結果($\gamma = 1, 2, 3, 4$)

	提案手法 C ($\gamma = 1$)	提案手法 C ($\gamma = 2$)	提案手法 C ($\gamma = 3$)	提案手法 C ($\gamma = 4$)
MAE(°)	10.47	**10.16**	10.25	10.41

表 ５．１より，提案手法 C において自然数γに対しては，$\gamma = 2$が最適であることが

わかった．また，表 5.1 の MAE の値およびそれらの値を二次関数で近似した場合のグラフを図 5.1 に示す．

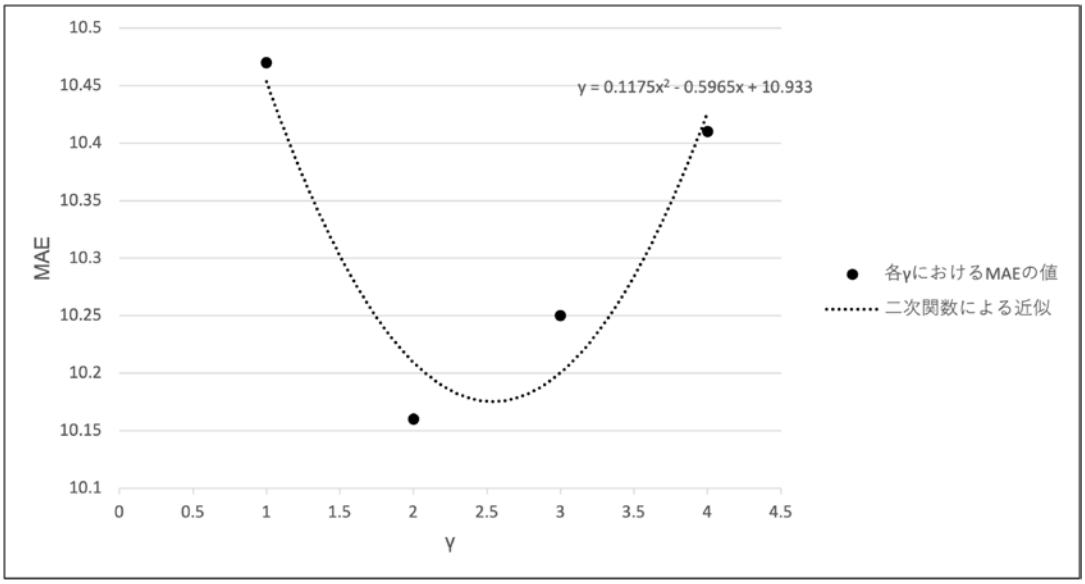

図 5.1 提案手法 C の MAE の値および近似した二次関数

図 5.1 で近似した二次関数が極小値となるのは，$\gamma \cong 2.5$のときである．したがって，提案手法において MAE が最小となるγの値は，およそ 2.5 であると予想される．この予想のもと，$\gamma = 2.5$とした場合の提案手法 C の実験結果を表 5.2 に示す．

表 5.2 提案手法 C の実験結果($\gamma = 2.5$)

	L2CS-Net	提案手法 C ($\gamma = 2.5$)
MAE(°)	10.41	**10.31**

図 5.1 より，$\gamma = 2.5$のとき MAE は最小となると期待されたが，実験結果より$\gamma = 2.5$で最小とならない．したがって，図 5.1 の近似が正しくない可能性が示唆される．

次に，提案手法 C および提案手法 D の結果の比較を行う．上記の実験結果より，MAE が最小となるγの値の前後の自然数と考えられる，$\gamma = 2,3$とした場合の提案手法 C および提案手法 D の実験結果を表 5.3 に示す．

表 5.3　提案手法 C および提案手法 D の実験結果

	L2CS-Net	提案手法 C ($\gamma = 2$)	提案手法 D ($\gamma = 2$)	提案手法 C ($\gamma = 3$)	提案手法 D ($\gamma = 3$)
MAE(°)	10.41	**10.16**	10.30	10.25	10.20

表 5.3 より，提案手法 C は平均角度誤差を最大で0.25°，提案手法 D は平均角度誤差を最大で0.21°，従来手法である L2CS-Net より改善できることがわかった．また，提案手法 C と提案手法 D の実験結果の比較により，γの値によっては提案手法 C と提案手法 D の優劣が異なることから，左右の目の構造の差異は視線推定に影響を与えない可能性を示唆している．

