

個人間の視線パターンに基づくFew-Shotパーソナライズ顕著性予測

諸戸 祐哉[†]、前田 圭介^{††}、小川 貴弘[†]、長谷山 美紀[†]

本研究では、対人視線パターンを活用した数ショットのパーソナライズされた顕著性予測手法を提案する。一般的な顕著性マップとは異なり、パーソナライズされた顕著性マップ(PSM)は個人の視覚的注意を捉え、個人の視覚的嗜好に関する洞察を提供する。しかし、視線パターンが複雑で、個人から広範な視線追跡データを収集することが困難であるため、PSMを予測することは困難である。限られたデータからPSMを予測するための効果的な戦略は、他の人の視線追跡データを使用することである。他者のPSMを効率的に扱うために、本研究では、視線追跡データを取得する画像の選択と、PSMの構造情報の保存に焦点を当てる。提案手法では、これらの画像は人物により多様な視線パターンをもたらすように選択され、テンソルベースの回帰を用いて構造情報が保存される。実験結果は、これら2つの要因が少数ショットのPSM予測に有益であることを示している。

キーワード：顕著性予測、パーソナライズ顕著性マップ、テンソル回帰、人物類似度、適応的画

1. まえがき

人間は視覚システムを通じて、複雑な現実世界にある豊富な視覚情報から選択的に重要な情報を得ることができる。多くの研究者が、このような人間のメカニズムを画像処理モデルに導入することを試みている¹⁾²⁾³⁾。具体的には、近傍の部位よりも顕著な部位を表す顕著性マップは、人間の本能的な視覚知覚¹⁾⁴⁾⁵⁾⁶⁾⁷⁾を再現すると予測される。パーソナライズなしで、各画像に対して顕著性マップを予測する。しかし、同じシーンを見たとしても、個人によって異なる領域に焦点が当てられる、つまり、個人差が存在する⁸⁾⁹⁾¹⁰⁾。個々の視覚的注意をモデル化するために、顕著性マップはここ数年パーソナライズされている¹¹⁾¹²⁾¹³⁾¹⁴⁾¹⁵⁾。従来の顕著性マップとそのパーソナライゼーションは、それぞれユニバーサル・サリエンシー・マップ(USM)とパーソナライズド・サリエンシー・マップ(PSM)と呼ぶことで区別される。USMは個人差を省略し、PSMは各個人について予測する。

パーソナライズされた視覚的嗜好は、PSM間の違いによって反映される¹⁶⁾¹⁷⁾¹⁸⁾；したがって、個性は多くの状況で有用である。このような個性は、個人の視覚的嗜好が重要な役割を果たすアプリケーションにおいて、大きな可能性を秘めている。例えば、PSM予測はターゲット広告に適用することができ、個人の視覚的嗜好を理解することで広告の配置を決定したり、UIデザインを作ったりすることができます¹⁹⁾²⁰⁾。もう一つの例は、画像や商品に対するユーザの興味に関連する領域やコンテンツを強調するレコメンダーシステムである²¹⁾²²⁾。さらに、PSMは、言葉にすることは難しいが習慣的な行動や内面化された専門知識に現れる知識である暗黙知に関するスキルの伝達に貢献することができる²³⁾²⁴⁾。これらのアプリケーションは、USMを超えるPSMの予測の実用的な妥当性を示している。ここで、未視認画像に対するPSMを事前に求めるためには、個々の視線パターンの傾向からPSMを予測する必要がある。

個々の視線パターンをモデル化するために、視覚刺激、例えば画像と個々のPSMの関係は、過去に各人から得られた視線追跡データに基づいて分析されるべきである。そして、画像に現れる視線パターンは複雑で異なるため、これらの特徴がPSM予測の難しさにつながる。個々の視線パターン傾向を抽出するために、いくつかの研究者が数千の画像の視線追跡データを収集している¹¹⁾¹²⁾¹⁴⁾¹⁶⁾。

Received February 3, 2025; Revised May 24, 2025; Accepted July 18, 2025

[†] Graduate School of Information Science and Technology, Hokkaido University
(Sapporo, Japan)

^{††} Faculty of Information Science and Technology, Hokkaido University
(Sapporo, Japan)

【機械翻訳コンテンツの著作権について】

当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
引用の際には、必ず原著論文の書誌情報をご記載ください。詳細は映像情報メディア学会著作権規定をご覧ください。

Few-shot Personalized Saliency Prediction Based on Interpersonal Gaze Patterns

Yuya Moroto[†], Keisuke Maeda^{††}, Takahiro Ogawa (member)^{††},
Miki Haseyama (member)^{††}

Abstract This study proposes a few-shot personalized saliency prediction method that leverages interpersonal gaze patterns. Unlike general saliency maps, personalized saliency maps (PSMs) capture individual visual attention and provide insights into individual visual preferences. However, predicting PSMs is challenging because of the complexity of gaze patterns and the difficulty of collecting extensive eye-tracking data from individuals. An effective strategy for predicting PSMs from limited data is the use of eye-tracking data from other persons. To efficiently handle the PSMs of other persons, this study focuses on the selection of images to acquire eye-tracking data and the preservation of the structural information of PSMs. In the proposed method, these images are selected such that they bring more diverse gaze patterns to persons, and structural information is preserved using tensor-based regression. The experimental results demonstrate that these two factors are beneficial for few-shot PSM prediction.

Key words: Saliency prediction, personalized saliency map, tensor-based regression, person similarity, adaptive image selection.

1. Introduction

Humans can selectively obtain vital information from the abundant visual information in the complex real world through their visual system. Many researchers have attempted to introduce such human mechanisms into image processing models¹⁾²⁾³⁾. Specifically, a saliency map, which represents the salient parts that are more noticeable than the neighboring parts, is predicted to reproduce the human instinctive visual perception¹⁾⁴⁾⁵⁾⁶⁾⁷⁾. A saliency map is predicted for each image without personalization. However, different individuals focus on different areas even when viewing the same scene, that is, individual differences exist⁸⁾⁹⁾¹⁰⁾. To model individual visual attention, saliency maps have been personalized over the past few years¹¹⁾¹²⁾¹³⁾¹⁴⁾¹⁵⁾. A traditional saliency map and its personalization are distinguished by referring to them as a universal saliency map (USM) and a personalized saliency map (PSM), respectively. A USM omits individual differences, whereas a PSM is predicted for each person. Personalized visual preferences can be reflected

by differences between PSMs¹⁶⁾¹⁷⁾¹⁸⁾; thus, individuality can be useful in many situations. Such individuality has significant potential in applications where personal visual preferences play a crucial role. For instance, PSM prediction can be applied to targeted advertising, where understanding individual visual preference helps determine advertisement placements or make UI designs¹⁹⁾²⁰⁾. Another example is a recommender system, which highlights the regions or contents related to users' interests in images or products²¹⁾²²⁾. Moreover, PSMs can contribute to transferring skills related to tacit knowledge, which is knowledge that is difficult to verbalize but manifests in habitual behaviors or internalized expertise²³⁾²⁴⁾. These applications demonstrate the practical relevance of predicting PSMs beyond USMs. Here, to obtain a PSM for unseen images in advance, the PSM should be predicted from the individual gaze pattern tendency.

To model individual gaze patterns, the relationship between visual stimuli, e.g., images, and the individual PSM should be analyzed based on eye-tracking data obtained from each person in the past. Then, the gaze patterns emerging in the images are complex and different, and these characteristics lead to the difficulty of PSM prediction. To extract individual gaze pattern tendencies, several researchers have collected eye-tracking data for thousands of images¹¹⁾¹²⁾¹⁴⁾¹⁶⁾. The prediction mod-

Received February 3, 2025; Revised May 24, 2025; Accepted July 18, 2025

[†] Graduate School of Information Science and Technology, Hokkaido University
(Sapporo, Japan)

^{††} Faculty of Information Science and Technology, Hokkaido University
(Sapporo, Japan)

これらの研究で採用された予測モデルは、ディープラーニングに基づいており、各人のために大量の学習データを必要とする。大規模なPSMデータセットは公開されているが、大量の個々のアイトラッキングデータの取得は、アプリケーション内の新規人物にとって大きな負担となり、時間のかかる作業となる可能性がある。その結果、限られた量の学習用視線追跡データでPSM予測を行う必要がある。

限られた量のデータからPSMを予測するためには、対象人物と類似した視線パターンを持つ人物から得られたPSMを利用することが有効な戦略である。ある人物が対象人物の視線パターンと類似した視線パターンを持つかどうかを判断するためには、同じ画像に対する複数の視線追跡データのペアが必要である。しかし、このようなペアを大量に取得することはできず、アイトラッキングデータを取得するための画像選択は重要なプロセスである。以前の研究²⁵⁾では、視線の散乱を誘発する画像を適応的画像選択(AIS)を用いて選択し、対象者と他者(本研究では訓練人物と呼ぶ)との間の視線パターンの類似性を効率的かつ着実に得ることができた。さらに、先行研究¹³⁾では、協調多出力ガウス過程回帰(CoMOGP)²⁶⁾を、PSM予測のための訓練人から得られたPSMと共に用いた。しかし、このような回帰に基づく方法では、ベクトル形式の入力が必要であり、PSMの構造情報を効果的に利用することはできない。

構造情報は、人間の視覚系における顕著な領域を検出するための重要な手がかりとなる¹⁾。

先行研究¹⁾で提案された方法は、手作業で作成された画像特徴を抽出し、それらの中心-周囲差分を取る。本手法では、特徴抽出の結果、画素位置とその関係を考慮した複数の特徴マップを算出する、つまり、2次元の空間構成が保存される。このプロセスに触発され、我々は、このような2次元空間構成がPSM予測に有用であり、画像内の顕著な領域の相対的な位置と分布を捉えるという仮説を立てた。このように、本論文では、PSM予測の性能向上のための構造情報として、2次元空間構成に着目している。次に、複数の訓練者に対して予測されたPSMを効果的に利用するために、構造情報を考慮するPSM予測手法を構築する必要がある。

したがって、少数ショットのPSM予測を改善するためには、視線追跡データを取得するための適応的な画像選択と、訓練者のために予測されたPSMの構造情報の保存を協調的に取り入れることが望ましい。

対人視線パターンに基づく数ショットのパーソナライズされた顕著性予測法を提案する。提案手法では、AIS²⁵⁾とテンソルベース回帰モデル²⁷⁾を協調的に用いる。AISスキームは、選択された画像の多様性と、画像を選択するためのトレーニング人物から得られるPSMの多様性に焦点を当てる。AISスキームにより、対象者と訓練者の間の視線パターンの類似性を効率的かつ着実に得ることができる。また、テンソルベース回帰モデル²⁷⁾の入力と出力はマルチアレイテンソル形式であるため、構造情報を保持したまま、学習者のPSMから対象者のPSMを予測する。したがって、アイトラッキングデータを取得するための効果的な画像選択と、トレーニング対象者のために予測されたPSMの構造情報の保存を実現する。

2. 関連作品

2.1 USM予測

画像処理の分野では、USM予測は伝統的な研究対象である。具体的には、ディープラーニング手法¹⁾⁴⁾⁵⁾が開発されるまで、手作業で作成した画像特徴量に基づいて初期のUSMモデルを構築した。対照的に、深層学習手法、例えば畳み込みニューラルネットワーク(CNN)、生成的敵対ネットワーク(GAN)、視覚変換器は、学習フェーズを必要としないこれらのモデルを凌駕している⁶⁾⁷⁾。多くのUSM予測手法が提案されているが、個人差を考慮していないため、PSM予測の性能向上という点では限界がある。

2.2 PSM予測

測定器の進歩は、過去10年間で、PSM予測への関心を喚起してきた。オープンな大規模データセットは、PSM予測モデルの構築に大きく貢献する¹¹⁾。具体的には、マルチタスクCNNベースのモデルは、USMとPSMの違いに注目することで、高精度なPSM予測を実現した¹¹⁾。また、先行研究¹⁶⁾の拡張版として、人物固有情報符号化フィルタを用いたCNNモデル(CNNPIEF)が提案された。CNN-PIEFでは、人物固有の情報を埋め込むことで、予測モデルのパーソナライズが可能になる。

els adopted in these studies are based on deep learning, which requires a massive amount of training data for each person. The large-scale PSM dataset is openly available; however, the acquisition of a massive amount of individual eye-tracking data can be a significant burden and time-consuming task for new persons in the application. Consequently, a PSM prediction method with a limited amount of training eye-tracking data is required.

To predict a PSM from a limited amount of data, an effective strategy is to use PSMs obtained from persons with similar gaze patterns to the target person. To determine whether a person has gaze patterns similar to those of the target person, several pairs of eye-tracking data for the same images are required. However, such pairs cannot be acquired in large quantities, and the selection of images to acquire eye-tracking data is an important process. In a previous study²⁵⁾, images that induce the scattering of gazes were selected using adaptive image selection (AIS) to efficiently and steadily obtain the similarity of gaze patterns between the target and other persons (called training persons in this study). Additionally, in a previous study¹³⁾, the collaborative multi-output Gaussian process regression (CoMOGP)²⁶⁾ was used with the PSMs obtained from training persons for PSM prediction. However, such regression-based methods require vector-format input, and the structural information of PSMs cannot be effectively used.

