Do Deep-Learning Saliency Models Really Model Saliency?

Phutphalla Kong[1,2], Matei Mancas[1], Nimol Thuon[2], Seng Kheang[2], Bernard Gosselin[1]

[1]TCTS Lab, FPMs, University of Mons, Belgium

[2]DICE, Institute of Technology of Cambodia, Cambodia

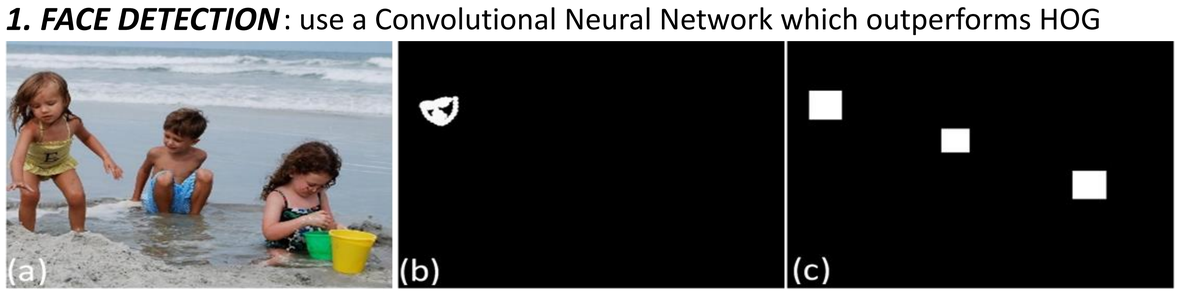
{matei.mancas,bernard.gosselin}@umons.ac.be, {phutphalla,sengk}@itc.edu.kh, thuonnimol1@gmail.com

ABSTRACT

This paper investigates the importance of bottom-up vs. top-down attention. We enrich with top-down info. classical bottom-up models of attention. Then, the results are compared with DNN-based models. Our provocative question is: "do deep-learning saliency models really predict saliency or they simply detect interesting objects?". We found that if DNN saliency models very accurately detect top-down features, they neglect a lot of bottom-up info. which is surprising and rare, thus by definition difficult to learn.

GENERIC TOP-DOWN FRAMEWORK

1. FACE DETECTION: use a Convolutional Neural Network which outperforms HOG

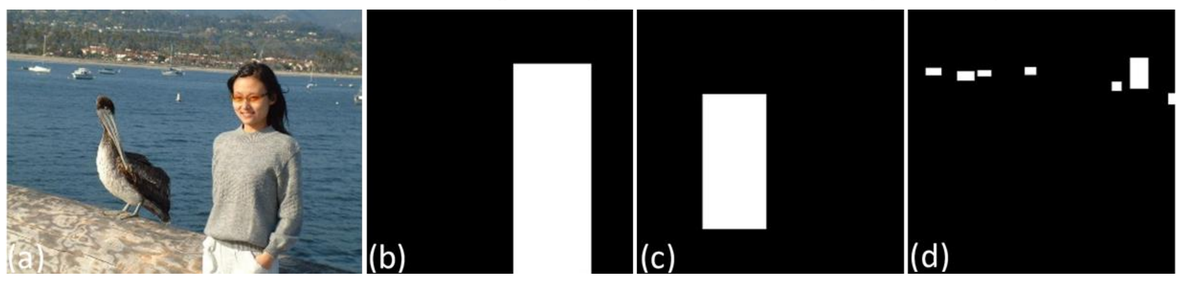


Comparison results between HOG and CNN-based face detectors. (a) Input image, (b) Result of HOG-based face detector, and (c) Result of CNN-based face detector.

2. TEXT DETECTION: use Connectionist Text Proposal Network (detect a text line in seq.)

Result of text detection. (a) Input image, (b) Text detection (green bounding-boxes), and (c) Binary text masks.

3. OBJECT DETECTION: select 3 categories (person, animal, transportation) from YOLO2



Result of object detection. (a) Input image, (b) Person detection, (c) Animal detection, and (d) Transportation detection (here the small boats in the back are detected).

