Crowd-Powered Experts

Helping Surgeons Interpret Breast Cancer Images

Carsten Eickhoff
Department of Computer Science
ETH Zurich, Switzerland
ecarsten@inf.ethz.ch

ABSTRACT

Crowdsourcing is often applied for the task of replacing the scarce or expensive labour of experts with that of untrained workers. In this paper, we argue, that this objective might not always be desirable, but that we should instead aim at leveraging the considerable work force of the crowd in order to support the highly trained expert. In this paper, we demonstrate this different paradigm on the example of detecting malignant breast cancer in medical images. We compare the effectiveness and efficiency of experts to that of crowd workers, finding significantly better performance at greater cost. In a second series of experiments, we show how the comparably cheap results produced by crowdsourcing workers can serve to make experts more efficient *AND* more effective at the same time.

Categories and Subject Descriptors

Models and Principles [**User/Machine Systems**]: Human Information Processing

Keywords

Crowdsourcing; Experts; Image Annotation; Cancer Recognition; Breast Cancer.

1. INTRODUCTION

In recent years, crowdsourcing has been established as an integral component in many academic and industrial projects. Typical applications for the crowd include data collection, annotation or evaluation [8]. A topic of particular interest in the research community, is replacing expensive domain experts with untrained crowd labour (see for example [1] and [5]). For a wide array of tasks, the *Crowd-replaced Expert* paradigm has been shown to hold. When tasks are sufficiently simplified and broken down into atomic building blocks, the crowd's lack of professional training and extensive experience can be remedied by aggregating submissions across several workers. The objective that most of these applications optimise for is efficiency in terms of time and money at comparable quality.

On the other hand, there are approaches in which the crowd is used to support experts by reducing the decision space before experts start their work. Following the *Crowd-powered Expert* paradigm, the massive work force of the crowd can be combined with the professional's training and experience. A popular example is given by Harris' CV pre-screening [6], in which the author relies on crowd workers to sort and filter large numbers of applications submitted towards popular job openings. After this pre-processing, trained HR experts selected the most suitable candidates from the crowd-created short list. Many of the gamified protein folding [2] or medical structure annotation projects [9] can be counted in this class as well.

In this paper, we will for the first time explicitly compare the crowd-powered expert to the individual performances of crowd or expert. Concretely, we look at the classification of cell mass biopsy images. Based on this image material (see Figure 1), pathologists can determine the malignancy of tissue samples. The lab-based diagnosis is a lengthy process that involves careful inspection of many such samples. In this domain, misses and false alarms, understandably, introduce high emotional strain and potentially physical risk for the patient. In the course of this study, we try to employ a crowd of untrained workers to support medical experts.

Our investigation is guided by three fundamental research questions. **(RQ1)** Firstly, we suspect that there are task-inherent differences in how expert labour compares to crowd-sourced annotations. For the task at hand, we will contrast the performance of a single medical professional to that of an untrained crowd of varying size. **(RQ2)** Instead of replacing the expert, can we use crowd-labour in order to support medical experts such, that they reach greater accuracy or efficiency? **(RQ3)** Finally, we are interested in which way the crowd benefits from the same support that was previously given to domain experts.

2. METHODOLOGY

In the following, we will briefly introduce the data and procedures that underlie our experiments.

2.1 Classification of biopsy images

Pathological identification of cancer cells involves a multitude of factors. There are, however, a range of image-level indicators that make most cancer cells stand out among regular tissue [10]. (1) The **size of cells** tends to be homogeneous given a specific type of tissue. The presence

GamifIR '14, April 13 2014, Amsterdam, Netherlands

http://dx.doi.org/10.1145/2594776.2594788.