Structural information is an important clue to detecting salient areas in the human visual system¹⁾. The method proposed in the previous work¹⁾ extracts hand-crafted image features and takes the center-surround differences of them. In this method, as a result of feature extraction, several feature maps are calculated with considering the pixel positions and their relationships, that is, the two-dimensional spatial configuration is preserved. Inspired by this process, we hypothesized that such two-dimensional spatial configuration is useful for PSM prediction and captures the relative positioning and distribution of salient regions within an image. In this way, this paper focuses on the two-dimensional spatial configuration as the structural information for performance improvement of PSM prediction. Then, it is necessary to construct a PSM prediction method that considers structural information that is compatible with the effective use of PSMs predicted for several training persons. Therefore, to improve few-shot PSM prediction, it is desirable to collaboratively

incorporate the adaptive selection of images to acquire eye-tracking data and preservation of the structural information of PSMs predicted for training persons.

We propose a few-shot personalized saliency prediction method based on interpersonal gaze patterns. In the proposed method, we collaboratively use AIS²⁵⁾ and the tensor-based regression model²⁷⁾. The AIS scheme focuses on the variety of selected images and the variation in PSMs obtained from the training persons for selecting images. Through the AIS scheme, we can efficiently and steadily obtain the similarity of gaze patterns between the target and training persons. In addition, the input and output of the tensor-based regression model²⁷⁾ are in a multi-array tensor format; thus, it predicts the PSMs of the target person from the PSMs of training persons while preserving the structural information. Therefore, we realize the effective selection of images to acquire eye-tracking data and preservation of the structural information of PSMs predicted for training persons.

2. Related Works

2.1 USM Prediction

In the field of image processing, USM prediction is a traditional research subject. Specifically, early USM models were constructed based on hand-crafted image features until the development of deep learning methods¹⁾⁴⁾⁵⁾. In contrast, deep learning methods, e.g., convolutional neural networks (CNNs), generative adversarial networks (GANs), and vision transformers, have outperformed these models, which do not require training phase⁶⁾⁷⁾. Although many USM prediction methods have been proposed, they have limitations in terms of the performance improvement of PSM prediction because they do not account for individual differences.

2.2 PSM Prediction

The advancement of measurement instruments has sparked interest in PSM prediction over the last decade. The open large-scale dataset significantly contributes to the construction of PSM prediction models¹¹⁾. Specifically, a multi-task CNN-based model achieved highly accurate PSM prediction by focusing on the difference between USMs and PSMs¹¹⁾. In addition, a CNN model with person-specific information encoded filters (CNN-PIEF) was proposed as an extended version of the previous study¹⁶⁾. In CNN-PIEF, the embeddings of person-specific information enable the personalization of the prediction model. In addition, in a previous work²⁹⁾, a model based on conditional GANs with person clusters

表1 USMとPSMの代表的な予測手法の比較。

方法	パーソナライゼーション	Order of 学習データ	Additional Data
USM予測			
計算モデル ¹⁾⁴⁾⁵⁾	-	-	-
ディープラーニングに基づくUSM予測 ⁶⁾⁷⁾	-	10 ⁴	-
PSM予測			
マルチタスクCNN ¹¹⁾	✓	10 ³ / Person	-
CNN-PIEF ¹⁶⁾	✓	10 ³ / Person	人特有の情報
Sherkati et al. ²⁹⁾	✓	10 ³ / Person	人特有の情報
Strohmら ³⁰⁾	✓	10 ³ / Person	事前学習済み人物埋め込みネットワーク
人物類似度ベースのアプローチ ¹³⁾²⁵⁾³¹⁾	✓	10 ² / Person	他人のアイトラッキングデータ
(Our setting)			

また、先行研究²⁹⁾では、人物固有情報から構築された人物クラスターを持つ条件付きGANに基づくモデルが提案され、別の研究³⁰⁾では、ユーザ埋め込みを学習するためのシャムCNNに基づくモデルが提案された。これらのモデルは、ディープラーニングを用いてPSMを予測することに成功した。しかし、ディープラーニングはCNNを訓練するのに十分な量の訓練視線追跡データを必要とし、新しい人物の予測を行うには相当量の訓練データが必要である。

この点に関して、いくつかの研究では、対象人物と類似した視線パターンを持つ人物から得られたアイトラッキングデータを用いて、学習データの量を減らすことが試みられている¹³⁾²⁵⁾³¹⁾。しかし、これらの方法では、PSMの構造情報を効果的に利用することはできない。構造情報は、人間の視覚系における顕著な領域を検出するための重要な手がかりとなる¹⁾。したがって、複数の訓練者に対して予測されたPSMを効果的に利用するためには、構造情報を考慮するPSM予測手法を構築する必要がある。

USM予測とPSM予測の代表的な手法の比較を表1に示す。表に示すように、USM予測手法は様々な人から視線データを収集することで、大量の学習データで学習することができるが、PSM予測手法はパーソナライズされた予測のため、各人の限られた学習データで学習する必要がある。さらに、深層学習ベースのPSM予測手法¹¹⁾¹⁶⁾²⁹⁾³⁰⁾は、依然として1人当たり約1,000個のアイトラッキングデータという大量の学習データを必要とし、これは新規人物からの学習データの収集に大きな負担となる。CNN-PIEF¹⁶⁾やSherkatiら²⁹⁾に関しては、PSMは年齢や性別などの人物固有の情報を用いて予測できるが、実世界のアプリケーションではそのような情報は必ずしも利用できない。

Strohmらの研究³⁰⁾では、人物埋め込みネットワークは大量の学習データで事前学習される。一方、我々の設定は、1人当たり100個の訓練データという中程度の量の訓練データを必要とするが、パーソナライゼーションとスケーラビリティのバランスが適切である。そのため、学習データや研究目的が異なるため、性能比較が困難である。

3. 提案する数ショットPSM予測

提案する少数ショットPSM予測は3つのフェーズから構成され、そのフロー全体を図1に示す。ここでは、大量の視線追跡データを持つ訓練者P人と、限られた量の視線追跡データを持つ対象人がいると仮定する。まず、マルチタスクCNN²⁸⁾を学習させ、先行研究¹¹⁾¹⁶⁾を参照して、学習者のPSMを予測する。次に、AISスキーム²⁵⁾に基づいて、対象者が注視する共通画像を選択する。共通画像は、より多様な視線パターンを人物にもたらしように選択されている。最後に、提案手法は、訓練者のPSMを用いてテンソル対行列回帰²⁷⁾を用いてPSMを予測する。したがって、対人視線パターンを効率的に利用することで、効果的に画像を選択し、視線追跡データを取得し、人物の訓練用に予測されたPSMの構造情報を保持することができる。

3.1 トレーニング用マルチタスクCNN

マルチタスクCNNモデル²⁸⁾を学習するために、学習画像 $X_n \in \mathbb{R}^{d^1 \times d^2 \times d^3}$ ($n = 1, 2, \dots, N$; N は学習画像の数)とそのUSM $U(X_n) \in \mathbb{R}^{d^1 \times d^2}$ 、ここで $d_1 \times d_2$ と d_3 はそれぞれ画像とカラーチャンネルのサイズを表す。

Table 1 Comparison of representative methods for USM and PSM prediction.

Method	Personalization	Order of Training data	Additional Data
USM Prediction			
Computational models ¹⁾⁴⁾⁵⁾	-	-	-
Deep learning-based USM prediction ⁶⁾⁷⁾	-	10 ⁴	-
PSM Prediction			
Multi-task CNN ¹¹⁾	✓	10 ³ / Person	-
CNN-PIEF ¹⁶⁾	✓	10 ³ / Person	Person-specific information
Sherkati et al. ²⁹⁾	✓	10 ³ / Person	Person-specific information
Strohm et al. ³⁰⁾	✓	10 ³ / Person	Pretrained person embedding network
Person similarity-based approach ¹³⁾²⁵⁾³¹⁾ (Our setting)	✓	10 ² / Person	Other person's eye tracking data

constructed from person-specific information was proposed, and in another study³⁰⁾, a siamese CNN-based model for learning user embedding was proposed. These models successfully predicted PSMs using deep learning. However, deep learning requires a sufficient amount of training eye-tracking data to train CNNs, and significant amounts of training data are required to make predictions for a new person.

In this regard, several studies have attempted to reduce the amount of training data by using eye-tracking data obtained from persons with similar gaze patterns to the target person¹³⁾²⁵⁾³¹⁾. However, in these methods, the structural information of PSMs cannot be effectively used. Structural information is an important clue to detecting salient areas in the human visual system¹⁾. Thus, it is necessary to construct a PSM prediction method that considers structural information that is compatible with the effective use of PSMs predicted for several training persons.

We show the comparison of the representative methods for USM and PSM prediction in Table 1. As shown in the table, while the USM prediction methods can be trained on a large amount of training data by collecting gaze data from various people, the PSM prediction methods should be trained on a limited amount of training data for each person due to personalized prediction. Furthermore, deep learning-based PSM prediction methods¹¹⁾¹⁶⁾²⁹⁾³⁰⁾ still require a large amount of training data, about 1,000 eye-tracking data per person, which is a heavy burden on the collection of training data from new persons. With regard to CNN-PIEF¹⁶⁾ and Sherkati et al.²⁹⁾, PSMs can be predicted by using the person-specific information, e.g., age, and gender, but such information is not always available in the real-world applications. In the study by Strohm et al.³⁰⁾,

the person embedding network is pretrained on a large amount of training data. On the other hand, our setting requires a moderate amount of training data, 100 training data per person, but offers an adequate balance between personalization and scalability. Therefore, the performance comparison is difficult due to the difference of the training data and the research objective.

3. Proposed Few-shot PSM Prediction

The proposed few-shot PSM prediction comprises three phases, and the entire flow is depicted in Fig. 1. Here, we assume that there are P training persons with a massive amount of eye-tracking data and a target person with a limited amount of eye-tracking data. First, the multi-task CNN²⁸⁾ is trained to predict the PSMs of the training persons by referring to the previous studies¹¹⁾¹⁶⁾. Next, we select common images that the target person gazes at based on the AIS scheme²⁵⁾. The common images are selected such that they bring more diverse gaze patterns to persons. Finally, the proposed method predicts the PSM using tensor-to-matrix regression²⁷⁾ with the PSMs of the training persons. Therefore, by efficiently using the interpersonal gaze patterns, we can effectively select images to acquire eye-tracking data and preserve the structural information of PSMs predicted for training persons.

3.1 Multi-Task CNN for Training Persons

To train the multi-task CNN model²⁸⁾, we prepare the training images $\mathbf{X}_n \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ ($n = 1, 2, \dots, N$; N being the number of training images) and their USMs $\mathbf{U}(\mathbf{X}_n) \in \mathbb{R}^{d_1 \times d_2}$, where $d_1 \times d_2$ and d_3 denote the size of the image and the color channel, respectively. To effectively obtain the predicted PSMs of training persons, previous studies¹¹⁾¹⁶⁾ have adopted the specific approach of predicting the difference map

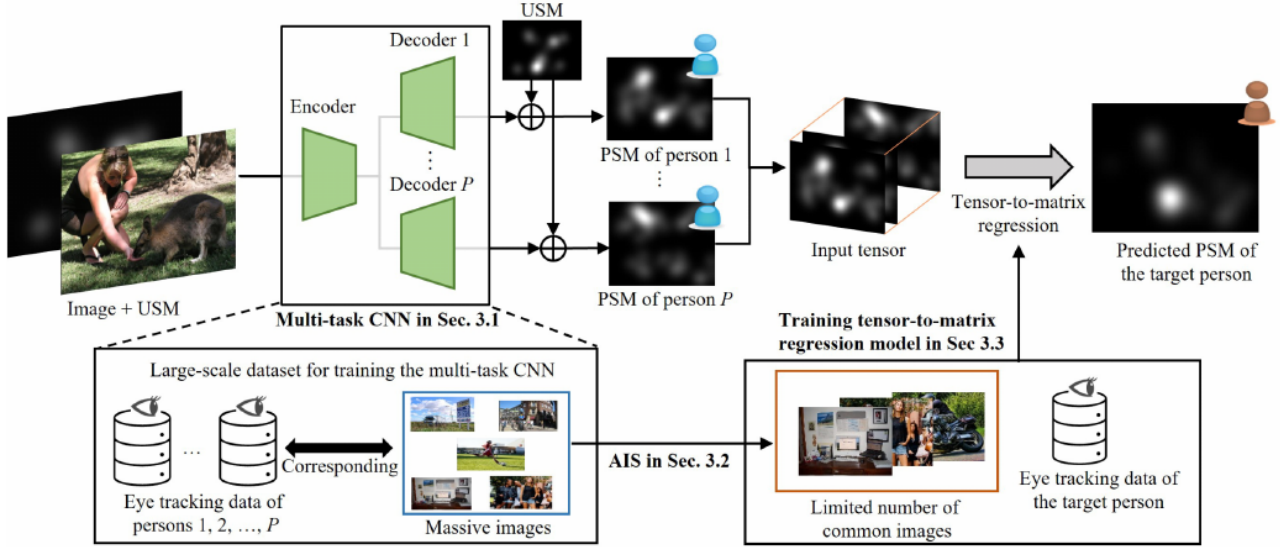


Fig. 1 提案するPSM予測手法の全フローは、3つのフェーズから構成される。第1段階では、マルチタスクCNN²⁸⁾がP人の訓練者のPSMを予測する。次に、AISスキーム²⁵⁾を用いて、対象者が注視する共通画像としてI画像を選択する。最後に、訓練者のPSMを用いてテンソル対行列回帰²⁷⁾を用いてPSMを予測する。

