Do you need experts in the crowd? A case study in image annotation for marine biology

Jiyin He Jacco van Ossenbruggen Arjen P. de Vries

{j.he, jacco.van.ossenbruggen, arjen.de.vries}@cwi.nl

Centrum Wiskunde en Informatica, Science Park 123
1098XG, Amsterdam, the Netherlands

ABSTRACT

Labeled data is a prerequisite for successfully applying machine learning techniques to a wide range of problems. Recently, crowd-sourcing has shown to provide effective solutions to many labeling tasks. However, tasks in specialist domains are difficult to map to Human Intelligence Tasks (or HITs) that can be solved adequately by "the crowd". The question addressed in this paper is whether these specialist tasks can be cast in such a way, that accurate results can still be obtained through crowd-sourcing. We study a case where the goal is to identify fish species in images extracted from videos taken by underwater cameras, a task that typically requires profound domain knowledge in marine biology and hence would be difficult, if not impossible, for the crowd. We show that by carefully converting the recognition task to a visual similarity comparison task, the crowd achieves agreement with the experts comparable to the agreement achieved among experts. Further, non-expert users can learn and improve their performance during the labeling process, e.g., from the system feedback.

Categories and Subject Descriptors

[**Human computer interaction (HCI)**]: User studies; Laboratory experiments

Keywords

Image labeling, Crowdsourcing, User studies

1. INTRODUCTION

Creating ground truth data for video-based retrieval and computer vision research is often a time consuming task done by humans using dedicated tools such as those presented in [6]. Recently, crowd-sourcing as a collaborative problem solving strategy has received much attention. In particular, within the computer vision communities, where *large scale* ground truth data are needed, the wisdom of the crowd was shown to provide effective solutions in a wide range of problems [1, 5, 7–9]. Typically, the annotation/labeling tasks the crowd is asked to perform are relatively easy, that is, little or no expert knowledge is required.

Instead, this paper studies an image labeling task that requires highly specialized domain knowledge. The ground truth obtained serves as training material for machine learning approaches that aim to classify fish species on video footage of Taiwanese coral reefs. Correctly identifying fish species on this footage requires expertise from marine biologists, which is highly localized: biologists specialized on the Australian reefs perform not as good as those specialized on the Taiwanese coral reef fish species. Further, since experts are a scarce and expensive resource, it is unlikely that they would provide the amount of image labels needed for the purpose of training and evaluating the fish classification models. The question is then, can we create a ground truth set of sufficient *quantity* with sufficient *quality* by taking advantage of the collaborative problem solving ability of the crowd, while solving the problem that the crowd generally lacks the domain knowledge required by the task?

A smart way of presenting a problem or decomposing a complicated problem into simpler sub-problems may greatly reduce its difficulty and makes an infeasible task feasible. Typical examples include Foldit [2] that uses a puzzle solving game for protein structure prediction. Another example is Galaxy zoo [3] that uses “citizens’ wisdom” to contribute to morphological classification of galaxies. For our labeling problem, we use the expertise of marine biologists to transform the fish identification task into a game based on a visual similarity comparison task that can be performed by a large number of non-experts. We then conduct a user study and seek the answers to the following questions: (i) Can non-expert players of this game achieve acceptable performance evaluated with the labels provided by the experts? and (ii) Can players learn and improve their performance during the game? We find that after the task conversion, non-expert players achieve an agreement with the experts comparable to that achieved among the experts themselves. Further, players improve their performance while playing the game: they are able to recognize a fish better not only when they see the same fish again, but also when they see a different fish from the same species.

Our contributions are two fold. First, we propose a task conversion approach to solve an image recognition problem that requires highly specialized domain knowledge with non-expert users. Second, our study on the learning behavior of the non-expert users provides insights into the ability and limits of the crowd.

2. LABELING WITH EXPERTS

Experts are expensive and a scarce resource. Therefore we ask experts to label only a small subset of our data as ground truth

OAIR'13, May 22-24, 2013, Lisbon, Portugal.

Table 1: Cohen's kappa for measuring expert agreement.

