J-027

深層学習を用いたモバイル UI の視覚的重要性予測手法の提案
Predicting Visual Importance of Mobile UI Using Deep Learning

山本 愛海† 清 雄一† 田原 康之† 大須賀 昭彦†
Ami YAMAMOTO Yuichi SEI Yasuyuki TAHARA Akihiko OHSUGA

1 はじめに

UI を設計する際は，どのような要素がユーザにとって重要であると感じられるかを理解することが必要である．UI 設計プロセスでは，デザイナーが作成した UI に対するフィードバックやアイトラッキングの結果をもとに UI を反復的に改良していくが，これらの反復プロセスには時間とコストがかかる．

そこで本研究では，深層学習を用いてモバイル UI の視覚的重要性を予測する手法を提案する．モバイル UI の視覚的重要性を正確に予測することで，デザイナーへのリアルタイムフィードバックやデザインの最適化を行うことができる．

本研究では，モバイル UI のスクリーンショット画像だけでなく，ボタンや画像，テキストなどの UI 要素のカテゴリや位置を表すセマンティックセグメンテーション画像も入力として用いた．特徴量を組み合わせる手法について 2 種類検討し，予測された視覚的重要性マップについて客観評価，主観評価を行った結果，提案手法においてベースラインよりも高い評価が得られた．

2 視覚的重要性

本研究では，広く研究されている視覚的顕著性ではなく視覚的重要性に着目した．視覚的顕著性は，アイトラッキングにより得られた実際の視線情報から推定されるのに対し，視覚的重要性のデータは，ユーザの視線に関わらず，ユーザがデザインを見たときに重要だと感じた部分をマッピングすることにより作成されている．そのため，視覚的重要性は位置や色合いだけでなく，テキスト，画像といった意味的なカテゴリと強く関連している [1]．

デザインに対する視覚的重要性の予測手法はいくつか研究されている [1][2]．しかし，モバイル UI の視覚的重要性予測に特化した手法は検討されておらず，モバイル UI 特有の要素である UI 要素の位置やそのカテゴリなどを加味した研究はまだ行われていない．前述の通り，視覚的重要性はボタンや画像，テキストといった UI 要素に関連しているため，UI 要素の位置やカテゴリを表すセマンティックセグメンテーション画像を用いることで，予測性能が向上すると考えた．

3 提案手法

本研究では，自然画像に対する顕著性予測モデルである MSI-Net[3] をベースとした．MSI-Net はエンコーダ・デコーダ構造をとり，マルチスケール特徴を抽出するために異なる拡張率で複数の畳み込み層をもつ Atrous Spatial Pyramid Pooling(ASPP) モジュール [4] を組み込んでいる．

また，UI 特有の特徴を抽出するために，UI 要素のカテゴリや位置を表すセマンティックセグメンテーション画像のオートエンコーダを活用した．MSI-Net と同じ構造のエンコーダとデコーダを繋げることでオートエンコーダを構築し，Rico[5] 中の，Imp1k[2] に含まれている UI に対応するセマンティックセグメンテーション画像を再構成するように学習させた．本研究では，セマンティックセグメンテーション画像の特徴量を抽出するためにオートエンコーダのエンコーダ部分のみを用い，重みは固定した．提案手法の全体像を図 1 に示す．

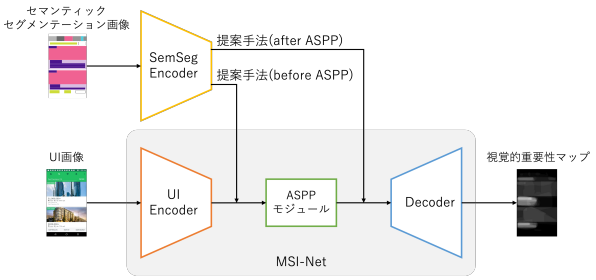

図 1 提案手法の全体像

提案手法では，セマンティックセグメンテーション画像のエンコーダ (SemSeg エンコーダ) と，UI 画像のエンコーダ (UI エンコーダ) で同じアーキテクチャを使用している．そのため，それぞれのエンコーダの出力の空間的特徴はある程度共通している部分があると考えた．そこで提案手法 (before ASPP) では，SemSeg エンコーダの出力と，MSI-Net の ASPP モジュール前のプーリング層の出力を組み合わせ，視覚的重要性予測モデルを構築している．ASPP モジュールの入力にはその前の 2 つのプーリング層の出力も含まれているため，組み合わせる特徴量の影響範囲は狭くなると考えられる．

そこで，よりセマンティックセグメンテーションの影響を強くするために，提案手法 (after ASPP) では SemSeg エンコーダの出力と，ASPP モジュール後の畳み込み層の出力を組み合わせた．

特徴量を組み合わせる方法として，今回はリストを平均する Average レイヤーを用いた．単純に足し合わせるのではなく 2 つの特徴量の平均をとっているため，重要性マップの値が大きくなりすぎず，セマンティックセグメンテーションの特徴をバランス良く取り入れることができるのではないかと考えられる．本研究では，SemSeg エンコーダを使用しないモデル (MSI-Net) をベースラインモデルとして，提案手法との比較を行う．

4 実験

ベースとするモデルの学習については，ImageNet で学習済みの重みを初期値として，SALICON[6] で公開されているトレーニングセット 10,000 枚，テストセット 5,000 枚を使用して事前学習を行った．提案手法においては，その後，学習済みの SemSeg エンコーダの出力を Average レイヤーによって組み合わせるようにモデルを変更し，Imp1k のモバイル UI データ（トレーニングセッ

† 電気通信大学 大学院情報理工学研究科
Graduate School of Informatics and Engineering,
The University of Electro-Communications