5.4　提案手法のまとめと考察

第 5.3 節に示した提案手法の実験結果より，提案手法 C および提案手法 D がアピアランスベースの視線推定において有効であることを示した．続いて，予備実験における提案手法 A および提案手法 B より精度改善を図れたことから，予備実験で明らかになった，目の詳細な分析である目の特徴量が視線推定に十分に活用できていないという課題に対するアプローチの有効性も示した．さらに，提案手法 C に関して視線残差のロスの重み係数γの値の分析を行い，予想と実際の結果が異なることから今後より詳細な分析が必要であると考える．また，提案手法 C と提案手法 D の実験結果の比較により，左右の目の構造の差異が視線推定に影響を与えない可能性を示した．

5.5　むすび

本章では，提案手法の評価実験を行い，提案手法の有効性を示した．

第６章　結論と今後の課題

６．１　結論

本論文では，アピアランスベースの視線推定のモデルである L2CS-Net の問題点として，同手法が顔画像のみを入力に用いていることから，詳細な視線の推定に必要な目の分析が不十分である点を挙げた．この問題点に対して，顔画像に対応する両目の画像も入力に用いることで，視線推定の精度改善が期待される．そのため，本論文では予備実験として，顔画像と両目の画像を入力に用いた，L2CS-Net の改善手法を提案した．さらに，予備実験より明らかになった課題を解決するような手法を提案した．最後に，提案手法の評価実験により，顔画像からおおよその視線方向，両目の画像からおおよその視線方向を正しい視線方向に近づけるような残差を推定する手法がアピアランスベースの視線推定，ひいては予備実験で明らかになった，目の特徴量を視線推定に十分に活用できていないという課題に対して有効であることを示した．

６．２　今後の課題

今後の課題は引き続き，本論文の研究目的である，アピアランスベースの視線推定精度の改善となる．評価実験により提案手法の有効性は示した．しかし，提案手法の平均角度誤差は，最小でも10.16°存在している．すなわち，本論文冒頭で述べた，日常生活で視線推定のモデルを使用する際に誤差の影響で支障をきたすという問題が依然として残る．したがって，アピアランスベースの精度改善は今後も必要であると考える．

また，提案手法における視線残差のロスの係数であるγの値の分析に関して，評価実験では予想通りの結果が得られなかったことから，今後の課題として MAE が最小となるγの値の検討も必要である．

さらに，本論文で述べた予備実験と提案手法は，顔の特徴量と目の特徴量をそれぞれ独立して用いている．しかし，顔と目の位置関係や顔の向きに対する目の向きなどの相互関係を考慮する必要があると考える．そのため，この考察に基づく顔と目の関係性を考慮可能なモデルの作成が必要である．

謝辞

本研究に際して，丁寧かつ素晴らしいご指導をしてくださり，実験環境および快適な研究環境を与えてくださった渡辺教授に心より感謝いたします．また，本論文の校閲を始めとし，多くのご指導を頂いた早稲田大学国際情報通信センターの石川孝明様に深く感謝申し上げます．

日頃から貴重な意見をくださり，研究室における温かい環境を提供してくださった同研究室の皆様に感謝いたします．

最後に，私をここまで育ててくださり，常に心を支えてくださり，生活を支えてくださっている家族に感謝いたします．

参考文献

[1] Xucong Zhang, Yusuke Sugano, Mario Fritz, Andreas Bulling : “Appearance-Based Gaze Estimation in the Wild”, CVPR, pp.4511-4520, Jun. 2015.

[2] Ahmed A.Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub AL-Hamadi : “L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments”, arXiv preprint arXiv:2203.03339, Mar. 2022.

[3] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, Antonio Torralba : “Gaze360: Physically Unconstrained Gaze Estimation in the Wild”, ICCV, pp.6912-6921, Oct. 2019.

[4] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, Geng Lu : “A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation”, AAAI, pp.10623-10630, Jan. 2020.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin : “Attention Is All You Need”, NeurIPS, pp.5998-6008, Jun. 2017.

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio : “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”, EMNLP, pp.1724-1734, Oct. 2014.

