Do You Need Experts in the Crowd?

A case study in image annotation for marine biology

Jiyin He, Jacco van Ossenbruggen, and Arjen P. de Vries
Centrum Wiskunde & Informatica
An image labeling problem that requires specialists’ knowledge
An image labeling problem that requires specialists’ knowledge

What is in the picture?

[Image of a fish]
An image labeling problem that requires specialists’ knowledge

What is in the picture?
- A fish
An image labeling problem that requires specialists’ knowledge

What is in the picture?
- A fish

Which species is it?
An image labeling problem that requires specialists’ knowledge

What is in the picture?
- A fish

Which species is it?
- Chaetodon trifascialis
Some background

- Underwater cameras
- Videos
- Computer vision systems
  - Detection
  - Tracking
  - Recognition

Ground truth Needed!
Fish species recognition

- Large set of labeled images/videos needed
- Expert knowledge needed
  - Non-experts often lack the knowledge needed to recognize a fish
  - Non-experts may not be able to map the common name of a fish to its scientific name
- Experts are expensive, rare resources
  - Even experts can have their expertise in different types of fish or fish in different areas
What can non-experts (not) do?

- Assumptions
  - Non-experts are not able to *actively* name fish species
  - But may able to *passively* judge if two fish are visually similar

- Possible tasks
  - Manual clustering
  - Classification with textbook images as category labels
An interface to support fish recognition with experts - collecting ground truth

*Bad image*: images with no fish, multiple fishes of different species, or fish partially behind other underwater objects.

- Step 1: Enter the scientific name that applies to the majority of the fishes below. **Scolopsis lineata** (Note: please enter "unknown" if the species is unrecognizable)
- Step 2: Find fishes that do not belong to **Scolopsis lineata**: select "other species" and enter the correct species name.

Finished & Go to questionnaire
An interface to support fish recognition with non-experts

Fish4Label

Welcome! cactus Achievements Instructions Log out Change password

2/50

Query image

Scores

Session score: 4
Total score: 176

Candidate images

Ostracion immaculatus
Halichoeres ornatissimus
Xanthichthys lineopunctatus
Scarus chameleon

Bolbometopon muricatum
Plectrogyphidodon dickii
Stephanolepis cirrhifer
Others

Fish4Knowledge 2012
## Experts vs. non-experts

<table>
<thead>
<tr>
<th>Candidate source</th>
<th>Verification source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experts</td>
<td>From their knowledge</td>
</tr>
<tr>
<td></td>
<td>Text book</td>
</tr>
<tr>
<td>Non-experts</td>
<td>Given by the system</td>
</tr>
<tr>
<td></td>
<td>System feedback</td>
</tr>
</tbody>
</table>
A study of non-expert annotators

• Can non-experts effectively separate similar species given the current setup?

• Can non-experts learn during the labeling process, e.g., from the system feedback?
A study of non-expert annotators

- Controlled experiments
- 190 expert labeled images
- 3 experts provided ground truth
- 2 simulated labeling conditions

<table>
<thead>
<tr>
<th>Exp</th>
<th>Candidate type</th>
<th>#Users</th>
<th># Labels/image</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True label is present together with similar but incorrect labels</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>In 25% of the cases, true labels were removed, while similar but incorrect labels are present</td>
<td>32 (28 +4)</td>
<td>13</td>
</tr>
</tbody>
</table>
Reliability of non-expert labels

• Compared to expert labels
  • agreement in terms of Cohen’s kappa;
  • non-experts labels aggregated by simple majority voting

<table>
<thead>
<tr>
<th>Expr.</th>
<th>Expert vs.</th>
<th>Species level</th>
<th>Family level</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>expert</td>
<td>0.55~0.67</td>
<td>0.75~0.85</td>
</tr>
<tr>
<td>1</td>
<td>non-experts</td>
<td>0.55~0.65</td>
<td>0.72~0.83</td>
</tr>
<tr>
<td>2</td>
<td>non-experts (new)</td>
<td>0.45~0.65</td>
<td>0.68~0.73</td>
</tr>
<tr>
<td>2</td>
<td>non-experts (old)</td>
<td>0.53~0.68</td>
<td>0.74~0.80</td>
</tr>
</tbody>
</table>
Do non-experts learn?

- Two types of learning
  - Memorization
  - Generalization

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Memorization</th>
<th>Generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>labels</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>0.30</td>
</tr>
<tr>
<td>2 (new)</td>
<td>0.30</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Average user scores that are normalized by the maximum score one can achieve at each label.
Conclusions

• Converting an active labeling task to a passive image comparing task allows non-expert users to perform image labeling task that requires highly specialized knowledge

• In ideal case, non-experts can achieve an agreement with experts comparable to that achieved between experts

• In the more confusing case, novice non-experts are more likely get confused compared to experienced users

• Non-expert users are able to learn in terms of both memorization and generalization
Reliability of non-expert labels

- Accuracy of aggregated labels
- Novice users are likely to be confused when correct labels are not present

<table>
<thead>
<tr>
<th>Expr.</th>
<th>User type</th>
<th>Species level</th>
<th>Family level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ndcg@1</td>
<td>ndcg@5</td>
<td>ndcg@1</td>
</tr>
<tr>
<td>1</td>
<td>22 new users</td>
<td>0.84</td>
<td>0.88</td>
</tr>
<tr>
<td>2</td>
<td>28 new users</td>
<td>0.72(&lt;=)</td>
<td>0.77(&lt;=)</td>
</tr>
<tr>
<td>2</td>
<td>4 old users</td>
<td>0.88</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Main findings

• When expert feedback is available
  • In ideal case, non-experts can achieve an agreement with experts comparable to that achieved between experts
  • In the more confusing case, novice non-experts are more likely to get confused
  • Implication: It’s important to select good candidates

• When expert feedback is not available
  • Can aggregation on noisy feedback generate reasonable results?
    • If not:
      • More sophisticated aggregation method
      • More users - reach sufficient confidence
      • Training session with expert feedbacks before labeling
Main findings (2)

- Non-experts learn while playing the game
  - memorizing - performance on same image improves
  - generalization - performance on same species improves
- When there is no feedback (3 users)
  - 3 users set the initial labels for the peer-agree runs - work independently
  - User score with experts:
    - each judgement gets 0, 1, 2, 3 points if agree with 0, 1, 2, or 3 experts
    - 50 images per session
- Users seem to be able to improve without feedback (Need more evidence), to what limit?

<table>
<thead>
<tr>
<th></th>
<th>session 1</th>
<th>session 2</th>
<th>session 3</th>
<th>session 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92</td>
<td>99</td>
<td>116</td>
<td>101</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>94</td>
<td>90</td>
<td>99</td>
</tr>
<tr>
<td>3</td>
<td>83</td>
<td>81</td>
<td>93</td>
<td>90</td>
</tr>
</tbody>
</table>
Some images are more confusing than others

- Let clarity score = \#majority vote/\#vote
- Per image clarity score in Experiment 1

<table>
<thead>
<tr>
<th>Clarity scores</th>
<th>4/23 votes</th>
<th>4/23 votes</th>
<th>4/22 votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image ordered by clarity scores</td>
<td>![Fish image]</td>
<td>![Fish image]</td>
<td>![Fish image]</td>
</tr>
</tbody>
</table>