第 26 回日本バーチャルリアリティ学会大会論文集（2021 年 9 月）

360° 画像からの注目部分の予測とその評価

澤邉裕紀 [1)]，池畑諭 [2)]，相澤清晴 [1)]

1) 東京大学 情報理工学系研究科（〒 113-0033 東京都文京区本郷 7-3-1）

2) 国立情報学研究所（〒 101-8430 東京都千代田区一ツ橋 2-1-2）

概要: 360° 画像は閲覧者の視野角よりも広い情報を持つので，視野外の情報の見落としが発生する．そこで本研究では，顕著性マップの予測により，360° 画像における最適な注目部分領域 (Region of Interest, RoI) の集合を予測する手法を提案する．360° 画像ドメインに適したデータ拡張に基づく顕著性予測 CNN を学習し，領域内の顕著性と領域同士の IoU (Interaction-Over-Union) に基づいた最適な RoI 集合を獲得する．提案した手法により得られた注目部分画像をクラウドワーカーによる主観評価によって評価し，提案手法が手動で領域を選択するよりも優れた領域選択が可能である事を示す．

キーワード： 顕著性マップ，360° 画像，RoI, 機械学習

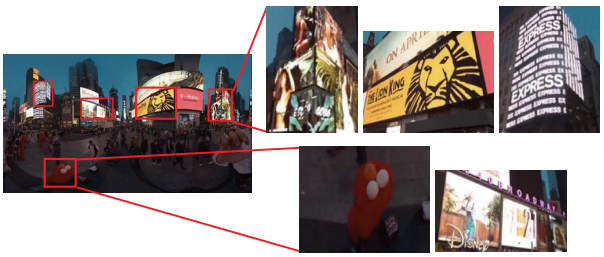

図 1: 本研究で提案する課題．左の 360° 画像から右の注目部分画像を検出し，RoI を抽出する．

1. 序論

360° カメラに撮影される 360° 画像は，通常のカメラの透視投影画像の視野角が normal field-of-view(NFoV) [9] である 65° に限定されるのに対し，全方位の情報を持つという利点がある．しかし，人間の視野では 360 度の視野を同時に全て観測できないため，オブジェクトやイベントを閲覧者が見逃す可能性がある．この問題を解決するため，本研究では，単一の 360° 画像を入力として，360° 画像中の RoI を，視野角 65° 以下の複数の RoI(Region of Interest,RoI) として出力する手法を提案する．各部分領域に対応する視野角を可変にすることによって，図 1 で示されているように各部分領域のサイズやアスペクト比を自在に変化させることが可能となり，過不足なく RoI を提示することが可能となる．本研究の第一の貢献として，単一の 360° 画像から複数の RoI を予測する課題を提案する．単一 360° 動画から単一透視投影動画を検出する Su らの手法 [9] と異なり，単一画像から複数 RoI を抽出する課題はこれまで扱われてこなかったが，VR 映像において重要なイベントの見逃しを防いだり，360° 画像を一般画像でサマライズするには重要な課題である．

提案手法では，360° 画像中の顕著性 [3] が高い領域を RoI とする．顕著性とは人間の視覚的注意の集めやすさを定量的に表したもので，画像認識，物体検出，ロボティクスや広告デザインなど多様な分野で利用されている．顕著性の予測は Itti らの手法 [4] を発端として，近年ではラベル付けされた注視情報を学習する CNN モデルが主流となっている [6]．しかし，既存の顕著性マップ（各画素の顕著性の値で構成される 2 次元画像）予測は主に一般的な透視投影カメラに基づいており，360° 画像に対する手法は限られている [2]．本研究では，天球回転を利用したデータ拡張を提案し，360 度画像に対する顕著性の予測精度を向上させる．本研究においては，360° 画像における顕著性マップの性能評価を Salient360! [5] [8] データセットによって行い，既存手法と比較による精度の向上を示す．顕著性マップを予測したのち．RoI を抽出するために，360° 画像を複数の重複する領域候補に分割し, 部分領域内での顕著性マップの値の和の最大化，各領域の重なりが最小化されるような評価関数（Salient-IoU）を最適化する．また，RoI 抽出については，先述した通り既存研究が存在しないため，Amazon Mechanical Turk [1] 上での実験参加者による評価用データの作成と，作成したデータを用いた主観評価によって，提案手法が手動で領域を選択するよりも優れた領域選択が可能である事を示す．

