(403) 画像処理

CNN を用いたウェブページに対する視覚的顕著性推定
Webpage Saliency Estimation Based on Convolutional Neural Network

新谷 浩平 † 滝本 裕則 †† 金川 明弘 ††
Kouhei Niiya † Hironori Takimoto †† Akihiro Kanagawa††
† 岡山県立大学大学院　情報系工学研究科　　†† 岡山県立大学　情報工学部

1　序論

インターネットが広く普及するにつれウェブページは我々にとって重要な情報源となっている．また，近年に実施された米国ベースのウェブユーザー調査でも，平均的なユーザが週に 23 時間オンラインになっていることが示された．このようなインターネット利用者の増大とサービスの多様化によって，ウェブユーザビリティ評価の必要性が高まっている．さらに，視線測定器の性能，及び操作性が向上したことにより視線情報がユーザビリティ評価に利用されるようになってきた．

一方，コンピュータビジョンの研究分野において古くから研究されている視覚的注意がある．これは，目から入力された信号の中から重要と思われる情報を判断し，効率的に情報を得るための機能である．近年では，深層学習に基づく CNN(Convolutional Neural Network) を用いた顕著性推定手法が多く提案されており，特徴量を手動で計算していた従来の手法と比べ，多くの評価指標でより良い結果を示している．[1]〜[4] なお，CNN ではモデル訓練のために大量の学習サンプルデータが必要である．

ところで，自然画像と比較してウェブページはテキストや写真，ロゴなどの視覚情報で構成されることが多い．つまり，ウェブページには自然画像と異なる顕著な刺激が多く含まれているため，自然画像を用いて学習を行った CNN ベースの顕著性モデルではウェブページに対して正確な顕著性マップ予測をすることは困難であると考えられる．この問題を解決するためには，大量のウェブページのデータベースを用いて CNN を学習させる必要がある．しかしながら，視線計測器を用いてのデータ収集やそのデータを分析するために多大な労力とコストがかかるなどの要因があるため，公開されているウェブページと顕著性マップを含むデータセットは小規模なものしかない．従来，学習のためのデータ量が少ない場合，データ拡張 (Data Augmentation) が学習のデータ量を増やすために用いられるが，顕著性推定に適用された例はない．その理由として，原画像の画素値をガンマ変換などによって変化させた場合，顕著な領域が明らかに変わるなどの悪影響が発生することが挙げられる．

本稿では，ウェブページに対する CNN を用いた高精度な顕著性推定を実現するため，データセットに対しデータ拡張を適用し，その有効性を検証する．また，トップダウン注意に影響を及ぼす顔や文字といった視覚的特徴に注目し，マスク情報として追加する顕著性推定モデルを提案し，その有効性を検証する．

2　提案手法

2.1　視覚的顕著性推定のためのデータ拡張

深層学習における顕著性推定の研究課題の一つとして画像データセットの拡大が挙げられるが，視覚的顕著性推定に関する研究の場合，大規模な視線計測器を用いたデータセットの作成には膨大な時間と労力を要する．一方，CNN には明るさの調整や画像を反転させるなどして学習データを増やすといったデータ拡張と呼ばれる技術がある．しかし，データ拡張を顕著性推定に適用した例はない．その理由として原画像をガンマ変換などによって画素値を変化させた場合，顕著な領域が明らかに変わるなどの悪影響が発生すると考えられていると推測される．なお， 人がウェブページを閲覧する際， グーテンベルク・ダイヤグラムのような視線移動のバイアスが存在する．しかし，空間的な注視の停留マップを求める顕著性推定問題においては，注視の順序に関するバイアスは大きな影響を与えないと考えている．

本稿では，既存の小規模なデータセットに対し，顕著な領域に対して悪影響を与えないと考えられる 3 つのデータ拡張を適用することで学習データの拡張を行う．拡張法として，水平方向に画像を反転させる「左右フリップ」，垂直方向に画像を反転させる「上下フリップ」，そして画像の一部分を切り取り原画像と同じサイズにリサイズする「拡大」を適用する． これらのデータ拡張を原画像と顕著性マップが対応するよう適用する．また，拡大後の顕著性マップに対して，線形な正規化と，Itti らによる非線形な正規化 [5] の 2 種をそれぞれ適用する．図 1 に各データ拡張の適用例を示す．