$M(X)_p \in \mathbb{R}^{d^1 \times d^2}$ ($p = 1, 2, \dots, P$) のようにUSMとPSMの間で:

$$M(X)_p = S(X)_p - U(X), \quad (1)$$

トレーニング人物の予測PSMを効果的に求めるために、先行研究¹¹⁾¹⁶⁾では、画像Xの視線追跡データに基づいて、 $S(X)_p$ がp番目のトレーニング人物のPSMを表す差分マップを予測する特定のアプローチを採用している。次に、トレーニング人物のPSMを同時に予測するために、1つの画像エンコーダとP個のPSMデコーダからなるマルチタスクCNNを構築し、以下の目的関数を最小化することにより、それらの学習可能なパラメータを最適化する:

$$\sum_{p=1}^P \sum_{n=1}^N \sum_{l=1}^L \|\hat{M}_l(X_n)_p - M(X_n)_p\|_F^2, \quad (2)$$

ここで、 $\hat{M}_l(X_n)_p$ ($l = 1, 2, \dots, L$; Lは1つのデコーダの畳み込み層の数)はl番目の層から計算された予測差分マップを表し、 $\|\cdot\|_F^2$ はフロベニウス・ノルムを表す。

テスト画像 X_{tst} が与えられたとき、p番目の人物の予測PSMは以下のように計算される:

$$\hat{S}(X_{tst})_p = \hat{M}_L(X_{tst})_p + U(X_{tst}). \quad (3)$$

したがって、マルチタスクCNNは、訓練者のPSMを同時に予測し、これらのPSMの関係を考慮することができる。

3.2 PSM予測のための適応的画像選択

N枚のトレーニング画像から数枚を選択し、ターゲットとトレーニング人物の傾向が類似していることを取得する。このような類似性を効果的に分析するために、AISスキーム²⁵⁾を用いて、人物により多様な視線パターンを生成するI共通画像を選択する。具体的には、AISスキームは、訓練者から得られた様々な共通画像とPSMのバリエーションに焦点を当てる。これらの要因を同時に考慮するために、AISスキームでは、各画像のオブジェクトのPSMのばらつきを利用する。

まず、学習画像 X_n の各オブジェクト $B_{n,j}$ ($j = 1, 2, \dots, J$; Jは学習画像中のオブジェクトカテゴリの数)についてPSMとその分散を計算する。次に、物体検出³²⁾を学習画像に適用し、i番目の画像中のj番目の物体に対して、 $d_{n,j}^h \times d_{n,j}^w$ の次元を持つ矩形を得る。物体 $B_{n,j}$ のPSM分散 $q_{n,j}$ は以下のように計算される:

$$q_{n,j} = \frac{1}{d_{n,j}^h d_{n,j}^w P} \sum_{p=1}^P \|\bar{S}(B_{n,j})_p \odot \bar{S}(B_{n,j})_p\|_F, \quad (4)$$

$$\bar{S}(B_{n,j})_p = S(B_{n,j})_p - \frac{1}{P} \sum_{p=1}^P S(B_{n,j})_p, \quad (5)$$

ここで、 $S(B_{n,j})_p$ は人物pの物体 $B_{n,j}$ に対するPSMを表し、ハダマード積の演算子を表す。

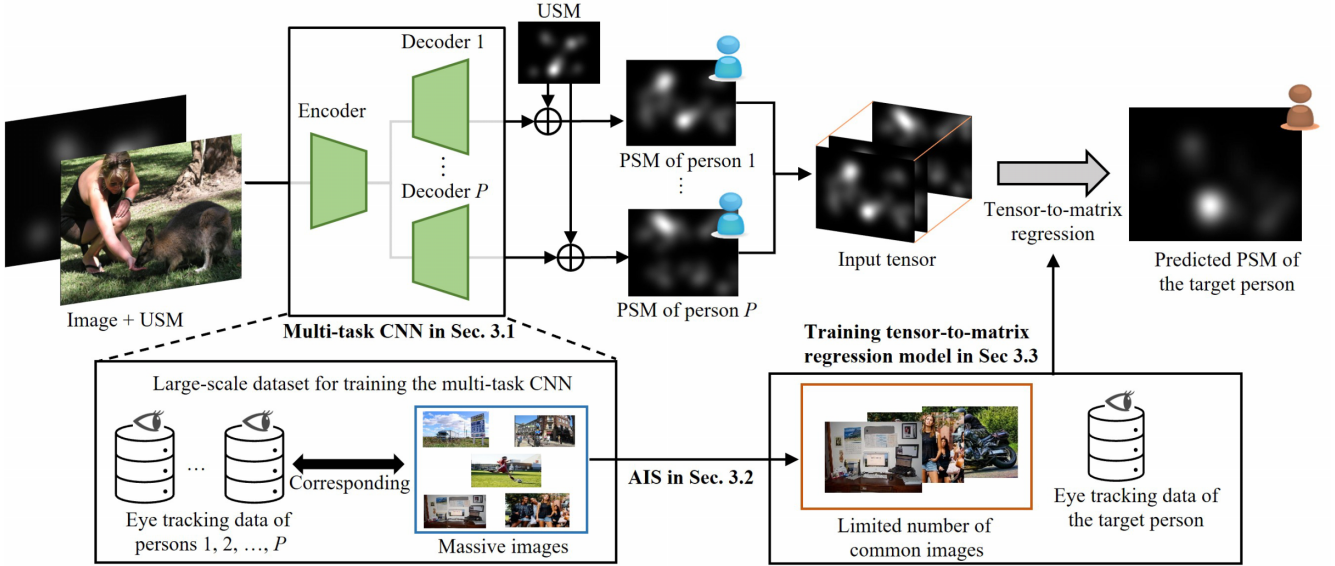


Fig. 1 Entire flow of the proposed PSM prediction method consisting of three phases. In the first phase, the multi-task CNN²⁸⁾ predicts the PSMs of P training persons. Next, using the AIS scheme²⁵⁾, I images are selected as common images that the target person gazes at. Finally, the PSM is predicted using tensor-to-matrix regression²⁷⁾ with the PSMs of training persons.

$M(\mathbf{X})_p \in \mathbb{R}^{d_1 \times d_2}$ ($p = 1, 2, \dots, P$) between USMs and PSMs as follows:

$$M(\mathbf{X})_p = \mathbf{S}(\mathbf{X})_p - \mathbf{U}(\mathbf{X}), \quad (1)$$

where $\mathbf{S}(\mathbf{X})_p$ denotes the PSM of the p th training person based on the eye-tracking data for the image $\mathbf{X}$. Next, to simultaneously predict the PSMs of training persons, we construct a multi-task CNN consisting of one image encoder and P PSM decoders and optimize their trainable parameters by minimizing the following objective function:

$$\sum_{p=1}^P \sum_{n=1}^N \sum_{l=1}^L \|\hat{M}_l(\mathbf{X}_n)_p - M(\mathbf{X}_n)_p\|_F^2, \quad (2)$$

where $\hat{M}_l(\mathbf{X}_n)_p$ ($l = 1, 2, \dots, L$; L being the number of convolution layers in one decoder) denotes a predicted difference map calculated from the l th layer, and $\|\cdot\|_F^2$ denotes the Frobenius norm.

Given the test image $\mathbf{X}_{\text{tst}}$, the predicted PSM of the p th person is calculated as follows:

$$\hat{\mathbf{S}}(\mathbf{X}_{\text{tst}})_p = \hat{M}_L(\mathbf{X}_{\text{tst}})_p + \mathbf{U}(\mathbf{X}_{\text{tst}}). \quad (3)$$

Therefore, the multi-task CNN can simultaneously predict the PSMs of the training persons and consider the relationship between these PSMs.

3.2 Adaptive Image Selection for PSM Prediction

We select a few images from the N training images to obtain the tendency of the target and training persons to be similar. To effectively analyze such similarity, the I common images that produce more diverse gaze patterns to persons are selected using the AIS scheme²⁵⁾. Specifically, the AIS scheme focuses on various common images and variations in PSMs obtained from the training persons. To simultaneously consider these factors, the AIS scheme uses the variation in PSMs for objects in each image.

First, we calculate the PSMs and their variance for each object $\mathbf{B}_{n,j}$ ($j = 1, 2, \dots, J$; J being the number of object categories in the training images) in the training images $\mathbf{X}_n$. Then, object detection³²⁾ is applied to the training images to obtain a rectangle with dimensions of $d_{n,j}^h \times d_{n,j}^w$ for the j th object in the i th image. The PSM variance $q_{n,j}$ for object $\mathbf{B}_{n,j}$ is calculated as follows:

$$q_{n,j} = \frac{1}{d_{n,j}^h d_{n,j}^w P} \sum_{p=1}^P \|\bar{\mathbf{S}}(\mathbf{B}_{n,j})_p \odot \bar{\mathbf{S}}(\mathbf{B}_{n,j})_p\|_F^1, \quad (4)$$

$$\bar{\mathbf{S}}(\mathbf{B}_{n,j})_p = \mathbf{S}(\mathbf{B}_{n,j})_p - \frac{1}{P} \sum_{p=1}^P \mathbf{S}(\mathbf{B}_{n,j})_p, \quad (5)$$

where $\mathbf{S}(\mathbf{B}_{n,j})_p$ denotes the PSM for object $\mathbf{B}_{n,j}$ of person p , and $\odot$ denotes the operator of the Hadamard

次に、 X_n がj番目のオブジェクトを含まない場合に $q_{n,j} = 0$ とし、画像 X_n が複数のm番目のオブジェクトを含む場合に最大の $q_{n,j}$ を設定する。次に、n番目の画像に対する $q_{n,j}$ の和を次のように求める：

$$\bar{q}_n = \sum_{j=1}^J q_{n,j}. \quad (6)$$

最後に、 $\bar{q}_n$ を用いて、これらの画像中のオブジェクトカテゴリ数を最大化する制約の下で、上位I枚の画像を共通画像として選択する。その結果、選択された共通画像は複数のオブジェクトカテゴリを持ち、これらの画像内のオブジェクトは高いPSM分散を示す。具体的には、AISスキームは、視線パターンと人物の視覚的嗜好の違いを分析するためにPSM分散に焦点を当て、オブジェクトカテゴリの数が多きほど、シーンの多様性が高くなる。画像に対して、AIS方式は、学習画像に対する画像選択が多様性を示す場合に、その強みを発揮する。また、AIS方式は物体レベルの視線解析に基づいており、視覚情報の多様性の維持に寄与している。一方、PSMの場合、同じような視線パターンを持つ人がいなくても、複数の人の視線パターンを組み合わせることで、対象者の視線パターンを表現することができ、この可能性は訓練人数が多いほど高くなる。また、一人のアイトラッキングデータにノイズや外れ値が含まれている場合、複数の人物のPSM分散に注目することで、AIS方式での影響を低減することができる。

3.3 テンソルから行列への回帰によるPSM予測

この小節では、数ショットPSM予測のためのテンソルから行列への回帰モデルを示す。テンソル入力回帰とベクトル入力回帰の比較を図2に示す。テンソル入力回帰はベクトル入力回帰よりも、構造情報を保持できる点で優れており、これは本論文の2次元空間構成がPSM内およびPSM間で優れていることを意味する。

提案手法では、個体間の視線パターンの類似性を明示的に定義したり計算したりすることはない。代わりに、類似性はテンソルから行列への回帰モデルの学習プロセスを通して暗黙的に捕捉される。具体的には、訓練対象者のPSMを入力として、対象者のPSMを予測するようにモデルを学習する。学習中、より類似した視線傾向を持つ個体が予測により多く寄与するように重みが調整される。

