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?

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

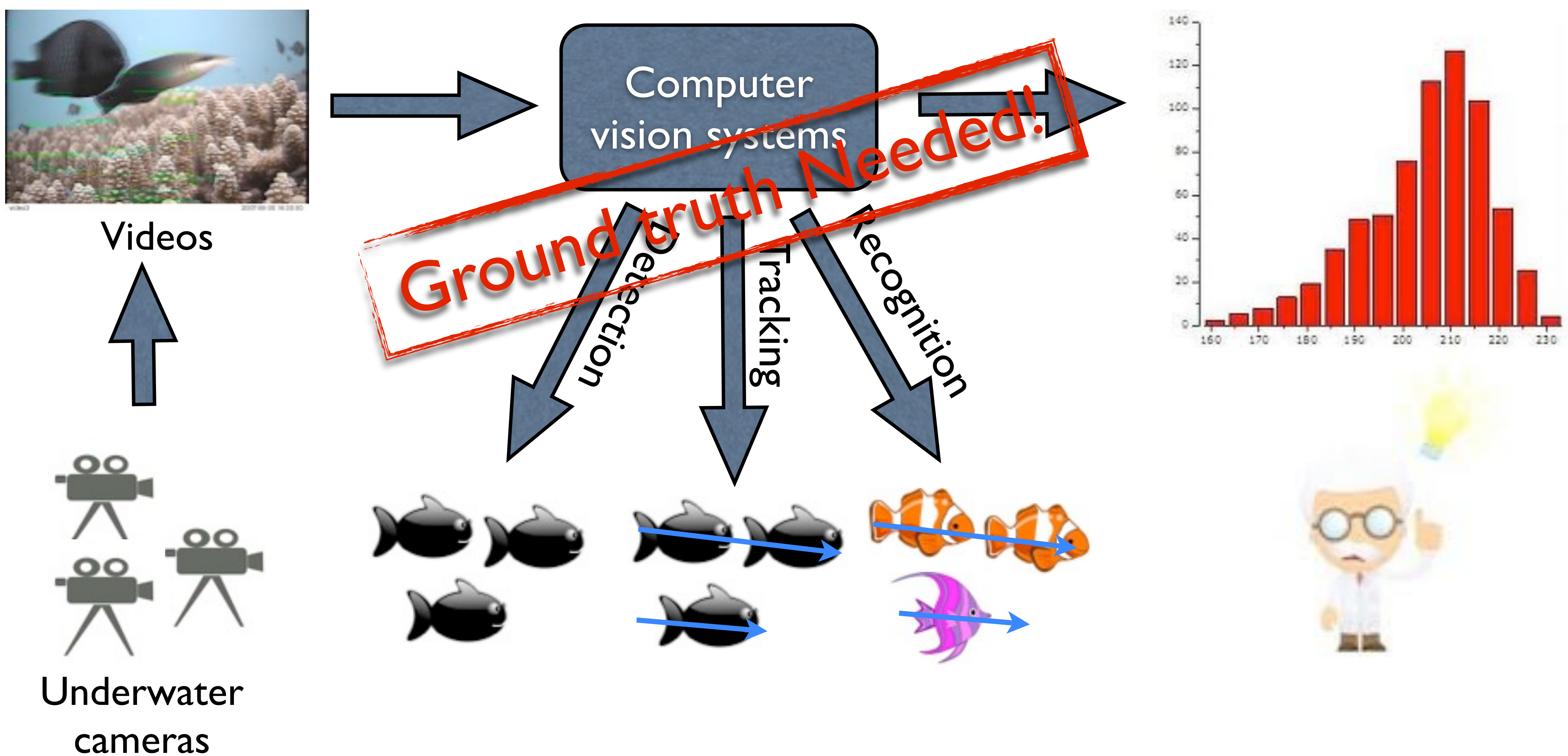
What is in the picture?

- A fish

Which species is it?