2.2　トップダウン注意を考慮した顕著性推定

トップダウン注意は事前知識や目的などの内発的刺激によって起こる視覚的注意である．また，人は無意識に顔や文字に対して注目する性質がある．[6] 本稿では，これらの特性を考慮し，ウェブページに対するトップダウン注意を考慮した顕著性推定法を提案する．具体的には，人が無意識に注目する顔や，注視する傾向にある文字の領域を特徴マスクとする．そして，アップサンプリングネットワークを用いた Xception モデル [**?**] に対して入力時にトップダウンに関するマスク情報を追加する．

モデルの入力層に対し，顔特徴のみを考慮する場合は

(a) 原画像　(b) 左右フリップ後の原画像　(c) 上下フリップ後の原画像

(d) 顕著性マップ　(e) 左右フリップ後の顕著性マップ　(f) 上下フリップ後の顕著性マップ

(g) 拡大後の原画像　(h) 線形な正規化後の顕著性マップ　(i) 非線形な正規化後の顕著性マップ

図 1: 各データ拡張の適用例

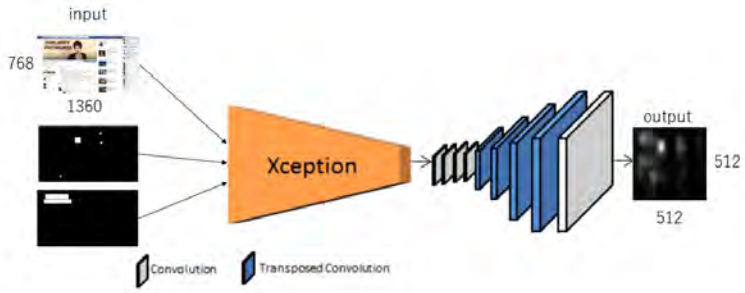


図 2: 顔と文字の特徴を考慮したモデルの概略図

原画像 RGB3 チャンネルと顔特徴マスク 1 チャンネルを合わせた 4 チャンネル入力とし，文字特徴のみを考慮する場合は原画像 RGB3 チャンネルと文字特徴マスク 1 チャンネルを合わせた 4 チャンネルとする．また，どちらも考慮する場合は原画像 RGB3 チャンネルと顔特徴マスク 1 チャンネル，文字特徴マスク 1 チャンネルの 2 つを合わせた 5 チャンネルとして同時に入力する．図 2 に 5 チャンネル入力時のモデルの概略図を示す．

3 評価実験

本章では，提案手法の有効性を確かめるために行った実験の条件，およびその結果について述べる．

3.1 データセット

本実験では，公開されているウェブページに関する顕著性のデータセット Webpage Saliency[9] を用いる．Webpage Saliency は UMN VIP(University of Minnesota Visual Information Processing Lab) が提供しているデータセットで，149 枚の原画像および対応する顕著性マップが含まれている．本実験では，データセットの中から 140 枚を用いており，モデル訓練のためにランダムな 70 枚を訓練データとして利用した．また，残り 70 枚を検証データとして訓練からは除外し評価のために用いた．

3.2 評価指標

視覚的顕著性の評価指標はいくつか提案されており，視覚的顕著性予測の研究の多くは複数の評価指標によってモデルを評価している．本実験では顕著性マップ予測タスクの評価指標には MAE(Mean Abusolute Error)，KLD(Kullback-Leibler Divergence) の 2 つを用い，有効性を検証する．評価指標と学習の損失関数は同じものを用いることとする．MAE は平均絶対誤差であり，2 つの画像間の対応する各画像の差を単純に最小化する．

表 1: 各データ拡張に対する結果

条件	評価指標	
データ拡張	MAE	KLD
CAT2000	0．1349	0．8546
データ拡張なし	0．1129	0．5592
左右フリップ	0．1075	0．5942
上下フリップ	0．1054	0．5228
拡大 (線形)	0．1075	0．5287
拡大 (非線形)	0．1049	0．5184
全適用	**0．1039**	**0．5030**

MAE は視覚的顕著性予測の分野で使用されることは少ないものの，画像比較では最もシンプルな一般的な方法である．KLD は視覚的顕著性予測でよく用いられる評価指標の 1 つである．KLD は 2 つの確率分布の差異を測る尺度として用いられる． また，両指標とも値が低いほど高精度であることを示す．

