Vol.2022-HCI-197 No.50
2022/3/16

ARグラスにおけるSaliency Mapを利用した「ながら見」用映像配置方法

田中 大[1,a)]　川喜田 裕之[1,b)]　半田 拓也[1,c)]

概要：近年，コンシューマーデバイスとしてのARグラスの需要が高まっている．本研究では，ARグラスによる映像の視聴のうち，日常生活と映像視聴を並行して行う「ながら見」を対象とする．スマートフォンやタブレットと異なり，ARグラスはハンズフリーで映像視聴することが出来る．しかし，ながら見においてARグラスの表示領域のどこに映像を配置するかについて，日常生活の中で邪魔にならないか，映像視聴の負担にならないかという課題点があり，議論が十分なされているとは言えない．本研究では，画像中の顕著性を表すSaliency Mapを用いて日常生活で邪魔にならない領域の推定を行い，ながら見における快適な映像表示領域を決定するアルゴリズムについて議論する．ながら見に適したUIをモデリングし，その有効性を検討した．

1. はじめに

近い将来コンシューマーデバイスとしてARグラスが浸透することが予想されている[1]．ARグラスは，スマートフォンのような従来の表示デバイスと異なり，現実空間中に映像を表示することができ，またその映像視聴は着用者のみに限られるという特徴がある．そのため，ARグラスによる視聴スタイルは多様化し，映像視聴体験は現在よりもさらにパーソナルな体験となる．ARグラスによる視聴スタイルの特徴としては，各ユーザーが自分好みの映像を長時間日常生活と並行して「ながら見」するようになると予想されるため，その際のUIについて議論を行う必要がある．

筆者らはARグラスで視聴する動画のフォーマットをVirtualTVと呼称し過去に議論した[2]．VirtualTVの表示位置について，現在市販のARグラスで実装されている主な方法を整理すると

- 視野内に固定する方法
- 実空間に固定する方法
- tagalong[3]

の３つに分類出来る．視野内に固定する方法はARグラスの表示面上に映像を固定する方法である．これが一般的な方法であり，SLAM(Simultaneous Localization and Mapping)技術を必要としないことから，多くのARグラスに実装されている．しかし，周りの環境によっては映像が視界を邪魔してしまう可能性がある．実空間に固定する方法はSLAMにより実空間の特定の位置に映像を固定する方法である．VirtualTVを実際のオブジェクトと同様に扱うことができるため，ユーザーにとって直感的に使いやすい．しかし，VirtualTVを別の場所へ移動するためにはユーザーによる操作が必要である．tagalongはMicrosoft社のARグラスHololensに実装されているアルゴリズムで，顔を動かし表示領域の外に表示オブジェクトが出ると，表示領域の近くまでそのオブジェクトを移動するというアルゴリズムである．これは作業中のマニュアル参照などに適した提示方法である一方で，表示領域外にVirtualTVが出てしまうことがあるため，「ながら見」として視界へ映像を入れ続けたいというニーズを満たすことが出来ない．

以上の現状を踏まえ，本論文では，ながら見における快適な映像配置のために，VirtualTVの配置用アルゴリズムを作成した．本手法は，人間の注視する可能性の高い領域を表したSaliency Mapを用いて日常生活で邪魔にならない領域を推定することを特徴とし，ARグラスにおけるながら見を想定する．

2. 関連研究

本研究ではながら見を対象とするため，ユーザーが実空間において注視する可能性の高い領域（Saliency Map）を推定する必要がある．Saliency Mapを推定する研究は古く

[1] NHK 放送技術研究所
NHK Science and Technology Research Laboratories 1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510 Japan
[a)] tanaka.m-oc@nhk.or.jp
[b)] kawakita.h-dq@nhk.or.jp
[c)] handa.t-es@nhk.or.jp

から盛んに行われている [4]．近年の Deep Neural Network の隆盛に伴い Convolutional Neural Network による推定法 [5] も多く提案されており，高精度を達成している．

