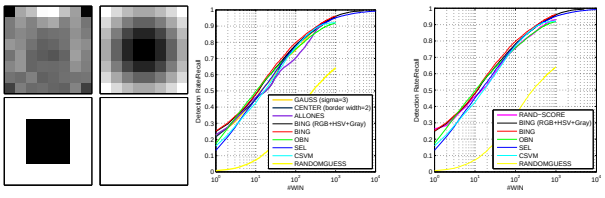
Cracking BING and Beyond

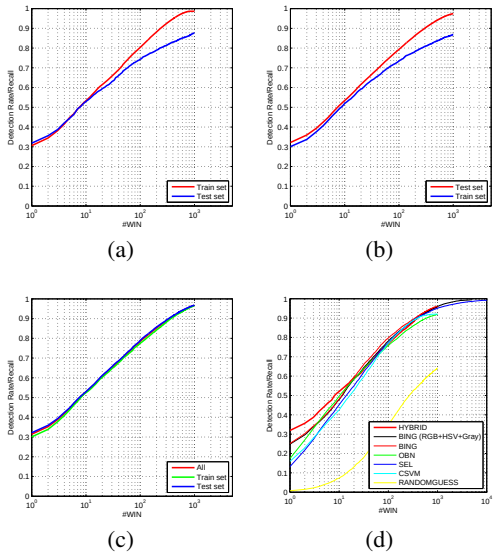
Qiyang Zhao
zhaoqy@buaa.edu.cn
Zhibin Liu
liuzhibin@nlsde.buaa.edu.cn
Baolin Yin
yin@nlsde.buaa.edu.cn

State Key Laboratory of
Software Development Environment,
Beihang University



(a) (b)

Figure 1: Experiments on templates in BING: (a) four learned/hand-tuned templates and their performances. (b) performance of RAND-SCORE.



(a) (b)

(c) (d)

Figure 2: Performance of greedy scheme and hybrid scheme: (a) cover set of training images, and its performance on test images; (b) cover set of test images, and its performance on training images; (c) cover set of all images, and its performance on two sets respectively; (d) comparison of hybrid scheme with other methods.

The problem, *generic objectness proposal*, aims to reduce the candidate windows for object detection tasks. The popular evaluation criterion for related methods is detection-rate/windows-amount(*DR-#WIN*), where *DR* is the percentage of groundtruth objects covered by proposal windows. An object is considered "covered" by a window only if the strict PASCAL-overall criterion [3] is satisfied (the intersection of a proposal window and the object rectangle is not smaller than half of their union, so we call it "0.5-criterion" for short). Under the *DR-#WIN* evaluation framework, BING [2] in CVPR2014, obtains the best performance on the VOC2007 test set. It recalls 96.2% objects with only 1,000 proposal windows. The more surprising is the method is totally a realtime one.

The authors of BING suggest that, after being resized to a fixed size (8×8), almost all annotated rectangle regions share a common characteristics in gradients [2]. This commonness is captured by a template W learned from training images with a linear SVM. Besides this, the subtle differences between diverse width/height configurations are captured in a re-weighting model. Therefore BING consists of two stages: calculating W in stage I, and learning the re-weighting model in stage II. Furthermore, BING uses smart bitwise operations to calculate the inner product of W and candidate windows, so to improve the efficiency.

We designed several templates by hands to substitute W, to verify whether templates play a key role in BING. These templates become less correlated to W in turn, but their performances on VOC 2007 test set are very close, see Fig.1.a. Next we discarded any templates and directly assigned the scores of stage I with uniformly random values (we call this method RAND-SCORE). Surprisingly, the performance of RAND-SCORE is even very close to BING, as shown in Fig.1.b. It is clear that these templates do not have as strong significance as suggested in [2]. Then what on earth makes BING performing so well?

To get the deep insight, we finished a theoretical analysis from the view of combinatorial geometry. We try to construct a small set of windows to "cover" all legal rectangles (we call it a *full cover set*). This is an atypical covering problem in combinatorial geometry [1]. We proposed four lemmas to solve it in the full paper. In conclusion, for an image of the width M and height N, we can use $s(i,j)$ windows of the width $2^i \cdot \sqrt{2}$ and height $2^j \cdot \sqrt{2}$ to cover all $2^i \le w \le 2^{i+1}, 2^j \le h \le 2^{j+1}$ rectangle regions, where $s(i,j) = \lceil \frac{M-2^i}{\left(1-\frac{\sqrt{2}}{2}\right)\cdot 2^i \cdot \sqrt{2}} \rceil \cdot \lceil \frac{N-2^j}{\left(1-\frac{\sqrt{2}}{2}\right)\cdot 2^j \cdot \sqrt{2}} \rceil$. Suppose the image size is $M = 2^m, N = 2^n$, and the object rectangles' widths and heights start from 2^k, then the amount of all windows in our *full cover set* is

$$\sum_{i=k}^{m-1}\sum_{j=k}^{n-1} s(i,j) = \sum_{i=1}^{m-k}\sum_{j=1}^{n-k} \lceil \frac{2^i-1}{\sqrt{2}-1} \rceil \cdot \lceil \frac{2^j-1}{\sqrt{2}-1} \rceil = O(2^{-2k} \cdot MN) \qquad (1)$$

Particularly, when the widths/heights of all object rectangles are at least 16, the amount is 19,600. While on the restriction of 32, we need only 4,225 windows. These amounts are far less than what people imagined before. We call it the Achilles' heel of the *DR-#WIN* evaluation framework. Recall that in BING, the widths/heights of proposal windows are doubled each time, in the same way as in Lemma 3.1-3.4. Its non-max suppression step, 0.25 relative to the normalized size 8, is very close to the step $(1-\frac{\sqrt{2}}{2}) \approx 0.29$ in Lemma 3.2. These two settings meet our analysis well and bring the success to BING.

In real applications, we should pay more attention to those "hot" locations/sizes instead of all possibilities. We designed a greedy scheme to pick the "hottest" window in each round to construct an identical cover set for all images. We also proposed a hybrid scheme to address the huge difference between low-probability sample spaces of the training and test sets. With the increase of the number of proposal windows, we replace the windows in the greedy set with those of RAND-SCORE with increasing probabilities.

In our experiments, our greedy scheme performs considerably well: all *full cover sets* are reduced to about 1,000 windows, and the first windows have 0.3+ *DR*'s in all experiments. In most time, the *DR*'s of our hybrid scheme are higher than OBN and CSVM, and close to SEL and BING. It recalls 95.68% objects with 1000 proposal windows. Especially, its *DR*'s are 13.99% $\sim$ 40.29% (relatively) higher than all other methods in average on the first ten windows. At last, the time consumptions are all nearly zero because the major computations are to resize proposal windows for specific images.

To sum up, what can we benefit from the two schemes for object detection researches? We argue it needs a bigger picture to answer this question because it depends on whether the 0.5-criterion is effective and objective. If the 0.5-criterion is still adopted in future, the baseline should be RAND-SCORE or our hybrid scheme instead of random guesses. Both of them bring more challenges to future researches.

[1] Pach J. and Agarwal P. *Combinatorial Geometry*. John wiley & Sons, ISBN: 9780471588900, 1995.

[2] Cheng M. M., Zhang Z. M., Lin W. Y., and Torr P. Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proc. CVPR*, 2014.

[3] Everingham M., Van Gool L., Williams C. K. I., Winn J., and Zisserman A. The pascal visual object classes (voc) challenge. *IJCV*, 88(2): 303-338, 2010.