3.3 データ拡張の有効性の検証

3.3.1 実験条件

顕著性マップのデータセットに対して，提案するデータ拡張が有効であるかを検証する．提案手法の有効性を検証するために，まず，自然画像で学習した際の予測精度を確認するため CAT2000[10] を用いて学習を行う．次に，訓練データ 70 枚，検証データ 70 枚を初期データ (拡張なし) として顕著性推定を行う．左右フリップ，上下フリップ，拡大 (線形な正規化)，拡大 (非線形な正規化)，全てのデータ拡張を適用した場合の各精度を確認する．なお，拡大を行う際，よりデータを増やすことを考え，原画像 1 枚につき 4 枚を生成している． また，どの場合においても検証データは 70 枚で固定である．

3.3.2 結果と考察

顕著性マップ予測モデルに対して，データ拡張の有効性について検証する．各データ拡張の結果を表 1 に示す．MAE について，拡張なしとデータ拡張を行った場合とを比較すると，全てにおいて良い結果が得られた．KLD について，左右フリップを除く全てのデータ拡張法において，データ拡張なしと比べ高精度な結果が得られた．MAE と KLD どちらの場合においても，全てのデータ拡張を適用した場合が最も高精度なより良い結果を示していることがわかる．

3.4 トップダウン注意を考慮した顕著性推定

3.4.1 実験条件

アップサンプリングネットワークを用いた Xception モデルに対して，トップダウン注意を考慮するため，入力時に顔や文字のマスク情報を用いる．具体的には，RGB3 チャンネルの原画像に新たにマスク情報のチャンネルを追加し，有効性について比較を行う．顔マスク，文字マスク，その両方をそれぞれ顕著性推定のための Xception モデルに入力する．本実験では，学習に使用するデータセットは表 1 の全てのデータ拡張を適用した場合を用いる．また，検証には 3．1 節で述べた検証データ 70 枚を用いる．比較対象は，マスクなしの結果と比較する．

表 2: 各マスクに対する結果

条件	評価指標	
マスク（チャンネル数）	MAE	KLD
顔（4ch）	0．1025	0．4899
文字（4ch）	0．1038	0．4757
顔と文字（5ch）	**0．1008**	**0．4646**

3.4.2 結果と考察

各マスク情報を用いた場合の結果を表 2 に示す．表 1 と比べ，MAE と KLD どちらの場合においても，マスク情報を入力した全ての場合においてよい結果が示された．また，どちらの評価指標においても顔と文字両方のマスク情報を用いた結果が最も高精度であった．

これらの結果から，顔や文字に対するトップダウン注意はウェブページにおいても有効であると考えられる．顔特徴マスクを用いたときの出力例を図 3 に，文字特徴マスクを用いたときの出力例を図 4 に，そして，両方のマスクを用いたときの出力例を図 5 に示す．

4 結論

本稿では，CNN を用いて，ウェブページに対する視覚的顕著性予測の精度を向上させるため顕著性推定に特化したデータ拡張の提案，トップダウン注意に影響を及ぼす視覚的特徴に注目した顕著性推定モデルを提案した．評価実験の結果，データ拡張において MAE ではすべての場合において予測精度が高くなった．KLD では左右フリップを除き，全てにおいて良い結果が得られた．また，顔，文字特徴を追加したモデルは MAE，KLD どちらの評価指標においても高い予測精度であることを確認した．

参考文献

[1] H. Takimoto, S. Katsumata, Sulfayanti F. Situju, A. Kanagawa, and A. Lawi : “Visual Saliency Estimation Based on Multi-task CNN”, Proc. of The Fourteenth Int. Conf. on Industrial Manag, 2018.

[2] M. Kümmerer, L. Theis, and M. Bethge: “Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet”, The International Conference on Learning Representations, 2015.

[3] X. Huang, C. Shen, X. Boix, and Q. Zhao:

"SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks",Proc. of IEEE International Conference on Computer Vision, pp. 262-270, 2015.

[4] E. Vig, M. Dorr, and D. Cox: "Large-scale optimization of hierarchical features for saliency prediction in natural images",Proc. of IEEE Computer Vision and Pattern Recognition, pp. 2798 - 2805, 2014.

[5] L. Itti, C. Koch, and E. Niebur: "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans. PAMI, Vol. 20, No. 11, pp. 1254-1259, 1998.