Saliency Map 推定を可視化したヒートマップの例を図 1 に示す．人物が Saliency の高い領域として可視化されている．人物が目の前にいた場合にはその領域を注視する可能性が高いため，直感に即した推定結果となっている．AR は実空間中へバーチャルなオブジェクトを表示する．Saliency Map の推定により顕著性の低い領域にテキストラベルを配置する方法 [6] 等，ラベルの配置方法については多数議論されているが，AR グラスを含む HMD（head-mounted display）における視聴用の映像の配置のモデリングを行う議論は十分になされていない．

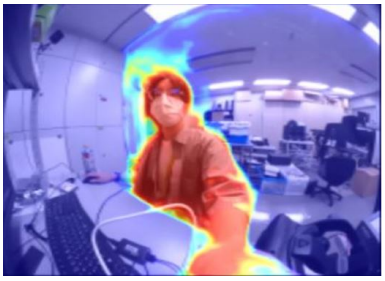

図 1 CNN による Saliency Map 推定結果の一例

3. ながら見用 VirtualTV 配置決定法

本研究の目的は AR グラス内に表示する VirtualTV のながら見用配置決定アルゴリズムを作成することである．将来的に AR グラスは視野内の広い領域に情報を提示できると期待できることや，VirtualTV のながら見を想定し，VirtualTV の配置の要件を

- 連続視聴に負荷のない領域にある (人間工学的制約)
- 急激な移動がなく時間的に連続性がある (時間的制約)
- 生活の中で注視する領域を邪魔しない

という 3 点に整理した．

この要件を踏まえて，今回はある時刻 t における，VirtualTV の AR グラス視野内における中心座標 X_t について議論する．X_t はユーザーの状況により変化する確率変数であり，ここではユーザーの状況が与えられた時の最適な X_t の分布を考える．

映像という連続視聴の体験のためには極端に高い位置や視界の端に表示することは好ましくない．このような人間工学的に連続視聴に負荷のない領域を H と定義する．

急激な移動がなく時間的に連続性があるとは，ユーザーの視界の中で VirtualTV が認識できない速度で移動することがなく，連続的に移動することである．表示される VirtualTV は時間的に連続性のある映像であり，表示領域は一時刻前の座標 X_{t-1} に依存する．

生活の中で見たいと思う領域を邪魔しないためには，AR グラスの外向きカメラにより状況を認識する必要がある．状況を認識し邪魔にならないように配慮する．これを目的として人間が注視する可能性の高い領域を示す Saliency Map を利用する．Saliency Map の示す人間が注視する確率の高い領域を S とすると，注視する可能性の低い領域を $\bar{S}$ と書ける．

以上の議論から，X_t の分布は $p(X_t|H, X_{t-1}, \bar{S})$ のように，3 つの要素による条件付き分布と捉えることが出来る．H，X_{t-1}，$\bar{S}$，はそれぞれ独立であり，$p(X_t)$ に一様分布を仮定すると，ベイズの定理から

$$p(X_t|H, X_{t-1}, \bar{S}) \propto p(X_t|H)p(X_t|X_{t-1})p(X_t|\bar{S}) \quad (1)$$

のように必要な分布は積の形に分解することが出来る．よって，以上 3 つの分布を設計し，かけ合わせることで X_t の推定を行うことが出来る．$p(X|H)p(X|X_{t-1})p(X|\bar{S})$ のイメージを図 2 に示す．

$p(X_t|H)$ は人間工学的制約条件である．AR グラスの広い視野内で連続視聴に負荷のかからない領域を矩形として設計する．矩形のパラメータは，視界における中心座標 X_H，高さ h_H，幅 w_H である．

$p(X_t|X_{t-1})$ は一時刻前の座標周辺に留まる制約条件である．これをパラメータ r を用いて，X_{t-1} を中心とする半径 r の円の中に留まると読み替え，半径 r の一様分布と設計する．

$p(X_t|\bar{S})$ は Saliency Map Detection のアルゴリズム [5] を用いる．推定された Saliency Map の各画素における値の符号を反転し正規化することで $p(X_t|\bar{S})$ を設計する．

