視覚障碍者のための周辺情報選択システム

河合 隆哲[1,a)]　岩村 雅一[,1,b)]　黄瀬 浩一[,1,c)]

概要：近年，ディープラーニングの発展から文字認識や物体認識の精度が著しく向上し，人と同程度の認識性能に近づきつつある．それにより，認識技術を視覚障碍者の「目」として扱う，以前から行われてきた試みが盛んになっている．特に，ディープラーニングを用いた物体認識により，撮影された画像に映った物体を読み上げる試みがよく行われている．そして，その多くは画像中に映った物体が単一または数個であるという前提がある．しかし，撮影された画像に多くの物体が映った場合，物体認識により得られた全ての結果を，視覚障碍者に音声で伝えることは困難であり，また不要な情報を多く含んでしまう．そこで，重要な情報を選択する情報選択システムが必要である．本稿では，画像中に存在する重要な情報の位置を示す手法を提案する．その際に参考にしたのは，多くの情報に囲まれいても効率的に情報を処理できる晴眼者の認知機能である．この晴眼者の認知機能の根幹にあるボトムアップ型注意とトップダウン型注意を模倣することで情報選択システムを構築する．

キーワード：ボトムアップ型注意, トップダウン型注意, 顕著性マップ, 選択的注意, 視覚障碍者支援

1. はじめに

視覚に障害を持たない人々（晴眼者）は，目から多くの情報を得ることができるのに対し，視覚障碍者は目から得ることのできる情報が限られる．近年，ディープラーニングの発展により，視覚障碍者の目の代わりとして，認識技術を用いる試みが盛んになっている [1,2]．特に，目の前に何があるのかを知るために，撮影された画像に対してディープラーニングの物体認識を用いて，映った物体を読み上げる試みがよく見られる．Seeing AI [*1]や，TapTapSee [*2]などのスマートフォンのカメラを用いた物体認識アプリも開発されている．そして，その多くは図 1 のように，撮影される画像に映った物体が，単一または数個であるという前提がある．

しかし，少し離れた周囲に何があるかを知るために，物体認識を用いる場合，画像には多くの物体が映ることが予想される．図 2 のように，撮影画像に多くの情報が含まれていた場合，認識したすべての情報を視覚障碍者であるユーザに音声で伝えるには時間が掛かるうえに，ユーザにとって不必要な情報も多く含まれる．よって，周囲の情報をむやみにすべて伝えることはかえって混乱を招く．そこで，ユーザに伝えるべき重要な情報を選択するシステムが必要である．

図 1: TapTapSee の使用例[*3]

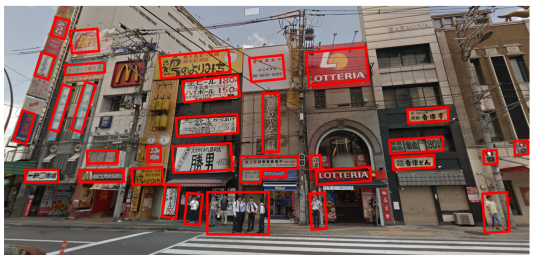

図 2: 画像中に大量かつ多様な情報が含まれており，物体認識・文字認識の結果をすべて伝えると不都合が生じる例

画像から重要な情報を選択する研究は，我々の知る限りあまり行われていない．類似した研究に，視線予測が存在

[1] 大阪府立大学大学院工学研究科
Osaka Prefecture Uniersity
[a)] kawai@m.cs.osakafu-u.ac.jp
[b)] masa@cs.osakafu-u.ac.jp
[c)] kise@cs.osakafu-u.ac.jp
[*1] https://www.microsoft.com/en-us/seeing-ai
[*2] https://taptapseeapp.com/
[*3] https://www.cbsnews.com/news/seeing-eye-phone-app-helps-blind-know-what-theyre-looking-at/