	Species level		Family level	
Comparison	Avg.κ	Sdv.	Avg.κ	Stv.
E1 vs. E2	0.55	0.008	0.85	0.004
E1 vs. E3	0.48	0.008	0.75	0.000
E2 vs. E3	0.67	0.006	0.76	0.0001

and developed a cluster-based interface shown in Figure 1(a) to facilitate their labeling process. First, the expert is asked to enter the species name that applies to the majority of the images in a cluster, which automatically assigns the same species name to all images in the cluster. Then, he/she manually correct the species names of those images that do not belong to the same cluster. In the worst case, the expert will have to manually assign a species name to each of the images, while in the best case (i.e., the cluster is pure), the expert only needs to enter the species name once.

To obtain clusters with relatively good quality, two students manually clustered 3000 images randomly sampled from our video data into 28 clusters. To limit the amount of effort experts need to examine the clusters, at most 30 images are randomly selected from each cluster and shown to the experts. As the size of the clusters is unevenly distributed, we obtain a total of 190 labeled images. Three marine biologists, having a research experience of 30, 10 and 25 years in Taiwanse coral reef fish respectively, were invited to create the ground truth labels.

We make the following observations about the obtained ground truth. (i) Biologists are sometimes not sure which species a fish should belong to: a) one of the experts assigns labels such as "A or B" to 3 images, and b) in 45 cases[1], a family or higher level label is assigned. In the former case, we consider both labels mentioned; in the latter case, we consider all species under a higher level label as possible target labels. Thus it is possible that an image has multiple labels assigned by a single expert. In total 288 species and 20 families were mentioned as labels for the 190 images. (ii) Biologists do not always agree. Table 1 shows the agreement between biologists in terms of Cohen's κ[2], assuming the complete category set consists of all unique species mentioned in the labels provided by the experts. No perfect agreement was achieved, neither at species nor family level. This result suggests that our labeling task is not trivial even for experts.

Further, a post-labeling questionnaire with the experts reveals that some species are visually very similar and not distinguishable based on available information. For example, biologists normally distinguish Chromis Chrysura and Chromis Margaritifer by their body size, while in video footage the size of a fish depends on its distance to the camera, and therefore it is hard to distinguish them based on the observations made from the images/video footage.

3. LABELING WITH NON-EXPERTS

3.1 Interface

With the labeling interface presented in Fig. 1(a), it is very hard, if not impossible, for those who do not have knowledge with coral reef fish species to effectively provide labels. Therefore for non-experts a labeling interface as shown in Figure 1(b) is developed.

[1]A case is a {*image, expert label*} pair, thus 190x3 cases in total.

[2]When there exist multiple labels for an image assigned by one expert, we randomly draw one of them to be evaluated; we repeat this process 100 times and report the averaged κ and its standard deviation over the 100 runs. Agreement calculated in this way is rather conservative

The players are asked to compare a *query image*, i.e., the image to be labeled, to a set of *candidate labels*, i.e., textbook images of candidate species. They click a candidate label if they believe that the fish in that candidate label and the query image belong to the same species, or "others", if none of the candidates is similar enough to be considered as the correct answer. A feedback score for the chosen label is provided. Ideally, players can learn from the feedback and improve their performance.

The labeling process is divided into sessions of 50 query images. This gives a break as well as a goal for the players. It typically takes 5 to 10 minutes to complete a session. To avoid overloading the players with too many candidates, we limit the number of candidates to 7. To increase user engagement with the labeling task, we show the top 10 scorers. This competition element is meant to encourage people to achieve higher scores and play more sessions.

3.2 Experiment setup

3.2.1 Two simulated situations

During the manual clustering stage, we find that 53 out of the 190 images were assigned to "wrong" clusters. That is, there exist many fish that look similar but belong to different species. Our first research question thus boils down to *Can non-experts distinguish between similar species when examples of these species are displayed next to each other?* To find answers to this question, we conduct two experiments that simulate two situations.

Experiment 1. We first assume an ideal situation, where the *target label(s)* (labels suggested by the biologists) of the query image is always among the candidates. The primary goal of the experiment is to investigate whether the players can identify the target label when there exist very similar species.