以上の分布の積として求めた分布 $p(X|H)p(X|X_{t-1})p(X|\bar{S})$ により，X_t の推定量 $\hat{X}_t$ を

$$\hat{X}_t = \sum_X X p(X|H)p(X|X_{t-1})p(X|\bar{S}) \quad (2)$$

のように期待値として求める．

本手法は X_t の決定に AR グラスの外向きカメラのみを用いる．各種センサーや SLAM 機能を用いないため，より多くの AR グラスで使用可能であるという特徴がある．

4. 数値実験

本手法によりユーザーの注視する可能性が高い領域を避けて映像を表示出来ることを，一人称視点映像のデータセットを使ったシミュレーションにより確認する．視野内に VirtualTV を固定する方法を比較対象として，視界に映像が固定されて表示している状況を想定し，図 3 のように映像を視界の中心および左右それぞれの座標に固定して配置した場合を設定した．データセットの映像 D の中心を原点とし，D の幅を w_D として，それぞれ $(-\frac{1}{4}w_D, 0)$ に固定する場合を P_{left}，$(0, 0)$ に固定する場合を P_{center}，$(\frac{1}{4}w_D, 0)$ に固定する場合を P_{right} とした．

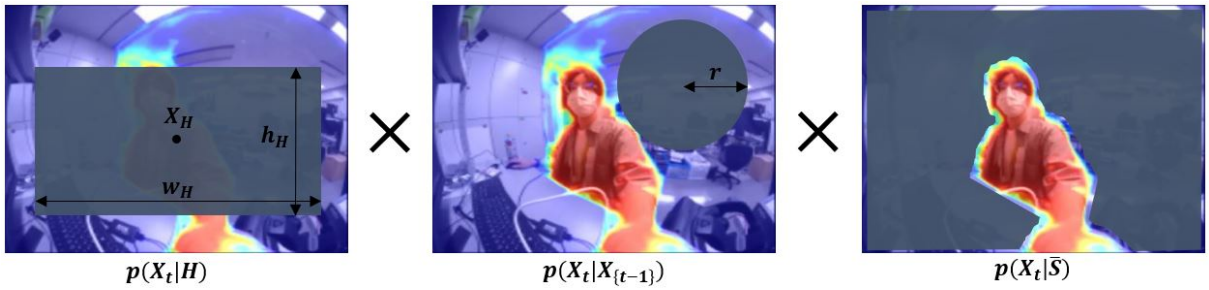

図 2 $p(X|H)p(X|X_{t-1})p(X|\bar{S})$ イメージ図

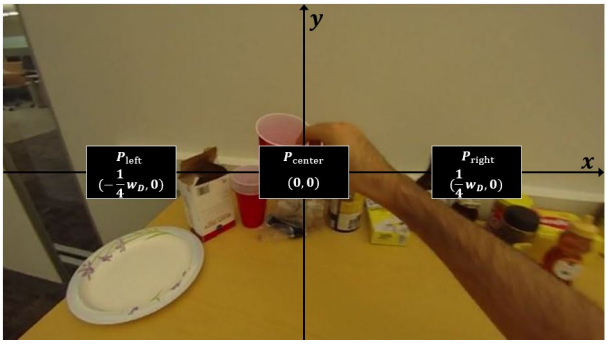

図 3 比較対象として映像を固定した位置

また，実験には表 1 のパラメータを用い，アルゴリズムの初期値は P_{center} と等しくした．ここで，h_D は D の高さとする．AR グラス着用時の視点映像として，室内で作業をしている一人称視点の映像である GTEA データセット [7] を用いた．

表 1 実験パラメータ

r	X_{H}	h_{H}	w_{H}
$0.2h_D$	(0,-0.1h_D)	$0.4h_D$	$0.8w_D$

4.1 実験設定

視野に対応するデータセットの映像を D，表示する VirtualTV を V とする．V の高さ h_V 及び幅 w_V はスマートフォンを 30cm ほど顔から離して置いた際に視野内における大きさに相当するように設定した．手法を評価するために，データセットに対するスコアを

$$\text{score} = 1 - \frac{S_V}{h_V w_V} \tag{3}$$

と定義した．ここで S_V は V のなかで Saliency が高いと推定された領域の面積をとする．このスコアは，表示する VirtualTV の映像領域と Saliency Map の重ならない割合を意味する．