したがって、類似性は、個人の類似性や事前のクラスタリングを測定することなく、回帰プロセスの一部として学習される。

第3章1で予測したPSMを用いて、対象人物のPSMを予測する。いくつかのPSMが入力として扱われる。第3.2節で選んだ画像 X_i に対応する入力テンソル $S(X_i) \in \mathbb{R}^P \times d^1 \times d^2$ ($i = 1, 2, \dots, I$)は以下のように構成される：

$$S(X_i) = [\hat{S}(X_i)_1, \hat{S}(X_i)_2, \dots, \hat{S}(X_i)_P]. \quad (7)$$

さらに、入力テンソル $S(X_i)$ に対して、対象人物 p^{tst} の教師付きPSM $S(X_i)_p^{tst}$ を用意する。ここでは、対象人物が第3.2節で選択した共通画像のみを注視すると仮定すると、教師ありPSM $S(X_i)_p^{tst}$ が得られる。テンソル行列回帰シナリオでは、重みテンソル $W \in \mathbb{R}^P \times d^1 \times d^2 \times d^1 \times d^2$ を用いて、新たに与えられた画像のPSMを以下のように予測する：

$$S_{TReg}(X_{tst})_{p^{tst}} = \langle S(X_{tst}), W \rangle_3, \quad (8)$$

ここで、 $\cdot$, $\cdot_Q$ はテンソル積、 Q は入力配列の数を表す。

重みテンソル W を最適化するために、 L_2 正則化を用いて二乗誤差の和を最小化する：

$$\min_{\text{rank}(W) \leq R} \sum_{i=1}^I \|S(X_i)_{p^{tst}} - \langle S(X_i), W \rangle_3\|_F^2 + \lambda \|W\|_F^2. \quad (9)$$

入力と出力がマルチ配列形式であるため、この最小化問題を解くのは難しいことに注意してください。したがって、先行研究²⁷⁾を参照して、 W は $\text{rank}(W) \leq R$ となるような縮小PARAFAC/CANDECOMP(CP)-rankを持つと仮定し、この制約の下で式(9)を解く。テンソルベースの回帰は高次元の問題やオーバーフィッティングに悩まされる傾向があるが、低ランク近似はそのような問題を軽減する。また、 L_2 正則化によりオーバーフィッティングを抑制している。削減されたCP-rank制約は次元削減戦略として機能し、汎化と予測効率を向上させるために多方向構造を利用する。この定式化では、リッジ(L_2)正則化はさらにベイズ事前分布として解釈され、予測手順²⁷⁾の統計的動機付けとなる。したがって、テンソルから行列への回帰モデルを用いることで、入力テンソルと出力行列をベクトル化することなく、構造情報を保持することができる。

product. Then, we set $q_{n,j} = 0$ when $\mathbf{X}_n$ does not include the j th object and set the largest $q_{n,j}$ when the image $\mathbf{X}_n$ includes several m th objects. Then, we obtain the sum of $q_{n,j}$ for n th image as follows:

$$\bar{q}_n = \sum_{j=1}^J q_{n,j}. \quad (6)$$

Finally, using $\bar{q}_n$, we select the top I images as common images under the constraint to maximize the number of object categories in these images. Consequently, the selected common images have multiple object categories, and the objects in these images exhibit a high PSM variance. Specifically, the AIS scheme focuses on the PSM variance to analyze the differences in gaze patterns and persons' visual preferences, and a higher number of object categories leads to greater scene diversity. For images, the AIS scheme has demonstrated its strength when image selection for training images exhibits high diversity. In addition, the AIS scheme is based on object-level gaze analysis, and it contributes to maintaining the diversity of visual information. In contrast, for PSMs, the target person's gaze pattern can be represented by combining the gaze patterns of multiple persons even if no person has a similar gaze pattern, and this possibility increases with the number of training persons. In addition, if the eye-tracking data of a single person contain noise or outliers, the effect can be reduced in the AIS scheme by focusing on the PSM variance of several persons.

3.3 PSM Prediction via Tensor-to-Matrix Regression

This subsection presents the tensor-to-matrix regression model for the few-shot PSM prediction. The comparison of the tensor-input regression with the vector-input regression is shown in Fig. 2. The tensor-input regression is superior to the vector-input regression in the points that it can preserve the structural information, which means the two-dimensional spatial configuration in this paper, within and between PSMs.

In the proposed method, we do not explicitly define or compute gaze pattern similarity between individuals. Instead, the similarity is implicitly captured through the training process of the tensor-to-matrix regression model. Specifically, the model is trained to predict the target person's PSMs using the PSMs of training persons as input. During training, the weights are adjusted such that individuals with more similar gaze tendencies contribute more to the prediction. Therefore, similarity is learned as part of the regression process, without

measuring similarities or prior clustering of individuals.

The PSMs predicted in Sec. 3.1 are used to predict the PSM of the target person. Several PSMs are treated as input. The input tensor $\mathcal{S}(\mathbf{X}_i) \in \mathbb{R}^{P \times d_1 \times d_2}$ ($i = 1, 2, \dots, I$) corresponding to the image $\mathbf{X}_i$ chosen in Sec. 3.2 is constructed as follows:

$$\mathcal{S}(\mathbf{X}_i) = [\hat{\mathbf{S}}(\mathbf{X}_i)_1, \hat{\mathbf{S}}(\mathbf{X}_i)_2, \dots, \hat{\mathbf{S}}(\mathbf{X}_i)_P]. \quad (7)$$

In addition, we prepare the supervised PSM $\mathcal{S}(\mathbf{X}_i)_{p^{\text{tst}}}$ of the target person p^{tst} for the input tensor $\mathcal{S}(\mathbf{X}_i)$. Here, we assume that the target person gazes only at the common images selected in Sec. 3.2, and we can obtain the supervised PSM $\mathcal{S}(\mathbf{X}_i)_{p^{\text{tst}}}$. In a tensor-to-matrix regression scenario, the weight tensor $\mathcal{W} \in \mathbb{R}^{P \times d_1 \times d_2 \times d_1 \times d_2}$ is used to predict the PSM of a newly given image as follows:

$$\mathcal{S}_{\text{TReg}}(\mathbf{X}_{\text{tst}})_{p^{\text{tst}}} = \langle \mathcal{S}(\mathbf{X}_{\text{tst}}), \mathcal{W} \rangle_3, \quad (8)$$

where $\langle \cdot, \cdot \rangle_Q$ denotes the tensor product and Q denotes the number of input arrays.

To optimize the weight tensor $\mathcal{W}$, we minimize the sum of the squared errors using L_2 regularization as follows:

$$\min_{\text{rank}(\mathcal{W}) \leq R} \sum_{i=1}^I \|\mathcal{S}(\mathbf{X}_i)_{p^{\text{tst}}} - \langle \mathcal{S}(\mathbf{X}_i), \mathcal{W} \rangle_3\|_F^2 + \lambda \|\mathcal{W}\|_F^2. \quad (9)$$

Note that it is difficult to solve this minimization problem because the inputs and outputs are in a multi-array format. Thus, by referring to a previous study²⁷⁾, we assume that $\mathcal{W}$ has the reduced PARAFAC/CANDECOMP (CP)-rank such that $\text{rank}(\mathcal{W}) \leq R$ and solve Eq. (9) under this constraint. Although tensor-based regression tends to suffer from high-dimensional problems and overfitting, the low-rank approximation mitigates such problems. In addition, L_2 regularization suppresses overfitting. The reduced CP-rank constraint serves as a dimensionality reduction strategy, which utilizes a multi-way structure to improve generalization and prediction efficiency. In this formulation, the ridge (L2) regularization is further interpreted as a Bayesian prior, which provides statistical motivation for the prediction procedure²⁷⁾. Therefore, using the tensor-to-matrix regression model, the proposed method preserves structural information without vectorizing the input tensor and output matrix.

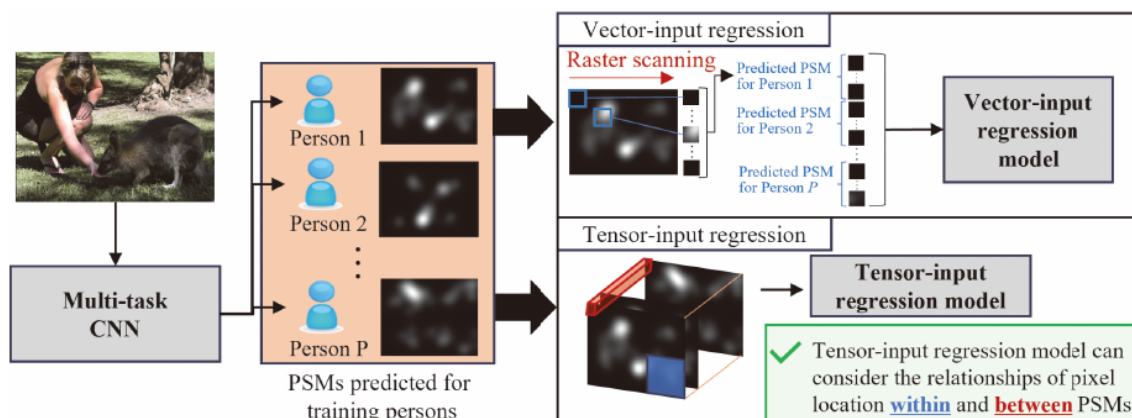


Fig. 2 提案手法におけるテンソル入力回帰とベクトル入力回帰の比較。テンソル入力回帰はベクトル入力回帰よりも構造情報を保持できる点で優れており、これは本論文の2次元空間構成がPSM内およびPSM間で優れていることを意味する。

4. 実験結果

4.1 データセット

本実験では、オープンソースの大規模PSMデータセット¹⁶⁾を用いた。PSMデータセットは、収録された画像が多様性を持ち、性能とロバスト性の評価に適するように構築された。PSMデータセットは、30人の参加者から得られた1,600枚の画像とそれに対応するアイトラッキングデータから構成される。参加者は視力が正常または矯正されており、自由視力条件下で1つの画像を3秒間注視した。予測されたPSMを評価するために、先行研究³³⁾に基づき、視線追跡データから全ての画像に対する各参加者のPSMをグランドトゥールズ(GT)マップとして構築した。提案手法で用いた USM と同様に、USM 予測誤差の影響を軽減するために、訓練者の平均 PSM を採用した。

提案手法では、マルチタスクCNNモデルを学習するために、視線追跡データを用いた学習画像を用い、テンソル対行列回帰モデルを学習するために、学習画像から選択する共通画像を用いた学習が必要であった。このように、1,100枚の画像をランダムに選択して学習させ、残りの500枚を実験のテスト画像として使用した。また、AIS方式に基づき、学習画像から共通画像を選択した。なお、AISで選択された一般的な画像は、予測モデルの学習にのみ使用される。評価は、画像選択メカニズムによって予測結果が偏らないようにするため、AISの影響を受けない別のテストセットで実行される。

また、無作為に20名をトレーニング対象者として抽出し、残りの10名をターゲット対象者として扱った。訓練データとテストデータの実用的なバランスをとるため、訓練被験者数と共通画像数を経験的に決定した。視線パターンの十分なばらつきを確保するため、20人のトレーニング対象者を無作為に選択した。1共通画像は、AISスキームにより学習画像から選択された。共通画像の数が少なすぎると、限られた視線パターンの集合に依存するため、予測が不安定になる可能性がある。一方、共通画像の数が多すぎると、対象者は膨大な数の画像を見るはずであり、予測は安定するかもしれないが、非現実的である。そこで、共通画像の数を実験的に決定し、その値を第4章で説明する。2. 対象者の視線追跡データは入手可能であったが、対象人物が共通画像を注視していると仮定しているため、対象人物の視線追跡データのみを共通画像としてPSM予測に用いた。

4.2 実験セットアップ

第3節でマルチタスクCNNモデルを最適化した。1. 第3節でテンソル対行列回帰モデルを最適化した。3. マルチタスクCNNモデルは、先行研究¹⁶⁾を参考に確率的勾配降下法³⁴⁾で最適化し、層数(L)、運動量、バッチサイズ、エポック、学習率をそれぞれ3、0.9、9、1,000、 3.0×10^{-5} に設定した。また、テンソルアンフォールディングで重みパラメータを微分するだけで、テンソルから行列への回帰モデルを最適化した。

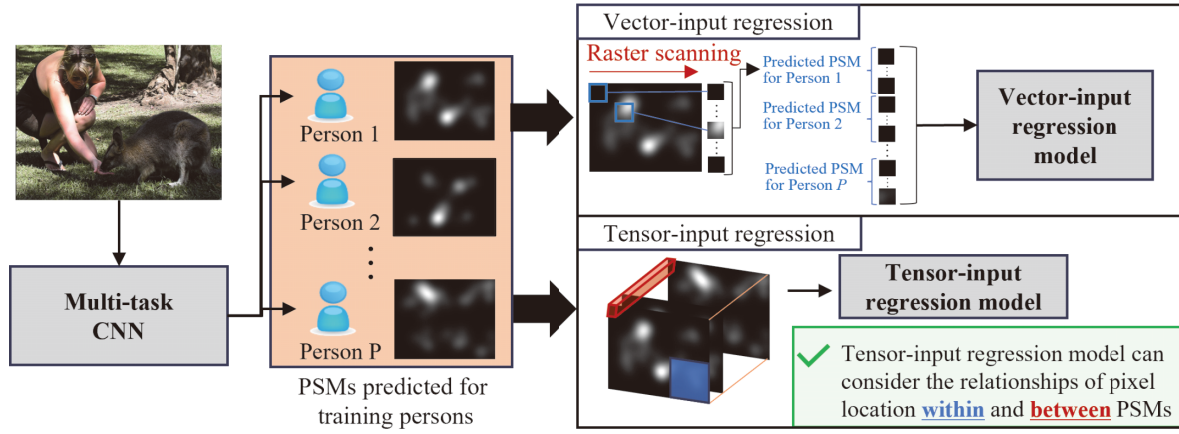


Fig. 2 Comparison of the tensor-input regression with the vector-input regression in the proposed approach. The tensor-input regression is superior to the vector-input regression in the points that it can preserve the structural information, which means the two-dimensional spatial configuration in this paper, within and between PSMs.