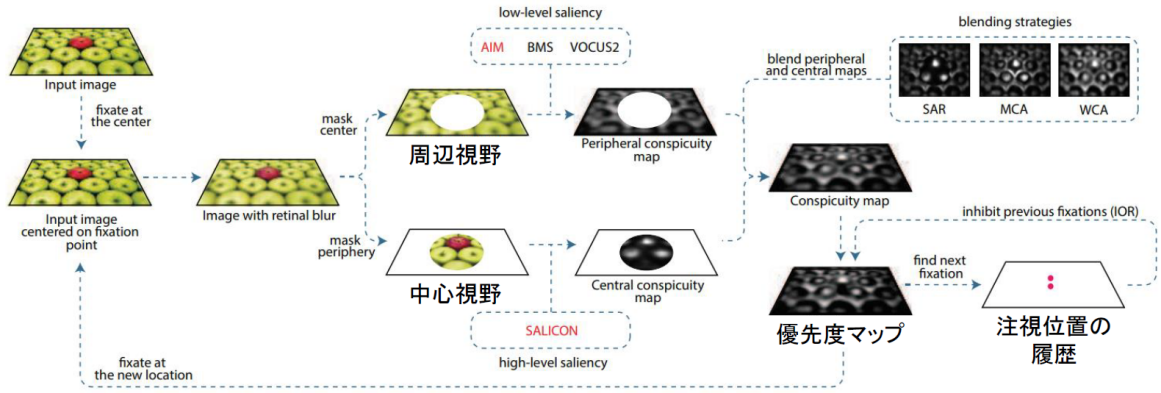

図 3: STAR-FC [3] のモデルの概要（一部改変）

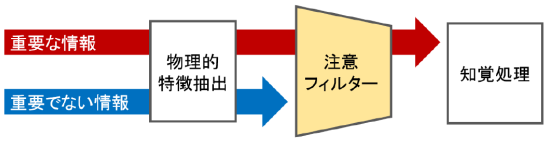

図 4: Broadbent のフィルター理論

するが，主な視線予測の研究は，画像中の顕著性の高い文字や物体を示しているに過ぎず，それだけでは重要な情報を選択できているとは言えない．

本稿では，情報選択システムを構築するにあたり，多くの情報に囲まれていても混乱しない晴眼者の認知機能に着目する．晴眼者の認知機能にはボトムアップ型注意とトップダウン型注意の 2 つの注意メカニズムが大きく関わる [4]．ボトムアップ型注意は外部から与えられる情報を，トップダウン型注意は内部状態に基づいた欲しい情報を取得する役割を担う．外部から与えられる情報とは，対向車などの障害物・案内や警告の看板といった私たち晴眼者が無意識に注意を向けるものを指す．晴眼者はこれらの情報を意識することなく処理しているが，視覚障碍者はこれらの情報を視覚から得られない．そのため視覚障碍者にとって，外部から与えられる情報は重要である．一方，内部状態に基づいた欲しい情報とは，人が意識的に注意を向けるものを指す．具体的には，空腹ならば食べ物や食事のとれる場所，道路を渡るならば信号機といった情報である．このように視覚障碍者の何かをしたい，何かを得たいという内部状態に応じた情報も重要である．本稿の情報選択システムでは，視線予測の最新手法 STAR-FC [3] とニューラルネットワークの Attention [5–7] を用いて，ボトムアップ型注意とトップダウン型注意を再現し，この 2 つの情報を取得する．そして，得られた情報から画像中の重要な情報がある位置を示す．

2. ボトムアップ型注意とトップダウン型注意

2.1 ボトムアップ型注意

ボトムアップ型注意とは，視野の中で他の物体と明らかに異なる特徴を持った物体（目立つ物体）を受動的に認識する注意メカニズムである [4]．晴眼者はボトムアップ型注意により，動いている物や目立つ物に対して注意を向けることができる．

ボトムアップ型注意を，推定する手法の一つに顕著性マップがある．顕著性マップとは，人が画像を見た際に注意を引きやすい場所を推定する計算モデルである．顕著性マップに関する研究は古くから行われており，初期の研究 [8] では色や輝度といった低レベルの特徴量から顕著性マップが算出されている．また近年の研究 [3, 7] では，ニューラルネットワークを用いて得られた特徴量から顕著性マップを算出する試みも行われている．視線予測の研究では，顕著性マップを求めることで注視の位置を推定している．

そこで提案手法では，顕著性マップを用いた視線予測の最新手法である Wloka らの STAR-FC [3] を，ボトムアップ型注意を再現したものとして用いる．Wloka らの STAR-FC [3] は，従来の視線予測の手法とは異なり，注視位置の時間的な順序を考慮している．一度マッピングされた注視位置の顕著性が低くなるため，既存の視線予測手法よりさらに人間らしい視線予測を可能にする．STAR-FC の構造は，図 3 のように入力画像に対し中心部分と周辺部分で異なる手法を用いて顕著性マップを作成し，それらを統合して最終的な顕著性マップを作成する．中心部分では高レベルオブジェクトベースの顕著性マップを求める SALICON [9] を用い，周辺部分では低レベル特徴ベースの顕著性マップを求める AIM [10] を用いる．顕著性マップの統合には，各ピクセルでの 2 つの顕著性マップの最大値を取る．統合した顕著性マップの中で最も値が大きい点を注視位置として