2. 提案手法

提案手法の概要を図 2 に示した．提案手法は 1 枚の 360° 画像を入力として，(1) 天球回転データ拡張に基づく顕著性予測および（2）Selective Search に基づく RoI 候補の抽出と Salient-IoU に基づく RoI の組み合わせ最適化で構成されている．

2.1 天球回転データ拡張に基づく顕著性予測

360° 画像から顕著性を予測する畳み込みニューラルネットワーク（CNN）には Martin ら [7] のモデルが存在するが，その学習に用いられた Salient360!データセット [5] [8] においては，360° 画像と対応する顕著性マップのペアが 85 組し

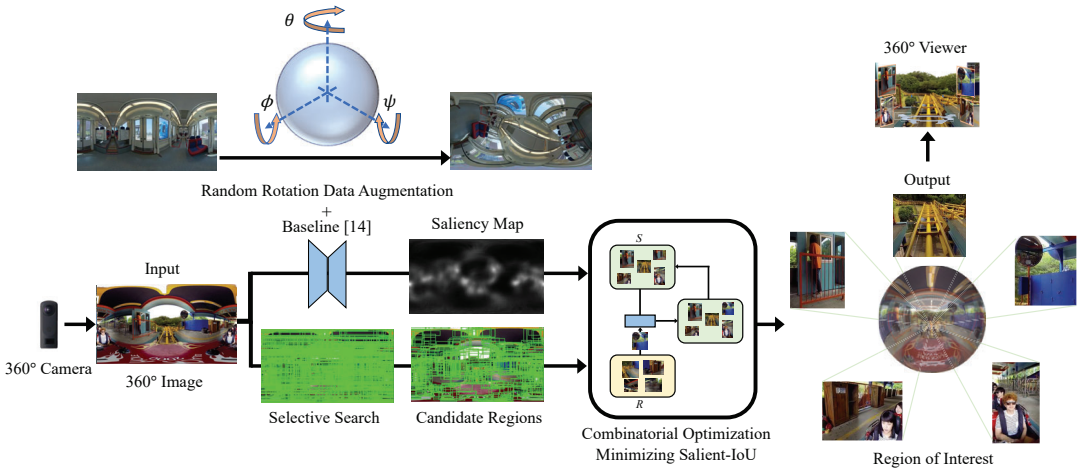

図 2: 提案手法により 360° 画像から注目部分画像を求めるフレームワーク

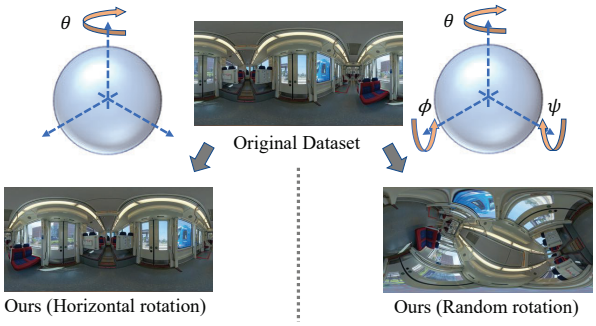

図 3: 天球回転によるデータ拡張

か存在せず，加えて高顕著領域が赤道付近に集中しており，学習されたモデルがそのバイアスを引き継いでしまう問題がある．本研究では学習時の緯度のバイアスを取り除くために，データセット内の ERP(Equirectangular Projection) 投影された 360° 画像を単位球面座標に逆投影し，球面をランダムに回転させて，再び ERP 上に再投影することで，学習データを拡張する手法を提案する．具体的には，図 3 に示したように，ERP 画像上で経度を回転させる軸に関して $\theta \in [-\pi, \pi]$, 赤道を含む面を傾ける軸に関して $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$,2 つの軸に独立した軸に関して $\phi \in [-\pi, \pi]$ の範囲で画像ごとにランダムで回転を与える操作を学習の epoch ごとに行った．評価実験においては，Martin ら [7] のモデルに対して提案手法を適用した結果を示す．