4. Experiments

4.1 Dataset

In this experiment, an open-source large-scale PSM dataset¹⁶⁾ was used. The PSM dataset was constructed such that the included images have high diversity and is suitable for performance and robustness evaluation. The PSM dataset comprises 1,600 images with corresponding eye-tracking data obtained from 30 participants. The participants had normal or corrected visual acuity and gazed at one image for three seconds under free viewing conditions. To evaluate the predicted PSMs, we constructed the PSMs of each participant for all images from the eye-tracking data as the ground truth (GT) map based on a previous work³³⁾. As the USM used in the proposed method, we adopted the mean PSMs of the training persons to reduce the influence of USM prediction errors.

In the proposed method, we required training images with eye-tracking data to train the multi-task CNN model and common images selected from the training images to train the tensor-to-matrix regression model. Thus, 1,100 images were randomly selected for training and the remaining 500 images were used as test images in the experiment. In addition, I common images were selected from the training images based on the AIS scheme. Note that the common images selected by AIS are used exclusively for training the prediction model. The evaluation is performed on a separate test set that is not influenced by the AIS for ensuring that the prediction results are not biased by the image selection mechanism. In addition, we randomly selected 20 participants

as training persons, and the remaining 10 persons were treated as target persons. The number of training subjects and common images were determined empirically to achieve a practical balance between training and test data. The 20 training persons were randomly selected to ensure sufficient variability in gaze patterns. The I common images were selected from training images via the AIS scheme. If the number of common images is too small, the prediction may become unstable due to depending on a limited set of gaze patterns. While, if the number of common images is too large, the target person should gaze at the massive number of images, which is unrealistic, although the prediction may become stable. In this way, we experimentally determined the number of common images, and its value is described in Sec. 4.2. Although the eye-tracking data of the target persons were available, we only used the eye-tracking data of the target persons as the common images for PSM prediction because we assumed that the target persons gazed at the common images.

4.2 Experimental Settings

We optimized the multi-task CNN model in Sec. 3.1 and the tensor-to-matrix regression model in Sec. 3.3. The multi-task CNN model was optimized via the stochastic gradient descent³⁴⁾ by referring to a previous study¹⁶⁾, and the number of layers (L), momentum, batch size, epoch, and learning rate were set to 3, 0.9, 9, 1,000, and 3.0×10^{-5} , respectively. In addition, the tensor-to-matrix regression model was optimized by simply differentiating the weight parameters with tensor unfolding. Furthermore, we set $I = 100$ and conducted additional experiments focusing on hy-

さらに、 $I = 100$ とし、テンソル対行列回帰モデルのハイパーパラメータを $R \in \{5, 10, \dots, 50\}$ と $\lambda \in \{0.01, 0.1, \dots, 10000\}$ で変化させ、性能への影響を調べた。

提案手法を客観的に評価するために、いくつかのUSM予測手法とPSM予測手法を比較手法として採用した。以下のUSM予測法を採用した: Signature⁵⁾, GBVS⁴⁾, Itti¹⁾, SalGAN⁶⁾, Contextual⁷⁾. Signature, GBVS, Ittiは、入力画像のみからUSMを予測する計算モデルである。SalGANとContextualは、大規模なアイトラッキングデータセットであるSALICONデータセット³⁵⁾で学習した深層学習ベースのモデルであり、パーソナライゼーションは考慮されていない。共通画像とその視線追跡データのみを用いた以下の2つの少数ショットPSM予測(FPSP)手法を採用した。

ベースライン1: ターゲット画像と共通画像間の視覚的類似性を用いたPSM予測³⁶⁾。

ベースライン2: Baseline1とUSM予測³⁷⁾に基づくPSM予測。

さらに、提案手法と同様の設定で、3つのPSM予測手法を比較した: 類似度ベースFPSP²⁵⁾、CoMOGPベースFPSP¹³⁾、物体ベース注視類似度(OGS)ベースFPSP³¹⁾。なお、他のPSM予測手法も存在するが¹¹⁾¹⁶⁾²⁹⁾³⁰⁾、第2節で述べたように少量の学習データから学習することはできない。そこで、本実験では前述の比較手法のみを採用した。

評価指標として、先行研究³⁸⁾に基づき、予測PSMとGTマップのKullback-Leibler divergence (KLdiv)と相互相関(CC)を採用した。具体的には、KLdivを用いて分布の類似度、すなわち構造的類似度を評価し、CCを用いて画素ベースの類似度を評価した。これらの2つの指標を用いることで、予測されたPSMとそのGTマップの間の大域的・局所的な類似性を評価することができる。

4.3 結果および考察

図3は予測された結果を示している。表2は定量的な評価結果である。図3に示すように、提案手法によって予測されたPSMはGTに近い分布を示し、構造情報の保存の有効性を示している。また、提案手法と比較手法を比較した表2に示すように、提案手法はKLdivの点で全ての比較手法を凌駕している。この結果から、構造情報を考慮したPSM予測にはテンソル-行列回帰が有効であることが確認された。

Table 2 KLdivに基づく定量的分布に基づく評価とCCに基づくピクセルに基づく評価。KLdivの値が低いほど性能が高いことを示し、CCの値が高いほど性能が高いことを示す。

Methods	KLdiv↓	CC↑
Signature ⁵⁾	8.04	0.413
GBVS ⁴⁾	6.89	0.437
Itti ¹⁾	9.04	0.322
SalGAN ⁶⁾	3.56	0.635
Contextual ⁷⁾	3.57	0.674
Baseline1 ³⁶⁾	7.64	0.401
Baseline2 ³⁷⁾	4.13	0.597
類似度ベースのFPSP ²⁵⁾ 。	1.82	0.735
CoMOGPベースのFPSP	1.38	0.765
OGSベースのFPSP	1.09	0.781
Proposed Method ($R = 50, \lambda = 1000$)	1.00	0.775

提案手法がUSM予測手法を上回ったことから、顕著性予測におけるパーソナライゼーションの有効性が確認された。また、提案手法を他のPSM予測手法と比較することで、構造情報に着目することの有効性が確認された。さらに、Baseline1とBaseline2は他者のPSMを使用しておらず、これらの手法と提案手法の比較から、他者から得られた視線追跡データを使用することの有効性が実証された。

OGSベースFPSP³¹⁾の "KLdiv" は提案手法と同様である。ここで、OGSベースのFPSPは、学習画像から入力画像中の物体を検索し、検索された物体の領域に対応するPSMを用いてPSMを予測する。入力画像に学習データセットに存在しないオブジェクトが含まれる場合、OGSベースのFPSPは既知のオブジェクトのPSMに依存するため、PSMの予測が困難になる可能性がある。その結果、入力画像に学習中に見られなかったオブジェクトが含まれる場合、その適用性は制限される可能性がある。一方、提案手法は、意味情報、例えば物体を含まない人物の視線パターン間の関係に着目している。このように、我々のアプローチは、入力画像に学習データセットに存在しないオブジェクトが含まれている場合でも、適用可能である。この特徴により、提案手法はより広範なシナリオで 사용할ことができる。

提案手法の "CC" は、OGSベースのFPSPのような他のPSM手法と同等であるが、最良ではない。

perparameter analysis by varying the hyperparameters of the tensor-to-matrix regression model over $R \in \{5, 10, \dots, 50\}$ and $\lambda \in \{0.01, 0.1, \dots, 10000\}$ to examine their effect on performance.

To objectively evaluate the proposed method, we adopted several USM and PSM prediction methods as comparison methods. We adopted the following USM prediction methods: Signature⁵⁾, GBVS⁴⁾, Itti¹⁾, SalGAN⁶⁾, and Contextual⁷⁾. Signature, GBVS, and Itti are computational models that predict USMs only from input images. SalGAN and Contextual are deep learning-based models trained on the SALICON dataset³⁵⁾, which is a large-scale eye-tracking dataset, without considering personalization. The following two few-shot PSM prediction (FPSP) methods using only common images and their eye-tracking data were adopted.

Baseline1: PSM prediction using visual similarity between target and common images³⁶⁾.

Baseline2: PSM prediction based on Baseline1 and USM prediction³⁷⁾.

In addition, we compared three PSM prediction methods with settings similar to those of the proposed method: similarity-based FPSP²⁵⁾, CoMOGP-based FPSP¹³⁾, and object-based gaze similarity (OGS)-based FPSP³¹⁾. Note that although other PSM prediction methods exist¹¹⁾¹⁶⁾²⁹⁾³⁰⁾, they cannot learn from a small amount of training data as discussed in Sec. 2.2. Therefore, we adopted only the aforementioned comparison methods in our experiment.

As the evaluation metrics, we adopted the Kullback–Leibler divergence (KLdiv) and cross-correlation (CC) between the predicted PSM and the GT map based on previous research³⁸⁾. Specifically, KLdiv was used to evaluate the similarity of the distribution, that is, structural similarity, and CC was used to evaluate pixel-based similarity. By using these two metrics, both the global and local similarities between the predicted PSMs and their GT maps can be evaluated.

4.3 Results and Discussion

Figure 3 presents the predicted results. Table 2 presents the quantitative evaluation results. As shown in Fig. 3, the PSMs predicted by the proposed method exhibit a distribution close to the GTs, demonstrating the effectiveness of preserving structural information. In addition, as shown in Table 2 that compares the proposed and comparison methods, the proposed method outperforms all comparison methods in terms of KLdiv. This result confirms that tensor-to-matrix regres-

Table 2 Quantitative distribution-based evaluation based on KLdiv and pixel-based evaluation based on CC. A lower KLdiv value indicates higher performance, whereas a higher CC value indicates higher performance.

Methods	KLdiv↓	CC↑
Signature ⁵⁾	8.04	0.413
GBVS ⁴⁾	6.89	0.437
Itti ¹⁾	9.04	0.322
SalGAN ⁶⁾	3.56	0.635
Contextual ⁷⁾	3.57	0.674
Baseline1 ³⁶⁾	7.64	0.401
Baseline2 ³⁷⁾	4.13	0.597
Similarity-based FPSP ²⁵⁾	1.82	0.735
CoMOGP-based FPSP ¹³⁾	1.38	0.765
OGS-based FPSP ³¹⁾	1.09	0.781
Proposed Method ($R = 50, \lambda = 1000$)	1.00	0.775

sion is effective for PSM prediction considering structural information. The effectiveness of personalization for saliency prediction was confirmed because the proposed method outperformed the USM prediction methods. In addition, by comparing the proposed method with other PSM prediction methods, the effectiveness of focusing on structural information is confirmed. Moreover, Baselines1 and 2 do not use the PSMs of other persons, and the comparison between these methods and the proposed method demonstrates the effectiveness of using eye-tracking data obtained from other persons.

The “KLdiv” of OGS-based FPSP³¹⁾ is similar to that of the proposed method. Here, OGS-based FPSP predicts PSMs by searching for objects in input images from the training images and using the PSMs corresponding to the regions of searched objects. In cases where an input image contains objects that were not present in the training dataset, OGS-based FPSP may be difficult to predict PSMs because it relies on PSMs of known objects. Consequently, its applicability may be limited when input images contain objects that were not seen during training. In contrast, the proposed method focuses on the relationships between persons’ gaze patterns without semantic information, e.g., objects. Thus, our approach remains applicable even when the input images include objects that were not present in the training dataset. This characteristic allows the proposed method to be used in a broader range of scenarios.

