Saliency Detection Guided by Semantic Knowledges

Insert Subtitle Here

Cui Zhe†  
 School of Computer Science and Technology  
University of Chinese Academy of Sciences  
 Beijing, China  
 cuizhe18@mils.ucas. ac.cn

Su Li  
 School of Computer Science and Technology  
 University of Chinese Academy of Sciences  
 Beijing, China  
 suli@usas.ac.cn

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

ABSTRACT

Salient object detection has achieved great success thanks to the use of fully convolutional neural networks. Existing salient object detection networks rely only on visual features while ignore high-level semantic information. In order to detect salient object in a scene, networks need to learn what the scene is tell. To alleviate these issues, we propose to leverage to generate a caption of the scene with image caption network to help network understand the context of a scene. With the guidance of the caption’s description, salient object detection network can better capture the main context of the image and location the salient object.

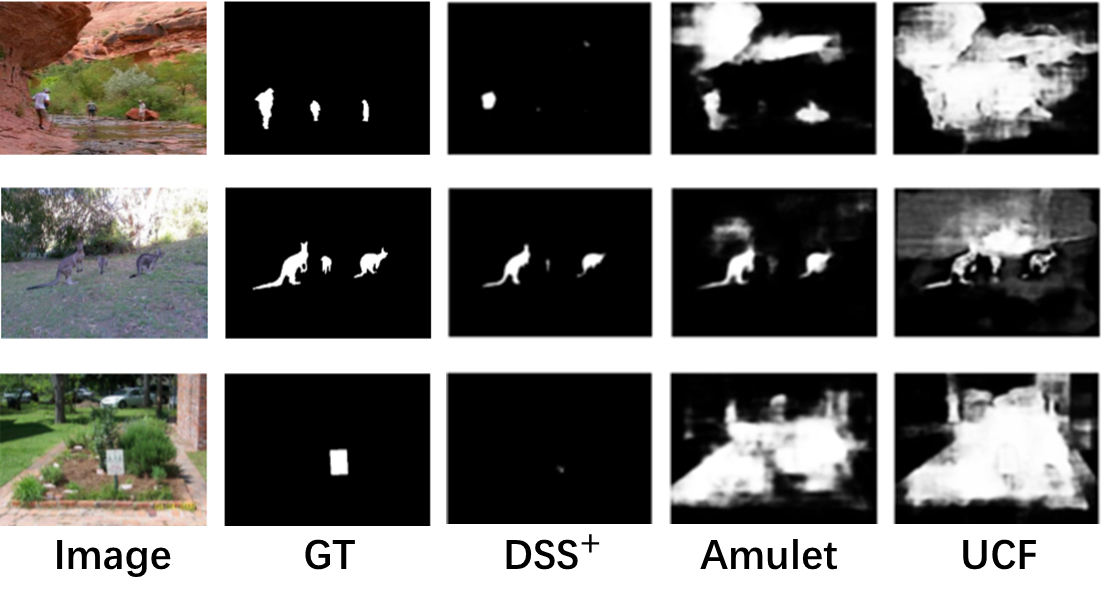
KEYWORDS

Salient detection, Image caption, Knowledge

1 Introduction

One reason people can focus attention on salient region when look at an image is the visual attention mechanism []. Inspired by the attention mechanism of people, salient object detection aims to find the most visual distinctive objects in an image. As an important part in image process, salient object detection has wide applications in many other visual tasks, such as semantic segmentation [], image classification [], and image caption []. Pick up the salient area in an image to process can reduce workload of this task.

CNN can extract multi-scale features, and many of CNN-based models [15, 36, 39, 7, 48, 11] have been proposed in the past decades. Most recent research efforts are devoted to study how to integrate this multi-scale features and good performance has been achieved. However, there is still a major drawback of this work as shown in Figure 1: the entire salient objects can hardly be detected in the clutter scene. The reason for the fail is that the visual cues contract of the salient object and the background is not significant. Salient contrast is not only manifested in the differences of visual cues, but also involves high-level cognition and understanding. High-level description of an image is important to aid the network to gain a global overview on an image.

**Figure 1: Examples of some saliency maps generated by state-of-the-art method.**

Existing salient object detection Lack high-level semantic information. To address the above-mentioned issues, we present a new network which simultaneously utilize two tasks, including salient object detection and image caption. Using image description generated by image caption to guide the salient detection.

