Vol.2013-CVIM-188 No.9
2013/9/2

データ駆動型アプローチによる視線・視覚的注意の推定

菅野 裕介[1,a)]

概要：人間の視覚を理解することは画像・映像メディア理解や人間自身の内部状態推定を行う上で重要な課題であり，視線計測や視覚的顕著性，視覚的注意に関する研究がますます広く行われるようになっている．中でも，低解像度目画像からの注視点推定や注視予測のための視覚的顕著性モデル最適化，人間の注視行動からの画像理解など多様な課題に対して，データを元にした学習ベースのアプローチが取られる事例が多くなりつつある．本発表では，こうしたデータ駆動型アプローチによる視線・視覚的注意推定の取り組みについて，最新の事例を紹介する．

1. はじめに

人間がどのように視覚情報を処理しているかを理解することは古くからコンピュータビジョン研究の重要な目標であり続けてきた．視線による注意のメカニズムは人間の視覚情報処理において中心的な役割を果たすものと考えられており，視線の計測と解析は様々な研究分野から幅広く取り組まれてきた研究課題となっている．近年では，視線計測を行うための基本的な技術自体はある程度成熟しつつあり，工学的には更に簡便な視線計測を行うための手法や，計測した視線を実応用するための方法論が重要な課題になりつつあると言える．

カメラベースの非装着型視線計測手法は，幾何的な眼球モデルと角膜反射光などの眼球内特徴を用いて視線方向を推定するモデルベース手法と，目画像自体を特徴として推定を行うアピアランベース手法の大きく二つに分けられ，それぞれ過去に様々な手法が提案されてきた [10]．アピアランベース手法には，外部光源などの特殊な計測装置を必要としないという大きな利点があり，特に低解像度目画像からの推定を目標とする場合は，事前に収集した目画像のデータセットを元に推定器の学習を行うのが一般的である．一方，比較的多数の学習データが必要となることや，頭部姿勢変動へ対応することなどが技術的課題として挙げられる．

また，注意の推定を行う別の手法として，視覚的顕著性モデルがある [15]．視覚的顕著性は人間のボトムアップな視覚的注意メカニズムに相当し，Itti ら [13] により計算モデルとしての定式化が行われて以降，自然画像から視覚的顕著性マップを計算するための手法が数多く提案されている [3]．このような視覚的顕著性の計算モデルを用いることで，視野内において人間がどの領域を注視しやすいかを顕著度のマップとして表現することができるため，人の側を観測する視線計測手法とは逆に，動画像側から注意の予測を行うことができる．視覚的顕著性の計算モデルの構築にあたっては，視覚的注意のはたらきに対応する何らかの仮説を元にしたルールベースのアプローチが取られることが多かったが，近年では実際の画像に対する人間の注視点データセットから最適な注視点分布予測モデルを学習するようなデータ駆動型のアプローチが一定の成功を収めつつある [39]．

さらに，こうして獲得した視線や注意を実際の応用に生かす上でも，学習による手法が有効である場合は多い．視覚的注意推定同様，従来は仮説に基づく検証を行うのが視線運動解析においては一般的であったが [12]，特に視線や注意の動きを動画像や人間の内部状態の認識タスクに応用する場合，視線や注意と目的とする対象の関係性が明確でない場合も多いため，機械学習的なアプローチが取られる場合が増えつつある．

以上を踏まえ，本稿では大きく分けて次の二点，計測と応用の観点から機械学習と視線・視覚的注意の接点について議論する．

(1) 学習アプローチによる視線・視覚的注意の推定
(2) 視線・視覚的注意の認識タスクへの応用

それぞれ近年の取り組みを紹介しながら，各分野が抱える課題を整理していく．

[1] 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo
[a)] sugano@iis.u-tokyo.ac.jp

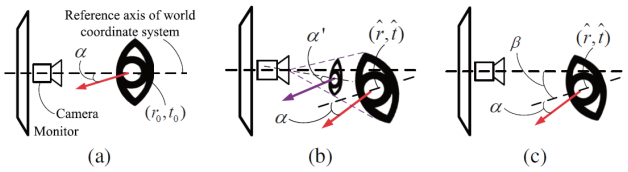

図 1 アピアランスベース視線推定における頭部姿勢変動対応

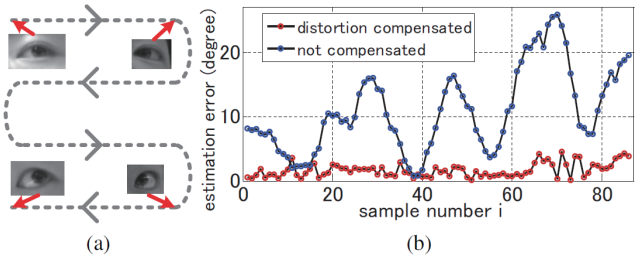

図 2 (a) 頭部姿勢変動に伴う画像歪みに起因する誤差補償のための学習データ (b) 頭部姿勢変動補償の有無による推定誤差の比較

2. 視線・視覚的注意の推定

2.1 アピアランスベース視線推定のための学習

アピアランスベースの視線推定における基本的な問題設定は，目画像と注視点との関係性を学習することにあり，これまでに，ニューラルネットワークを用いた手法 [1], [36] や近傍点補間による手法 [31]，ガウス過程回帰を利用した手法 [35] などが提案されてきた．文献 [19] では，L1 ノルム最小化の枠組みを用いて線形補間を行うためのデータを適応的に選択することで，少数の学習データから視線推定を行う手法を提案している．このようにして得られた結合係数と同一の結合係数を用いて学習データの注視点座標を線形結合することにより，入力目画像に対応する未知の注視点座標を高い精度で出力することが可能となる．さらに同様の枠組みを入力目画像切り出しの位置合わせにも用いることで，簡便で高精度な視線推定を実現している．

