様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １６ 日現在

機関番号：２５３０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１４０２

研究課題名（和文）一人称視点映像に対する視覚的注意推定技術と注視誘導可能な情報提示の実現

研究課題名（英文）Visual Attention Estimation and Attention Retargeting for First Person Vision

研究代表者
滝本　裕則（Takimoto, Hironori）

岡山県立大学・情報工学部・准教授

研究者番号：１０４１３８７４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：人と調和する情報環境の実現に寄与する新しいデバイスとして、インテリジェントグラスなどの小型カメラを備えたウェアラブル透過型ディスプレイが注目されている。我々は、ウェアラブルカメラにより撮影される一人称視点映像より、人の内部状態（意図・興味）を推定するための重要な手掛かりである視覚的注意を高精度に推定する技術の実現に取り組んだ。また、人の日常的な活動を支援する情報提示空間の実現を目指し、ウェアラブル透過型ディスプレイを対象とした注視誘導技術を実現するため、視覚的注意と視覚的な閾下情報提示に基づく注視誘導技術の確立についても検討を行った。

研究成果の学術的意義や社会的意義
ウェアラブル透過型ディスプレイを用いたより自然なインタラクションを実現するため、状況によって特定の領域に人の注視を自然に誘導する技術が望まれている。注視誘導の実現によって必要な情報へアクセスしやすくなることで、インタフェースのユーザビリティ向上が期待できるだけではなく、情報が記憶に残りやすくなるというメリットがある。したがって、ウェアラブルデバイスを用いたより効果的なインタフェースの実現に向けて、視覚的顕著性に基づく一人称視点映像から内部状態推定技術に加え、ウェアラブル透過型ディスプレイを対象とした視覚的顕著性に基づく注視誘導技術の実現が望まれている。

研究成果の概要（英文）：Wearable transmissive displays equipped with small cameras, such as smart glasses, are attracting attention as new devices that contribute to realizing an information environment in harmony with people. We have worked on realizing the technique to precisely estimate visual attention, which is an important cue for estimating a person's internal state (intention and interest) from the first-person video captured by a wearable camera. We also studied the establishment of a gaze guidance technology based on visual attention and visual threshold information presentation to realize a gaze guidance technology for wearable transmissive displays, aiming to recognize an information presentation space that supports daily activities for humans.

研究分野：画像工学，知覚情報処理

キーワード：画像処理　視覚的顕著性　一人称視点映像　深層学習

１．研究開始当初の背景

Attentive user interface といった人中心のインタラクティブシステムにとって、人の内部状態を推定することは重要な要素技術であり、その手がかりとして注視情報や視覚的注意が注目されている。映像から空間的な視覚的注意、いわゆる視覚的顕著性マップを推定する技術は、人が五感から得られる 80%以上を占める視覚情報のうち「いつ・どの領域に・どの程度注目しているか」といった情報を得ることが可能であるため、人の内部状態を推定するための重要な手掛かりとなる。特に、ウェアラブルカメラにより撮影される一人称視点映像は、装着者の内部状態・行動意図推定や行動支援などへの応用が期待されている。よって、一人称視点映像に対する視覚的注意推定技術の実現が望まれている。

人が注意を向けやすい映像中の領域とその度合いを推定する視覚的顕著性モデルはこれまでに多く提案されている。しかし、色や輝度のコントラストなどの低次の視覚特徴から構成される従来の視覚的顕著性モデルは、実環境下において高精度な注意推定が困難である。また、一人称視点映像に含まれている「自己動作・運動によって生じる視覚的運動刺激などの処理機能」が考慮されておらず真に実用的なものは未だ無い。したがって、ウェアラブルカメラを用いた人中心のインタラクティブシステムを実現するためには、高次の視覚的顕著性特徴と自己動作・運動に基づく一人称視点に対する視覚的注意推定技術の実現が望まれている。