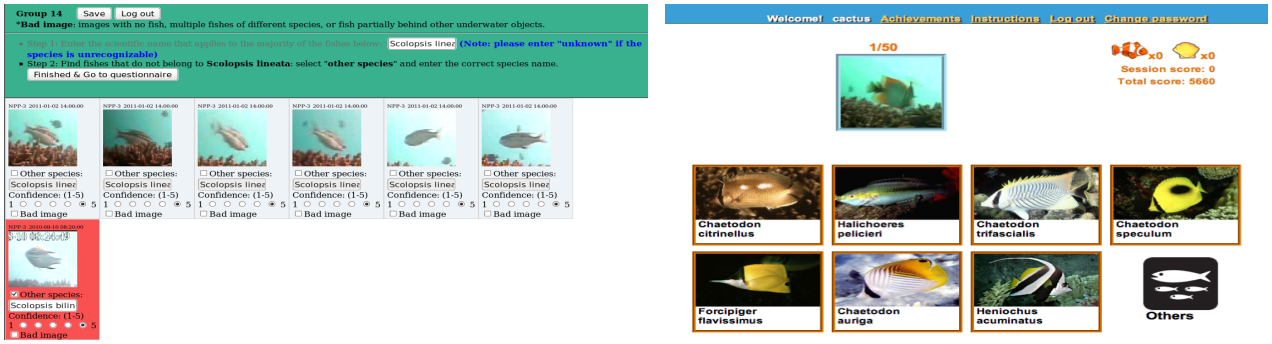
We select candidates that are similar to the target labels as follows. Let $c = \{i_n\}_{n=1}^{N}$ be a cluster containing N query images, and $f(i)$ maps an image to one of the species $S = \{s_m\}_{m=1}^{M}$. We compute a relevance score between an image $i \in c$ and a species as $\text{score}(i, s) = \text{count}(\{f(j) = s, j \in c\})/N$. All species with a non-zero score are the ones that were clustered together, which means that they are visually similar. We select top 7 species as candidates. If less than 7 species were available, we fill the remaining slots with random images. If more than 7 species have non-zero scores, we make sure that the target labels are in the candidates.

Experiment 2. We then consider a more realistic situation when some target labels are not in the candidates. In practice, we do not have information about the target labels of the query images. We need to select candidates based on certain similarity measures computed with automatic methods, which are most likely imperfect. It is then important to know whether the non-expert players can still make right choice, that is, select "others", when similar species are displayed as candidates. We use the same setting as in Expr.1 to select candidates and deliberately remove the target labels from the candidates for a set of randomly selected query images. Notice, if too many target labels are removed, users may expect that "others" is always the safe bet when they are not sure. With a few trial runs, we decide to remove the target labels for 25% of the query images.

3.2.2 System feedback

Our second research question was: *Can non-experts learn and improve their performance during the game?*

Players may be able to improve their performance for different reasons: learning from the system feedback, getting used to the quality of the images, etc. In this study, we do not aim to identify *why* and *how* users improve, but focus on whether they *can* learn and improve. Here we only consider the simplest and ideal system



(a) Species recognition interface for experts

(b) Game interface for non-experts

Figure 1: Expert and game interfaces for labeling fish species.

feedback, namely feedback from the expert labels. Specifically, we assign scores to each click on an candidate label based on the biologists' voting, i.e., a click on an option can receive 0, 1, 2, or 3 points. In practice when expert labels are not available, other types of feedback should be used, e.g., peer-agreement, automatic similarity measures. We leave questions such as how these feedbacks influence the user learning behavior to future investigation.

3.2.3 *Aggregation of obtained labels*

We use convenience sampling to collect players. We launched the game in our own social network and in public events, e.g., demo exhibitions. Our users have a diverse background and age groups, including school age children as well as university students and researchers. We collect labels for the 190 images labeled by the experts. 22 players contributed 72 sessions in Expr. 1 and 32 players contributed 49 sessions in Expr. 2. On average each image received 19 and 13 labels, respectively. Notice that in Expr. 2 we have more players but less sessions. This is because most of the sessions of Expr. 2 were done in a public event, where people typically try out for just one session. Four players have participated both experiments and in total played 9 sessions in Expr. 2. In our evaluation of Expr. 2, we will treat their contributions separately, as they may have been trained in Expr. 1 and their performance is not comparable with those who were new to the game.

To aggregate the labels from multiple players into a single label assignment for evaluation, we use a simple majority voting strategy. Since experts may give multiple labels to an image (as ground truth), we do not simply take the winner of the majority voting as the chosen label, but rank the candidates in descending order of their votes. In Expr. 2 when target labels are not displayed and "others" are *correctly* chosen, they do not provide information about which label should be assigned to the image. We ignore these cases for aggregating as they neither hurt or help the performance.