さらに，アピアランスベースの視線推定におけるもう一つの重要な課題として，頭部姿勢変動への対応が挙げられる．上に挙げた従来手法は頭部姿勢固定の条件下で注視点の推定を行うため，基本的には目画像特徴と注視点座標との間に 1 対 1 の対応関係が存在する．しかし，実際にはカメラから見た頭部姿勢が異なれば目画像の見え方も大きく変化するため，単純なマッピングを学習するだけでは対応できない．

文献 [18] では，目画像と注視点座標のマッピング関数に頭部位置による注視点の変動と頭部回転による目画像の歪みの補償項を組み合わせ，頭部姿勢の変動に頑健な視線推定を実現する手法を提案した．ここでは頭部姿勢の情報はカメラベースの頭部追跡手法により得られていると仮定し，頭部姿勢と目画像の組が入力で与えられた時に 3 次元視線方向ベクトルを出力することが課題となる．この時，図 1 に示す通り，頭部姿勢変動に対応するためには従来手法で考えられていた学習時頭部姿勢での視線推定結果（図 1(a)）と頭部位置変化に対する幾何的な補償項（図 1(b)）に加え，画像の変化に起因する誤差の補償項（図 1(c)）を考慮する必要がある．

提案手法ではこの画像歪み補償項を，図 2(a) のような視線を固定して顔向き方向のみを変化させた映像を新たな学習データとして獲得する．すなわち，基準となる頭部姿勢下で獲得された学習データを用いて推定した視線方向と実際の視線方向のずれ $\Delta\boldsymbol{\phi}$ と頭部姿勢のずれ $\Delta\boldsymbol{v}$ の組を上記ビデオから獲得し，この関係性をガウス過程回帰を利用して学習している．図 2(b) に示す通り，評価実験では，頭部位置が変動するテストデータに対して従来手法を適用した場合に発生する 10 度程度の推定誤差を，提案手法により 2 度程度に抑えることができることを確認した．

また，既存の注視推定技術が共通して抱える欠点の一つに，推定のためのパラメータ学習がユーザーごとに必要となる点がある．学習を行うためには対象ユーザーの正しい注視位置を何らかの形で知る必要があるため，通常はあらかじめ定義された複数の参照点を注視するようユーザーに指示し，眼球情報を獲得する作業を事前に行うことになる．ユーザーの能動的な参加を要求するこのようなキャリブレーション作業は，利用シナリオによっては大きな制限になる．

文献 [28] では，顕著性モデルを学習に利用した，明示的なキャリブレーション作業を全く必要としない注視点推定手法を提案した．鍵となるアイデアは，顕著性マップを注視点の存在確率と捉えることにある．提案手法では，ある映像と，頭部姿勢固定の条件下でそれを鑑賞している人物の目画像のみを入力とし，顕著性マップと目画像が同期したデータセットを獲得する．目画像の類似度を元に，ここから少数の代表目画像とその注視点存在確率分布を生成し，最終的に目画像と注視点座標の関係性を学習する．これにより，キャリブレーション動作や特定の推定モデル，アプリケーションシナリオに依存しない注視点推定が可能になる．

提案手法は主に 4 つのステップで構成される．顕著性抽出ステップでは，まず入力の動画フレームから顕著性マップが算出される．マップ統合ステップでは，より精度の高い確率分布を得るために目画像の類似度に基づいてマップを統合する．ステップの出力として，少数の代表目画像とその注視点座標の存在確率分布が生成される．推定器構築ステップでは，これを元に回帰学習を行い，注視点推定器を構成する．さらに，ここで構成した注視点推定器の推定結果を元に再度顕著性マップモデルのパラメータ最適化を行うことで，更に注視点推定精度の向上を図る．

基本となるボトムアップの顕著性算出モデルとしては，Harel らによる GBVS (Graph-Based Visual Saliency) [11]

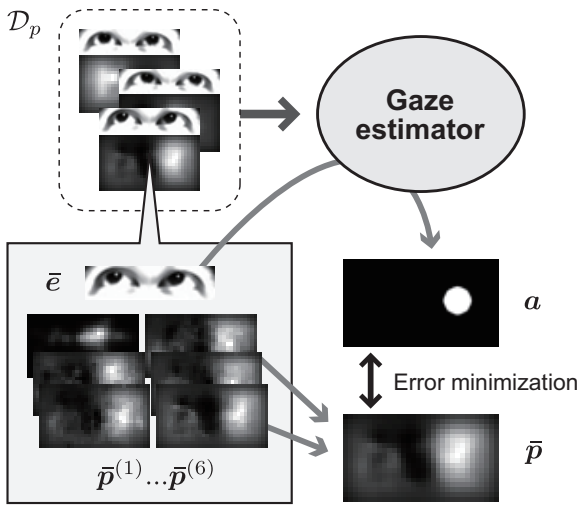

図 3 フィードバック学習による顕著性モデルの最適化

モデルを用い，5 つの低レベル画像特徴（色，輝度，方向，点滅，動き）が抽出される．これに加え，顔検出に基づく顕著性マップ [6] を導入する．これら 6 種類のマップの平均を取ったものが顕著性マップとして用いられる．

これらの顕著性マップは注視点座標のおおよその目安になるが，注視点が必ずしも顕著性のピーク座標に対応する訳ではなく，さらに複数のピークを持つ平坦なマップが算出されることも少なくないため，注視点座標を特定する際の精度は十分とは言い難い．これに対処するために，マップ統合ステップでは注視点座標の存在確率分布として使うことの出来るような，より精度の高いマップを生成する．すなわち，類似した目画像は注視点も類似していると考えられるため，これらの目画像に対応する顕著性マップは，正しい注視点座標の周辺で高い顕著性を共有し，かつそれ以外の領域ではランダムな値を持つ可能性が高い．従って，目画像の類似度に基づく加重平均を取ることで，正しい注視点位置周辺に強い顕著性のピークを持つようなマップが生成出来る．推定器構築ステップではここまでの処理により得られたデータを用い，ガウス過程回帰 [24] により注視点推定器を構成する．