4.2 結果

図 4 に本手法をデータセットに適用している様子を示す．画像中の黒い矩形が VirtualTV の表示される位置であり，Saliency の高い領域を避けて VirtualTV を表示出来ることが確認できた．

データセット 28 本の映像において，本手法と P_{left}，P_{center}，P_{right} のスコアを図 5 にまとめる．

初期値である P_{center} のスコアと比較し，本手法は全ての動画において本手法のスコアが上昇していることから，Saliency に合わせて邪魔になる領域を避けて動いていることが確認できた．しかし，P_{left}，P_{right} と比較すると，スコアは同程度となっていた．

4.3 考察

P_{left}，P_{right} のスコアが高くなった原因としては，今回使用したデータセットでは Saliency の高い領域が中心に集まる傾向にあることが考えられる．以下でその理由を説明する．図 6 はデータセットの Saliency Map を足し合わせたヒートマップである．

一人称視点で作業を行う映像は，視野の中心に注視する可能性の高い物体が集まる．また，自身の手も注視する可能性の高い領域として検出されるため，映像の中心に Saliency が集まっていることがわかる．よって，P_{left}，P_{right} のスコアは P_{center} と比較して大きく上がっている．今回使用したデータセットのような状況では，画面の端に表示することで邪魔をする可能性が低いと言える．Saliency の高い領域が視野の中心付近に集まるのは自然ではあるが，実際の日常生活においても同様であるか，ながら見を行う場合の視線の動きはどのようになるか，さらなる検証が必要である

また，Saliency の高い領域を避けるアルゴリズムであるにもかかわらずスコアが 1 にならない原因は以下の 2 点が考えられる．

1 点目は，本手法は V の中心座標を求めており，その過程で V のサイズがモデルに入っていない点である．Saliency の高い領域が V の一部に重なっているが，中心座標は重なっていないため大きく避けないというケースがあった．完全に V と Saliency の高い領域の重なりを少なくするためには，推定に V のサイズを考慮するモデルに改良する必要があることが分かった．ただし，注視する可能性の高い領域の近くに表示することで，V を見やすくする効果もあるため，パラメータの細かい調整が必要である．

図 4 GTEA に対して VirtualTV が Saliency の高い領域を避けている様子

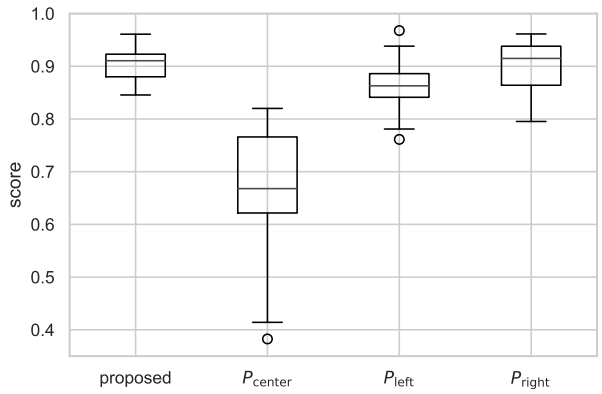

図 5 スコアの比較

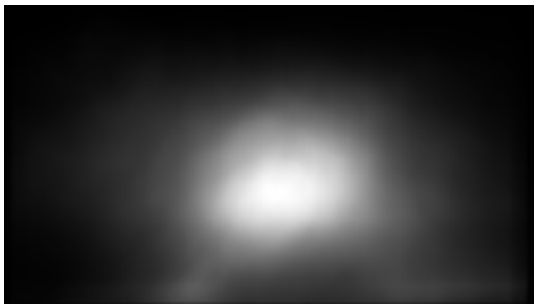

図 6 全データセットにおける Saliency Map の加算平均

2 点目は，$p(X|X_{t-1})$ により，周辺が全て Saliency の高い状態の時，動かなくなってしまう点である．図 7 に，本手法が Saliency の高い領域を避けきれない場合を示す．