[7] GazeHub@Phi-ai Lab., “Datasets”, http://phi-ai.buaa.edu.cn/Gazehub/3D-dataset/, (2022 年 12 月参照)

図一覧

表一覧

研究業績

[1] 杉山秀治，渡辺裕 :“目の特徴量を付与したアピアランスベースの視線推定モデルの検討”，電子情報通信学会総合大会, Mar. 2023. （発表予定）

目の特徴量を付与したアピアランスベースの視線推定モデルの検討

A Study of Appearance-Based Gaze Estimation Models with Eye Features

杉山秀治[*1]　　渡辺裕[*1]
Hideharu Sugiyama　　**Hiroshi Watanabe**

[*1] 早稲田大学基幹理工学部
School of Fundamental Science and Engineering, Waseda University

1. まえがき

近年，人間の視線情報はエンゲージメント調査やメタバースなど多くのアプリケーションで活用されている．アピアランスベースの視線推定手法は，人物の顔画像や目の画像から視線方向を推定する手法である．

本稿では，アピアランスベースの視線推定手法の一つである **L2CS-Net[1]**の精度改善手法を提案する．まず **L2CS-Net** の問題点を述べ，目の特徴量を用いた視線推定モデルをする．さらに評価実験により，提案手法の有効性を示す．

2. 従来手法

L2CS-Net は，対象人物の顔画像を入力に用いた人物に依存しないアピアランスベースの視線推定手法の一つである．三次元視線推定のためのデータセットである，**Gaze360[2]** に対して，**CNN** を用いたモデルの中では最良の手法となっている．

L2CS-Net は顔画像のみを入力に用いているため，目の詳細な分析が不十分である．より正確な視線方向を推定するため粒度の高い目の情報を考慮できるモデルが必要である．

3. 提案手法

顔画像だけでなく対応する両目の画像も **CNN** を用いたネットワークに入力する．それぞれの画像に対し，**ImageNet** で学習済みの **ResNet-50** を用いて特徴量抽出を行い，それらの特徴量から視線方向を推定する．

また左右の目の構造の相違が推定精度にどの程度影響を及ぼすかを確認するために，左右の目のバックボーンを分離する場合と分離しない場合の結果を比較する．バックボーンを分離する場合を提案手法 **A**，分離しない場合を提案手法 **B** とする．提案手法 **B** のネットワークを図 **1** に示す．

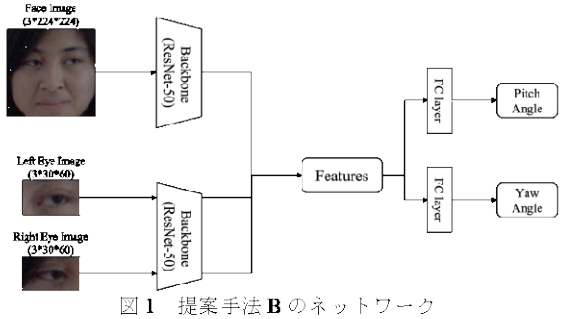

図 1　提案手法 **B** のネットワーク

図 **1** に示すように，提案手法では顔画像と両目の画像の特徴量を全て結合し，一つの特徴量とする．この結合した特徴量を全結合層に入力し，ピッチ角とヨー角をそれぞれ独立に推定することで推定視線方向を決定する．なお，顔固定座標系におけるオイラー角のうち，ピッチ角とヨー角が求まれば視線方向は一意に決定できる．

4. 実験

提案手法の学習及び評価には **Gaze360** をデータセットとして用いる．評価指標には平均角度誤差，**Mean Angular Error(MAE)**を用いる．角度誤差は，正解視線方向と推定視線方向のなす角であり，誤差指標として用いる．

実験条件として，**L2CS-Net** との比較のためにエポック数を **50**，学習率を **0.00001**，バッチサイズを **16** と統一した．

提案手法による実験結果を以下の表 **1** に示す．

表 **1** より，提案手法が従来手法である **L2CS-Net** より平均角度誤差を最大で**0.11°**改善できることがわかった．

表 1　提案手法と従来手法の比較

	L2CS-Net	提案手法 A	提案手法 B
MAE(°)	**10.41**	**10.32**	**10.30**

また，提案手法 **A** と **B** の比較により，左右の目のバックボーンの分離が視線推定結果には影響しないことがわかった．

5. むすび

本稿では，アピアランスベースの視線推定精度の改善手法として，目の特徴量を追加するモデルを提案し，有効性を確認した．

参考文献

[1] Ahmed A. Abdelrahman *et al.*, "L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments", arXiv preprint arXiv:2203.03339, 2022.

[2] Petr Kellnhofer *et al.*, "Gaze360: Physically Unconstrained Gaze Estimation in the Wild", ICCV, pp. 6912-6921, Oct. 2019.

[3] Yihua Cheng *et al.*, "A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation", AAAI, pp.10623-10630, Jan. 2020.