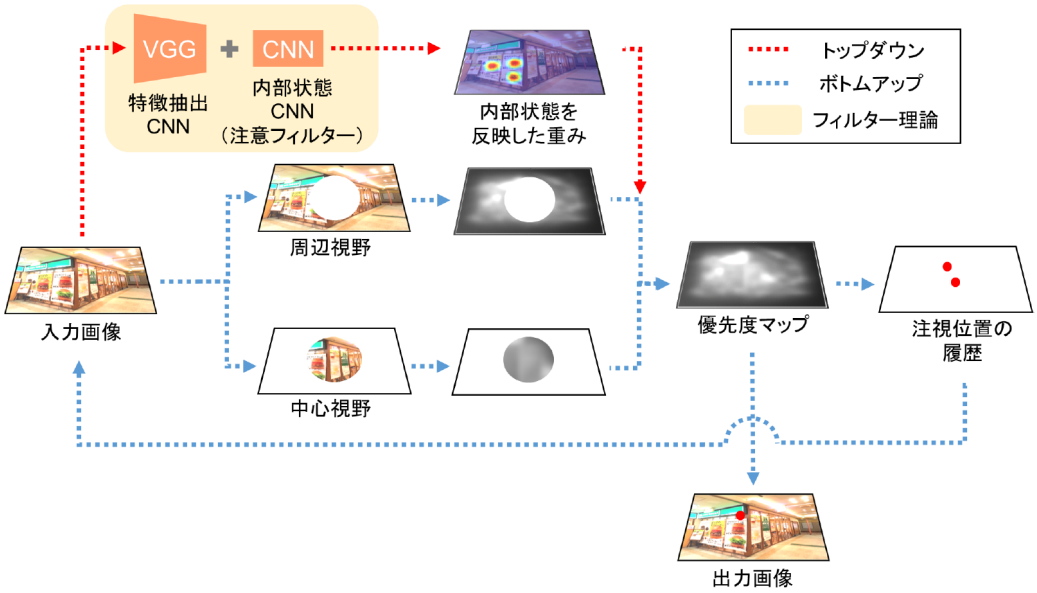

図 5: 提案手法の概要

保存する．そして，注視位置を中心として再度処理を行い，一度記録した注視位置の顕著性マップの値を減衰させ，次の注視位置を推定する．

2.2 トップダウン型注意

トップダウン型注意とは，事前知識に基づいて特定の物体を能動的に認識する注意メカニズムである [4]．晴眼者はトップダウン型注意により，自身が欲している情報に注意を向けることができる．

トップダウン型注意と類似した注意メカニズムに選択的注意がある．選択的注意は，一度に処理できる情報量には限界があるため，周囲の必要な情報に対しては注意を向け，不必要な情報に対しては無視をするという考え方である．選択的注意は視覚に限ったものではなく，初期の研究では聴覚を対象にしたカクテルパーティ効果が主に扱われてきた [11, 12]．カクテルパーティ効果とは，一つの部屋でたくさんの人間が話していても，自分にとって必要な声だけを聴きとることができる，という現象である．この現象に対し，いくつかの実験が行われ，その結果をもとに選択的注意をモデル化した理論の一つにフィルター理論 [11] がある．フィルター理論の一連の処理の流れを図 4 に示す．図 4 のように，フィルター理論では，まず入力情報の物理的特徴が抽出され，次に知覚処理が行われる．しかし，知覚処理が行われる際には，処理容量に限界があるため，すべての情報を同時には処理できない．そのため，最初に処理された物理的特徴の中から必要なものを注意フィルターが選択し，その後に知覚処理が行われる．

この注意フィルターと同様の機能を持つものとして，ニューラルネットワークの Attention [5–7] がある．Attention は，入力情報のより有効な要素に対して重みをかける手法の総称である．既存の視線予測の研究 [7] では，Attention をボトムアップ型注意を再現するための手段として用いており，トップダウン型注意を再現するためには用いられていない．本稿では，Attention を用いて選択的注意を計算機上で再現し，トップダウン型注意を情報選択システムに反映させる．

3. 提案手法

提案手法の概要を図 5 に示す．提案手法では，ボトムアップ型注意を最新の視線予測手法 STAR-FC を用いて再現し，外部から与えられる情報を取得する．またトップダウン型注意をニューラルネットワークの Attention を用いて再現し，内部状態に基づいた欲しい情報を取得する．得られた 2 つの情報を合わせることで重要な情報がある位置を画像上にマッピングし示す．