- Chaetodon trifascialis

Some background



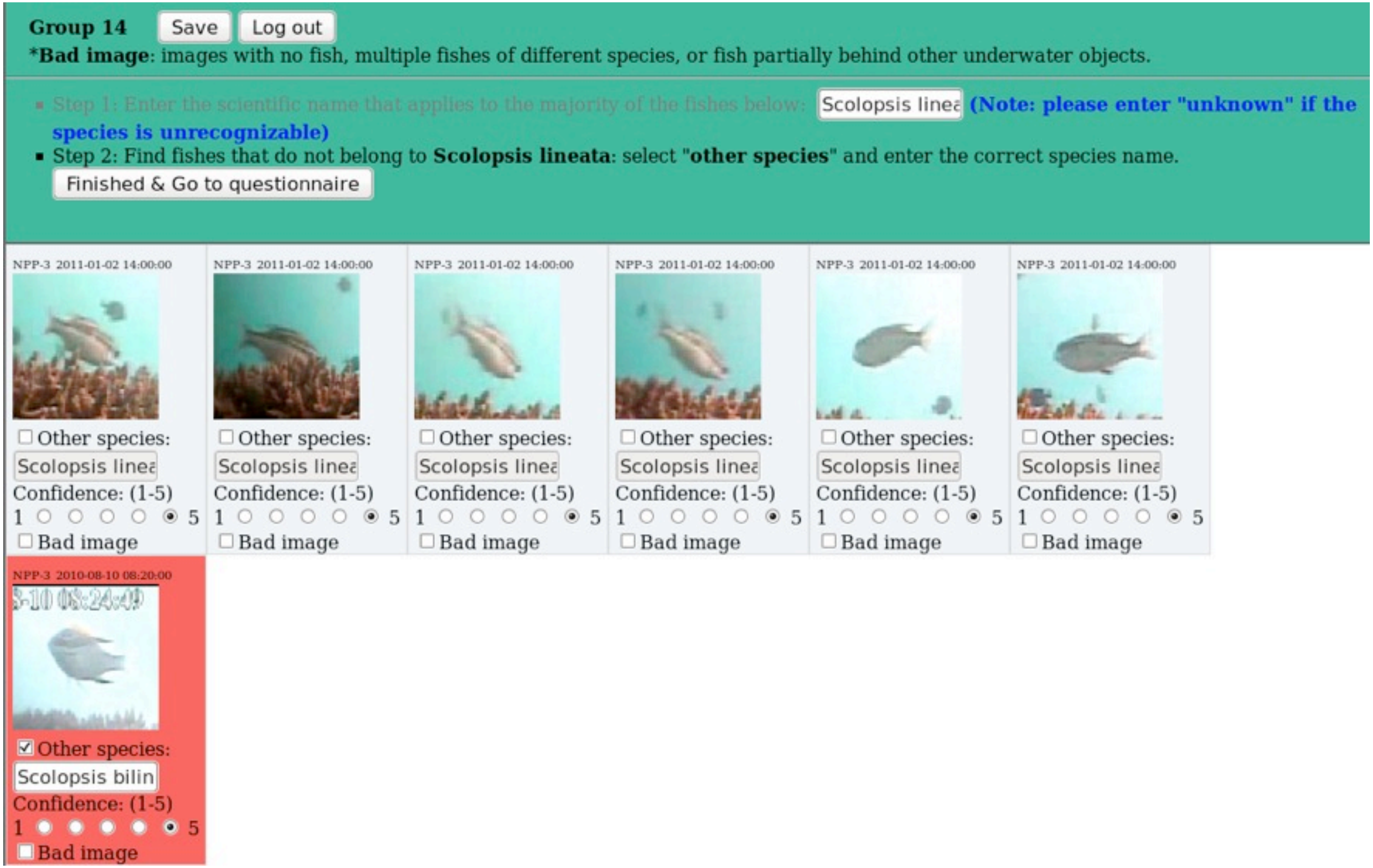
Fish species recognition