2.2 顕著性と IoU を考慮した RoI の予測

次に，予測された顕著性を 360° 画像中の RoI の判断根拠として，選出された領域候補群から画像中の最も重要な領域を表す n 個（本研究では $n = 5$）の RoI を選択する．領域候補の選出には，物体認識で一般的に利用されている Selective Search [10] を用いる．これは色・テクスチャ・サイズの類似度に基づき，画像を小領域に分割する手法である．選出された各領域候補から顕著性の高い n 個の領域を RoI として抽出するが，単純に対応する顕著性の平均値を判断基準としてしまうと特定の領域に集中してしまったり，重複した領域や小領域が検出されやすくなり，広い FoV を持つ RoI を広範な範囲から獲得する事ができない．一方で，領域内の顕著性の合計値を判断基準とすると，単純に領域候補の中で最も大きな領域から順番に検出してしまう．そのため本研究では，各領域候補の顕著性のみを考慮するのでなく，RoI 間の重なり合いも考慮して，領域集合が全体として最適となるような RoI の集合を求める．具体的には，領域内の顕著性の合計値と，領域同士の IoU を考慮した評価関数を定義し，評価値を最小化する組み合わせを選ぶ問題を解く．

m 個の領域候補から選ばれる任意の n 個の RoI 候補の集合 S が RoI として適切であるかを評価する関数として，次式で表される Salient-IoU (SIoU) を提案する．

$$\mathrm{SIoU} = \frac{a}{n}\sum_{i=1}^{n}\frac{1}{g(I_i)} + (1-a)\frac{1}{{}_nC_2}\sum_{i,j\in[1,n],i\neq j}\mathrm{IoU}(I_i, I_j) \tag{1}$$

ここで，$I_{k\in[1,n]} \in S$ は部分集合に含まれる領域候補であり，$g(I_k)$ は予測された顕著性の領域候補内の合計値である．ネットワークによって予測された顕著性マップは画像中の全ての和が 1 になるように正規化されている．また，単位球上で画素の密度が一定になるように，和は緯度 $\lambda \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ に応じた重み付け $w = \cos\lambda$ による重み付け和とする．$\mathrm{IoU}(I_i, I_j)$ は領域候補同士の重複率を表す IoU (Intersection-Over-Union), a は調整重み ($a \in [0, 1]$) である．a を 1 に近づけると顕著性を重視し，a を 0 に近づけると領域間の重なりを重視する．顕著性を考慮しつつ広範な範囲から部分領域候補を選出するためのパラメータ選択については後述の実験によって検討する．

Salient-IoU に基づき，最適な RoI を推定するアルゴリズムは以下の通りである．まず Selective Search [10] によっ

表 1: 天球観点によるデータ拡張の評価

	AUC_Judd ↑	AUC_Borji ↑	NSS ↑	CC ↑	SIM ↑	KLD ↓
Baseline [7]	0.728	0.693	0.874	0.432	0.545	1.074
Ours w/ random rotation	**0.774**	**0.735**	**1.183**	**0.584**	**0.609**	**0.842**
Ours w/ horizontal rotation	0.772	0.730	1.105	0.553	0.592	0.949

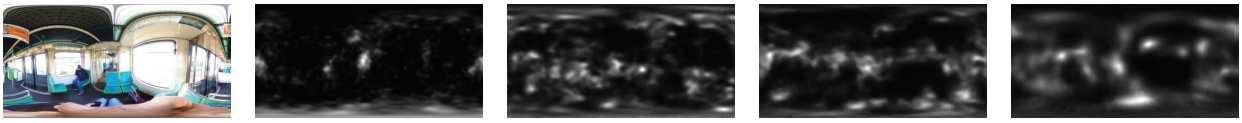

図 4: 左から，テストデータの入力画像，顕著性マップの ground truth，baseline [7] のデータ拡張手法で学習した結果，提案手法のデータ拡張で水平線上で回転させてデータ拡張し学習した結果，提案手法のデータ拡張で 3 自由度で回転させてデータ拡張し学習した結果