以下，3.1 で Attention を用いたトップダウン型注意の再現について，3.2 でボトムアップ型注意とトップダウン型注意の出力の統合について詳細に説明する．

3.1 トップダウン型注意の再現

既存の視線予測の研究 [7] では，ボトムアップ型注意を再現するために Attention を用いるのに対し，本稿ではトップダウン型注意を再現するために Attention を用いる．選択的注意のモデルである図 4 のフィルター理論 [11] を，ニューラルネットワークの Attention を用いて計算機上で実装する．図 5 のように，フィルター理論の物理的特徴抽出を特徴抽出 CNN，注意フィルターを内部状態 CNN（Attention）として提案手法に組み込む．

特徴抽出 CNN には，ImageNet で事前学習済みの Si-

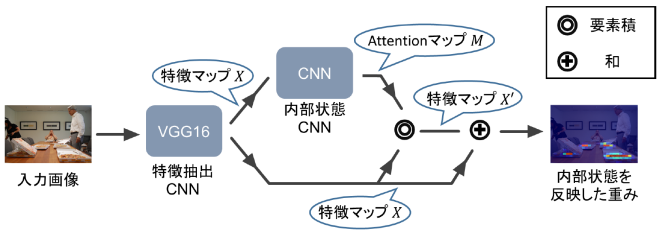

図 6: 特徴抽出 CNN と内部状態 CNN の概要

monyan らの VGG16 [13] を用いる．この VGG16 は特徴抽出器として用いるために全結合層を取り除く．また，より大きな特徴マップを得るために，最後の二つのプーリング層も取り除く．内部状態 CNN は，convolution 層と pooling 層，batchnormalization 層で構成される．学習時には特徴抽出 CNN は学習させず，内部状態 CNN のみ学習させる．このときに学習させるデータセットは，ユーザの内部状態を完全に推定する技術が存在しないため，事前にいくつかの内部状態とそれに対応する重要な情報の対象を定義しておく（例，空腹ならば食べ物が重要な情報）ことで用意する．特徴抽出 CNN と内部状態 CNN の概要を図 6 に示す．まず 224×224 の解像度にした入力画像を特徴抽出 CNN に通すことで，特徴マップ $X \in \mathbb{R}^{28\times28\times512}$ を得る．これを内部状態 CNN に通すことで，特徴マップ内での内部状態に応じた重要度を示す Attention マップ $M \in [0,1]^{28\times28}$ を得る．特徴マップ X と Attention マップ M の要素積を取ることで，内部状態に合わせて強調された特徴マップ $X' \in [0,1]^{28\times28}$ ができる．しかし，Attention によって元々の特徴マップ X から有用な情報が失われる可能性がある．そこで，特徴マップ X と内部状態に合わせて強調された特徴マップ X' を足し合わせる．これは Residual Connection [5] と呼ばれる手法である．これにより，内部状態を反映した重み $W \in [0,1]^{28\times28}$ を得る．

3.2 ボトムアップ型注意とトップダウン型注意の出力の統合

ボトムアップ型注意の再現のために用いる STAR-FC は，まず入力画像に対して中心視野と周辺視野で求めた顕著性マップを統合する．統合した顕著性マップと内部状態を反映した重み W の要素積を，図 5 の赤い点線と青い点線が交わる位置で求める．このとき，内部状態を反映した重みの各要素が 0 より大きく 1 より小さい最低値 α を取るように調整する．これによって顕著性の高い情報と内部状態に応じた情報を考慮した優先度マップを作ることができる．この優先度マップの中で最も値が大きい位置を重要な情報がある点として，入力画像に対しマッピングする．さらにその位置を保存し，入力画像の中心がその位置に来るようして再び優先度マップを求める．画像の中心を変えることでできる空白は画像の平均画素値で穴埋めする．そして，再度作成された優先度マップでは，重要情報があるとして既に保存された位置の値が減衰するようにし，次の重要情報の位置を求める．これを繰り返し，入力画像に重要な情報の位置を複数回マッピングしたものを情報選択システムの出力結果とする．

4. 実験

4.1 実験条件

内部状態を空腹として定義し，内部状態 CNN には MS COCO データセット [14] の 5 種類（pizza, hotdog, cake, sandwich, donut）のラベルが付いた食品画像を学習させた．学習データには 10000 枚，テストデータには 360 枚を用いた．また，内部状態を反映した重みの各要素の最低値 α は 0.4 とした．