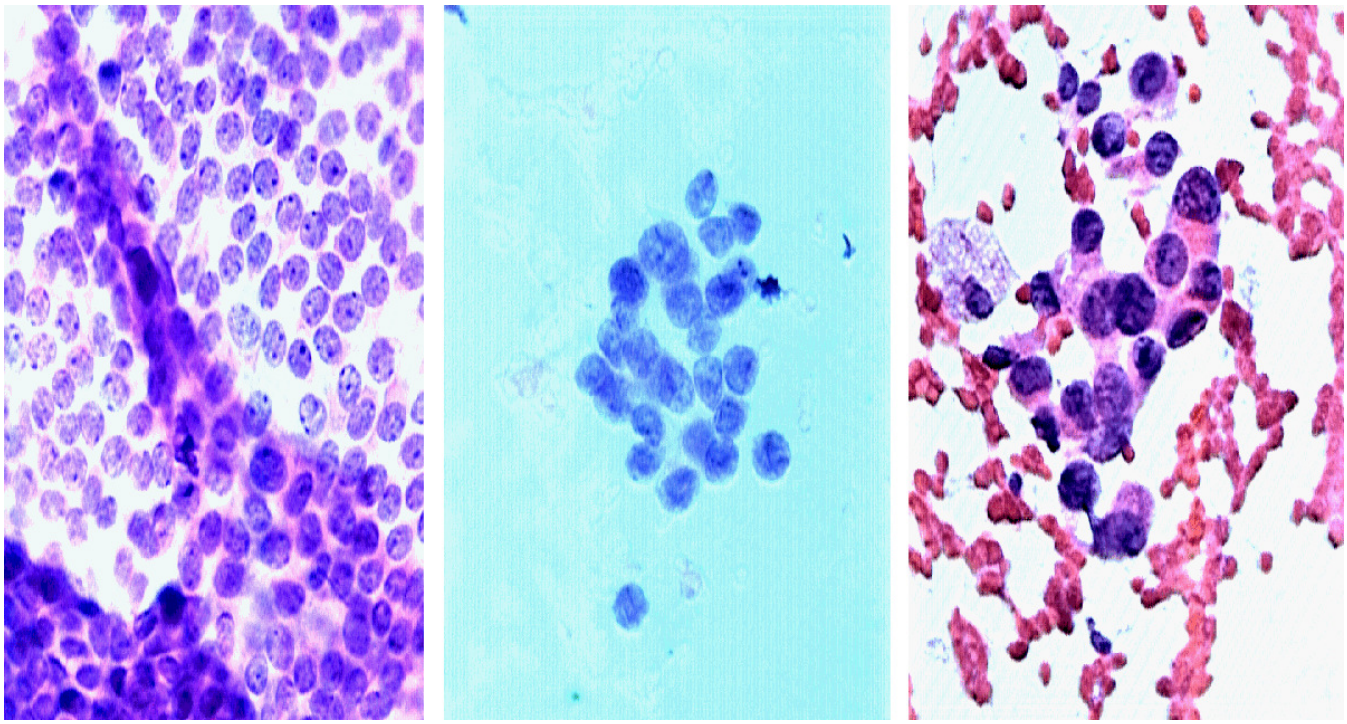

Figure 1: FNA biopsy samples of benign (left) and malignant (center and right) breast tumor cells.

of significantly larger cells is evidence for the uncontrolled growth that is indicative of malignant tumors. (2) Similarly, the **shape** of benign cells usually shows only limited variance, whereas malignant cells can develop arbitrary structures that do not conform with the general pattern of their surroundings. (3) The **color of the cell nucleus** should be identical for regular cells of the same type. Cancer cells often have significantly larger and darker nuclei that are more densely packed with DNA. (4) Regular cells show similar **texture**. Malignant tumors, on the other hand, can range from smooth surfaces to ragged or lumpy textures for neighbouring cells. (5) Finally, for healthy tissue, **cell arrangement** tends to be orderly, with regular distances between cells. Cancer cells can spread out or clutter almost arbitrarily.