The “CC” of the proposed method is comparable to that of other PSM methods such as the OGS-based

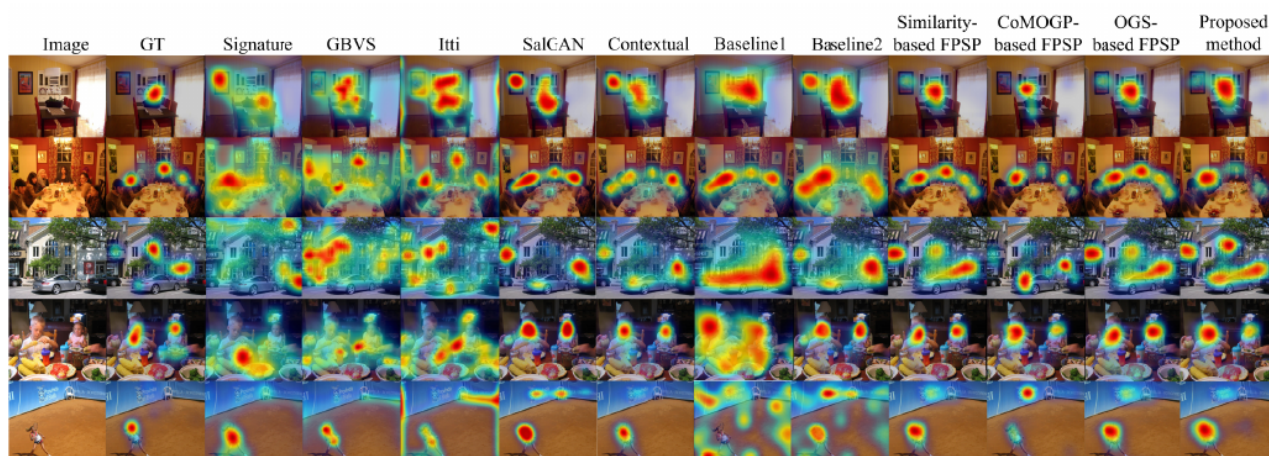


図3 予測されたPSMの例

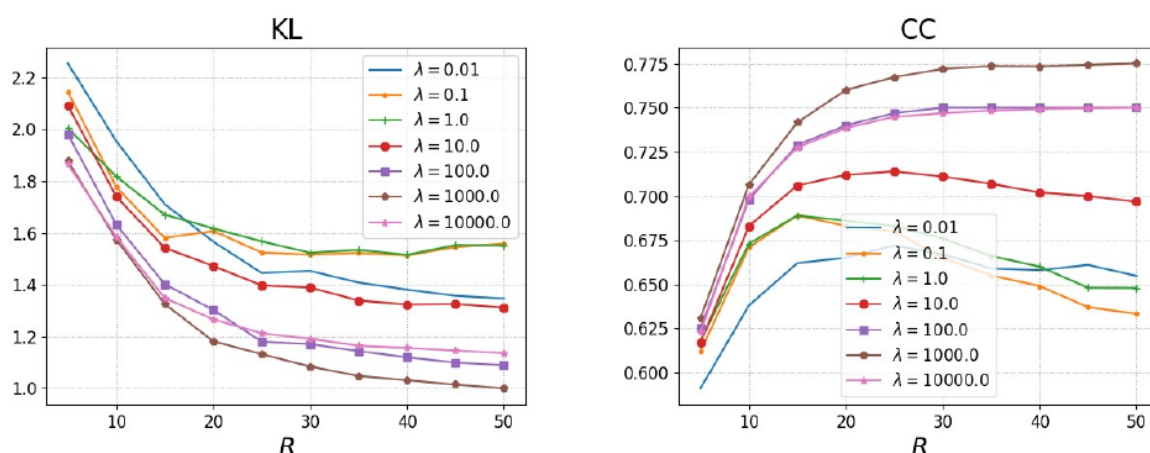


Fig. 4 ハイパーパラメータの変化に対する評価指標の値の変化
 テンソルから行列への回帰の

その理由として考えられるのは、第3節の重みテンソルの低ランクテンソル近似において、モデルの表現力がランクRによって制限されることである。一般に、Rを大きくすることで、きめ細かい局所情報を含む複雑なパターンを捉えるモデルの能力が向上するが、計算コストも大幅に増加する。テンソルから行列への回帰モデルにおけるハイパーパラメータの議論を以下に述べる。提案手法の「CC」がOGSベースのFPSPの「CC」よりも低いもう一つの理由は、提案手法が物体カテゴリのような明示的な意味情報を組み込んでいないためである。一方、OGSベースのFPSPは、オブジェクトレベルのマッチングを利用して顕著性情報を伝達するため、CCスコアが高くなる。さらに、我々の実験で使用された画像は、多種多様な汎用画像を含んでおり、テスト画像内の多くのオブジェクトもトレーニング画像に現れる。これにより、OGSベースのFPSPは、そのオブジェクトベースのメカニズムを効果的に利用することができる。

しかし、上述したように、OGSベースのFPSPは、テスト画像に未見のオブジェクトが含まれる場合、適用範囲が限定される。このようなシナリオでは、オブジェクト依存のメカニズムは適用できない。一方、提案手法は、物体固有の情報ではなく、対人視線類似性に依存しており、特に未見の物体が現れる実世界の設定において、提案手法の適用性を支持するものである。

ここで、「CC」はピクセルベースの評価、「KLdiv」は分布ベースの評価である。このように、提案手法は「KLdiv」が高いため、構造情報を保持することができる。ここで、第1節で述べたように、構造情報は人間の視覚系における顕著な領域を検出するための重要な手がかりとなる¹⁾。この観点から、提案手法は「CC」の点でOGSbased FPSPに大きく劣らず、「KLdiv」の点で優れているため、有効であることを確認した。

【機械翻訳コンテンツの著作権について】

当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
 引用の際には、必ず原著論文の書誌情報をご記載ください。詳細は映像情報メディア学会著作権規定をご覧ください。

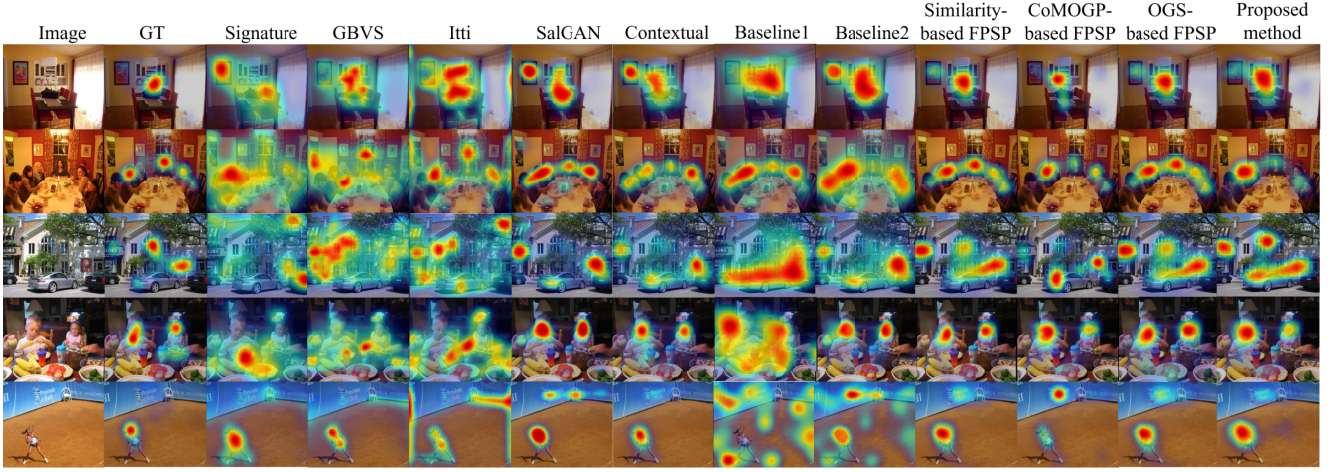


Fig. 3 Examples of predicted PSMS.

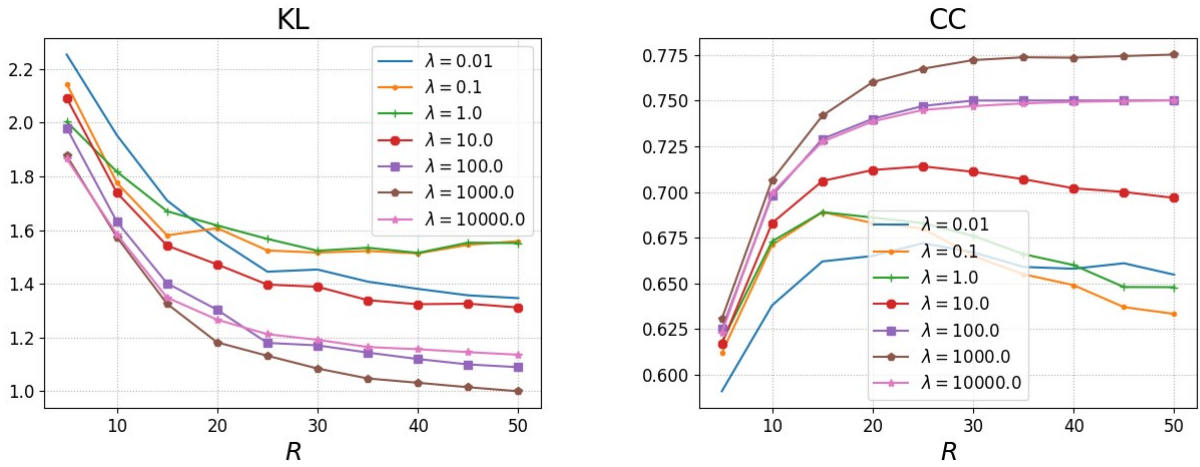


Fig. 4 Changes in the values of the evaluation metrics in response to changes in hyperparameters of tensor-to-matrix regression.

FPSP; however, it is not the best. One possible reason is that the expressive capacity of the model is limited by the rank R in the low-rank tensor approximation of the weight tensor in Sec. 3.3. In general, increasing R enhances the model’s ability to capture complex patterns, including fine-grained local information; however, it also substantially increases computational cost. The discussion of hyperparameters in the tensor-to-matrix regression model is described below. Another reason why the “CC” of the proposed method is lower than that of OGS-based FPSP is that the proposed method does not incorporate explicit semantic information such as object categories. In contrast, the OGS-based FPSP utilizes object-level matching to transfer saliency information, which results in higher CC scores. Additionally, the images used in our experiments contain a wide variety of general-purpose images, and many objects in the test images also appear in the training images. This enables the OGS-based FPSP to utilize its object-based

mechanism effectively. However, as discussed above, the OGS-based FPSP has limited applicability when the test images include unseen objects. In such scenarios, its object-dependent mechanism cannot be applied. In contrast, the proposed method relies on interpersonal gaze similarity rather than object-specific information, which supports the applicability of the proposed method, particularly in real-world settings where unseen objects appear.

Here, “CC” is a pixel-based evaluation, whereas “KLdiv” is a distribution-based evaluation. Thus, the proposed method can preserve structural information because of its high “KLdiv.” Here, as mentioned in Sec. 1, structural information is an important clue to detecting salient areas in the human visual system¹⁾. From this viewpoint, we confirmed that the proposed method is valid because it is not significantly inferior to OGS-based FPSP in terms of “CC” and superior in terms of “KLdiv.” Therefore, we emphasize the effectiveness of

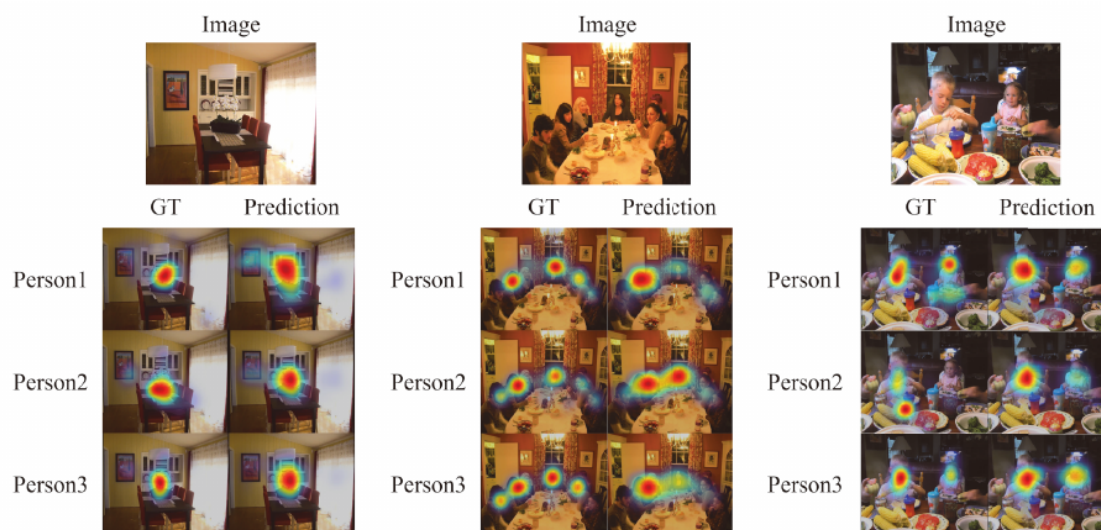


図5 提案手法を用いて複数人について予測されたPSMの例。

したがって、構造情報を保存することで、PSM予測における提案手法の有効性を強調する。

ハイパーパラメータ分析に焦点を当てた追加実験により、テンソルから行列への回帰モデルのハイパーパラメータの変化に対する評価スコアを確認した。図4は、 R と λ に対する評価スコアである。図に示すように、 R が増加するにつれて性能は向上するが、 $\lambda=1000$ は R に関係なく最高の性能を達成する。これらの結果は実験的に得られたが、傾向は安定していた。この値を超えると、 λ をさらに増加させると性能が低下した。そこで、性能の局所最大値に対応する設定として、 $\lambda=1000$ を採用した。また、 R が増加するにつれて性能は向上するが、 $R=50$ 付近で性能の向上が飽和することが示された。この点を超えると、 R をさらに増加させると、予測性能に意味のある向上は見られず、主に計算時間が長くなる。そこで、モデルの性能と効率の実用的な妥協点として、 $R=50$ を採用した。また、 λ は正則化のハイパーパラメータであるため、極めて高い λ 値は必要ない。そこで、提案手法におけるテンソル-行列回帰の望ましいハイパーパラメータを確認する。本研究では、探索を R に限定しているが、計算コストを削減するために、より効率的なモデル設計による更なる研究は、今後の研究の有望な方向性である。また、テンソル-行列回帰モデルの重みテンソルには、先行研究²⁷⁾を参考にCP-rank分解を採用した。他のランク分解法の検証も今後の課題である。