ト 160 枚，テストセット 40 枚）のみを使用してファインチューニングを行った．ベースライン手法においては，モデルの変更は行わず，Imp1k のモバイル UI データのみを使用してファインチューニングを行った．

本研究では，損失関数に KL ダイバージェンスを使用した．KL ダイバージェンスは予測が外れた場合に大きなペナルティを与えるため，顕著なターゲット検出を目的とするモデルに適している．また，ファインチューニングにおいて，バッチサイズ 4，学習率は 1e-4 とし，最適化関数には Adam を用いた．

4.1 評価

顕著性や視覚的重要性マップの予測モデルの性能を評価するために，様々な評価指標が用いられている．今回の研究では，視覚的重要性の先行研究 [1][2] で用いられている，決定係数 (R^2)，相関係数 (CC)，平均平方二乗誤差 ($RMSE$)，KL ダイバージェンス (KL) の 4 つの指標を用いて客観評価を行った．

評価結果を表 1 に示す．

表 1　視覚的重要性予測モデルの性能

	$R^2 \uparrow$	$CC \uparrow$	$RMSE \downarrow$	$KL \downarrow$
ベースライン	0.505	0.841	0.102	0.151
before ASPP	**0.523**	**0.852**	**0.0993**	**0.145**
after ASPP	0.466	0.847	0.103	0.149

また，ベースライン手法と提案手法により予測された視覚的重要性マップの例を図 2 に示す．

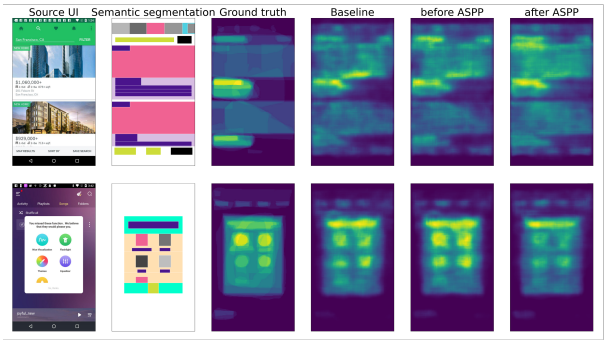

図 2　予測された視覚的重要性マップの例

5 考察

表 1 より，全ての評価指標において提案手法 (before ASPP) が最も良い結果となった．また，提案手法 (after ASPP) は，CC と KL の指標でベースラインよりも優れた結果となったが，R^2 の指標ではベースラインよりも大きく劣る結果となった．

図 2 より，ベースライン手法と比較して，提案手法で予測した視覚的重要性マップは全体的にやや滑らかになっている．また，上段の例において，提案手法 (before ASPP) では画像内の建造物のエッジを残しながらも UI 全体の重要性を予測することができているが，提案手法 (after ASPP) では画像内の特徴が消えてしまっているように見える．提案手法 (after ASPP) ではセマンティックセグメンテーションの影響が強くなりすぎて，ベースの特徴量を壊してしまう場合があると考えられる．

また，下段の例において，提案手法 (before ASPP) ではベースライン手法よりもアイコンの重要性が高くなっている．セマンティックセグメンテーション画像には，UI 要素の位置だけでなく，アイコンやテキストなどのカテゴリ情報も含まれているため，UI 要素のセマンティックセグメンテーション画像の特徴量を適切に用いることで，視覚的重要性マップの予測性能が向上することがわかった．

6 おわりに

本研究では，モバイル UI の視覚的重要性を正確に予測することを目的とし，深層学習を用いて UI 要素のセマンティックセグメンテーション画像の特徴を組み合わせる手法を 2 つ提案した．客観評価では，UI 要素のセマンティックセグメンテーション画像を用いないベースライン手法と比較を行い，提案手法 (before ASPP) においてベースラインよりも高い評価が得られた．提案手法 (before ASPP) で予測された視覚的重要性マップはベースライン手法よりも滑らかになり，より Ground truth に近いマップを予測することができた．

本研究では，SemSeg エンコーダの重みを固定し，ベースとなるモデルに後付けすることでモデルを構築している．そのため，特に提案手法 (after ASPP) では事前学習した重みの影響が薄くなり，効果的な学習が行えていないと考えられる．そこで，今後の展望として，事前学習の段階でセマンティックセグメンテーション画像を用いて学習を進めることを検証したい．また，本研究で提案した視覚的重要性予測モデルを活用し，モバイル UI デザインの最適化や，デザイナーへのフィードバックツールへの応用も検討したい．

謝辞

本研究は JSPS 科研費 JP21H03496, JP22K12157 の助成を受けたものです．

参考文献

[1] Z. Bylinskii, N. W. Kim, P. O’ Donovan, S. Alsheikh, S. Madan, H. Pfister, F. Durand, B. Russell, and A. Hertzmann, “Learning Visual Importance for Graphic Designs and Data Visualizations,” Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, pp. 57-69, 2017.

[2] C. Fosco, V. Casser, A. K. Bedi, P. O’Donovan, A. Hertzmann, and Z. Bylinskii, “Predicting Visual Importance Across Graphic Design Types,” Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, pp. 249–260, 2020.

[3] A. Kroner, M. Senden, K. Driessens, and R. Goebel, “Contextual encoder–decoder network for visual saliency prediction,” Neural Networks 129, pp. 261–270, 2020.

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4), pp. 834–848, 2018

[5] B. Deka, Z. Huang, C. Franzen, J. Hibschman, D. Afergan, Y. Li, J. Nichols, and R. Kumar, “Rico: A mobile app dataset for building data-driven design applications,” Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, pp. 845–854, 2017.

[6] M. Jiang, S. Huang, J. Duan, and Q. Zhao, “SALICON: Saliency in Context,” IEEE conference on computer vision and pattern recognition, pp. 1072–1080, 2015.