て生成された領域候補集合 R から，球面座標系において緯度に対応する FoV，経度に対応する FoV のどちらかが NFoV=65° を超えた要素は局所的な情報を与えるべき RoI として不適切であるとして削除する．その後，RoI 集合 S を領域内の顕著性合計値が高い上位 n 個の領域候補によって初期化し，それらを R から削除する．次に R から顕著性合計値が最も高い要素を取り出し，それを集合 S のいずれかの要素と交換した場合に生成される RoI 集合 S' の評価関数 (式 1) の値が，交換前の評価値よりも小さい場合 S' と S を置き換える．尚，複数の入れ替え候補が存在した場合最も評価値が小さい集合を採用する．以上の取り出し，計算，入れ替えと削除の一連の操作を R の要素がなくなるまで繰り返し行い, 最終的な S を最適な RoI 集合とする．

3. 評価実験

3.1 データセット

顕著性予測ネットワークの訓練データセットとして Salient360! [5] [8] を利用した．Salient360!は，ICME'17 と ICME'18 の Grand Challenge で用いられた 360° 画像に対する顕著性予測の評価データセットであり，2.1 節で述べた通り 85 組の入力・正解画像が含まれている．我々の実験では，これを訓練データ 78 枚とテストデータ 7 枚と分けて評価を行った．RoI 予測の評価については既存のデータセットがないため，評価用データセットを作成した．まず YouTube [11] 上の 5 本の 360° 屋外映像からそれぞれから 45 フレームを抽出した．次に，Amazon Mechanical Turk [1] 上で 14 人のクラウドワーカーに各 360° 画像を 30 秒間ブラウザ上で閲覧してもらい，その後最も印象に残った 5 つの領域をバウンディングボックスによって指定することを指示した．各 360° 画像につき 3 人のクラウドワーカーがこの作業を行い，得られたバウンディングボックスを元に 360° 画像から透視投影画像を 5 枚ずつ抽出し，これを正解の RoI とした．これにより計 135 セットの評価用データセットが作成された．

3.2 顕著性マップ予測における天球回転データ拡張の効果

360° 画像に対する顕著性予測ネットワークを提案した近年の Martin らによる手法 [7] をベースラインとし，天球回転によるデータ拡張なしで学習させた場合 (Baseline)，赤道上で水平方向のみにランダム回転させてデータ拡張した場合 (Baseline w/ horizontal)，3 自由度でランダム回転させてデータ拡張した場合 (Baseline w/ 3DoF) の 3 通りを比較し，天球回転データ拡張の顕著性マップ予測への効果を検証した．アーキテクチャはベースラインから変更せず，学習条件は epoch と batchsize のみを変更した．評価指標としては Salient360!の ICME'17 と ICME'18 の Grand Challenge で用いられた 5 種類に加えて AUC_Borji を加えた 6 種類とした．

評価結果を表 1 に示した．全ての評価指標において天球回転に基づくデータ拡張，その中でも自由度の高い回転が性能向上を確認した．着目すべきは，テストデータには回転が加えられていないにも関わらず，提案手法においてより正確な顕著性を予測した点である．得られた顕著性を可視化した図 4 においても，ベースライン手法では，学習データのセンターバイアスにより，赤道付近の顕著性を高く予測する傾向があり，提案手法は高緯度領域においても頑健に顕著性を予測できている事が確認できる．

3.3 顕著性と IoU を考慮した RoI の検出

本研究で作成した評価用データセットを用いて，提案手法の有効性についてユーザースタディを行った．クラウドワーカーに 360° 画像を 30 秒間見せ，その後 (1) ワーカー選択の RoI，(2) 提案手法で予測された RoI， (3)Selective Search で求めた領域候補よりランダムで 5 つ選んだもの, の 3 種類の画像群を見せ,「もしあなたが 360° 画像を 5 枚の一