個人間のPSMの違いを示すために、提案手法を用いて複数人について予測したPSMの例を図5に示す。これらの例のほとんどは、提案手法によって正しく予測された。図5で予測されたPSMの中には、実際の視線位置と完全に一致しないものもあった。例えば、右の列のPerson2のように、特定の個人については、主な焦点が少しずれるかもしれない。しかし、提案手法は、全体的な有人領域を妥当な性能で捉えている。人間の視線行動は、視覚刺激だけでなく、短期的な意図や長期的な精神状態といった一過性の要因によっても変化する可能性があることに注意することが重要である。したがって、個人の注意を完全に予測することは本質的に困難である。このようなばらつきがあるにもかかわらず、提案手法は各人の注意傾向を近似することができ、パーソナライズされた顕著性のモデリングに有効であることが実証された。

5. 結論

本研究では、対人視線パターンに基づく数ショットPSM予測法を提案した。提案手法は、適応的な画像選択スキームとテンソルから行列への回帰を組み込み、それぞれ画像の効果的な画像選択と構造情報の保存を実現する。ベクトル化せずに入出力PSMを扱うことで、提案手法は構造情報を保持する。オープンデータセットでの実験により、これらの要素を取り入れることの有効性が実証された。

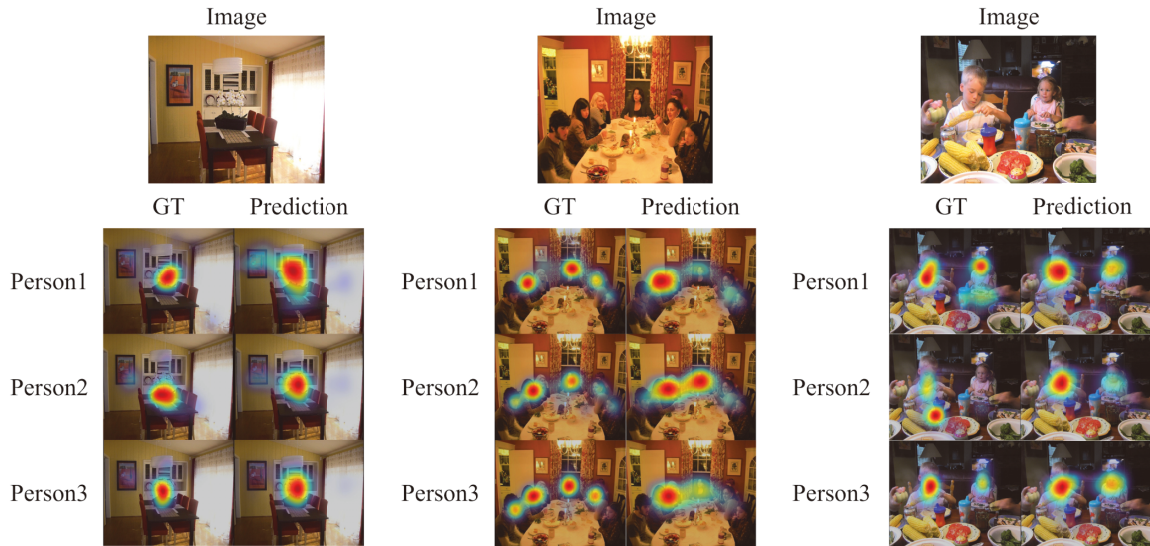


Fig. 5 Examples of PSMs predicted for several persons using the proposed method.

the proposed method for PSM prediction by preserving the structural information.

We confirmed the evaluation scores in response to changes in the hyperparameters of the tensor-to-matrix regression model through additional experiments focusing on hyperparameter analysis. Figure 4 shows the evaluation scores in response to R and λ . As shown in the figure, as R increases, the performance improves, whereas $\lambda = 1000$ achieves the best performance regardless of R . Although these results were obtained experimentally, the trend was stable. Beyond this value, further increases in λ led to performance degradation. Therefore, we adopted $\lambda = 1000$ as a setting corresponding to a local maximum in performance. The results also indicate that the performance improves as R increases, but the improvement saturates around $R = 50$. Beyond this point, further increases in R lead primarily to longer computation times without meaningful gains in prediction performance. Therefore, we adopted $R = 50$ as a practical compromise between model performance and efficiency. In addition, an extremely high λ value is not required because λ is a regularization hyperparameter. Therefore, we confirm the desirable hyperparameters of tensor-to-matrix regression in the proposed method. While this study limits the exploration to R , further investigation with more efficient model designs to reduce computational cost is a promising direction for future work. In addition, we adopted the CP-rank decomposition for the weight tensor in the tensor-to-matrix regression model by referring to a previous study²⁷⁾. The validation of other rank decomposition methods is also a future work.

To show the difference in PSMs across individuals, we present the examples of PSMs predicted for several persons using the proposed method in Fig. 5. Most of these examples were correctly predicted by the proposed method. While some predicted PSMs in Fig. 5 did not perfectly align with the actual gaze locations. For example, the main focus might be slightly shifted for certain individuals, such as Person2 in the right column. However, the proposed method still captured the overall attended regions with reasonable performance. It is important to note that human gaze behavior can vary not only due to the visual stimulus but also based on transient factors such as short-term intentions or long-term mental states. Therefore, perfectly predicting individual attention is inherently difficult. Despite this variability, the proposed method was able to approximate the attention tendencies of each person, which demonstrates its effectiveness in modeling personalized saliency.

5. Conclusions

This study has proposed a few-shot PSM prediction method based on interpersonal gaze patterns. The proposed method incorporates the adaptive image selection scheme and tensor-to-matrix regression for effective image selection of images and the preservation of structural information, respectively. By treating the input and output PSMs without vectorization, the proposed method preserves structural information. The experiments on the open dataset demonstrate the effectiveness of incorporating these factors.

Acknowledgement

This work was partly supported by the JSPS KAKENHI Grant Numbers JP24K02942, JP23K21676, and JP23K11211.

References

- 1) Laurent Itti, Christof Koch, and Ernst Niebur : “A Model of Saliency-based Visual Attention for Rapid Scene Analysis,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259 (1998)
- 2) Inam Ullah, Muwei Jian, Sumaira Hussain, Jie Guo, Hui Yu, Xing Wang, and Yilong Yin, “A Brief Survey of Visual Saliency Detection,” Multimedia Tools and Applications, vol. 79, pp. 34605–34645 (2020)
- 3) Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li, “Salient Object Detection: A Survey,” Computational Visual Media, vol. 5, pp. 117–150 (2019)
- 4) Jonathan Harel, Christof Koch, and Pietro Perona, “Graph-based Visual Saliency,” Advances in Neural Information Processing Systems, pp. 545–552 (2007)
- 5) Xiaodi Hou, Jonathan Harel, and Christof Koch, “Image Signature: Highlighting Sparse Salient Regions,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 34, no. 1, pp. 194–201 (2012)
- 6) Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel O’Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i-Nieto, “Salgan: Visual Saliency Prediction with Generative Adversarial Networks,” arXiv preprint arXiv:1701.01081 (2017)
- 7) Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel, “Contextual Encoder–decoder Network for Visual Saliency Prediction,” Neural Networks, vol. 129, pp. 261–270 (2020)
- 8) Aoji Li and Zhenzhong Chen, “Personalized Visual Saliency: Individuality Affects Image Perception,” IEEE Access, vol. 6, pp. 16099–16109 (2018)
- 9) Evan F Risko, Nicola C Anderson, Sophie Lanthier, and Alan Kingstone, “Curious Eyes: Individual Differences in Personality Predict Eye Movement Behavior in Scene-viewing,” Cognition, vol. 122, no. 1, pp. 86–90 (2012)
- 10) Olivier Le Meur, Antoine Collot, Zhi Liu, Rămä, Adrien Le Roch, and Andrea Helo, “Visual Attention Saccadic Models Learn to Emulate Gaze Patterns from Childhood to Adulthood,” IEEE Trans. Image Processing, vol. 26, no. 10, pp. 4777–4789 (2017)
- 11) Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao, “Beyond Universal Saliency: Personalized Saliency Prediction with Multi-task CNN,” in Proc. Int’l Joint Conf. Artificial Intelligence, pp. 3887–3893 (2017)
- 12) Sikun Lin and Pan Hui, “Where’s Your Focus: Personalized Attention,” arXiv preprint arXiv:1802.07931 (2018)
- 13) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “Few-shot Personalized Saliency Prediction Using Person Similarity Based on Collaborative Multi-output Gaussian Process Regression,” in Proc. IEEE Int’l Conf. Image Processing, pp. 1469–1473 (2021)
- 14) Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, and Junfeng He, “Learning from Unique Perspectives: User-aware Saliency Modeling,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, pp. 2701–2710 (2023)
- 15) Qiong Wang, Meriem Outtas, Julie Fournier, Elise Etchamendy, Myriam Chérel, and Lu Zhang, “Predicting Personalized Saliency Map for People with Autism Spectrum Disorder,” in Proc. Int’l Conf. Content-based Multimedia Indexing, pp. 34–40 (2023)
- 16) Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu, “Personalized Saliency and its Prediction,” IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 2975–2989 (2018)
- 17) Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool, “The Interestingness of Images,” in Proc. IEEE Int’l Conf. Computer Vision, pp. 1633–1640 (2013)
- 18) Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam, “Towards Measuring and Inferring User Interest from Gaze,” in Proc. Int’l Conf. World Wide Web Companion, pp. 525–533 (2017)
- 19) Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O’Donovan, Aaron Hertzmann, and Zoya Bylinskii, “Predicting Visual Importance Across Graphic Design Types,” in Proc. Annual ACM Symposium on User Interface Software and Technology (UIST), pp. 249–260 (2020)
- 20) Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson W.H. Lau, “Task-driven Webpage Saliency,” in Proc. European Conference on Computer Vision (ECCV), pp. 287–302 (2018)
- 21) Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim, “A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction,” Expert Systems with Applications, vol. 23, no. 3, pp. 329–342 (2002)
- 22) Mohamed Nader Jelassi, Sadok Ben Yahia, and Engelbert Mephu Nguifo, “A Personalized Recommender System Based on Users’ Information in Folksonomies,” in Proc. Int’l Conf. World Wide Web (WWW), pp. 1215–1224 (2013)
- 23) Jun Nakamura, and Sanetake Nagayoshi, “The Pottery Skills and Tacit Knowledge of a Maser: An Analysis Using Eye-tracking Data,” Procedia Computer Science, vol. 159, pp. 1680–1687 (2019)
- 24) Weiwei Yu, Dian Jin, Wenfeng Cai, Feng Zhao, and Xiaokun Zhang, “Towards Tacit Knowledge Mining within Context: Visual Cognitive Graph Model and Eye Movement Image Interpretation,” Computer Methods and Programs in Biomedicine, vol. 226, pp. 107107 (2022)
- 25) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “Few-shot Personalized Saliency Prediction Based on Adaptive Image Selection Considering Object and Visual Attention,” Sensors, vol. 20, no. 8: 2170, pp. 1–15 (2020)
- 26) Trung V Nguyen, Edwin V Bonilla, “Collaborative Multi-output Gaussian Processes,” in Proc. Association for Uncertainty in Artificial Intelligence, pp. 643–652 (2014)
- 27) Eric F Lock, “Tensor-on-tensor Regression,” Journal of Computational and Graphical Statistics, vol. 27, no. 3, pp. 638–647 (2018)
- 28) Xi Yin and Xiaoming Liu, “Multi-task Convolutional Neural Network for Pose-invariant Face Recognition,” IEEE Trans. Image Processing, vol. 27, no. 2, pp. 964–975 (2017)
- 29) Rezvan Sherkati and James J Clark, “Clustered Saliency Prediction,” arXiv preprint arXiv:2207.02205 (2022)
- 30) Florian Strohm, Mihai Băce, and Andreas Bulling, “Learning User Embeddings from Human Gaze for Personalized Saliency Prediction,” in Proc. ACM on Human-Computer Interaction, vol. 8, pp. 1–16 (2024)
- 31) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “Few-shot Personalized Saliency Prediction with Similarity of Gaze Tendency Using Object-based Structural Information,” in Proc. IEEE Int’l Conf. Image Processing, pp. 3823–3827 (2022)
- 32) Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, “Yolox: Exceeding Yolo Series in 2021,” arXiv preprint arXiv:2107.08430 (2021)
- 33) Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, “Learning to Predict Where Humans Look,” in Proc. IEEE Int’l Conf. Computer Vision, pp. 2106–2113 (2009)
- 34) Léon Bottou, “Large-scale Machine Learning with Stochastic Gradient Descent,” in Proc. Int’l Conf. Computational Statistics, pp. 177–186 (2010)
- 35) Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, “Salicon: Saliency in Context,” in Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, pp. 1072–1080 (2015)
- 36) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “User-centric Visual Attention Estimation Based on Relationship between Image and Eye Gaze Data,” in Proc. Global Conf. Consumer Electronics, pp. 44–45 (2018)
- 37) Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “User-specific Visual Attention Estimation Based on Visual Similarity and Spatial Information in Images,” in Proc. IEEE Int’l Conf. Consumer Electronics-Taiwan, pp. 479–480 (2019)
- 38) Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand, “What Do Different Evaluation Metrics Tell Us about Saliency Models?,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 740–757 (2018)

Acknowledgement

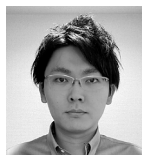
This work was partly supported by the JSPS KAKENHI Grant Numbers JP24K02942, JP23K21676, and JP23K11211.