これら一連の処理によって得られる推定器を用いることで入力動画における注視点座標位置が得られるが，提案手法ではこの結果を元に，視覚的顕著性モデルのパラメータ最適化を行うことで更なる全体の精度向上を目指す．すなわち，注視点存在確率分布 $\bar{\boldsymbol{p}}$ は前述の通り 6 種類のマップの平均として算出されていたが，実際には各特徴マップの貢献度には違いがあり，次のように特徴毎に異なる適切な重みを用いて加重平均を取ることで性能向上が期待できる．

$$\bar{\boldsymbol{p}}_i = \sum_{f=1}^{6} \omega_f \bar{\boldsymbol{p}}_i^{(f)}, \tag{1}$$

ここでは，学習データセット $\mathcal{D}_p$ に含まれる目画像 $\bar{\boldsymbol{e}}$ に対応する注視点座標とそれに対応する注視点存在確率分布

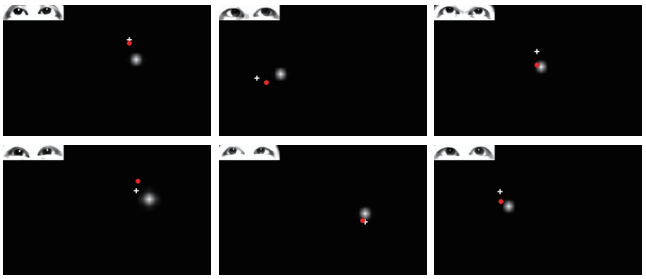

図 4 注視点推定結果

$\bar{\boldsymbol{p}}$ がなるべく一致するように，$\bar{\boldsymbol{p}}$ を算出する特徴重みを再学習する．具体的には図 3 に示すように，各目画像 $\bar{\boldsymbol{e}}$ に対して注視点の leave-one-out 推定結果を獲得し，推定された注視点の周辺に単一のピークを持つターゲットマップ $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M\}$ を算出する．これを元に，注視点存在確率分布とターゲットマップとの二乗誤差が最小になるように，非負制約を伴う最小二乗法 [17] により重みベクトル $\boldsymbol{\omega} = {}^t(\omega_1, \ldots, \omega_6)$ を最適化している．

$$\boldsymbol{\omega} = \arg\min_{\boldsymbol{\omega}} \sum_{i=1}^{M} ||\boldsymbol{a}_i - \sum_{f=1}^{6} \omega_f \bar{\boldsymbol{p}}_i^{(f)}||^2, \text{s.t. } \boldsymbol{\omega} \geq \mathbf{0}. \tag{2}$$

図 4 に，推定結果の例を示す．提案手法による推定結果は，背景のガウス分布として描画されている．左上の目画像がそれぞれの推定のために使われた入力画像であり，重ねて描画されている円はキャリブレーションデータに基づく注視点推定結果，および Tobii 社製視線推定装置 TX300 による推定結果を示す．明示的なキャリブレーションを一切行っていないにもかかわらず，提案手法によりキャリブレーションに基づく結果に非常に近い推定結果が得られていることがわかる．

2.2 視野特性を考慮した視覚的顕著性モデル

機械学習的に視覚的顕著性モデルを構築する研究は近年盛んに行われるようになったが [39]，これら既存のモデルは，そのほとんどが視覚刺激を空間的に一様なものとして扱っている．しかしながら，人間の視細胞は網膜全体に不均一に分布しており，網膜部位の中心窩からの距離に応じて異なる特性を持つことが知られている．Tatler らの研究 [32] では，空間周波数の異なる基本特徴マップから計算された顕著性マップについて，サッカードの大きさの違いによってその予測性能がどのように変化するかが調査されている．例えばエッジ特徴では，高周波数特徴に基づくマップにおいて長いサッカードの予測性能が低下するなど，それぞれの特徴マップに対して，適切なスケールが変化することが報告されており，視野内での視覚特性を考慮することで注視移動予測の高精度化が可能になることが示唆されている．

空間的な視野特性を考慮した手法はこれまでいくつか提案されているが [23], [33], [34], [37]，これらの手法で導入

されている視野特性は比較的単純なモデルであり，人間の複雑な視野特性を適切にモデル化するには至っていないと言える．文献 [16] では，データを元にした学習を行うことで，全てをルールとして明示化するのは困難な課題である特徴毎に異なる視野特性と視覚的顕著性の関係を，適切にモデル化することを目指している．

本研究では，網膜を中心窩からの距離を基準としたいくつかの領域に分割し，顕著性モデルの特性が各領域で異なると仮定することで人間の視野特性のモデル化を試みる．提案モデルではまず，視覚刺激となる入力画像に対して，特徴毎に既存の顕著性マップモデルを適用し，F 個の特徴マップを算出する．次に，中心窩からの距離と方向に応じた特性の違いを捉えるために，各特徴マップを現在の注視点 $\boldsymbol{p}_{\mathrm{cur}}$ を中心とした C 個の円環領域に分割し，さらにこれを上方向 90 度，下方向 90 度，左右方向 90 度の 3 つに分割する．最終的な顕著性マップはこれらの $F \times 3C$ 個のマップを線形統合することによって得られるが，f 番目の特徴マップの c 番目の領域に対する重み $w_{c,f}$ はそれぞれ異なるものとする．これらの重みの組み合わせ $\{w_{c,f}\}$ を学習データセットに対して最適化することで，人間の視野特性を考慮した視覚的顕著性モデルを獲得する．

本研究では頭部固定条件下で視線計測装置が許す最大範囲に近い水平視野角 57 度での視線計測実験を行い，学習データセットを作成した．データセットは，計 6000 (= 400 枚 ×15 人) の視線データから構成される．それぞれのデータに含まれる 240 (= 60Hz×4 秒) の視線座標から，毎秒 22 度を超える速さでの視線移動をサッカードとして抽出し，サッカードとサッカードの間の視線データの平均座標を注視点座標とした．