2.2 Human computation tasks

In order to evaluate the validity of the *Crowd-powered Experts* paradigm, we propose a two-step human computation approach. In the first, optional, annotation stage, a crowd of untrained workers is tasked with annotating each biopsy image in terms of the previously presented low-level criteria. All criteria are based on the notion of uniformity of the depicted cells. We ask the crowd to rate this homogeneity along the 5 discussed dimensions on a 5-point scale ranging from 1 (random) to 5 (uniform). High values indicate benign nature of the relevant cell mass while low values can be seen as evidence for malignant growths.

In the subsequent classification step, medical experts are presented with the same, previously annotated, images and are asked to make the binary decision between benign and malignant tissue. In addition to the original images, they have access to the annotations made by the crowd. We hypothesise that, in this way, the expert's attention can be efficiently guided towards the tell-tale signs of malignancy without having to scan the entire image.

2.3 Dataset

Our experimental dataset is a collection of 569 biopsy images. They were taken from *fine needle aspirations* (FNA) of breast masses, created by Dr. William Wolberg at the University of Wisconsin Hospital [12]. For each image, there are known binary labels of benign (63%) and malignant (37%), that were established in lab-based diagnoses. This corpus is an established standard resource that has been used in a number of computer vision and machine learning studies, as well as in the medical literature. Figure 1 shows examples of benign and malignant FNA samples from the collection (sample ids: 925277, 916799 and 926682).

3. EXPERIMENTS

In this section, we will describe the set-up and outcome of our experiments. All HITs were offered on the crowdsourcing platforms Amazon Mechanical Turk (AMT) and CrowdFlower (CF) between January and February 2014. Unless explicitly stated otherwise, we did not note any statistically significant differences in the result quality and general behaviour of workers between platforms. The assignments were offered at a pay rate of $ 0.05 per image.

For both stages, annotation and classification, we created user interfaces on the respective platforms, trying to make them resemble each other as closely as possible given the individual platform restrictions and styles. Figure 2 shows an example of the resulting UIs for the first task on AMT. It should be noted, that, in response to the outcome of a pilot test, we dropped Criterion 3 from the annotation interface. Most images in our dataset do not grant a clear

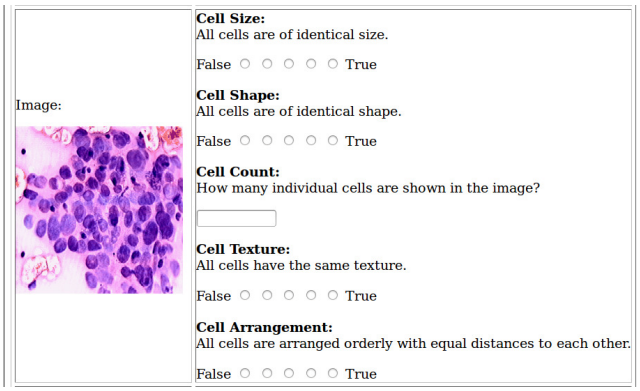

Figure 2: Annotation interface on AMT.

view on cell nuclei, which resulted in erratic and unreliable annotations along this dimension. Each set of 5 image annotations was accompanied by a brief survey to establish basic demographics of our crowd.

Previous work found significant amounts of inaccurate or automated submissions to crowdsourcing tasks and discussed a wide range of counter strategies [4]. We decided to aim for a simple method of ensuring the annotators attention by collecting multiple independent annotations for each image and including a question that asked the worker to count the cells in the current image. We reject all submissions for which the cell count deviates by more than 20% from the mean cell count across workers.

The classification interface for crowd workers contained a set of guidelines introducing the characteristics of cancer tissue as compared to regular cells. In this way, we tried to give the crowd a high-level understanding of how to distinguish benign from malignant cell mass. We recruited three medical experts who were remunerated at a rate of $80 per hour. This pay level appears to be representative of their professional field.