2 Related Work

The original work of salient object detection can trace back to the work of Liu et al. [29] and Achanta et al. [30]. Traditional salient object detection method [] rely on hand craft features to predict salient score, such as center prior [], background prior [], contrast prior [], spectrum based []. The hand craft features are low-level and cannot capture the high-level semantic features.

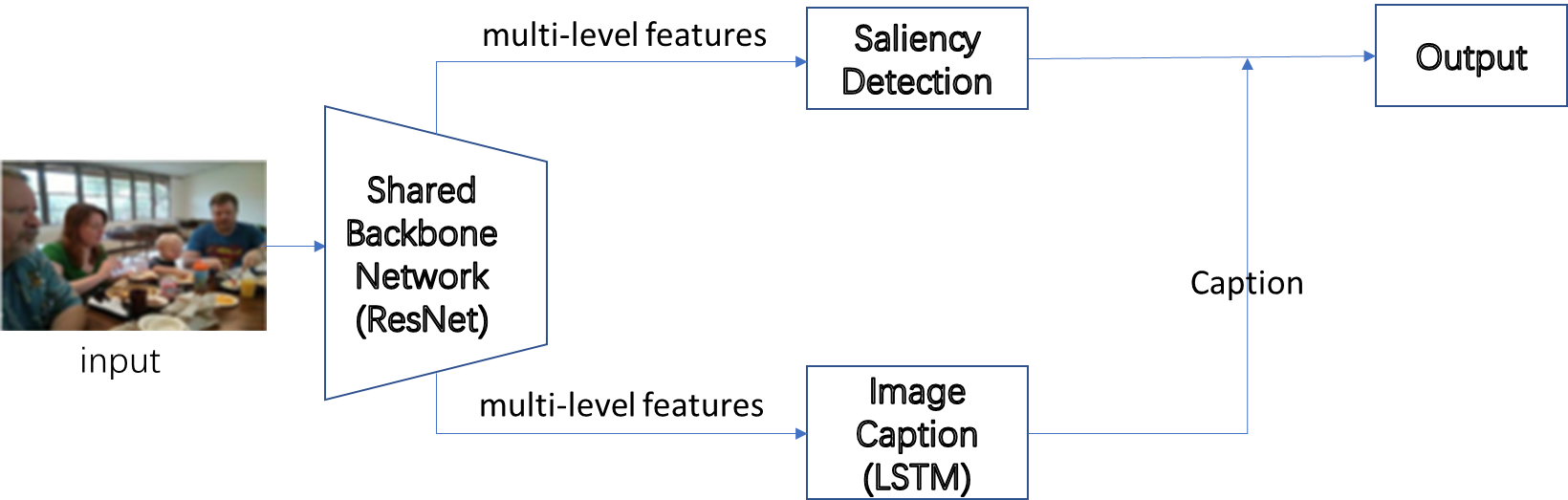
Thanks to the renaissance of the deep convolutional neural network (CNN), which has powerful ability to extract high-level features. The salient object detection began to use CNN to extract features and get great progress. Zhao et al. [] uses two pathways to extract local and global context, which are fed into an CNN classifier for salient classiﬁcation. This method fed each processing unit of an image into classiﬁer for saliency score prediction and has obvious disadvantages: time consuming, unable to use overall spatial information, and all pixels of each patch share the same salience score.

CNN will gradually reduce the size of the original image and is designed for classification task instead of pixel-wise task. Inspired by the great success of fully convolutional neural network (FCN) in semantic segmentation, salient object detection area has turned attention to FCN. FCN can generate salient map of the same size as the original image in an end-to-end way. Wang et al. [] use the salient priors to make the training of depth model easier and the prediction more accurate. Compared with the feedforward neural network, the output of the full convolution neural network in each step is provided as a feedback signal, so that the model can redetermine the salient prediction by correcting its previous errors until the final prediction is generated in the last time step. Hou et al. [] proposed a new method to fuse the lower-level features and the higher-level features by adding several short connections from deeper side-outputs to shallower ones based on the HED []. Chen et al. [] introduced residual learning to compensate the deviation between the prediction and the ground truth. In their work, reverse attention is proposed to focus on context information and get more detailed segmentation results.

In order to describe an image, image caption needs to selectively focus on regions of interest and salient object detection has shown usefulness in image caption. There are some works leverage salient object detection to aid image caption. Chen et al. [] propose a boosted attention method that incorporates stimulus-based human attention with existing top-down visual attention to guide image caption. Tavakoli et al. [] propose a saliency-boosted image captioning model in order to investigate benefits from low-level cues in language models.