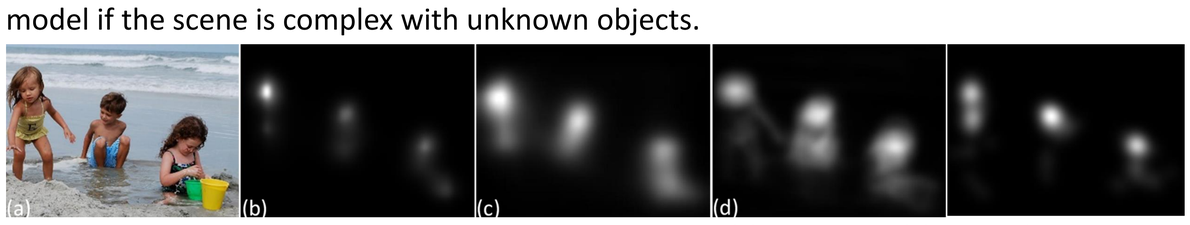
4. CONTEXT-BASED TOP-DOWN INFORMATION

A centered Gaussian function was also added into the image because it plays an important role for natural images. The OSIE and MIT300 datasets contain mainly natural images, so a centered Gaussian function is the best choice.

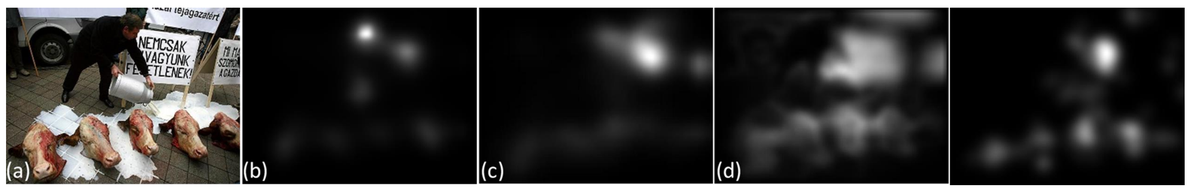
DNN-BASED VS. BOTTOM-UP MODELS

1. QUALITATIVE COMPARISON

DNN-based models such as SAM-ResNet and Salicon provide poorer results than RARE model if the scene is complex with unknown objects.



Result where DNN-based models are better than bottom-up models. (a) Input image, (b) Result of SAM-ResNet, (c) Result of Salicon, (d) Result of our model, and (e) Eye tracking map.



Result where DNN-base models are less good than bottom-up models. (a) Input image, (b) Result of SAM-ResNet, (c) Result of Salicon, (d) Result of our model, and (e) Eye tracking map.

2. QUANTITATIVE COMPARISON

Our experiment shows that on the OSIE dataset RARE bottom-up model alone is better than SAM-ResNet for **5.7%** of the images. RARE augmented with our generic framework is better than SAM-ResNet on **14.3%** of the images. According to MIT300 benchmark, our model has the best results compared to all bottom-up models. It is still surpassed by some DNN-based models, but a lot of those models are now less good than ours.

CONCLUSION

- ✓ Understand differences in visual attention computation between classical bottom-up saliency models and DNN-based saliency models
- ✓ Relative importance of bottom-up and top-down information
- ✓ Mixing a bottom-up model (our naïve top-down info. framework) → the best results among all bottom-up models on MIT300 saliency (esp. KLD)
- ✓ DNN-based models results cannot be explained (seem to neglect BU info.)
- ✓ **Future work**: how a DNN model can be mixed with bottom-up models

2018 IEEE International Conference on Image Processing (October 7-10 – Athens, Greece)

TOP-DOWN VS. BOTTOM-UP INFLUENCE

To evaluate our result, we use the Correlation Coefficient (CC), Kullback-Leibler Divergence (KLD), Normalized Scanpath Saliency (NSS), Similarity (SIM), and Judd Area Under the ROC curve (AUCJ). The smallest values represent the best results in KLD metric. For the other metrics, higher values are the best.

Table 1. Results using RARE model (OSIE dataset) on the number of images (on a total of 700) where at least an object is detected. The result with bold-fonts represents the best result in comparison.