ボトムアップの特徴マップの算出には，前述のものと同様に Graph-Based Visual Saliency (GBVS) モデル [11] を用いる．まず入力画像を 3 つのスケール (1/4，1/8，1/16) にダウンサンプリングし，それぞれのスケールで色，輝度，方向の特徴マップを算出する．さらにトップダウンの影響を考慮したモデルとして顔検出に基づく特徴マップ [6] を用いている．

ある画像とその画像に対する計測されたサッカードデータが与えられたとき，図 5 に示すように，サッカードの終点 $\boldsymbol{p}_{\mathrm{next}}$ を中心とした半径 1 度の領域に単一のピークをもつターゲットマップ $\boldsymbol{t}$ と，重みベクトル $\boldsymbol{w}$ によって計算される顕著性マップ $\boldsymbol{s}(\boldsymbol{w})$ との誤差が最小化されるように最適化を行う．

学習データセットにおいてサッカード前後の真値座標の組 $\boldsymbol{p}_{\mathrm{cur}}$，$\boldsymbol{p}_{\mathrm{next}}$ が与えられたとき，点 $\boldsymbol{p}_{\mathrm{next}}$ 周辺に単一のピークを持つターゲットマップ $\boldsymbol{t}$ を以下のように定義する．

$$t(\boldsymbol{p}) = \begin{cases} 1 & \text{if } |\boldsymbol{p} - \boldsymbol{p}_{\mathrm{next}}| \leq t', \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

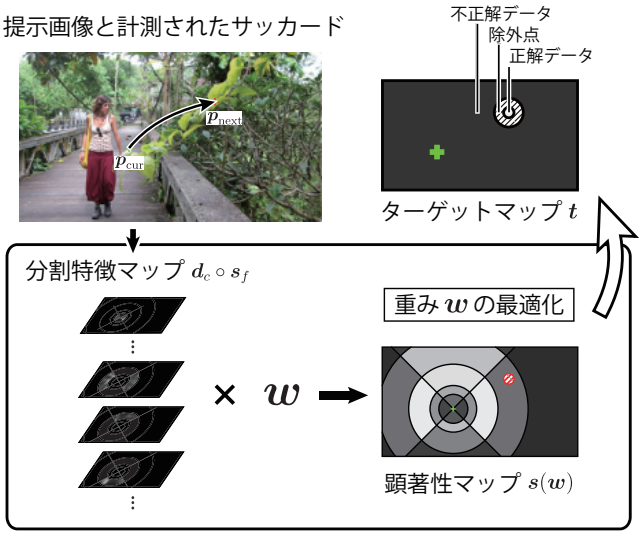

図 5 視野特性を考慮した顕著性モデルの学習

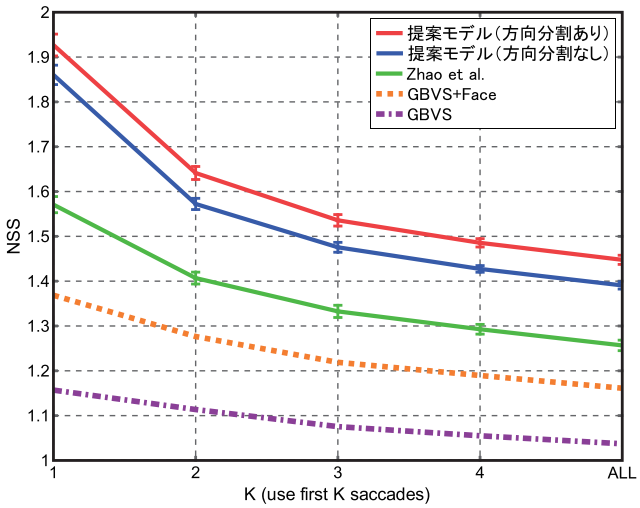

図 6 第 K 番目までのサッカードを使った各モデルの性能比較

境界値 t' は，人間の中心窩の大きさを考慮し，約 1 度となるよう設定する．

このターゲットマップに対し，本研究では前述の手法と同様に，非負制約を伴う最小二乗法 [17] により重みの最適化学習を行う．このとき，実際には計算の効率化のため，学習データの全画素からランダムに選択した画素に対して誤差関数の最小化を行った．ターゲットマップのうち，$t(\boldsymbol{p}) = 1$ の正解データとなる領域からあわせて 5 万点，$t(\boldsymbol{p}) = 0$ の不正解データとなる領域からあわせて 100 万点選択し，学習を行った．ただし，不正解データのうち，$\boldsymbol{p}_{\mathrm{next}}$ との距離が 1 度から 4 度の点は信頼度の低いデータとみなして除外した．

比較対象として，学習を行わず全ての特徴の重みを均一とした GBVS モデル，および顔特徴を追加した GBVS+Face モデルのほか，空間的に均一な学習を行うモデルのベースラインとして Zhao らのモデル [38] を用いた．Zhao らのモデルの学習には提案モデルと同様のデータセットを用い

表 1 画像選好識別に用いた視線移動統計量

Fixation	Position $\boldsymbol{p}$	Mean (×2)
		Variance (×2)
		Covariance
	Duration T	Mean
		Variance
		Sum
	Time t	Mean
		Variance
	Count	
Saccade	Direction $\boldsymbol{d}$	Mean (×2)
		Variance (×2)
		Covariance
	Length l	Mean
		Variance
		Sum
	Duration T	Mean
		Variance
		Sum
	Time t	Mean
		Variance
	Count	

た．すなわち，これは上述の提案モデルで分割数 $C = 1$ とした場合に相当する．学習を行わないモデルと比較すると学習を行うモデルのサッカード予測性能は常に高く，中でも方向による特性の違いを考慮した提案モデルが最も高い性能を示すことが確認できる．