3 Proposed Method

In this paper, we propose a knowledge guidance neural network for salient object detection. The network composes of two sub-networks: image caption network and salient object detection network, as shown in figure 2. The two sub-networks are used to extract visual features and semantic knowledges respectively. Our methods embed visual features and its semantic knowledges in a same space to detect salient object.

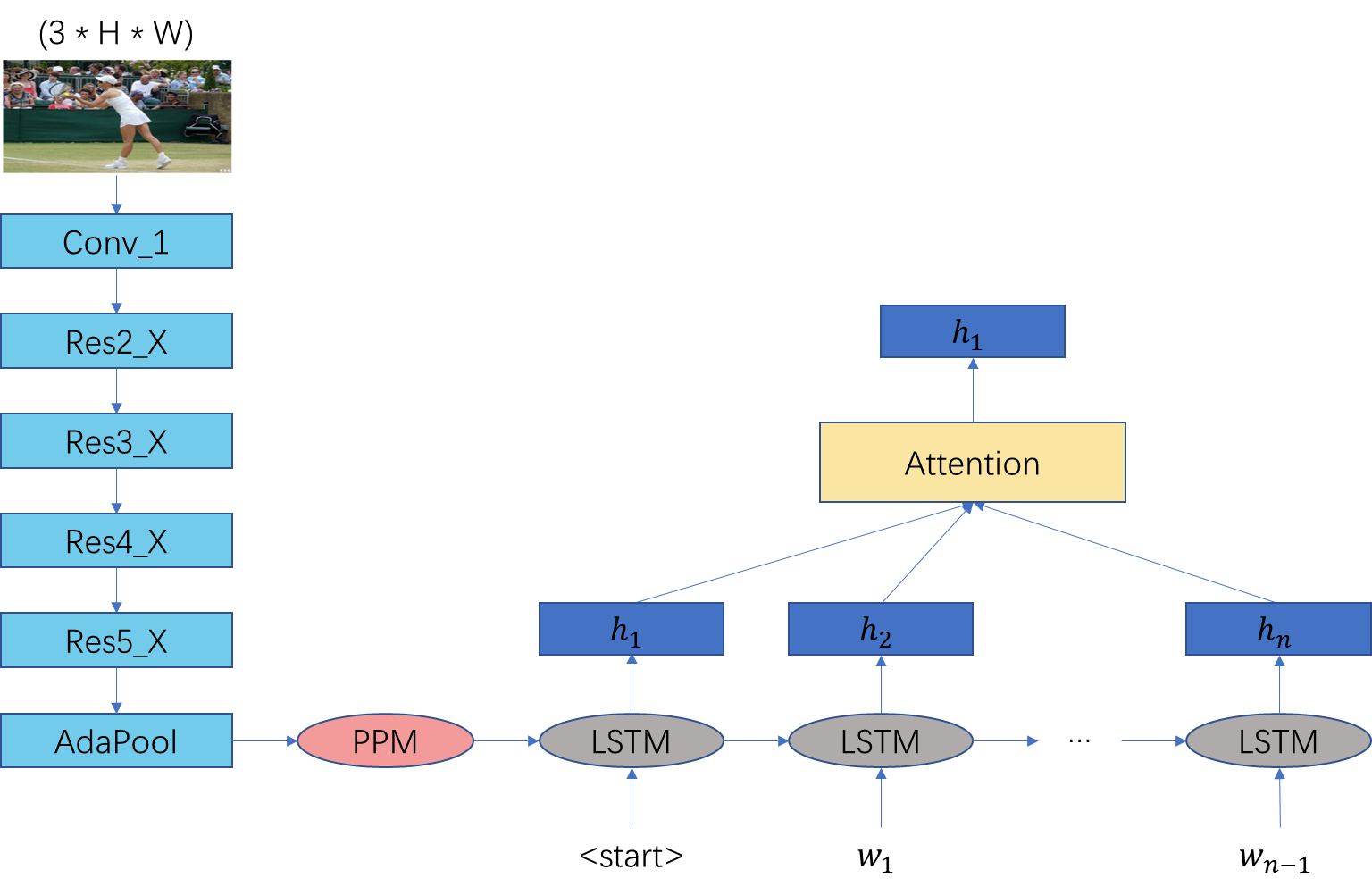


**Figure 2:** **The overall architecture of our proposed network. There are two sub-modules (Image Caption Network and Salient Object Detection Network) based on the shared backbone network** **ResNet-50.**

Our proposed network used ResNet-50 [] as a common feature extractor to extract multi-scale features and denote them as. The feature map extracted from Conv\_1 has the largest special size and extracted from Res\_5 has the smallest special size. We remove the last global pooling and fully connected layers, and use only the ﬁve residual blocks.