一方、人と調和する情報環境の実現に寄与する新しいデバイスとして、インテリジェントグラスなどの小型カメラを備えたウェアラブル透過型ディスプレイが注目されている。本デバイスを用いたより自然なインタラクションを実現するため、状況によって特定の領域に人の注視を自然に誘導する技術が望まれている。注視誘導の実現によって必要な情報へアクセスしやすくなることで、インタフェースのユーザビリティ向上が期待できるだけではなく、情報が記憶に残りやすくなるというメリットがある。したがって、ウェアラブルデバイスを用いたより効果的なインタフェースの実現に向けて、視覚的顕著性に基づく一人称視点映像から内部状態推定技術に加え、ウェアラブル透過型ディスプレイを対象とした視覚的顕著性に基づく注視誘導技術の実現が望まれている。

２．研究の目的

本課題では、ウェアラブルカメラにより撮影される一人称視点映像より、人の内部状態（意図・興味）を推定するための重要な手掛かりである視覚的注意を高精度に推定する技術の実現を目的としている。また、人の日常的な活動を支援する情報提示空間の実現を目指し、ウェアラブル透過型ディスプレイを対象とした注視誘導技術を実現するため、視覚的注意と視覚的な閾下情報提示に基づく注視誘導技術の確立についても検討を行う。具体的には、以下に示す４つの課題について取り組む。

- 深層学習に基づく一人称視点映像に対する視覚的顕著性推定モデルの提案
- 自己運動を考慮した視覚的顕著性推定の補正
- 光投影による注視誘導技術の提案
- 注視情報が付与された一人称視点映像データベースの構築

３．研究の方法

前述した４つの課題について、それぞれ研究の方法を示す。

(1) 深層学習に基づく一人称視点映像に対する視覚的顕著性推定モデルの提案

深層学習モデルを用いた一人称注視予測の高精度化に向けて、既存モデル[1]に対して、様々な CNN モデルを適用することによりその有効性を検証した。適用する様々な CNN モデルとして、マルチスケールで局所的から大域的までの顕著な特徴の抽出が可能な Dilated convolution、行動認識のモデルでよく用いられる空間方向の畳み込みに加えて時間方向に対しても畳み込みを行う 3D convolution に注目した。

深層学習モデルにおいて Multi-scale に特徴を抽出する Dilated convolution が提案されている。Dilated convolution では、通常の畳み込み演算に加え一定画素離れた画素に対して同様に畳み込み演算を行う。Dilated convolution を用いた Resnet の派生モデルに Dilated residual Networks(DRNs)[2]が提案されている。また、Dilated convolution の問題点である計算コストの増大を解決した手法として Joint Pryramid Upsampling(JPU)[3]が提案されている。一方、動画の畳み込みを行う手法として、Ji らによる 3D convolution が提案されている[4]。3D convolution とは、入力動画に対して空間方向のみの畳み込み演算である 2D convolution を拡張し、時間方向にも畳み込み演算を行うことにより時空間情報を考慮した特徴を抽出可能にしたものである。我々はこれらをモデルに導入し、視覚的顕著性推定に対する有効性を検証した。

(2) 自己運動を考慮した視覚的顕著性推定の補正

一人称視点映像に含まれる特有の情報として対象者の自己運動が挙げられる。この自己運動には、装着者の身体の動きに加え頭部の動きが含まれており、頭部の動きは対象を視野の中心で捉えようとする際に発生することが多い。したがって、一人称視点映像から頭部の動きを推定することで、より高精度な視覚的注意予測が可能であると考えられる。そこで、自己運動と視覚的注意の関係に基づいた自己運動注意マップを提案し、自己運動を考慮した視覚的顕著性推定の補正を行った。

まず、一人称視点映像の隣接する異なる 2 フレームに対してそれぞれ特徴点を検出した。特徴点検出と特徴点を表す局所特徴量記述のために Accelerated KAZE 法[5]を用いた。次に、検出された特徴点に対して注目する隣接フレーム間で対応点探索を行うことによりオプティカルフローを求めた。その後、カメラの回転行列と並進移動ベクトルを求めた。

