Object proposals: the discussion of future work

Yang Yi
Outline

• Previous work

• Future work
  • CVPR 2010: What is an object?
  • ECCV 2010: Category Independent Object Proposals
  • CVPR 2011: Proposal Generation for Object Detection using Cascaded Ranking SVMs
  • IJCV 2013: Selective Search for Object Recognition
  • CVPR 2014: BING: Binarized Normed Gradients for Objectness Estimation at 300fps
  • ECCV 2014: Edge Boxes: Locating Object Proposals from Edges

• Conclusion and Discussion
Previous Work

• Traffic sign detection
  • Color probability model (CPM)
  • Two methods for traffic sign detection
    • CPM + MSER + integral channel features detector [1]
    • CPM + MSER + SVM [2]
  • A traffic sign detection and recognition system
  • Two papers

[1] Real-time traffic sign recognition system via color probability model and integral channel features. CCPR 2014
Future Work

Object proposals

Generic object detection

Objectness
• Object?
  • Objects are stand-alone things with well-defined closed boundaries and centers.
  • Any object has at least one of the three distinctive characteristics:
    • A well-defined closed boundary in space;
    • A different appearance from their surroundings;
    • Sometimes it is unique within the image and stands out as salient.

• Background?
  • Amorphous background stuff.

• Objectness?
  • Objectness is usually represented as a value which reflects how likely an image window covers an object of any category.
CVPR 2010: What is an object?

Goal:

- Distinguish **objects with a well-defined boundary** in space, such as cows and telephones, from **amorphous background elements**, such as grass and road.

Object:

- A well-defined **closed boundary**
- A **different appearance** from their surroundings
- Sometimes it is **unique** within the image and stands out as salient

Cues:

- Multi-scale Saliency (MS)
- Color Contrast (CC)
- Edge Density (ED)
- Superpixels Straddling (SS)

Bayesian cue integration
Multi-scale Saliency (MS)

• This MS cue measures the uniqueness characteristic of objects.

• [1]: a global saliency measure based on the spectral residual of the FFT, which favors regions with an unique appearance within the entire image.

• MS provides a rough indication of where an object is as it is designed to find blob-like things.

Color Contrast (CC)

- This cue measures the **different appearance** characteristic of objects.
- The CC cue is a local measure of the dissimilarity of a window to its immediate surrounding area.
- CC provides more accurate windows, but sometimes misses objects entirely.

Fig. 3: CC success and failure. **Success:** the windows containing the objects (cyan) have high color contrast with their surrounding ring (yellow) in images (a) and (b). **Failure:** the color contrast for windows in cyan in image (c) is much lower.
Edge Density (ED)

- The ED cue captures the **closed boundary** characteristic of objects, as they tend to have many edges in the inner ring.

- The ED is computed as the density of edges in the inner ring.

- ED provides many false positives on textured areas.

---

Fig. 4: **ED success and failure.** Success: given images (a) and (b) the cyan windows covering the bus and the aeroplane score high as the density of edges is concentrated in these regions. **Failure:** in image (c) the cyan window along with many other windows covering the water score high determining a high rate of false positives. In particular the windows covering the boats have a low score. We show the Canny edge maps in (d), (e) and (f).
Superpixels Straddling (SS)

• A different way to capture the closed boundary characteristic of objects.
• A key property of superpixels is to preserve object boundaries: all pixels in a superpixel belong to the same object.
• An object is typically over-segmented into several superpixels, but none straddles its boundaries.
• A superpixel $s$ straddles a window $w$ if it contains at least one pixel inside and at least one outside $w$.
• SS is very distinctive but depends on good superpixels, which are fragile for small objects.

![Diagram of Superpixels Straddling](image)

Fig. 5: The SS cue. Given the segmentation (b) of image (a), for a window $w$ we compute $SS(w, \theta_{SS})$ (eq. 4). In (c), most of the surface of $w_1$ is covered by superpixels contained almost entirely inside it. Instead, all superpixels passing by $w_2$ continue largely outside it. Therefore, $w_1$ has a higher SS score than $w_2$. The window $w_3$ has an even higher score as it fits the object tightly.
Fig. 7: SS success and failure. **Success:** The cyan windows in (a) and (b) have high SS score computed from segmentations (d) and (e). **Failure:** Segmentation produces superpixels (f) not preserving the boundaries of the small objects in (c), resulting in low SS.
ECCV 2010: Category Independent Object Proposals

Key idea:
• Generates a set of segmentations by performing graph cuts based on a seed region and a learned affinity function.
• The regions are ranked using structured learning based on various cues.

Fig. 1. Our pipeline: compute a hierarchical segmentation, generate proposals, and rank proposed regions. At each stage, we train classifiers to focus on likely object regions and encourage diversity among the proposals, enabling the system to localize many types of objects. See section 3 for a more detailed overview.
Cascaded ranking SVM

- **First stage:** generates a number of proposal windows at each scale $k$ for image $I$. (multi-scale/aspect-ratios)

- **Second stage:** re-ranks these windows globally, so that the best proposals across scales are forwarded.