3.1 Image Caption Network

We use the classical image caption method NIC [] to generate the caption of given image. The overall process flow of our image caption module is shown in figure 3. We take the last residual blocks’ encoded feature maps as input of caption network to generate word step by step. In order to get fixed size feature map, an adaptive pool layer is added on the top of the ResNet-50. Consider that we only need the high-level semantic features and the generated specific word is not what we need. We only use the hidden state generated in each time step instead of the word.



**Figure 3: Illustration of our Image Caption Network. The extracted feature is sent to the LSTM cell to generate word step by step.**

Figure 4 shows the caption of some images. As we can see, not every word has a corresponding target in the image. Consider that not every word is equally contributed to the salient object, we introduce a filter machine giving each word a salient score to filter the non-salient word.



**Figure 4: Examples of some images’ Caption []. Red word contributed more to the salient object detection.**

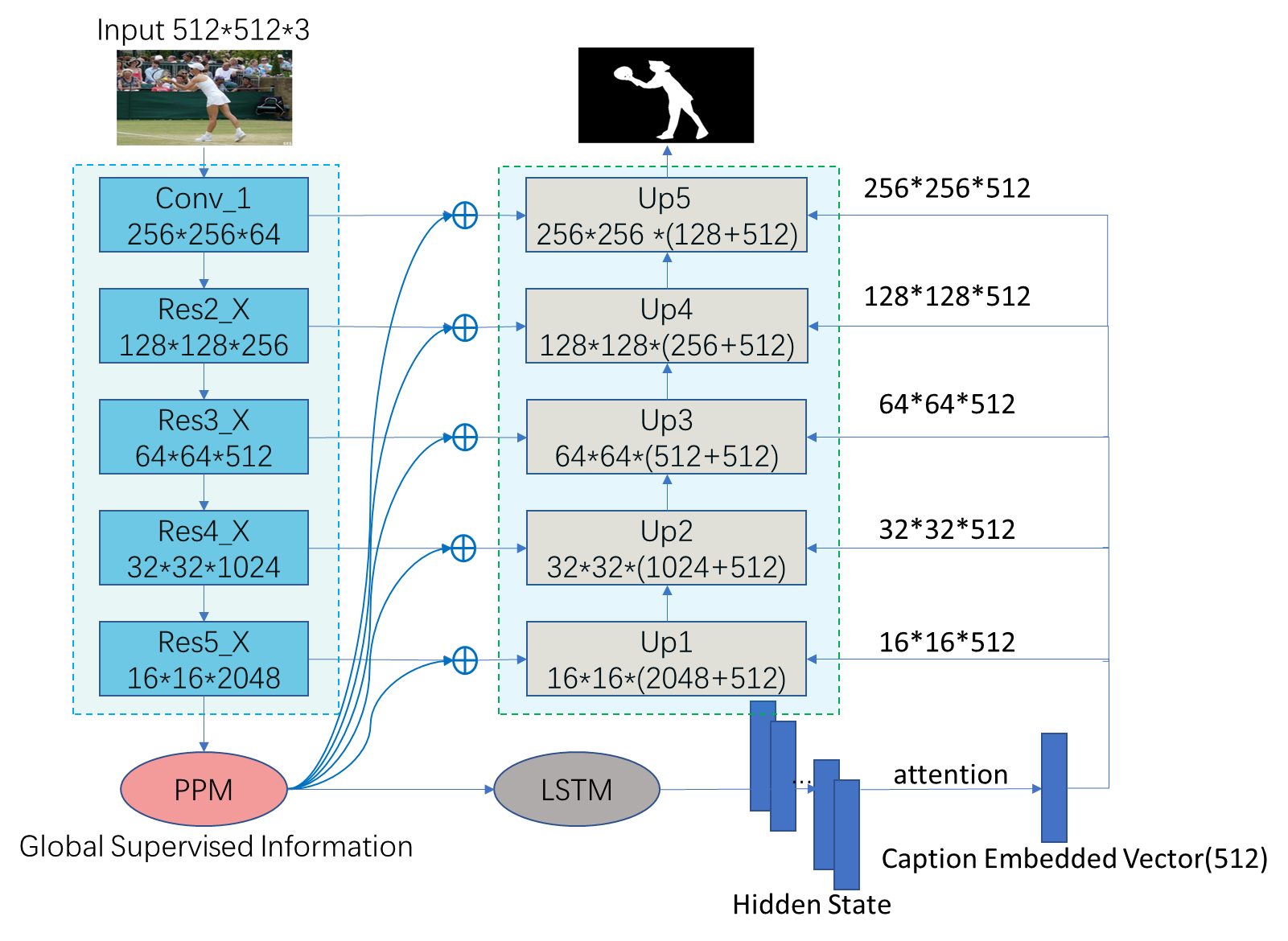
3.2 Salient Object Detection Network

In the salient object detection network, we combined semantic knowledges extracted by image caption network with U-shape models to guide the feature integration process. CNN is capable of extracting rich representation including low level visual features and high-level discriminative information. The base CNN network of our salient Object Detection is the ResNet-50 [].

In addition, we add an Pyramid Pooling Module (PPM) [] to get different receptive field of the image. PPM is first proposed in semantic segmentation task. We utilize PPM by using three pool channels with different kernel sizes (1, 2, 4, 8) followed by a up sample layer to resume to the initial size. Feature maps of different sizes contain global information of different scales. Feature bins of 1 \* 1 contains the roughest global information Then we concatenate these pool channels maps to form global information features of different scales. The PPM module can be formulated as (1), where Up represents the bilinear interpolate upsample.

Bottom up的每一层都融合全局特征和语义知识。

(1)



**Figure 5:** **The pipeline of the proposed method.**

temp

4 Experiments

4.1 Implementation Details

Following most salient object detection works [33,49,52], we use the DUTS dataset to train the proposed model. The backbone of our network is ResNet-50 pretrained on ImageNet and the image caption network is NIC pretrained on MSCOCO 2014. Our model is implemented in PyTorch. We use Adam optimizer to optimize the parameter and the hyper-parameters are set as followed: learning rate = 5e-5, weight decay = 0.0005.