ここで、人の視界は水平方向約 200 度に開けているが、その細部まで見ることができるのは中心視野とよばれる約 2 度に限られている。周辺視野の基本的な特性についてはこれまで多くの研究が行われており、周辺視の視力は離心角度 10 度で中心視の約 20%にまで低下することが知られている。そこで、2 次元ガウス分布により相対視力特性のモデル化を行った。また、得られた回転行列より各軸方向の角速度を求めた。得られた各軸方向の角速度に基づいて、2 次元ガウス分布の中心位置を変更することにより、頭部姿勢変動に応じた相対視力分布である自己運動注意マップを生成した。

最後に、課題(1)で得られた顕著性マップに対して自己運動注意マップを適用し、両マップの積を求めることにより自己運動を考慮した視覚的顕著性マップの補正を行った。

(3) 光投影による注視誘導技術の提案

ウェアラブル透過型ディスプレイを想定した注視誘導実現に向けて、光投影システムによる注視誘導の第一段階として、実空間の平面に対する外観制御法を提案した。まず、プロジェクタカメラ系を用いて外見制御が可能な実環境下の領域（平面）を推定した。視覚的顕著性マップにおいて、顕著度が高い領域は注目を引き付けやすい。よって、実環境の任意の対象領域に注視を誘導するため、視覚的顕著性に基づく画像再配色法を用いて対象領域とその周囲の理想的な外観を推定した。この推定は、ボトムアップ型視覚的注意のための視覚的顕著性モデルのリバースエンジニアリングに基づくものである。その後、プロジェクタとカメラの動的フィードバックシステムを用いて、注視誘導のための最適な投影パターンを求め平面に投影した。

(4) 注視情報が付与された一人称視点映像データベースの構築

撮影装置として Pupil labs 社製 Mobile Eye Tracking Headset を用いた。本装置は人の視線を計測するためのものであり、右目上部に 1 つの World カメラと両目を撮影するための 2 つの Eye カメラを備えている。World カメラの仕様は 1920×1080 画素、30fps、視野角は約 90 度である。一方、Eye カメラとして赤外カメラと赤外 LED をセットで備えており、その仕様は 640×480 画素、120fps である。なお、視線は暗瞳孔検出に基づいて検出される。

４．研究成果

４つの課題について、それぞれ研究成果を示す。

(1) 深層学習に基づく一人称視点映像に対する視覚的顕著性推定モデルの提案

ベースラインとなる Huang らのモデルに各 CNN モデルを適用し有効性の評価を行った。提案モデル 1 は FE-module の各特徴抽出ネットワークに用いられる VGG を画像認識などの分野で VGG よりも精度が良いとされる Resnet に変更したモデルである。提案モデル 2 と 3 はマルチスケールな特徴を抽出可能な Dilated convolution を使用した Resnet(DRNs)を FE-module に適用したモデルである。なお、提案モデル 2 では DRN-A を適用している。一方、提案モデル 3 では DRN-A における Gridding artifact と呼ばれる問題を解決し、DRN-A よりも精度が良いとされる DRN-D を FE-module に適用したモデルである。提案モデル 4 は DRN-D を基幹に Dilated convolution を高速化した JPU を FE-module に適用したモデルである。提案モデル 5 は FE-module の空間特徴の抽出を行うネットワークに対し、3D convolution を用いた Resnet(3D-Resnet)を適用したモデルである。提案モデル 6 は FE-module の空間特徴の抽出を行うネットワークに加え、時間特徴の抽出を行うネットワークに対しても 3D-Resnet を適用することで、I3D[6]と同じ構成にしたモデルである。

実験結果として、表 1 に各評価指標のスコアを示す。評価実験の結果、提案手法を適用したモデルは Dilated convolution を用いた場合は精度が低下したが、3D convolution を用いると精度が向上することを確認した。また、各評価指標に着目した場合、AUC、AAE 共に空間方向および時間方向両方に 3D convolution を適用したモデルが最も予測精度が高いことを確認した。