Our experimental set-up includes 2 key parameters: The classification C can be carried out either by an expert (C_E) or by a crowd of n workers (C_{C_n}), with n ranging from 1 to 5 who determine the final label in a majority voting scheme. Additionally, the classification can be aided by preliminary annotations A_m carried out by a crowd of m workers where $m = 1, 5, 10, 15, 20$. To accommodate for these individual experimental conditions without risking expert training effects, the dataset was divided into 6 stratified folds, each containing 94 images. Table 1 shows the results across all experimental conditions in terms of classification accuracy, time efficiency and cost efficiency. It should be noted that each time and cost item represents the total accumulated cost per image including crowdsourcing costs, expert salary and crowdsourcing platform overhead.

The different combinations of parameter settings aim at investigating the research questions proposed in Section 1. Statistical significance of gains and losses between crowd and experts were computed using a Wilcoxon signed rank test at $\alpha = 0.05$-level.

3.1 Experimental results

First of all, we were interested whether a group of untrained crowd workers would be able to replace a trained medical expert for the task of classifying biopsy images (Rows 1 to 6 in Table 1). We can note that the addition of further crowd judgements results in an initial accuracy gain which, however, stagnates at $n = 3$ workers. No combination of workers was able to match the trained professional's accuracy. Quite naturally, the monetary cost of crowd classification increases linearly in the number of workers. Due to the potentially parallel way in which crowdsourcing assignments can be submitted, there was only a mild increase in time per image when requiring a higher number redundant annotations. Finally, we took the test to the extreme for a small number of images (not shown in the table due to small sample size) and invested a crowdsourcing budget equal to the medical professional's effective rate per image. In this way, we collected $n = 21$ judgements per image, still failing to achieve the same level of accuracy as the medical professional. With respect to our first research question, this lets us conclude that the task at hand requires a degree of training that cannot easily be replaced by redundancy. Consequently, for all subsequent experiments, we will concentrate on the C_{C_3} case, for which we confirmed optimal crowd accuracy.

Our second research question was concerned with the effect that crowd-generated low-level annotations have on the accuracy of expert classification results (Rows 6, 8, ..., 14, 16). There seems to be a clear benefit in terms of accuracy as well as time efficiency when providing medical experts with low-level annotations generated by the crowd. While this improvement is not yet significant for annotations made by a single worker ($C_E + A_1$), crowds of 5 or more workers appear to be reliable enough to present significant merit, both in terms of accuracy as well as time efficiency. The speed-up that medical experts experience due to the introduction of low-level annotations serves for a mildly lower classification cost which even compensates for the crowdsourcing expenses. We did not note further statistically significant benefits for crowds of size $m > 5$. The optimum accuracy/cost cut seems to reside in the $2 \leq m \leq 9$ region. With respect to our second research question, we note that medical experts were able to perform their task faster and more reliably, when aided by crowd-powered pre-processing.

Finally, our third research question asked whether experts can be outperformed by crowds that have access to the previously generated low-level annotations. To make a fair comparison, we should finally also allow our crowd of annotators access to the output of the annotation step (Rows 6 to 16). As we can see, there is a significant benefit in using the crowd annotations. However, even in this setting, and for the best observed performance, the crowd was unable to achieve the same level of accuracy as a single unaided expert (C_E). Similarly as for the expert, the highest accuracy was achieved for $m = 5$ annotations per image. We conclude our third research question, much in the spirit of our previous findings with noting significant speed-ups and accuracy gains when adding low-level annotations (see RQ2). At the same time, there is still a significant accuracy gap between expert and untrained workers (see RQ1).

3.2 Crowd demographics

To conclude our experimental investigation of the *Crowd-powered Experts* paradigm, Table 2 shows a brief summary of the demographic information collected in the companion survey that accompanied our crowdsourcing assignments. By asking for their individual educational and medical back-

Table 1: Results of the classification task.