3. 視線・視覚的注意の応用

3.1 注意運動特徴による行動認識

ここまでの章では主に，視線や視覚的注意を推定する課題においてデータ駆動型の機械学習アプローチを取っている例を紹介したが，獲得した視線・注意を実際の応用に生かす際にも，同様のアプローチを取ることが重要となる場合が多い．

文献 [29] では，視線情報を用いた画像選好の識別手法を検討した．二つの画像が並んで表示されている状況を想定し，これを閲覧したユーザーの視線の動きのみを入力として，二枚のどちらを好んでいるかを推定することを目的とする．特定カテゴリに属する刺激を用いた選好と注視時間の関係についての研究は過去にも行われている [2], [9], [25] が，本研究ではこれを一般の自然画像を対象に拡張し，データ駆動型アプローチを取ることで識別性能の向上を図っている．

具体的には，左右画像それぞれの凝視・サッカードに関して表 1 に示すような複数の統計量を算出し，これらを結合したものを特徴量とした教師あり識別タスクとして左右どちらの画像が好みかを推定している．識別機の学習にはランダムフォレスト [4] を用いた．

図 7 に示す通り，画像選好のような主観的な価値は不特

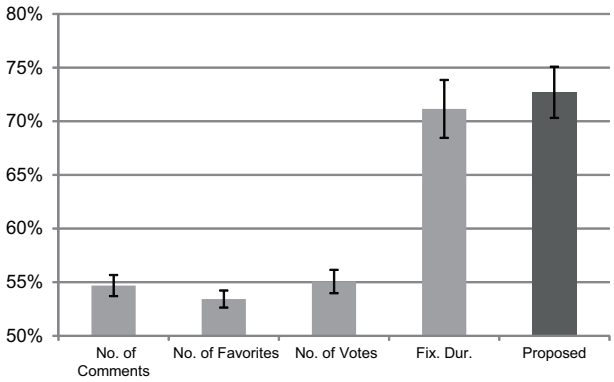

図 7 画像選好識別手法の比較

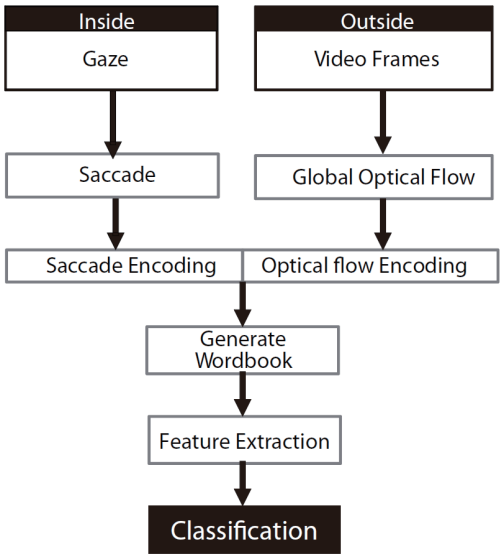

図 8 視線運動・頭部運動の基本運動語の抽出

定多数人物の評価（画像共有ウェブサイトにおけるコメント数，お気に入り数）や同じ画像対を見た他の被験者の評価の多数決ではほとんど識別できない．しかし，対象人物の視線の動き方を特徴として用いることで約 73% の精度で推定が可能となる．また，従来手法で用いられていたような注視時間の比較による単純な識別と比較しても，多数の特徴量を用いた学習アプローチを取ることで性能が向上することも確認できた．

また，人間の注視行動は，その時人物がどのような作業に従事しているかにも深く関連している．例えば一人称視点映像と視線運動が得られた場合，これらの情報から得られる注視の動きから，その時対象人物が行っている作業を識別することができる可能性がある．Bulling らは，視線運動の統計量を特徴量とした教師付き学習の問題として自己動作を認識することが出来ることを示した [5]．しかし，より一般的な，注意対象が大きく移るようなタスクを考えた場合，視覚的注意の移動は視線移動だけで表されるものではなく，頭部を主とした身体的な動作も加わって表されるものであると考えられる．

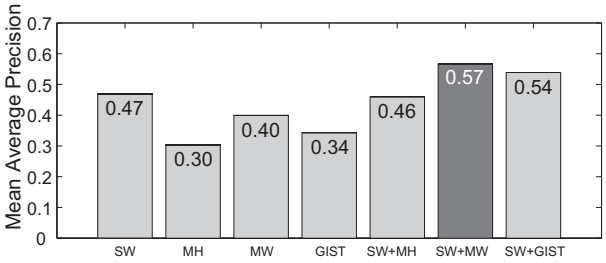

図 9 様々な特徴量における自己動作認識の Mean average precision

文献 [21] では，視線運動と頭部運動を共に用いた自己動作認識手法を提案した．提案手法では一人称視点映像と視線データのそれぞれから基本運動を抽出し，これを空間的な方向と強度により離散化したものを基礎特徴とする．従事している作業が違えば，こうした離散的な運動特徴が出現する時系列パターンにも違いが現れると考えられるため，運動特徴の出現パターンの統計量を動作識別のための最終的な特徴としている．

図 8 に示すように，提案手法では一人称視点映像と視線測定器それぞれのセンサからまず基本要素となる頭部運動（平均オプティカルフロー）とサッカードを抽出し，方向・強度を離散化することで視線運動文字と頭部運動文字を得る．これら二つの文字列に対して，基本運動文字それぞれの出現頻度である 1-gram wordbook から 4 文字の連続として観測される 244 種の基本運動語の出現頻度である 4-gram wordbook まで，計 4 つの Wordbook を集計する．ここから (1) 最大出現語の出現数，(2) 全出現語の平均出現数，(3) 出現語数，(4) 出現数の分散，(5) 最大出現数と最小出現数の差，の 5 つの統計量を算出することで，自己動作認識のための結合特徴量を抽出している．