- 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

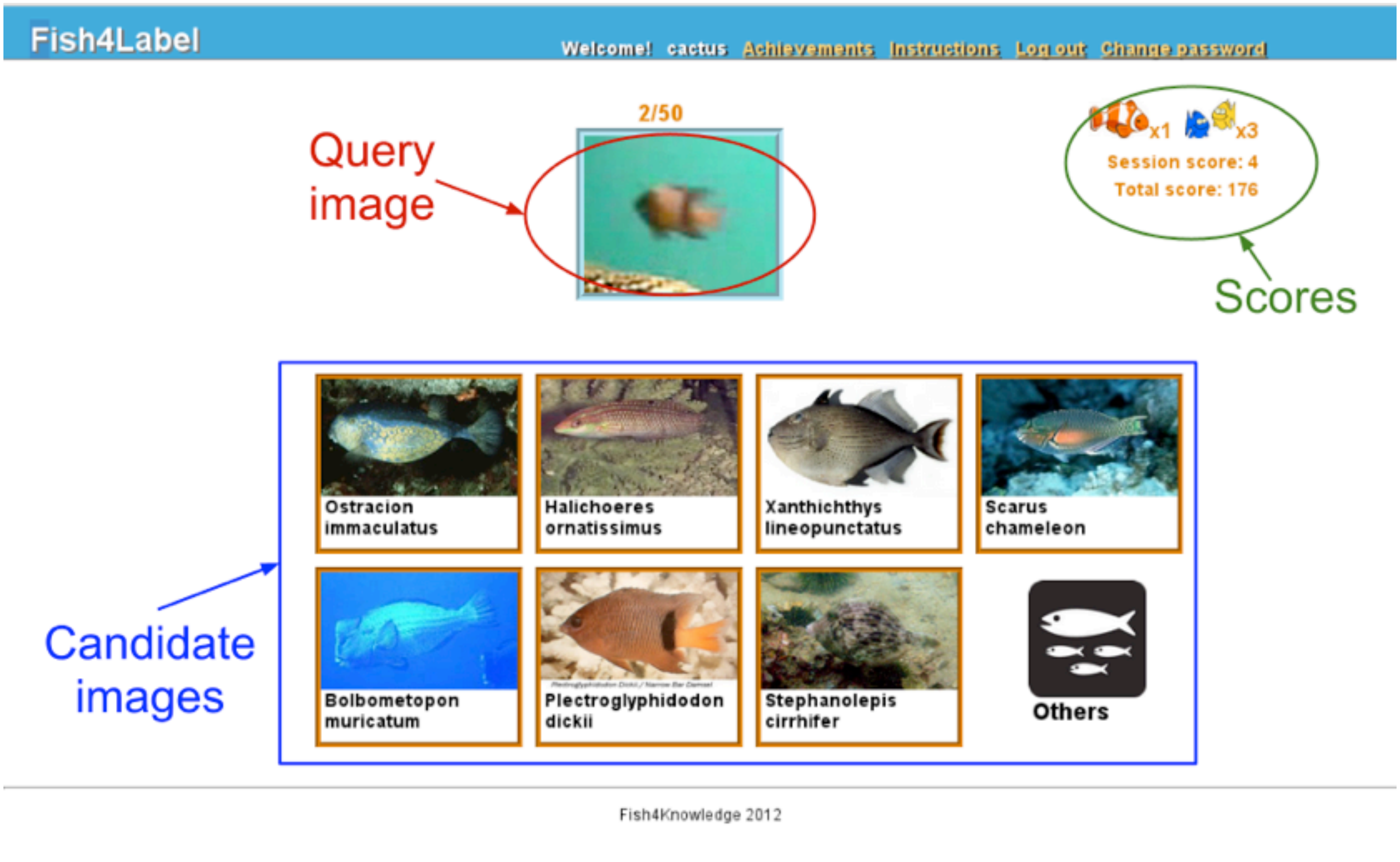
What can non-experts (not) do?