4.2 実験結果と考察

トップダウン型注意の出力である内部状態を反映した重みを，図 7 示す．図 7(a) は入力画像を，図 7(b) は入力画像に対する食べ物領域を示したマスク画像を示しており，図 7(c) は内部状態を反映した重みをヒートマップで可視化したものである．図 7(b) と図 7(c) を比較すると，重要な情報とした食べ物領域が，内部状態を反映した重みに表れていることが分かる．食べ物領域には示されていない部分が，内部状態を反映した重みに表れているものが一部見受けられる．これはラベルが与えられていない食べ物を特徴抽出 CNN と内部状態 CNN が検出したためである．

次に，ボトムアップ型注意のみの出力結果（STAR-FC）と提案手法の出力結果を図 8 に示し，比較する．図 8(a) のボトムアップ型注意の出力結果は，人の注視が集まりやすい人の顔部分に重要な情報があると示している．一方，図 8(b) のボトムアップ型注意とトップダウン型注意の両方を考慮した提案手法は，食べ物付近に重要な情報あると示している．このことから，提案手法が内部状態を正しく反映していることが分かる．

5. まとめと今後の展望

本稿では，晴眼者のボトムアップ型注意とトップダウン型注意を模倣し，外部から与えられる情報と内部状態に基づいた欲しい情報を考慮した情報選択システムを提案した．最新の視線予測の手法である STAR-FC を用いてボトムアップ型注意を再現し，外部から与えられる顕著性の高い情報が存在する領域の検出を行った．また VGG16 と Attention を用いてトップダウン型注意を再現し，内部状態に基づいた欲しい情報が存在する領域の検出を行った．これらの検出結果を統合し，画像中の重要な情報がある領域をマッピングした．

今後の展望としては，外部から与えられる情報の取得に，

(a) 入力画像　(b) 食べ物領域　(c) 内部状態を反映した重み

図 7: トップダウン型注意の出力結果

(a) ボトムアップ型注意のみの出力結果（STAR-FC）　(b) ボトムアップ型注意とトップダウン型注意を考慮した提案手法の出力結果

図 8: ボトムアップ型注意のみの出力結果と提案手法の出力結果

視覚的顕著性とは異なる要素を用いることを考えている．また，ウェアラブルデバイスを用いることで，生体データを取得し，ユーザの内部状態の推定とその内部状態に基づいた欲しい情報の検出を自動的に行うことを考えている．

謝辞　本研究は，JSPS 科研費 基盤研究 B #17H01803 と立石科学技術振興財団　研究助成 (A) #2181004 の補助による．

参考文献

[1] Kayukawa, S., Higuchi, K., Guerreiro, J., Morishima, S., Sato, Y., Kitani, K. and Asakawa, C.: BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians, *Proc. CHI Conference on Human Factors in Computing Systems* (2019).

[2] Kacorri, H., Kitani, K. M., Bigham, J. P. and Asakawa, C.: People with visual impairment training personal object recognizers: Feasibility and challenges, *Proc. CHI Conference on Human Factors in Computing Systems*, pp. 5839–5849 (2017).

[3] Wloka, C., Kotseruba, I. and Tsotsos, J. K.: Active Fixation Control to Predict Saccade Sequences, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[4] Egeth, H. E. and Yantis, S.: Visual attention: Control, representation, and time course, *Annual review of psychology*, Vol. 48, No. 1, pp. 269–297 (1997).

[5] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. and Tang, X.: Residual Attention Network for Image Classification, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[6] Hu, J., Shen, L. and Sun, G.: Squeeze-and-excitation networks, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141 (2018).
[7] Wang, W., Shen, J., Guo, F., Cheng, M.-M. and Borji, A.: Revisiting Video Saliency: A Large-Scale Benchmark and a New Model, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
[8] Itti, L., Koch, C. and Niebur, E.: A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 20, No. 11, pp. 1254–1259 (1998).
[9] Huang, X., Shen, C., Boix, X. and Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks, *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 262–270 (2015).
[10] Bruce, N. and Tsotsos, J.: Attention based on information maximization, *Journal of Vision*, Vol. 7, No. 9, pp. 950–950 (2007).
[11] Broadbent, D. E.: *Perception and communication*, Elsevier (2013).
[12] Treisman, A. M.: Contextual cues in selective listening, *Quarterly Journal of Experimental Psychology*, Vol. 12, No. 4, pp. 242–248 (1960).
[13] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *Proc. International Conference on Learning Representations (ICLR)* (2015).
[14] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft COCO: Common objects in context, *Proc. European Conference on Computer Vision (ECCV)*, pp. 740–755 (2014).