表 1: 各提案モデルの精度比較

モデル	FE-module の構造	$L_f \downarrow$	AUC ↑	AAE ↓
ベースライン [2]	RGB : 2D VGG16 Flow : 2D VGG16	0.9210	0.8915	10.9889
提案モデル 1	RGB : 2D Resnet18 Flow : 2D Resnet18	0.9281	0.8906	11.0545
提案モデル 2	RGB : 2D DRN(A)18 Flow : 2D DRN(A)18	0.9306	0.8913	11.0212
提案モデル 3	RGB : 2D DRN(D)22 Flow : 2D DRN(D)22	0.9470	0.8825	11.6734
提案モデル 4	RGB : 2D JPU(D)22 Flow : 2D JPU(D)22	0.9486	0.8858	11.4514
提案モデル 5	RGB : 3D Resnet18 Flow : 2D Resnet18	0.9309	0.8928	10.9060
提案モデル 6	RGB : 3D Resnet18 Flow : 3D Resnet18	**0.9196**	**0.8960**	**10.6269**

(2) 自己運動を考慮した視覚的顕著性推定の補正

課題(1)で得られた顕著性マップと推定した自己運動から求めた自己運動注意マップに基づいて、自己運動を考慮した視覚的顕著性マップを生成した。提案モデルによって得られた視覚的顕著性マップと実際の顕著性マップにて高い顕著度を持つ領域とは若干のずれはあるものの、頭部移動によって視線が移動しようとしているおおよその方向を顕著性マップが推定できていることを確認した。今後、課題(1)と同様の定量的な評価実験を行う必要がある。

なお、人は頭部運動だけではなく、眼球運動も利用して視線方向を変更している。視線方向を変える際、眼球と頭部は同じ速さで遷移先の注視方向へ直線的に動くといった単純な動きではなく、複雑に連動していることが示唆されている。具体的には、注視点を変更しようとする際、まず眼球が急速に注視の遷移方向に動き出す。その後、頭部は少し遅れて眼球の動きを追うように、かつ、眼球よりも遅い角速度で同じ方向に動く。また、頭部が注視の遷移方向に動くにつれて眼球はその頭部運動とは逆方向の運動をする、いわゆる前庭動眼反射と呼ばれる神経制御が働くとされている[7]。このような前庭動眼反射に関する挙動を考慮したモデルの提案が今後の課題である。

(3) 光投影による注視誘導技術の提案

提案手法の有効性を検証するためにプロジェクタカメラ系を用いた評価実験を行った。E 投影対象は灰色の布を貼り付けた板上に配置した写真とした。注視誘導の対象領域として、比較的顕著度の低い領域である領域を指定した。なお、既存のアプローチとして対象領域に対してのみスポットライト（白色光）を投影した場合との比較を行った。結果として、光投影による限られた範囲内での色成分制御であっても注視の誘導が可能であることを確認した[8]。今後、ウェアラブル透過型ディスプレイでの実装を想定した手法の改良が必要である。

(4) 注視情報が付与された一人称視点映像データベースの構築

COVID-19 感染拡大の影響により十分な被験者と実験時間を確保できず対応が遅れた。また、大規模な一人称視点映像データセットが公開されたため、結果として既存データセットを用いて実験を行った。

参考文献

[1] Y. Huang, M. Cai, Z. Li, and Y. Sato. ``Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition,'' Proc. of European Conference on Computer Vision, pp. 754-769, (2018).

[2] F. Yu, V. Koltun, and T. Funkhouser. ``Dilated Residual Networks,'' Proc. of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 472-480, (2017).

[3] W. Huikai, J. Zhang, K. Huang, K. Liang, and Y. Yu. ``FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation,'' arXiv preprint, arXiv:1903.11816, (2019).