図 8 に，上記の特徴を用いて 1 対多 2 クラス判別サポートベクターマシン [7] を用いてタスク識別を行なった際の Mean average precision を示す．読書，ビデオ鑑賞，紙への筆記，ウェブブラウジング，PC を用いたモニタ上の文書書き写しの 5 つのタスクを対象としており，提案手法（SW+MW）の他，カメラの自己運動特徴として Motion Histogram[14]，アピアランス特徴として GIST[22] を比較対象に加えている．それぞれの特徴を単独で用いた場合に比べて，自己運動と視線運動の両方を特徴とした場合に識別精度が向上することが示された．

3.2 物体領域クラスタリング

前章では視線や注意を認識タスクに応用する例をいくつか紹介したが，特に画像と視線の関係性を探る上での一つの課題として，注視対象領域の自動定義が困難である点が挙げられる．ウェブページや広告などを対象とした従来研究では，事前に定義された ROI (Region of Interest) を元に領域間の視線移動を解析するアプローチが取られることが多い [12]．画像の場合でも，セマンティックな意味が付

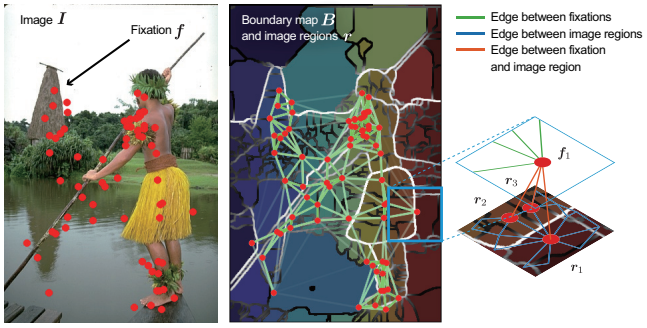

図 10 グラフに基づく画像領域と注視点のクラスタリング

与された ROI 間の視線移動頻度を特徴とすることで画像の意味理解が可能となることが先行研究により示されている [27]．しかし，このような画像 ROI を完全にボトムアップな形で検出・分割することはいまだ困難な課題である上に，多くの場合注視停留点のクラスタリングも曖昧さを含む．従って，分割された画像 ROI と注視点クラスタの組を自動で獲得することは容易な課題ではなく，上記の例でも注視対象領域の情報を無視した視線運動情報のみを特徴として用いていた．

これら二つの課題，画像の ROI 分割と注視点クラスタリングは相互に密接な関係がある．事前定義された ROI が与えられればこれを用いて注視点クラスタリングの曖昧性を解消することができる一方で [27]，注視点クラスタが事前に与えられればこれを画像セグメンテーションのシード情報として利用することができる [20], [26]．

文献 [30] では，このようなジレンマをグラフのラベル付問題として同時に解く手法を提案した．入力画像とそれを閲覧した複数人物の注視停留点座標が与えられ，全体のクラスタ数は未知の状態で，画像 ROI と注視点クラスタの組を出力することが目的となる．問題の曖昧性を解消するために，提案手法では画像 ROI と注視点のクラスタが常に一対一対応して存在していることを仮定する．この仮定の下で，画像空間においても注視点空間においても妥当なクラスタ対を発見するために，図 10 のような領域間重み付きグラフ$G = (\mathcal{V}, \mathcal{E})$を定義する．入力画像 $\boldsymbol{I}$ から算出した物体領域強度マップ $\boldsymbol{B}$ を元に過分割された画像領域$R = \{r_m\}$と，計測誤差をモデル化するためにそれぞれが微小な分散を持つガウス分布として表現された注視点座標$F = \{f_n\}$がともにグラフのノードとなる．

画像領域と注視点のクラスタリングは，グラフ G の i 番目のノードにクラスタラベル $l_i \in \mathcal{L}$ を付与する問題として捉えることができる．このラベル付け l のコスト関数 $E(l)$ を以下のように定義する．

$$E(l) = \sum_{i \in \mathcal{V}} D_i(l_i) + \sum_{(i,j) \in \mathcal{E}} V_{ij}(l_i, l_j) + \sum_{l^* \in \mathcal{L}} h_l \delta_{l^*}(l). \quad (4)$$