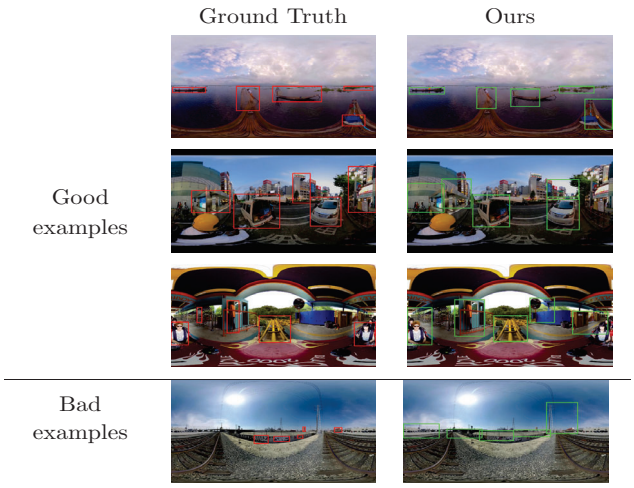

図 5: 予測された RoI と正解の比較

表 2: 各画像群の被選択率

	Random	User annotation	Our results	Tie
a=0	58/405	119/405**	**218**/405**	10/405
a=0.01	48/405	124/405**	**215**/405**	18/405
a=0.03	43/405	110/405**	**236**/405**	16/405
a=0.1	51/405	115/405**	**229**/405**	10/405
a=0.4	48/405	157/405*	**191**/405*	9/405
a=1	68/405	159/405	**164**/405	14/405

*:p_j0.1,**:p_j0.01

般画像にまとめるとしたら，あなたの直感に合致する選択肢はどれか」と質問した．また,「どの選択肢も同じ (tie)」という選択肢も用意した．それぞれの画像セットがどの条件に対応するかはワーカーには知らされず，かつ選択肢の順番も毎回入れ替えた上で，135 セットそれぞれについて 3 人ずつ回答を得た．実験は，Salient-IoU のパラメータを $n = 5$ として固定し，a のパラメータを変えて 6 通り行った．各画像群の被選択率は表 2 の通りである．興味深い事に，人力で作成された正解とされる RoI よりも，提案手法で予測された RoI の方が評価が高いという結果が得られた．また，顕著性をより重視して検出した RoI が，より選ばれやすくなっている事が明らかになった．

最後に，予測された RoI と評価用データセットの正解 RoI の定性的比較を図 5 に示した．多くの場合において予測結果は正解 RoI に近い結果を得たが，顕著性予測に用いたデータセットに含まれない文字がある画像や，物体が少ない画像を入力とすると，評価用データセットと離れた出力が得られる傾向があった．

4. 結論

本研究では，1 枚の 360° 画像から，視野角や位置関係が自由な透視投影画像群である注目部分画像を抽出する課題に取り組み，主観評価を行った．顕著性マップを予測するネットワークの学習では，学習データのセンターバイアスを克服する天球回転によるデータ拡張を提案し，ベースライン [7] に対する性能向上を示した．また，提案手法によって予測された RoI 群を人手でラベル付けした RoI と比較したユーザースタディの結果，提案手法が人間の直感的な結果と同等あるいはそれよりも優れた予測結果を与える事を示す統計的に優位な結果が得られた．

参考文献

[1] Amazon Mechanical Turk: https://www.mturk.com/.

[2] Chao, F.-Y., Zhang, L., Hamidouche, W. and Deforges, O.: Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks, *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, pp. 01–04 (2018).

[3] Desimone, R. and Duncan, J.: Neural mechanisms of selective visual attention, *Annual review of neuroscience*, Vol. 18, No. 1, pp. 193–222 (1995).

[4] Itti, L., Koch, C. and Niebur, E.: 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] J. Gutiérrez, E. David, A. C. M. P. D. S. P. L. C.: Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360 contents.

[6] Kummerer, M., Wallis, T. S. A., Gatys, L. A. and Bethge, M.: Understanding Low- and High-Level Contributions to Fixation Prediction, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).

[7] Martin, D., Serrano, A. and Masia, B.: Panoramic convolutions for 360° single-image saliency prediction, *CVPR Workshop on Computer Vision for Augmented and Virtual Reality* (2020).

[8] Rai, Y., Gutiérrez, J. and Le Callet, P.: A dataset of head and eye movements for 360 degree images, *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 205–210 (2017).

[9] Su, Y.-C., Jayaraman, D. and Grauman, K.: Pano2Vid: Automatic Cinematography for Watching 360° Videos, *Proceedings of the Asian Conference on Computer Vision (ACCV)* (2016).

[10] Uijlings, J. R., Van De Sande, K. E., Gevers, T. and Smeulders, A. W.: Selective search for object recognition, *International journal of computer vision*, Vol. 104, No. 2, pp. 154–171 (2013).

[11] YouTube: https://www.youtube.com/.