図 7 Saliency の高い領域を避けきれない場合

このようなケースでは Saliency の高い領域を避けきることが出来ないため，スコアの低下が発生したと考えられる．今回のスコアを目的とする場合，適切なパラメータ r の設定や，分布の見直しにより更にスコアが上昇する可能性が示唆された．

5. 実装

本手法を Microsoft 社の AR グラス Hololens2 を用いて実装した．実装において，ワークステーションで HoloLens2 に外付けしたウェブカメラの画像から Saliency Map の推定と本手法による座標の最適化を行い，通信により Hololens2 へ VirtualTV の座標を送り表示した．

実装に用いたワークステーションのスペックを表 2 にまとめる．

表 2 実験環境のワークステーションスペック

CPU	AMD Ryzen 9 5900X
GPU	Geforce RTX 3090 24GB
RAM	128GB
OS	Windows 10 Pro 64 ビット

表示速度は 30fps 以上で再生可能であり，VirtualTV がなめらかに動いていることを確認できた．また，視界の極端な端に移動する，視認性の悪い速度で移動するといった挙動も見られず，人物など分かりやすいオブジェクトが目の前に来た際にそれを避ける動きが実現できた．

一方，4 章で議論したように，Saliency の高い領域に囲まれた場合に VirtualTV が避けきれない現象が確認できた．具体的には，複数人で会話をする場合に，VirtualTV が人にかぶさってしまう場合が確認できた．このようなケースにおいて，本手法を用いても VirtualTV が日常生活の邪魔となってしまう可能性がある．状況に合わせて表示／非表示を切り替える，大きな場所の移動を許容するなど改善に努めていきたい．

6. まとめと展望

本稿では，AR グラスで映像を「ながら見」する体験を対象に，映像表示領域の動的な決定法について議論した．AR グラスで映像を視聴する際，視聴に負荷のかからない領域にあること，映像表示領域が時間的に連続的であること，ユーザーが注視する可能性の高い領域と重ならないことの 3 点が求められる．ユーザーが注視する可能性の高い

領域として Saliency Map の推定を行い，以上 3 点により映像表示領域の座標を確率分布として求めた．分布から逐次的に映像表示領域を最適化することでその座標を動的に決定するアルゴリズムを作成し，注視する可能性の高い領域を避けるように表示できることを一人称視点の映像データセットから確認した．また，アルゴリズムを Hololens2 に実装することでデモ環境を作成した．

更に快適性を追求するために，奥行，大きさ，移動速度など様々なパラメータを求める必要がある．将来的に「ながら見」の体験が増える可能性が高く，この体験のさらなる精緻なモデリングを行っていきたい．

参考文献

[1] Consumer ar revenues projected to reach $7.9 billion by 2023 - ar insider. https://arinsider.co/2019/08/08/consumer-ar-revenues-projected-to-reach-7-9-billion-by-2023/. (Accessed on 02/09/2022).
[2] 川喜田裕之, 吉野数馬, 小出大一, 久富健介. 3D 空間における 2D 動画の形態に関する検討. In *HCG2018-I-1-1*. HCG シンポジウム 2018, 2018.
[3] Billboard と tag-along - mixed reality — microsoft docs. https://docs.microsoft.com/ja-jp/windows/mixed-reality/design/billboarding-and-tag-along. (Accessed on 10/20/2021).
[4] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 20, No. 11, pp. 1254–1259, 1998.
[5] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3085–3094, 2019.
[6] Raphael Grasset, Tobias Langlotz, Denis Kalkofen, Markus Tatzgern, and Dieter Schmalstieg. Image-driven view management for augmented reality browsers. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 177–186. IEEE, 2012.
[7] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pp. 3281–3288. IEEE, 2011.