3.3 Evaluation

Quality of non-expert labels. We use Cohen's κ to measure the agreement between the aggregated non-expert labels and each of the three experts. We compare these to the pairwise agreement among the experts. When using majority voting, we take the top 1 candidate as the chosen label. In the case of ties, we use the same approach as described in Section 2 to calculate the agreement.

Further, NDCG [4] is used as it provides an intuitive interpretation of the correctness of the labels. For a query image, given the biologists' voting, each candidate can be rated as 0, 1, 2, or 3. The ranked list of candidates as a result of majority voting is then evaluated using these graded expert judgements.

Table 2: Agreement between experts and non-experts.

		E1		E2		E3	
		Avg. κ	Sdv.	Avg. κ	Sdv.	Avg. κ	Sdv.
Expr.1	Species	0.62	0.01	0.65	0.006	0.55	0.009
	Family	0.83	0.008	0.81	0.01	0.72	0.009
Expr.2	Species	0.65	0.009	0.50	0.008	0.45	0.009
(New)	Family	0.73	0.01	0.73	0.01	0.68	0.01
Expr.2	Species	0.53	0.01	0.68	0.01	0.64	0.02
(Old)	Family	0.80	0.02	0.78	0.02	0.74	0.01

Learning behavior of non-expert users. We study users' performance over time in terms of 1) memorization: when an image is shown again; and 2) generalization: when an unseen image that belongs to a seen species is shown.

We measure the performance of a single label as follows. Let $L = \{l_k\}_{k=1}^{K}$ be the candidate labels for an image, $J(l) = \{0, 1\}$ be a player's judgement, and $E(l) = \{0, 1, 2, 3\}$ be the expert votes of label l for the image. The performance of a judgement is defined as experts' votes for the chosen candidate normalized by the maximum votes one can achieve for the set of candidates: $s = \frac{\sum_{l \in L} J(l) \cdot E(l)}{\max_{l \in L} E(l)}$. Since scores achieved at a certain time point can be sensitive to players' random errors, we smooth the score at each time point with the scores achieved so far: $s_t = \sum_{i=1}^{t} s_i / t$. t refers to the tth time a player labels the same image (memorization), or a different image in the same species (generalization).

In a session, the first 12 images are randomly selected without repetition. After that, with a probability of 0.5 an image is selected from those that were already labeled in the current session. As images are selected randomly, the repetition of images (memorization) or species (generalization) do not happen the same number of times. In order to conduct reliable statistical testing for comparison (see Section 4), we consider repetitions of images/species that have more than 30 cases[3]. Specifically, we consider ≤ 4 repetitions of images for both experiments; ≤ 25 repetitions of species for Expr. 1, and ≤ 10 for Expr. 2. As fewer sessions were played in Expr. 2, less repetitions are available.

4. RESULTS AND DISCUSSION

Performance of non-experts. Table 2 shows the result of label agreement at both species and family level. In terms of Expr.1, if we compare Table 2 to Table 1, we see that the agreement between

[3]A case is a {*image(species), user*} pair.

Table 3: Non-experts' performance evaluated by NDCG. $^{\blacktriangle}(^{\blacktriangledown})$ indicates a significant difference (p-value< 0.01) tested using Wilcoxon signed-rank test.

Method	Species		Family	
	NDCG@1	NDCG@5	NDCG@1	NDCG@5
Expr.1	0.84	0.88	0.93	0.94
Expr.2.new	$0.72^{\blacktriangledown}$	$0.77^{\blacktriangledown}$	$0.86^{\blacktriangledown}$	0.94
Expr.2.old	0.88	0.86	0.91	0.94

Table 4: Comparing the performance in the first sessions under Expr. 1 and 2. Only "new" players are considered. Wilcoxon signed-rank test is used for significance testing.

Method	Species		Family	
	NDCG@1	NDCG@5	NDCG@1	NDCG@5
Expr.1	0.84	0.88	0.93	0.94
Expr.2	$0.72^{\blacktriangledown}$	$0.77^{\blacktriangledown}$	$0.86^{\blacktriangledown}$	0.94

Table 5: The impact of learning over time. Wilcoxon rank-sum test is used for significance testing. All comparisons are between the first label and the nth label.