式 (4) の第 1 項はデータコストに相当し，ノード i がクラスタ l_i に所属することの適切さを評価する．本研究では

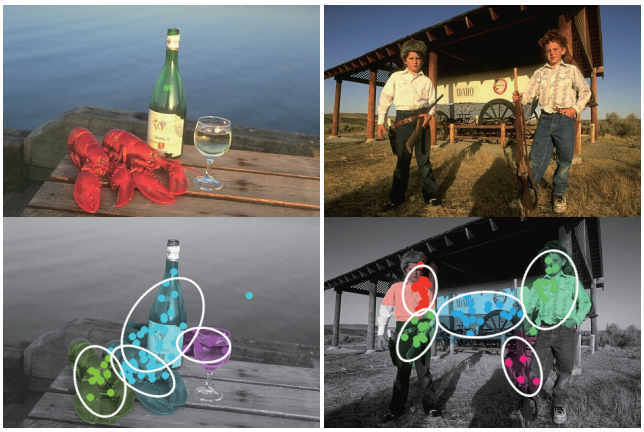

図 11 画像領域と注視点のクラスタリングの結果例

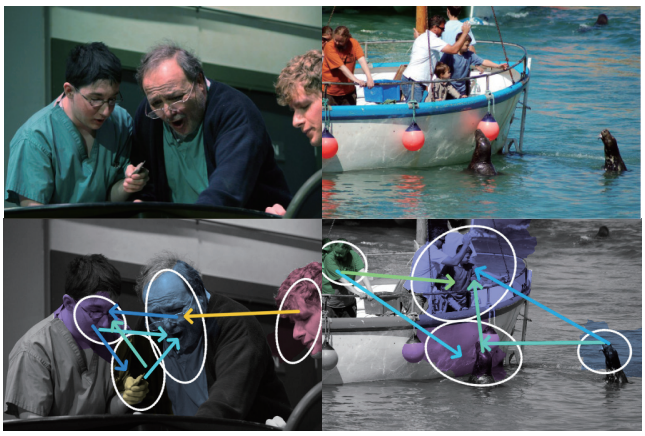

図 12 画像領域間視線移動の例

クラスタ l_i は二次元正規分布 $\mathcal{N}(\boldsymbol{\mu}_{l_i}, \boldsymbol{\Sigma}_{l_i})$ としてモデル化し，$D_i(l_i)$ はその対数確率密度関数を元に定義している．また，第 3 項は実際にノードに割り当てられたラベルの数に比例するラベルコストであり，クラスタ総数を間接的に調整する効果がある．

第 2 項が滑らかさコストに相当し，i，j の 2 ノード間に異なるラベルを割り当てた場合にエッジの種類に応じたコスト w_{ij} が加算される．

$$V_{ij}(l_i, l_j) = \begin{cases} w_{ij} & \text{if } l_i \neq l_j \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

ここでは注視点間，画像領域間，画像・注視点間の 3 種類すべてに次のような滑らかさ拘束を課している．注視点間に関しては，注視点座標のドロネー三角形分割により得られるノード間エッジに対して，その距離が短いほどコストが高くなるように設定する．同様に，画像領域間に関しては，隣接する画像領域間にエッジを設定し，2 領域を分ける物体境界強度が弱ければ弱いほど高いコストを設定する．最後に画像・注視点間に関しては，注視点の周囲にある画像領域に関して，その面積比に応じた滑らかさコストを設定している．すなわち，注視点が単一の画像領域に完全に包括される場合は強いコストが，複数の画像領域に跨っている場合はその全てに弱いコストが付与されることとなる．

式 (4) は未知数としてラベル付け l と各クラスタのパラメータ $\boldsymbol{\mu}_{l_i}, \boldsymbol{\Sigma}_{l_i}$ を持つため，反復アルゴリズムによる交互最適化を行う．まず l に関しては α 拡張アルゴリズム [8] を利用することで最適化が可能となる．クラスタパラメータ $\boldsymbol{\mu}_{l_i}, \boldsymbol{\Sigma}_{l_i}$ はデータコスト（第 1 項）のみに関連するため，クラスタに所属するノードの平均・分散をもって更新することで式 (4) の最小化を実現できる．

図 11 に提案手法により得られるクラスタリング結果を示す．図上に重ねられた点が注視点座標に対応し，画像領域との対応を色分けで示している．クラスタの個数が不明な状況でも，注視点座標と画像領域の分割両方の観点から妥当な結果が得られていることが確認できる．人手によるラベル付けと比較した実験では，注視点クラスタリングと画像分割を単独で行うよりも良好な性能が得られることが確認できている．

前述の通り，このようなクラスタ表現を行うことのメリットとして，画像領域間の視線移動がモデル化できるようになる点が挙げられる．Subramanian らの研究 [26] では 2 つのクラスタ l_i から l_j への視線の遷移確率

$$P(l_j|l_i) = \frac{n_s(l_i, l_j)}{n_f(l_i)} \tag{6}$$

を特徴として用いることを提案しており，これを提案手法の結果に適用した例を図 12 に示している．Subramanian らは一例として，画像中で何らかのインタラクションが発生している領域間において視線移動が頻繁に起こることを指摘しているが，図 12 の例でもこれが確認できる．提案手法を用いることで，こうした画像領域間視線移動傾向の特徴を完全にボトムアップに発見することが可能となる．

4. おわりに

本稿では，計測と応用の観点から機械学習と視線・視覚的注意の接点について整理した．アピアランスベースの視線推定は古くから学習に基づくアプローチが取られることが多く，実用上の課題としては頭部姿勢変動への対応，およびキャリブレーション負担の削減が挙げられる．視覚的顕著性モデルの構築においても，近年では人間の注視を教師データとした学習ベースのアプローチが多く取られるようになりつつあり，その重要度は今後さらに増していくものと考えられる．一方，視線や注意の情報を学習ベースの画像認識タスクに応用する上では，特に特徴設計手法の探求が今後の重要な課題の一つと言える．

謝辞 本稿で紹介した研究成果は，科学技術振興機構 CREST の助成により得られたものである．

参考文献

[1] Baluja, S. and Pomerleau, D.: Non-Intrusive Gaze Tracking Using Artificial Neural Networks, *Proc.*

NIPS1994, pp. 753–760 (1994).

[2] Bee, N., Prendinger, H., Nakasone, A., André, E. and Ishizuka, M.: Autoselect: What you want is what you get: Real-time processing of visual attention and affect, *Perception and Interactive Technologies*, pp. 40–52 (2006).

[3] Borji, A. and Itti, L.: State-of-the-Art in Visual Attention Modeling, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 35, No. 1, pp. 185 –207 (2013).

[4] Breiman, L.: Random forests, *Machine learning*, Vol. 45, No. 1, pp. 5–32 (2001).

[5] Bulling, A., Ward, J. A., Gellersen, H. and Troster, G.: Eye movement analysis for activity recognition using electrooculography, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 33, No. 4, pp. 741–753 (2011).

[6] Cerf, M., Harel, J., Einhauser, W. and Koch, C.: Predicting human gaze using low-level saliency combined with face detection, *Proc. NIPS2008*, pp. 241–248 (2008).

[7] Chang, C.-C. and Lin, C.-J.: LIBSVM: a library for support vector machines, *ACM Trans. on Intelligent Systems and Technology*, Vol. 2, No. 3, p. 27 (2011).