4.2 Datasets and Evaluation Metric

We evaluated our proposed network on six widely used benchmark datasets: DUTS [46], ECSSD [56], PASCAL-S [30], HKU-IS [27], DUT-OMRON [57], SOD [36], SED1 [], SED2 [].

**DUTS** contains 10553 images for training and 5019 images for testing. It is the largest salient object detection benchmark.

**ECSSD** contains 1000 semantically meaningful and structure complex images.

**PASCAL-S** contains 850 images built on the PASCAL VOC 2010 segmentation dataset.

**HKU-IS** contains 4447 images with high quality annotations. Images in this dataset have multiple disconnected salient objects or objects touching the image boundary.

**DUT-OMRON** contains 5168 high-quality images which is difficult and challenging. Images in this dataset contain one or more salient objects with a relatively complex background.

**SOD** contains 300 images which is one of the most challenging datasets currently.

**SED1** has 100 images containing only one salient object an image.

**SED2** has 100 images containing two salient object an image.

In order to compare with other methods, we adopt three widely used metrics, F-measure, mean absolute error (MAE) and S-measure [].

**F-measure** is a balanced mean of average precision and average recall which can be calculated by (2), where is set to 0.3 to weigh precision more than recall as in []. In computing F-measure, we normalize the predicted saliency maps into the range of [0, 255] and binarize the saliency maps with a threshold sliding from 0 to 255 to compare the binary maps with ground-truth maps. At each threshold, Precision and Recall can be computed.

(2)

**MAE** is a metric used to evaluate the average difference between prediction map and the ground-truth map. Let P denote the prediction map and Y denote the ground-truth map, MAP can be calculated by (3).

(3)

**S-measure** is proposed by Fan et al. [8], and it can be used to evaluate non-binary foreground maps. This measurement simultaneously evaluates region-aware and object-aware structural similarity between the saliency map and ground truth.

5 Conclusion

In this paper we propose to use image caption to guide salient object detection network to find salient object.

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: <https://doi.org/>10.1145/1188913.1188915.

[2] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226-236. DOI:https://doi.org/10.1145/567752.567774

[3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..

[29] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.

[30] R.Achanta,S.Hemami,F.Estrada,andS.Susstrunk,“Frequencytuned salient region detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1597–1604.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00