Fine-grained Visual Analysis:
From Classification to Retrieval

Yi-Zhe Song
SketchX Lab, CVSSP, University of Surrey, UK
http://sketchx.ai
Why fine-grained?

I am not just a “dog” 😞 😞 😞
Why fine-grained?

Husky

Chihuahua

Bulldog

Better 😊

At the very heart of human and computer vision!!
What is fine-grained?

- Surveys + Seminars exist
  - a good survey [1]
  - First Edition of 见微知著 (2019年12月11日)

- Classification + Retrieval most studied
  - Classification being the favourite child
  - Images → video, 3D, text
  - Recent branching to generation, transfer learning, hashing...

Classification vs. Retrieval

• “The Curse of the Labels”
  - Classification $\rightarrow$ hard to obtain expert labels
  - Retrieval $\rightarrow$ one can not retrieve without knowing the label
Problem with Classification

- Dataset! Dataset! Dataset! $\rightarrow$ Label! Label! Label!

- Obsession with parts
  - **Explicit** to start with
  - Now **implicit** as well $\rightarrow$ part is not everything

Explicit Models:
- **MA-CNN (ICCV17)**
- **NTS-Net (ECCV18)**

Implicit Models:
- **B-CNN (ICCV15)**
- **MC-Loss (TIP20)**
- **PMG (ECCV20)**
- **Pairwise confusion (ECCV18)**
Problem with Retrieval

• Ill-posed to start with $\rightarrow$ where do we get the labels?
  • Retrieval **dictates expert knowledge** to start with!

• Best input modality?
  • Yes, there is image (but is it the only choice?)
  • Human subjectivity $\rightarrow$ text best for that (?)

• There is just not enough work!
All about Retrieval

• Is the old “fine-grained” enough? \(\rightarrow\) more than just names (labels)!
  • Pose, instance-level details
    • “a Labrador standing on two feet, looking at the camera with a smile”
  • Latent sub-classes
    • Labrador \(\rightarrow\) English Labrador and American Labrador

• Flexibility to meet human subjectivity
  • as flexible as text?

• What would be the best input modality?

• More practical with real application scenarios?
Sketch for Retrieval

- IMPRECISE
  - Text: Many irrelevant results

- NO FLEXIBILITY
  - Image: Lots of very similar images

- FLEXIBLE & EXACT
  - Sketch: Customised list of closely relevant images

To be explored
Sketch for Retrieval

• Specific challenges
  • Cross-modal
  • Human subjectivity
  • Learning under small data
FG-SBIR: Fine-Grained Sketch-Based Image Retrieval

**FG-SBIR 1.0** – pose correspondence
(BMVC’15)

**FG-SBIR 2.0** – instance correspondence
(CVPR’16 Oral, SIGGRAPH’16, ICCV’17, 3xECCV’18, CVPR’19 Oral, CVPR’20)

**FG-SBIR 3.0** – on-the-fly retrieval
(CVPR’20 Oral)
**FG-SBIR: Fine-Grained Sketch-Based Image Retrieval**

- Dataset usually very small
  - ImageNet pre-training is thus a must + fine-tuning.
- Triplet Ranking Network
  - pushing positive sketch-photo pairs near, and negatives apart.

FG-SBIR: The Role of Jigsaw

- Jigsaw puzzles helps with fine-grained [1]
- See also [2] for classification


Solving a mixed-modality jigsaw model requires learning to:
- Bridge the domain discrepancy
- Understand holistic object configuration
- Encode fine-grained detail.

A permutation inference problem
- Normalisation via Sinkhorn iterations

Great performance boost to long standing practice of ImageNet pre-training.
## FG-SBIR: The Role of Jigsaw