[8] Delong, A., Osokin, A., Isack, H. and Boykov, Y.: Fast Approximate Energy Minimization with Label Costs, *Int. J. Comput. Vision*, Vol. 96, pp. 1–27 (2012).

[9] Glaholt, M. G., Wu, M.-C. and Reingold, E. M.: Predicting preference from fixations, *Psychology Journal*, Vol. 7, No. 2, pp. 141–158 (2009).

[10] Hansen, D. W. and Ji, Q.: In the Eye of the Beholder: A Survey of Models for Eyes and Gaze, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 32, No. 3, pp. 478 –500 (2010).

[11] Harel, J., Koch, C. and Perona, P.: Graph-based visual saliency, *Proc. NIPS2007*, pp. 545–552 (2007).

[12] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. and Van de Weijer, J.: *Eye tracking: A comprehensive guide to methods and measures*, Oxford University Press (2011).

[13] Itti, L., Koch, C. and Niebur, E.: A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 20, No. 11, pp. 1254–1259 (1998).

[14] Kitani, K. M., Okabe, T., Sato, Y. and Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos, *Proc. CVPR2011*, pp. 3241–3248 (2011).

[15] Koch, C. and Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry, *Human neurobiology*, Vol. 4, No. 4, pp. 219–227 (1985).

[16] Kubota, H., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A. and Hiraki, K.: Incorporating visual field characteristics into a saliency map, *Proc. ETRA2012*, pp. 333–336 (2012).

[17] Lawson, C. L. and Hanson, R. J.: *Solving least squares problems*, Society for Industrial Mathematics (1987).

[18] Lu, F., Okabe, T., Sugano, Y. and Sato, Y.: A Head Pose-free Approach for Appearance-based Gaze Estimation., *BMVC2011*, pp. 1–11 (2011).

[19] Lu, F., Sugano, Y., Okabe, T. and Sato, Y.: Inferring human gaze from appearance via adaptive linear regression, *ICCV2011*, pp. 153–160 (2011).

[20] Mishra, A., Aloimonos, Y. and Fah, C. L.: Active segmentation with fixation, *Proc. 12th IEEE International Conference on Computer Vision (ICCV 2009)*, pp. 468–475 (2009).

[21] Ogaki, K., Kitani, K. M., Sugano, Y. and Sato, Y.: Coupling eye-motion and ego-motion features for first-person activity recognition, *Proc. IEEE Workshop on Egocentric Vision*, pp. 1–7 (2012).

[22] Oliva, A. and Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope, *Int. J. Comput. Vision*, Vol. 42, No. 3, pp. 145–175 (2001).

[23] Parkhurst, D., Law, K. and Niebur, E.: Modeling the role of salience in the allocation of overt visual attention, *Vision Research*, Vol. 42, No. 1, pp. 107–123 (2002).

[24] Rasmussen, C. E. and Williams, C. K. I.: *Gaussian processes for machine learning*, The MIT Press (2006).

[25] Shimojo, S., Simion, C., Shimojo, E. and Scheier, C.: Gaze bias both reflects and influences preference, *Nature Neuroscience*, Vol. 6, No. 12, pp. 1317–1322 (2003).

[26] Subramanian, R., Katti, H., Sebe, N., Kankanhalli, M. and Chua, T.-S.: An Eye Fixation Database for Saliency Detection in Images, *Proc. ECCV 2010*, pp. 30–43 (2010).

[27] Subramanian, R., Yanulevskaya, V. and Sebe, N.: Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements, *Proc. ACM MM 2011*, pp. 33–42 (2011).

[28] Sugano, Y., Matsushita, Y. and Sato, Y.: Appearance-Based Gaze Estimation Using Visual Saliency, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 35, No. 2, pp. 329–341 (2013).

[29] Sugano, Y., Kasai, H., Ogaki, K. and Sato, Y.: Image Preference Estimation from Eye Movements with A Data-driven Approach, *Proc. PETMEI2013* (2013).

[30] Sugano, Y., Matsushita, Y. and Sato, Y.: Graph-based joint clustering of fixations and visual entities, *ACM Trans. Applied Perception (TAP)*, Vol. 10, No. 2, p. 10 (2013).

[31] Tan, K. H., Kriegman, D. J. and Ahuja, N.: Appearance-based eye gaze estimation, *Proc. WACV 2002*, pp. 191–195 (2002).

[32] Tatler, B. W., Baddeley, R. J. and Vincent, B. T.: The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task, *Vision Research*, Vol. 46, pp. 1857–1862 (2006).

[33] Vincent, B. T., Troscianko, T. and Gilchrist, I. D.: Investigating a space-variant weighted salience account of visual selection, *Vision Research*, Vol. 47, No. 13, pp. 1809–1820 (2007).

[34] Wang, W., Chen, C., Wang, Y., Jiang, T., Fangand, F. and Yao, Y.: Simulating Human Saccadic Scanpaths on Natural Images, *Proc. CVPR 2011*, pp. 441–448 (2011).

[35] Williams, O., Blake, A. and Cipolla, R.: Sparse and Semi-supervised Visual Mapping with the S^3GP, *Proc. CVPR2006*, pp. 230–237 (2006).

[36] Xu, L. Q., Machin, D. and Sheppard, P.: A novel approach to real-time non-intrusive gaze finding, *Proc. BMVC1998*, pp. 428–437 (1998).

[37] Zelinsky, G. J., Zhang, W., Yu, B., Chen, X. and Samaras, D.: The Role of Top-down and Bottom-up Processes in Guiding Eye Movements during Visual Search, *Proc. NIPS 18*, pp. 1569–1576 (2005).

[38] Zhao, Q. and Koch, C.: Learning a saliency map using fixated locations in natural scenes, *Journal of Vision*, Vol. 11, No. 3, pp. 1–15 (2011).

[39] Zhao, Q. and Koch, C.: Learning saliency-based visual attention: A review, *Signal Processing*, Vol. 93, No. 6, pp. 1401–1407 (2013).