Maps (images)	Metrics				
	CC	KLD	NSS	SIM	AUCJ
SM (279)	0.4179	1.1548	1.4118	0.4115	0.8291
F (279)	**0.5631**	**0.939**	**1.8914**	**0.5165**	**0.8525**
SM (425)	0.4637	1.0492	1.4626	0.4390	0.8311
TX (425)	**0.5478**	**0.9011**	**1.7870**	**0.4995**	**0.8544**
SM (138)	0.4754	1.1183	1.7178	0.4202	0.8516
Ani (138)	**0.5111**	**1.0425**	**1.8565**	**0.4716**	**0.8629**
SM (484)	0.4587	1.0971	1.5700	0.4262	0.8412
Per (484)	**0.4699**	**1.0594**	**1.6185**	**0.4626**	**0.8433**
SM (98)	**0.5152**	**0.9998**	**1.8336**	0.4471	**0.8636**
Tra (98)	0.4902	1.0135	1.7608	**0.4748**	0.8579
SM (all)	0.4683	1.0597	1.5364	0.4364	0.8365
CG (all)	**0.5001**	**0.9738**	**1.6231**	**0.4679**	**0.8472**

Table 2. Correlation result using several models (OSIE dataset)

Model		Metrics				
		CC	KLD	NSS	SIM	AUCJ
AIM	SM	0.3251	1.5241	1.0717	0.3454	0.7733
	FAPTTX	0.5392	1.1186	1.7311	0.4070	0.8496
AWS	SM	0.4583	1.1171	1.4855	0.4268	0.8219
	FAPTTX	0.6161	0.8313	2.0290	0.4995	0.8708
GBVS	SM	0.4380	1.0880	1.3496	0.425	0.8159
	FAPTTX	0.5608	0.9379	1.8104	0.4828	0.8488
RARE	SM	0.4683	1.0597	1.5364	0.4364	0.8365
	FAPTTX	**0.6235**	**0.8162**	**2.0868**	**0.5192**	**0.8719**

MIXING TOP-DOWN AND BOTTOM-UP INFORMATION

CSM = $(a*SM*CG^b) + (1-a)*SM$ SM: bottom-up saliency maps; a, b are 2 para. (a=0.75 and b=4)

CTSM = $(Tra*\text{CSM}) + \text{CSM}$ CG is the centered Gaussian image; *Tra* is smoothed masks of transportation

CASM = $(Ani*\text{CSM}) + \text{CSM}$ *Ani* is smoothed masks of animal

CPSM = $(Per*\text{CSM}) + \text{CSM}$ *Per* is smoothed masks of person

COSM = (CTSM + CASM + CPSM)/3

FAPTTX = $(\text{COSM} + F + w*T)/3$ *F,T* is smoothed masks of face and text detection; *w* is weight (=0.6)

RESULTS

Table 3. Comparing result between bottom-up models and ours

Model	Metrics				
	CC	KLD	NSS	SIM	AUCJ
Ours	**0.6166**	**0.7179**	**1.6762**	**0.5472**	**0.8388**
BMS	0.55	0.81	1.41	0.51	0.83
OS	0.54	0.84	1.41	0.51	0.82
GBVS	0.48	0.87	1.24	0.48	0.81

Table 4. Comparing result between DNN-based models and ours

Model	Metrics				
	CC	KLD	NSS	SIM	AUCJ
DSCLRCN	0.8	0.95	2.35	0.68	0.87
SALICON	0.74	0.54	2.12	0.6	0.87
SAM-Rest	0.78	1.27	2.34	0.68	0.87
Ours	**0.6166**	**0.7179**	**1.6762**	**0.5472**	**0.8388**
SalNet	0.58	0.81	1.51	0.52	0.83
eDN	0.45	1.14	1.14	0.41	0.82
GoogLeNet	0.49	0.99	1.26	0.45	0.81
JuntingNet	0.54	0.96	1.43	0.46	0.80

REFERENCES

[1] Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013), "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," Signal Processing: Image Communication, 28(6), 642-658.

[2] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," arXiv preprint arXiv:1612.08242, 2016.

[3] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting Text in Natural Image with Connectionist Text Proposal Network," in ECCV, 2016.

[4] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model," arXiv preprint, arXiv:1611.09571v3, 2017.