- 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



An interface to support fish recognition with non-experts



Experts vs. non-experts

	Candidate source	Verification source
Experts	From their knowledge	Text book
Non-experts	Given by the system	System feedback

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

Exp	Candidate type	#Users	# Labels/image
1	True label is present together with similar but incorrect labels	22	19
2	In 25% of the cases, true labels were removed, while similar but incorrect labels are present	32 (28 +4)	13

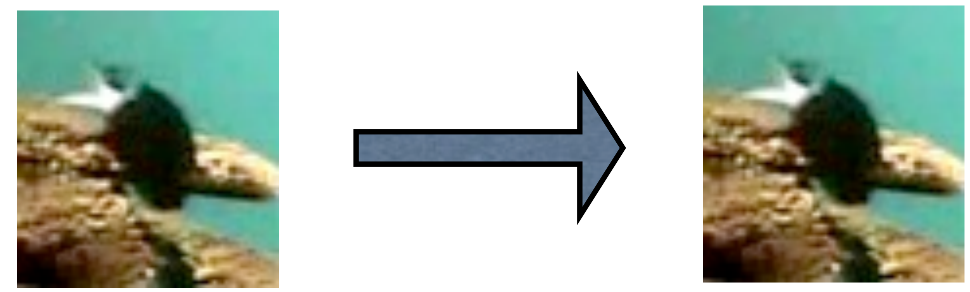
Reliability of non-expert labels

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

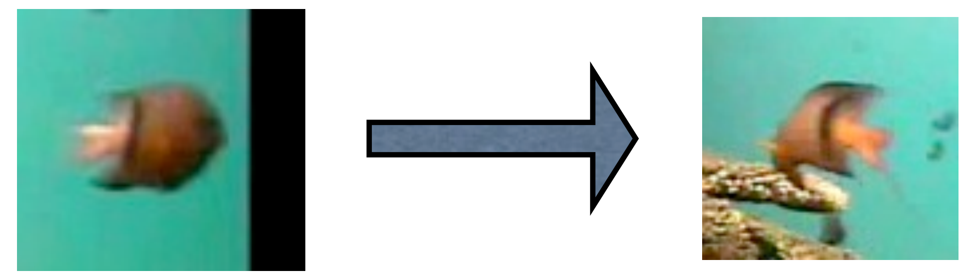
Expr.	Expert vs.	Species level	Family level
-	expert	0.55~0.67	0.75~0.85
1	non-experts	0.55~0.65	0.72~0.83
2	non-experts (new)	0.45~0.65	0.68~0.73
2	non-experts (old)	0.53~0.68	0.74~0.80

Do non-experts learn?

- Two types of learning
  - Memorization



  - Generalization



Exp.	Memorization			Generalization		
labels	1	2	3	1	5	10
1	0.30	**0.38**	**0.46**	0.42	**0.51**	**0.59**
2 (new)	0.30	**0.4**	**0.44**	0.37	**0.58**	**0.62**
Average user scores that are normalized by the maximum score one can achieve at each label						

CWI

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

CWI

Reliability of non-expert labels

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

Expr.	User type	Species level		Family level	
		ndcg@1	ndcg@5	ndcg@1	ndcg@5
1	22 new users	0.84	0.88	0.93	0.94
2	28 new users	0.72(<)	0.77(<)	0.86(<)	0.94
2	4 old users	0.88	0.86	0.91	0.94

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 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?

user	session 1	session 2	session 3	session 4
1	92	99	116	101
2	69	94	90	99
3	83	81	93	90

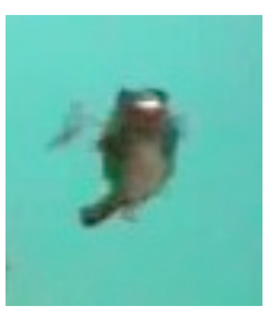
Some images are more confusing than others

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

4/23 votes

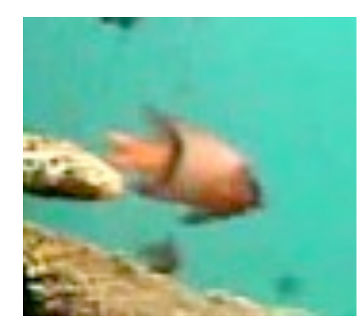
4/23 votes

4/22 votes



25/25 votes

24/24 votes



24/24 votes