<table>
<thead>
<tr>
<th>Pre-training</th>
<th>Self-supervised?</th>
<th>QMUL Shoe V1$^{4\times4}$</th>
<th>QMUL Shoe V2$^{3\times3}$</th>
<th>QMUL Chair$^{3\times3}$</th>
<th>QMUL Handbag$^{4\times4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting [16]</td>
<td>✓</td>
<td>41.74%± 2.30</td>
<td>30.42%± 0.54</td>
<td>72.78%± 4.35</td>
<td>54.05%± 2.77</td>
</tr>
<tr>
<td>Rotation [9]</td>
<td>✓</td>
<td>32.17%± 2.68</td>
<td>28.83%± 0.40</td>
<td>70.31%± 3.45</td>
<td>38.33%± 1.86</td>
</tr>
<tr>
<td>CPC [17]</td>
<td>✓</td>
<td>21.91%± 1.69</td>
<td>8.65%± 0.34</td>
<td>35.24%± 0.42</td>
<td>15.36%± 0.69</td>
</tr>
<tr>
<td>Matching [20]</td>
<td>✓</td>
<td>39.13%± 0.87</td>
<td>31.05%± 0.84</td>
<td>75.69%± 1.53</td>
<td>50.36%± 0.68</td>
</tr>
<tr>
<td>ImageNet [29]</td>
<td>X</td>
<td>43.48%± 1.74</td>
<td>33.99%± 1.09</td>
<td>85.16%± 1.56</td>
<td>52.62%± 2.04</td>
</tr>
<tr>
<td>Ours/1000-way</td>
<td>✓</td>
<td>42.78%± 3.75</td>
<td>30.24%± 1.74</td>
<td>79.59%± 1.53</td>
<td>49.40%± 3.97</td>
</tr>
<tr>
<td>Ours/ImageNet</td>
<td>✓✓</td>
<td>48.00%± 2.91</td>
<td>31.26%± 0.65</td>
<td>79.59%± 1.34</td>
<td>61.07%± 1.50</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>56.52%± 2.75</td>
<td>36.52%± 0.84</td>
<td>85.98%± 2.01</td>
<td>62.97%± 2.04</td>
</tr>
</tbody>
</table>

**NOTE:** opposite conclusions for category-level task!
Effect of jigsaw modality

- mixed-modal Jigsaw is the best
- granularity of jigsaw not crucial
FG-SBIR: On-the-Fly

Problem – “I can’t sketch”

- **Time** taken to draw a *complete* sketch
- **Drawing skill** of the user

FG-SBIR: On-the-Fly

**Old Setup:** sketch first, *then* retrieve

**New On-the-fly Setup:** retrieve *as you sketch*

Less is more!
FG-SBIR: On-the-Fly

• **Natural**: incomplete sketches can *already* retrieve!

• **Faster**: *no need* to sketch the whole thing

• **More accurate**: modelling the *sketching process* does help

In most cases, we can retrieve with ~30% less strokes!
FG-SBIR: On-the-Fly

- **Reinforcement Learning** (RL) for cross-modal modelling.
- **Reward design** to encourage early retrieval
- **Rank optimization** over a complete sketch drawing episode
Quantitative Results vs Different Baselines (A@q, m@A, and m@B)

<table>
<thead>
<tr>
<th></th>
<th>Chair-V2</th>
<th></th>
<th></th>
<th>Shoe-V2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m@A</td>
<td>m@B</td>
<td>A@5</td>
<td>A@10</td>
<td>m@A</td>
<td>m@B</td>
<td>A@5</td>
</tr>
<tr>
<td>B1</td>
<td>77.18</td>
<td>29.04</td>
<td>76.47</td>
<td>88.13</td>
<td>80.12</td>
<td>18.05</td>
<td>65.69</td>
</tr>
<tr>
<td>B2</td>
<td>80.46</td>
<td>28.07</td>
<td>74.31</td>
<td>86.69</td>
<td>79.72</td>
<td>18.75</td>
<td>61.79</td>
</tr>
<tr>
<td>B3</td>
<td>76.99</td>
<td>30.27</td>
<td>76.47</td>
<td>88.13</td>
<td>80.13</td>
<td>18.46</td>
<td>65.69</td>
</tr>
<tr>
<td>B4</td>
<td>81.24</td>
<td>29.85</td>
<td>75.14</td>
<td>87.69</td>
<td>81.02</td>
<td>19.50</td>
<td>62.34</td>
</tr>
<tr>
<td>TS</td>
<td>76.01</td>
<td>27.64</td>
<td>73.47</td>
<td>85.13</td>
<td>77.12</td>
<td>17.13</td>
<td>62.67</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>85.44</strong></td>
<td><strong>35.09</strong></td>
<td>76.34</td>
<td><strong>89.65</strong></td>
<td><strong>85.38</strong></td>
<td><strong>21.44</strong></td>
<td><strong>65.77</strong></td>
</tr>
</tbody>
</table>

Percentage-wise Results for Shoe-V2 (m@A, and m@B)

Percentage-wise Results for Chair-V2 (m@A, and m@B)

FG-SBIR: On-the-Fly
Classification ▷ Retrieval

• Classification → Retrieval
  • Obvious

• Retrieval → Classification
  • Cure for web data?
  • Sub-class discovery?

Conclusion

- Fine-grained is important!
- Classification bottlenecked
- Retrieval needs more work
  - Unique challenges
  - Practical applications
  - Can help classification
- Beyond 2D!