Figure 1. Summary of our method. An image (a) is first convolved with a set of linear classifiers at varying scales/aspect-ratios (b) producing response images (c). Local maxima are extracted from each response image, and the corresponding windows with top ranking scores are forwarded to the second stage of the cascade. Each proposed window is associated with a feature vector (d), and a second round of ranking orders these proposals (e) so that the true positives (marked as black) are pushed towards the top during training. Our method outputs the top ranking windows in this final ordering.
Selective search:

- Exhaustive search: aims to capture all possible object locations.
- Segmentation: uses the image structure to guide the sampling process.

**Fig. 1** There is a high variety of reasons that an image region forms an object. In (b) the cats can be distinguished by colour, not texture. In (c) the chameleon can be distinguished from the surrounding leaves by texture, not colour. In (d) the wheels can be part of the car because they are enclosed, not because they are similar in texture or colour. Therefore, to find objects in a structured way it is necessary to use a variety of diverse strategies. Furthermore, an image is intrinsically hierarchical as there is no single scale for which the complete table, salad bowl, and salad spoon can be found in (a).
Fig. 2 Two examples of our selective search showing the necessity of different scales. On the left we find many objects at different scales. On the right we necessarily find the objects at different scales as the girl is contained by the tv.
CVPR 2014: BING: Binarized Normed Gradients for Objectness Estimation at 300fps

BMVC 2014: Cracking BING and Beyond

Figure 1. Although object (red) and non-object (green) windows present huge variation in the image space (a), in proper scales and aspect ratios where they correspond to a small fixed size (b), their corresponding normed gradients, i.e. a NG feature (c), share strong correlation. We learn a single 64D linear model (d) for selecting object proposals based on their NG features.
BMVC 2014:
Cracking BING and Beyond
ECCV 2014:
Edge Boxes: Locating Object Proposals from Edges

![Graphs showing detection rate vs. number of proposals for different IoU thresholds (0.5, 0.7, 0.9). Each graph compares various methods: Objectness, Selective Search, CPMC, BING, Rahtu, Randomized Prim's, and Rentalankila. The x-axis represents the number of proposals ranging from 1 to 1000, and the y-axis represents the detection rate.](image)
ECCV 2014: Edge Boxes: Locating Object Proposals from Edges

Key idea:

• The number of contours that are **wholly enclosed** by a bounding box is indicative of the likelihood of the box containing an object.

• A contour is wholly enclosed by a box: if all edge pixels belonging to the contour lie within the interior of the box.

• Edges tend to correspond to object boundaries, and as such boxes that tightly enclose a set of edges are likely to contain an object.
Fig. 1. Illustrative examples showing from top to bottom (first row) original image, (second row) Structured Edges [16], (third row) edge groups, (fourth row) example correct bounding box and edge labeling, and (fifth row) example incorrect boxes and edge labeling. Green edges are predicted to be part of the object in the box ($w(s) = 1$), while red edges are not ($w(s) = 0$). Scoring a candidate box based solely on the number of contours it wholly encloses creates a surprisingly effective object proposal measure. The edges in rows 3-5 are thresholded and widened to increase visibility.
Table 2. Results for our approach, Edge Boxes 70, compared to other methods for IoU threshold of 0.7. Methods are sorted by increasing Area Under the Curve (AUC). Additional metrics include the number of proposals needed to achieve 25%, 50% and 75% recall and the maximum recall using 5000 boxes. Edge Boxes is best or near best under every metric. All method runtimes were obtained from [26].

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>N@25%</th>
<th>N@50%</th>
<th>N@75%</th>
<th>Recall</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>BING [11]</td>
<td>.20</td>
<td>292</td>
<td>–</td>
<td>–</td>
<td>29%</td>
<td>.2s</td>
</tr>
<tr>
<td>Rantalankila [10]</td>
<td>.23</td>
<td>184</td>
<td>584</td>
<td>–</td>
<td>68%</td>
<td>10s</td>
</tr>
<tr>
<td>Objectness [4]</td>
<td>.27</td>
<td>27</td>
<td>–</td>
<td>–</td>
<td>39%</td>
<td>3s</td>
</tr>
<tr>
<td>Rand. Prim’s [8]</td>
<td>.35</td>
<td>42</td>
<td>349</td>
<td>3023</td>
<td>80%</td>
<td>1s</td>
</tr>
<tr>
<td>Rahtu [7]</td>
<td>.37</td>
<td>29</td>
<td>307</td>
<td>–</td>
<td>70%</td>
<td>3s</td>
</tr>
<tr>
<td>Selective Search [5]</td>
<td>.40</td>
<td>28</td>
<td>199</td>
<td>1434</td>
<td>87%</td>
<td>10s</td>
</tr>
<tr>
<td>CPMC [6]</td>
<td>.41</td>
<td>15</td>
<td>111</td>
<td>–</td>
<td>65%</td>
<td>250s</td>
</tr>
<tr>
<td>Edge boxes 70</td>
<td>.46</td>
<td>12</td>
<td>108</td>
<td>800</td>
<td>87%</td>
<td>.25s</td>
</tr>
</tbody>
</table>
Conclusion

- Methods
  - Multi-cues integration
  - Segmentation (graph-cut)
  - Cascaded ranking SVM
  - Selective search (exhaustive search + segmentation)
  - BING (300fps)
  - Edge-boxes

- My opinion:
  - BING-like or Edge-boxes-like
    - Simple idea
    - Fast
    - High recall rate
    - High quality
    - **Small number of proposals**
  - Selective-search-like
    - Bottom-up grouping

- **Suggestions?**
Thanks!