Salient Object Subitizing

Jianming Zhang[1], Shugao Ma[1], Mehrnoosh Sameki[1], Stan Sclaroff[1], Margrit Betke[1], Zhe Lin[2], Xiaohui Shen[2], Brian Price[2], Radomír Měch[2]
[1]Department of Computer Science, Boston University. [2]Adobe Research.

Figure 1: How many salient objects are in each image?

How quickly can you tell the number of **salient** objects in each image in Fig. 1? It was found over a century ago that people are equipped with a remarkable capacity to effortlessly and consistently identify 1, 2, 3 or 4 items by a simple glance [2]. This phenomenon, later coined by Kaufman, et al. as *Subitizing* [3], has been observed under various measurements. It is shown that apprehension of small numbers up to three or four is highly accurate, quick and confident, while beyond this subitizing range, the feeling is lost. Accumulating evidence also shows that infants and even certain species of animals can differentiate between small numbers of items within the subitizing range, suggesting that subitizing may be an inborn numeric capacity of humans and animals. It is speculated that subitizing is a preattentive and parallel process, and that it can help humans and animals make prompt decisions in basic tasks like navigation, searching and choice making.

In this paper, we propose a subitizing-like approach to estimate the number (0, 1, 2, 3 and 4+) of salient objects in a scene, without resorting to any object localization process. Solving this *Salient Object Subitizing* (SOS) problem can benefit many computer vision tasks and applications.

Knowing the existence and the number of salient objects without the expensive detection process (*e.g.*, sliding window detection) can enable a machine vision system to select different processing pipelines at an early stage, making it more intelligent and reducing computational cost. For example, SOS can help a computer vision system suppress the object detection process, until the existence of salient objects is detected, and it can also provide cues for selecting among search strategies and early stopping criteria based on the predicted number. Differentiating between scenes with zero, a single and multiple salient objects can also facilitate applications like robot vision, egocentric video summarization, snap point prediction, iconic image detection and image thumbnailing, *etc*.

To study the SOS problem, we provide a new image dataset of about 7000 images, where the number of salient objects in each image has been annotated by Amazon Mechanical Turk (AMT) workers. The dataset is available on our project website: http://www.cs.bu.edu/groups/ivc/Subitizing/. In Fig. 2, we show some sample images in the proposed SOS dataset with the collected ground-truth labels. Although there are no bounding box annotations accompanying the numbers, it is usually pretty straightforward to see which objects these numbers refer to. The annotations from the AMT workers are further analyzed in a more controlled offline setting, which shows a high inter-subject consistency in subitizing salient objects.

Our ultimate goal is to develop a fast and accurate computational method to estimate the number of salient objects in natural images. The trivial counting-by-detection approach is quite challenging in this scenario, due to cluttered background, occlusion, and large appearance, position and scale variations of the objects in everyday images. Instead, inspired by the psychological observation that the subitizing is likely to be accomplished by recognizing holistic patterns [1], the proposed SOS method bypasses the challenging salient object localization process by using global features.

An implementation of our SOS method using an end-to-end Convolutional Neural Network (CNN) classifier attains 94% accuracy in detecting the existence of salient objects, and 42-82% accuracy (chance is 20%) in predicting the number of salient objects (1, 2, 3, and 4+) on our dataset (see Fig. 3), without resorting to any intermediate saliency map computation or salient object detection. These results are quite encouraging, considering that our CNN-based SOS method is capable of processing an image in a couple of milliseconds.

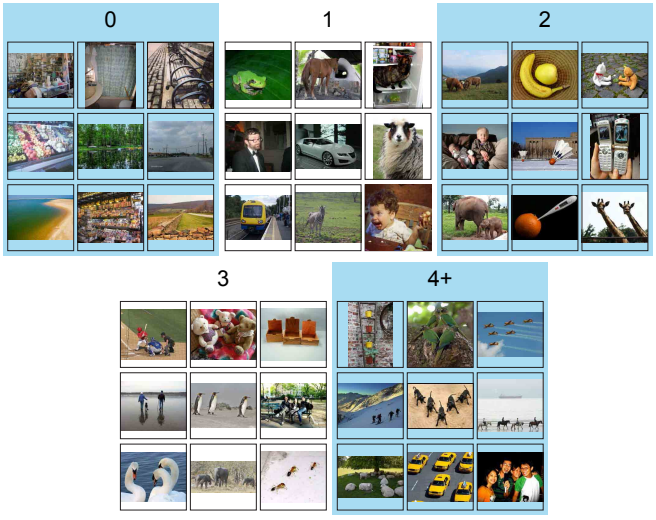

Figure 2: Sample images of the proposed SOS dataset. These images cover a wide range of content and object categories.

	0	1	2	3	4+
0	94% (319)	4% (15)	0	0% (1)	1% (3)
1	7% (41)	82% (507)	8% (50)	2% (11)	1% (8)
2	3% (7)	31% (68)	42% (91)	18% (40)	6% (13)
3	6% (8)	17% (23)	17% (23)	45% (61)	16% (22)
4+	6% (4)	3% (2)	10% (7)	14% (10)	67% (46)

Figure 3: Confusion matrix of our method using the fine-tuned CNN model. Each row corresponds to a ground-truth category. The percentage reported in each cell is the proportion of images of category A (row number) labeled as category B (column number).

We demonstrate applications of the SOS technique in guiding salient object detection and object proposal generation, resulting in state-of-the-art performance. In the task of salient object detection, we demonstrate that SOS can help improve accuracy by identifying images that contain no salient object. In the task of object proposal generation, we present a simple content-aware proposal allocation approach using SOS, and show consistent improvement over state-of-the-art.

[1] Brenda RJ Jansen, Abe D Hofman, Marthe Straatemeier, Bianca MCW Bers, Maartje EJ Raijmakers, and Han LJ Maas. The role of pattern recognition in children's exact enumeration of small numbers. *British Journal of Developmental Psychology*, 32(2):178–194, 2014.

[2] W Stanley Jevons. The power of numerical discrimination. *Nature*, 3: 367–367, 1871.

[3] EL Kaufman, MW Lord, TW Reese, and J Volkmann. The discrimination of visual number. *The American Journal of Psychology*, pages 498–525, 1949.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.