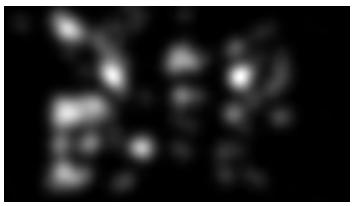
[6] M. Cerf , J. Harel, W. Einhaeuser, and C. Koch : "Predicting human gaze using low-level saliency combined with face detection", in Advances in Neural Information Processing Systems, 2008.

[7] K. He, X. Zhang, S. Ren, and J. Sun: "Deep residual learning for image recognition", Computer Vision and Pattern Recognition, 2016.

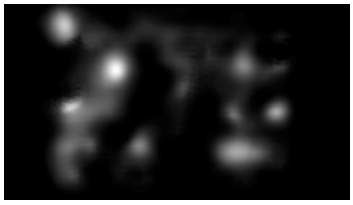
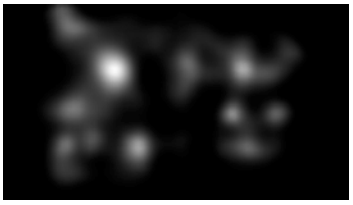
[8] F. Chollet,"Xception: Deep Learning with Depthwise Separable Convolutions", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1800 - 1807, 2017.

[9] Chengyao Shen, and Qi Zhao: "Webpage Saliency", European Conference on Computer Vision, 2014.

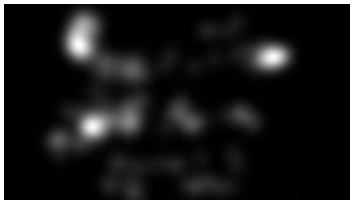
[10] A. Borji and L. Itti: "Cat2000: A large scale fixation dataset for boosting saliency research", Computer Vision and Pattern Recognition, 2015.

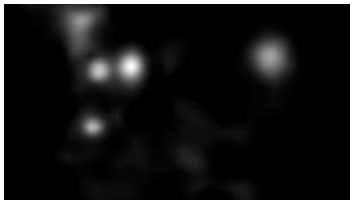
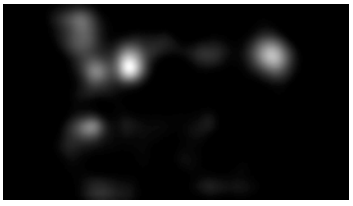

(a) 原画像 1　(b) 正解データ 1　(c) 顔マスク 1

(d) MAE1　(e) KLD1

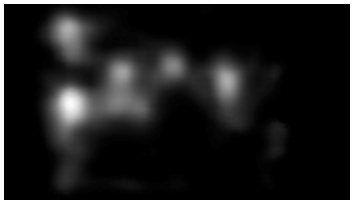
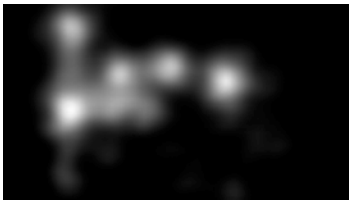

(f) 原画像 2　(g) 正解データ 2　(h) 顔マスク 2

(i) MAE2　(j) KLD2

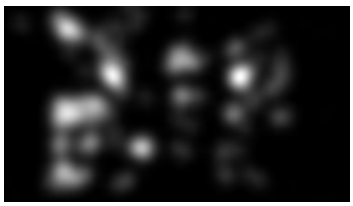


(k) 原画像 3　(l) 正解データ 3　(m) 顔マスク 3

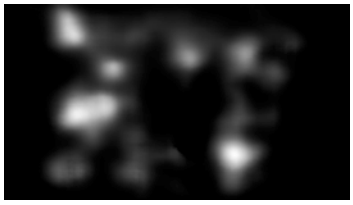
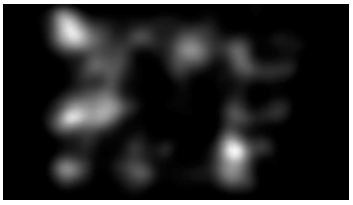