[4] S. Ji, W. Xu, M. Yang, and K. Yu. ``3D Convolutional Neural Networks for Human Action Recognition,'' IEEE Transactions on Pattern Analysis and Machine

Intelligence, Vol. 35, pp. 221-231, (2013).
[5] P. F. Alcantarilla, J. Nuevo, and A. Bartoli: ``Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces,'' Proc. of British Machine Vision Conference, (2013).
[6] J. Carreira and A. Zisserman: ``Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,'' Proc. of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 6299-6308, (2017).
[7] 沖中 大和, 満上 育久, 八木 康史: ``人の眼球と頭部の協調運動を考慮した視線推定,'' 情報処理学会研究報告, Vol. 2010-CVIM-202, pp. 1-8, (2016).
[8] H. Takimoto, K. Yamamoto, A. Kanagawa, M. Kishihara, and K. Okubo: ``Attention Retargeting Using Saliency Map and Projector Camera System in Real Space,'' IEEJ Transactions on Electrical and Electronic Engineering, Vol. 14, Issue 6, pp. 853-861, (2019).

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 H. Takimoto, F. Omori, and A. Kanagawa	４．巻 Vol. 35, Issue 1
２．論文標題 Image Aesthetics Assessment Based on Multi-stream CNN Architecture and Saliency Features	５．発行年 2021年
３．雑誌名 Applied Artificial Intelligence	６．最初と最後の頁 25-40
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/08839514.2020.1839197	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Takimoto Hironori、Yamamoto Katsumi、Kanagawa Akihiro、Kishihara Mitsuyoshi、Okubo Kensuke	４．巻 14
２．論文標題 Attention retargeting using saliency map and projector-camera system in real space	５．発行年 2019年
３．雑誌名 IEEJ Transactions on Electrical and Electronic Engineering	６．最初と最後の頁 853～861
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/tee.22874	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計6件（うち招待講演　1件／うち国際学会　2件）

１．発表者名 新谷 浩平, 滝本 裕則, 金川 明弘
２．発表標題 CNNを用いたウェブページに対する顕著性推定に関する研究
３．学会等名 第21回IEEE広島支部学生シンポジウム（21th HISS）
４．発表年 2019年

１．発表者名 泉水 長門, 滝本 裕則, 山内 仁, 金川 明弘, 遠部 雅弘
２．発表標題 顔特徴を用いたポートレート写真に対する審美的品質推定
３．学会等名 第21回IEEE広島支部学生シンポジウム（21th HISS）
４．発表年 2019年

１．発表者名
泉水 長門，滝本 裕則，大森 史耶，山内 仁，金川 明弘，遠部 雅弘

２．発表標題
視覚的顕著性を考慮したCNNに基づく写真の審美的品質推定

３．学会等名
精密工学会IAIPサマーセミナー2019

４．発表年
2019年

１．発表者名
滝本 裕則

２．発表標題
深層学習に基づく画像処理の取り組み紹介

３．学会等名
AI・IoTに関わる異業種研究会(オープンイノベーション促進事業　第4回　技術研究会)（招待講演）

４．発表年
2020年

１．発表者名
H. Takimoto, S. Katsumata, Sulfayanti F. Situju, A. Kanagawa, and A. Lawi

２．発表標題
Visual Saliency Estimation Based on Multi-task CNN

３．学会等名
The Fourteenth International Conference on Industrial Management (ICIM2018)（国際学会）

４．発表年
2018年

１．発表者名
F. Omori, H. Takimoto, H. Yamauchi, A. Kanagawa, T. Iwasaki, and M. Ombe

２．発表標題
Aesthetic Quality Evaluation using Multi-task CNN

３．学会等名
The Fourteenth International Conference on Industrial Management (ICIM2018)（国際学会）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	満倉　靖恵 (Mitsukura Yasue) (60314845)	慶應義塾大学・理工学部（矢上）・教授 (32612)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関