References

- Laurent Itti, Christof Koch, and Ernst Niebur : “A Model of Saliency-based Visual Attention for Rapid Scene Analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259 (1998)
- Inam Ullah, Muwei Jian, Sumaira Hussain, Jie Guo, Hui Yu, Xing Wang, and Yilong Yin, “A Brief Survey of Visual Saliency Detection,” *Multimedia Tools and Applications*, vol. 79, pp. 34605–34645 (2020)
- Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li, “Salient Object Detection: A Survey,” *Computational Visual Media*, vol. 5, pp. 117–150 (2019)
- Jonathan Harel, Christof Koch, and Pietro Perona, “Graph-based Visual Saliency,” *Advances in Neural Information Processing Systems*, pp. 545–552 (2007)
- Xiaodi Hou, Jonathan Harel, and Christof Koch, “Image Signature: Highlighting Sparse Salient Regions,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201 (2012)
- Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel O’Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i-Nieto, “Salgan: Visual Saliency Prediction with Generative Adversarial Networks,” *arXiv preprint arXiv:1701.01081* (2017)
- Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel, “Contextual Encoder–decoder Network for Visual Saliency Prediction,” *Neural Networks*, vol. 129, pp. 261–270 (2020)
- Aoqi Li and Zhenzhong Chen, “Personalized Visual Saliency: Individuality Affects Image Perception,” *IEEE Access*, vol. 6, pp. 16099–16109 (2018)
- Evan F Risko, Nicola C Anderson, Sophie Lanthier, and Alan Kingstone, “Curious Eyes: Individual Differences in Personality Predict Eye Movement Behavior in Scene-viewing,” *Cognition*, vol. 122, no. 1, pp. 86–90 (2012)
- Olivier Le Meur, Antoine Coutrot, Zhi Liu, Rămä, Adrien Le Roch, and Andrea Helo, “Visual Attention Saccadic Models Learn to Emulate Gaze Patterns from Childhood to Adulthood,” *IEEE Trans. Image Processing*, vol. 26, no. 10, pp. 4777–4789 (2017)
- Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao, “Beyond Universal Saliency: Personalized Saliency Prediction with Multi-task CNN,” in *Proc. Int’l Joint Conf. Artificial Intelligence*, pp. 3887–3893 (2017)
- Sikun Lin and Pan Hui, “Where’s Your Focus: Personalized Attention,” *arXiv preprint arXiv:1802.07931* (2018)
- Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “Few-shot Personalized Saliency Prediction Using Person Similarity Based on Collaborative Multi-output Gaussian Process Regression,” in *Proc. IEEE Int’l Conf. Image Processing*, pp. 1469–1473 (2021)
- Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, and Junfeng He, “Learning from Unique Perspectives: User-aware Saliency Modeling,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 2701–2710 (2023)
- Qiong Wang, Meriem Outtas, Julie Fournier, Elise Etchamendy, Myriam Chérel, and Lu Zhang, “Predicting Personalized Saliency Map for People with Autism Spectrum Disorder,” in *Proc. Int’l Conf. Content-based Multimedia Indexing*, pp. 34–40 (2023)
- Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu, “Personalized Saliency and its Prediction,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 2975–2989 (2018)
- Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool, “The Interestingness of Images,” in *Proc. IEEE Int’l Conf. Computer Vision*, pp. 1633–1640 (2013)
- Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam, “Towards Measuring and Inferring User Interest from Gaze,” in *Proc. Int’l Conf. World Wide Web Companion*, pp. 525–533 (2017)
- Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O’Donovan, Aaron Hertzmann, and Zoya Bylinskii, “Predicting Visual Importance Across Graphic Design Types,” in *Proc. Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 249–260 (2020)
- Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson W.H. Lau, “Task-driven Webpage Saliency,” in *Proc. European Conference on Computer Vision (ECCV)*, pp. 287–302 (2018)
- Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim, “A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction,” *Expert Systems with Applications*, vol. 23, no. 3, pp. 329–342 (2002)
- Mohamed Nader Jelassi, Sadok Ben Yahia, and Engelbert Mephu Nguifo, “A Personalized Recommender System Based on Users’ Information in Folksonomies,” in *Proc. Int’l Conf. World Wide Web (WWW)*, pp. 1215–1224 (2013)
- Jun Nakamura, and Sanetake Nagayoshi, “The Pottery Skills and Tacit Knowledge of a Maser: An Analysis Using Eye-tracking Data,” *Procedia Computer Science*, vol. 159, pp. 1680–1687 (2019)
- Weiwei Yu, Dian Jin, Wenfeng Cai, Feng Zhao, and Xiaokun Zhang, “Towards Tacit Knowledge Mining within Context: Visual Cognitive Graph Model and Eye Movement Image Interpretation,” *Computer Methods and Programs in Biomedicine*, vol. 226, pp. 107107 (2022)
- Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “Few-shot Personalized Saliency Prediction Based on Adaptive Image Selection Considering Object and Visual Attention,” *Sensors*, vol. 20, no. 8: 2170, pp. 1–15 (2020)
- Trung V Nguyen, Edwin V Bonilla, “Collaborative Multi-output Gaussian Processes,” in *Proc. Association for Uncertainty in Artificial Intelligence*, pp. 643–652 (2014)
- Eric F Lock, “Tensor-on-tensor Regression,” *Journal of Computational and Graphical Statistics*, vol. 27, no. 3, pp. 638–647 (2018)
- Xi Yin and Xiaoming Liu, “Multi-task Convolutional Neural Network for Pose-invariant Face Recognition,” *IEEE Trans. Image Processing*, vol. 27, no. 2, pp. 964–975 (2017)
- Rezvan Sherkati and James J Clark, “Clustered Saliency Prediction,” *arXiv preprint arXiv:2207.02205* (2022)
- Florian Strohm, Mihai Băce, and Andreas Bulling, “Learning User Embeddings from Human Gaze for Personalized Saliency Prediction,” in *Proc. ACM on Human-Computer Interaction*, vol. 8, pp. 1–16 (2024)
- Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “Few-shot Personalized Saliency Prediction with Similarity of Gaze Tendency Using Object-based Structural Information,” in *Proc. IEEE Int’l Conf. Image Processing*, pp. 3823–3827 (2022)
- Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, “Yolox: Exceeding Yolo Series in 2021,” *arXiv preprint arXiv:2107.08430* (2021)
- Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, “Learning to Predict Where Humans Look,” in *Proc. IEEE Int’l Conf. Computer Vision*, pp. 2106–2113 (2009)
- Léon Bottou, “Large-scale Machine Learning with Stochastic Gradient Descent,” in *Proc. Int’l Conf. Computational Statistics*, pp. 177–186 (2010)
- Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, “Salicon: Saliency in Context,” in *Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition*, pp. 1072–1080 (2015)
- Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “User-centric Visual Attention Estimation Based on Relationship between Image and Eye Gaze Data,” in *Proc. Global Conf. Consumer Electronics*, pp. 44–45 (2018)
- Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, “User-specific Visual Attention Estimation Based on Visual Similarity and Spatial Information in Images,” in *Proc. IEEE Int’l Conf. Consumer Electronics-Taiwan*, pp. 479–480 (2019)
- Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand, “What Do Different Evaluation Metrics Tell Us about Saliency Models?,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757 (2018)



Yuya Moroto received his B.S. degree in Electronics and Information Engineering from Hokkaido University, Japan in 2019, and his M.S. and Ph.D. degrees in Information Science and Technology from Hokkaido University, Japan in 2021 and 2024. His research interests include computer vision, biological information analysis, multimodal signal processing and its applications. He is a member of IEEE.



Keisuke Maeda received his B.S., M.S., and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2015, 2017, and 2019. At present, he is currently a Specially Appointed Associate Professor in the Data-Driven Interdisciplinary Research Emergence Department, Hokkaido University. His research interests include multimodal signal processing and machine learning and its applications. He is an IEEE and IEICE member.



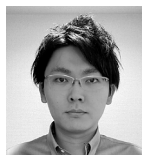
Takahiro Ogawa received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2003, 2005 and 2007, respectively. He joined Graduate School of Information Science and Technology, Hokkaido University in 2008. He is currently a professor in the Faculty of Information Science and Technology, Hokkaido University. His research interests are AI, IoT and big data analysis for multimedia signal processing and its applications. He was a special session chair of IEEE ISCE2009, a Doctoral Symposium Chair of ACM ICMR2018, an organized session chair of IEEE GCCE2017-2019, a TPC Vice Chair of IEEE GCCE2018, a Conference Chair of IEEE GCCE2019, etc. He has been also an Associate Editor of ITE Transactions on Media Technology and Applications. He is a senior member of IEEE and a member of ACM, IEICE and ITE.



Miki Haseyama received her B.S., M.S. and Ph.D. degrees in Electronics from Hokkaido University, Japan in 1986, 1988 and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University as an associate professor in 1994. She was a visiting associate professor of Washington University, USA from 1995 to 1996. She is currently a professor in the Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), an Editor-in-Chief of ITE Transactions on Media Technology and Applications, a Director, International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE). She is a member of the IEEE, IEICE, Institute of Image Information and Television Engineers (ITE) and Acoustical Society of Japan (ASJ).



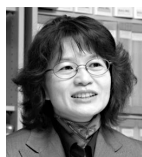
Yuya Moroto received his B.S. degree in Electronics and Information Engineering from Hokkaido University, Japan in 2019, and his M.S. and Ph.D. degrees in Information Science and Technology from Hokkaido University, Japan in 2021 and 2024. His research interests include computer vision, biological information analysis, multimodal signal processing and its applications. He is a member of IEEE.



Keisuke Maeda received his B.S., M.S., and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2015, 2017, and 2019. At present, he is currently a Specially Appointed Associate Professor in the Data-Driven Interdisciplinary Research Emergence Department, Hokkaido University. His research interests include multimodal signal processing and machine learning and its applications. He is an IEEE and IEICE member.



Takahiro Ogawa received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2003, 2005 and 2007, respectively. He joined Graduate School of Information Science and Technology, Hokkaido University in 2008. He is currently a professor in the Faculty of Information Science and Technology, Hokkaido University. His research interests are AI, IoT and big data analysis for multimedia signal processing and its applications. He was a special session chair of IEEE ISCE2009, a Doctoral Symposium Chair of ACM ICMR2018, an organized session chair of IEEE GCCE2017-2019, a TPC Vice Chair of IEEE GCCE2018, a Conference Chair of IEEE GCCE2019, etc. He has been also an Associate Editor of ITE Transactions on Media Technology and Applications. He is a senior member of IEEE and a member of ACM, IEICE and ITE.



Miki Haseyama received her B.S., M.S. and Ph.D. degrees in Electronics from Hokkaido University, Japan in 1986, 1988 and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University as an associate professor in 1994. She was a visiting associate professor of Washington University, USA from 1995 to 1996. She is currently a professor in the Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), an Editor-in-Chief of ITE Transactions on Media Technology and Applications, a Director, International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE). She is a member of the IEEE, IEICE, Institute of Image Information and Television Engineers (ITE) and Acoustical Society of Japan (ASJ).