(n) MAE3　(o) KLD3

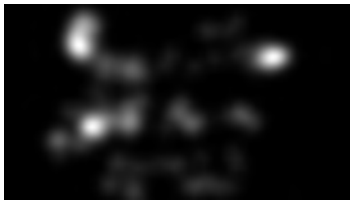
図 3: 顔特徴マスクを用いたときの出力例

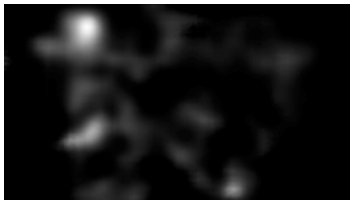
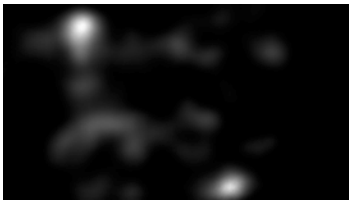

(a) 原画像 1　(b) 正解データ 1　(c) 文字マスク 1

(d) MAE1　(e) KLD1

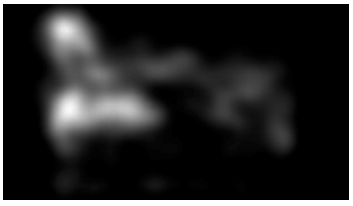

(f) 原画像 2　(g) 正解データ 2　(h) 文字マスク 2

(i) MAE2　(j) KLD2

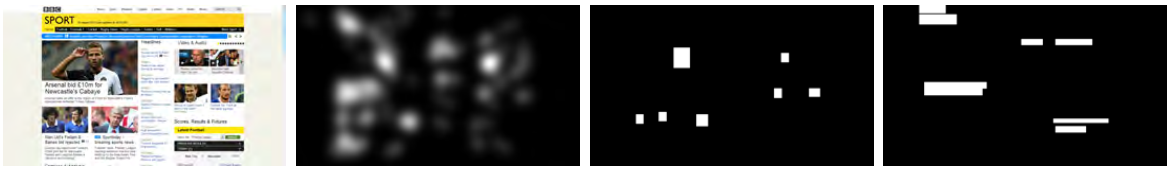


(k) 原画像 3　(l) 正解データ 3　(m) 文字マスク 3

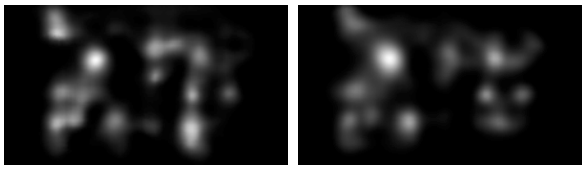



(n) MAE3　(o) KLD3

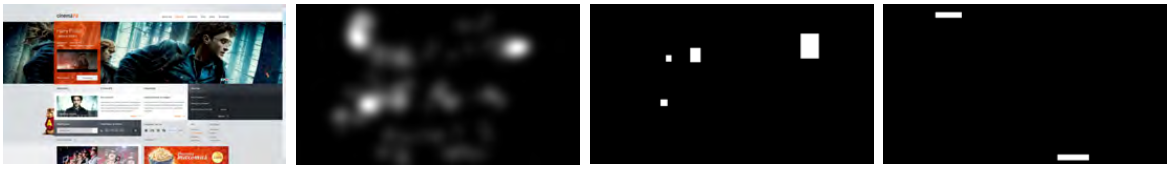
図 4: 文字特徴マスクを用いたときの出力例



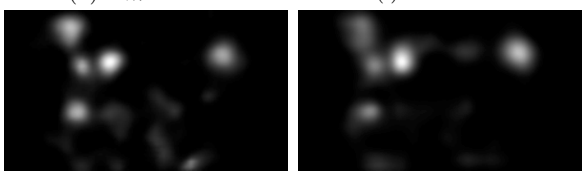
(a) 原画像 1　(b) 正解データ 1　(c) 顔マスク 1　(d) 文字マスク 1



(e) MAE1　(f) KLD1

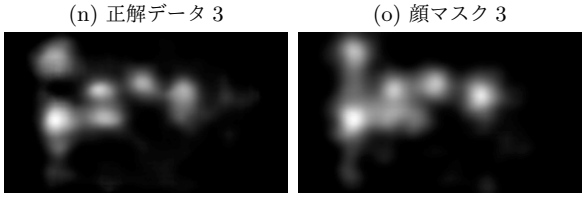


(g) 原画像 2　(h) 正解データ 2　(i) 顔マスク 2　(j) 文字マスク 2



(k) MAE2　(l) KLD2

(m) 原画像 3　(n) 正解データ 3　(o) 顔マスク 3　(p) 文字マスク 3



(q) MAE3　(r) KLD3

図 5: 両方のマスクを用いたときの出力例