	Memorizing				Generalization					
Labels	1	2	3	4	1	5	10	15	20	25
Expr.1	0.30	$0.38^{\blacktriangle}$	$0.46^{\blacktriangle}$	$0.51^{\blacktriangle}$	0.42	$0.51^{\blacktriangle}$	$0.59^{\blacktriangle}$	$0.63^{\blacktriangle}$	$0.67^{\blacktriangle}$	$0.70^{\blacktriangle}$
Expr.2.new	0.30	$0.40^{\blacktriangle}$	$0.44^{\blacktriangle}$	$0.52^{\blacktriangle}$	0.37	$0.58^{\blacktriangle}$	$0.62^{\blacktriangle}$	-	-	-

expert and non-expert labels are rather similar to that among the experts themselves. In terms of Expr.2, we see that the "new" players (those who only participated in Expr.2) achieve lower agreements with experts compared to players in Expr.1. while the performance of "old" players is comparable to that of Expr.1. This to some extent suggests that although the experimental condition has changed, the "training" the players received during Expr.1 has an influence on their performance in Expr.2.

Further, Table 3 shows the performance of non-expert labels in terms of NDCG. In practice, when using the collected labels as training data, often only the label(s) with the highest scores are considered as target labels. Thus it is important that the top ranked labels are correct according to experts' labels. We list the results of NDCG@1 and 5. Unlike the agreement comparison, here we do not have a baseline to compare to. However, we do see that the scores at least indicate that for a majority of the images, the non-experts have made correct choices. The new players in Expr.2 have a significant lower performance compared to Expr.1, while the performance of "old" users do not show significant difference compared to that achieved in Expr.1. We consider two potential explanations: 1) the set up of Expr.2 makes a more difficult task for novice players; or 2) since most of the new players did only one session, the general quality of the labels are not as good as that of Expr.1, where many played more than one session. To distinguish the two cases, we verify if the results from only the first session of each player in Expr.1 still outperform that of Expr.2. In Table 4 we see that indeed, a significant difference exists between the performance of the first session labels in the two experiments. That is, when target labels are absent while similar non-target labels are present, novice players are more likely to be confused. This suggests that selecting a good set of candidate labels is important.

Do non-experts learn? Table 5 shows the comparison of the averaged scores achieved at the first label for an image to that of the nth labels. These numbers confirms that there is a significant difference between the scores achieved with the first label and those achieved over time, in both experiments non-experts can learn and improve their labels over time. They do not only learn to provide more accurate labels for images that they have seen before, but also for similar images, i.e., different images that contain species that they have seen before.

5. CONCLUSION

We converted an image labeling task that requires extensive domain knowledge into an image matching game that is based on visual similarity comparison only. When the correct labels are always presented among the candidate labels, non-experts can play this game rather well: domain experts agree as often with the aggregated game labels as they agree with each other's labels. Users learn while playing to the extent that they perform better not only when they later see the same image again, but also when they later see different images from the same species. When the game is played under the more realistic condition that the correct label is not always presented, performance of novice users drops, but players that had played the game before still performed as good as under the ideal condition. Also under this condition, players still learned in terms of memorization and generalization.

A number of directions are left to be explored in the future. We used feedback from the experts, while in practice, the game will rely on automatic feedback or peer-agreement. The influence of feedback quality on users' performance and learning behavior is yet to be studied. Similarly, components within our labeling system such as the selection of candidates in practice will have to rely on automatic methods. While our user study have provided insights into how these components influence user performance, it remains unexplored how these should be integrated as a full fledged interactive system. Finally, we need to investigate how our approach can be extended to other domains such as medical image annotation.

Acknowledgements

This research was funded by European Commission FP7 grant 257024, in the Fish4Knowledge project (www.fish4knowledge.eu).

6. REFERENCES

[1] Y.-Y. Chen, W. H. Hsu, and H.-Y. M. Liao. Learning facial attributes by crowdsourcing in social media. In *WWW'11*, pages 25–26, 2011.

[2] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beene, A. Leaver-Fay, D. Baker, and Z. P. & Foldit players. Predicting protein structures with a multiplayer online game. *Nature*, pages 756 – 760, 2010.

[3] Galaxy Zoo. URL `http://www.galaxyzoo.org/`.

[4] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

[5] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int J Comput Vision*, 77(1-3):157–173, 2008.

[6] C. Spampinato, B. Boom, and J. He, editors. *VIGTA*, 2012.

[7] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI'04*, pages 319–326, 2004.

[8] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *CHI '06*, pages 55–64, 2006.

[9] J. Yuen, B. C. Russell, C. Liu, and A. Torralba. Labelme video: Building a video database with human annotations. In *ICCV*, pages 1451–1458, 2009.