	Method	Accuracy	Time/image	Cost/image
1	C_{C_1}	0.72	40 sec	$ 0.05
2	C_{C_2}	0.74	40 sec	$ 0.10
3	C_{C_3}	0.77	42 sec	$ 0.15
4	C_{C_4}	0.76	43 sec	$ 0.20
5	C_{C_5}	0.77	45 sec	$ 0.25
6	C_E	0.92	58 sec	$ 1.29
7	$C_{C_3} + A_1$	0.82	37 sec	$ 0.20
8	$C_E + A_1$	0.94	55 sec	$ 1.22
9	$C_{C_3} + A_5$	0.84	32 sec	$ 0.40
10	$C_E + A_5$	0.97	48 sec	$ 1.07
11	$C_{C_3} + A_{10}$	0.83	30 sec	$ 0.65
12	$C_E + A_{10}$	0.96	49 sec	$ 1.09
13	$C_{C_3} + A_{15}$	0.85	31 sec	$ 0.80
14	$C_E + A_{15}$	0.98	48 sec	$ 1.07
15	$C_{C_3} + A_{20}$	0.82	30 sec	$ 1.15
16	$C_E + A_{20}$	0.97	50 sec	$ 1.11

Table 2: Crowd demographics.

	AMT	CF	Overall
Platform split	72%	28%	100%
Gender split M/F	43%/57%	39%/61%	42%/58%
Native speaker	62%	57%	61%
College education	73%	68%	72%
Medical training	4%	3%	4%

ground, we wanted to control for the workers' existing domain knowledge. The results largely follow the insights gained by existing studies such as [7] and [11]. A total of 389 individual workers contributed to our experiments. The dominant share is given by college-educated English native speakers. Approximately 4% of the participants had prior medical training of some sort. The gender split shows a gentle tendency towards female crowds.

4. CONCLUSION

Recently, many academic and industrial research groups have been showing positive results when trying to replace expensive domain experts by a large number of non-expert judges hired via crowdsourcing platforms on the Web. Consistently, their finding is that, given appropriate instructions and interfaces, a sufficiently large group of non-experts performs as well, or even better than, a single expert at much lower cost. In this paper, we postulate that this spirit of making experts redundant might not be suitable or even desirable in all settings. We argue that the crowd can be much more effectively used to enhance the experts' performance and efficiency.

To this end, we investigated the concrete use case of breast cancer image classification into benign and malignant cell mass. Our experiments show that for this task, the crowd was unable to outperform trained medical personnel in any of the investigated settings. However, when we tasked untrained workers with making low-level annotations, the experts' accuracy and efficiency can be increased at comparable or even lower cost.

There are several promising directions for future work. In this paper, we inspected a use case in the medical domain, which requires a high amount of professional training. Additionally, cancer detection and classification is a task with high stakes. Each false positive or negative involves significant mental distress or lethal health risks. This is admittedly a beneficial setting for the *Crowd-powered Experts* paradigm. It would be interesting to investigate how well our findings and implications hold for different domains, such as assessing document relevance in TREC-like manner.

In this paper, we investigated a two-tier interaction between crowd and experts. Depending on the task and problem domain, it may however be interesting to introduce further layers of hierarchy in which workers and tasks interact. The output of each step is subsequently consumed by further human or automatic computation. With complex processing chains like this in mind, it will be interesting to phrase expertise on a more fine-grained scale than the binary expert/crowd separation that we regarded in this initial study. This could for example be done in the spirit of [3].

5. REFERENCES

[1] Omar Alonso and Stefano Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

[2] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307), 2010.

[3] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Pick-a-crowd: Tell me what you like, and i'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web*, pages 367–374. ACM, 2013.

[4] Carsten Eickhoff and Arjen P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, pages 1–17, 2013.

[5] Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880. ACM, 2012.

[6] Christopher G. Harris. You're hired! an examination of crowdsourcing incentive models in human resource tasks. In *WSDM Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, pages 15–18, 2011.

[7] Panagiotis G. Ipeirotis. The new Demographics of Mechanical Turk. `http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html`, 2010.

[8] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[9] Vivien Marx. Neuroscience waves to the crowd. *Nature methods*, 10(11):1069–1074, 2013.

[10] Roswitha Pfragner and R. Ian Freshney. *Culture of Human Tumor Cells*. Culture of Specialized Cells. Wiley, 2003.

[11] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. ACM, 2010.

[12] William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian. Breast cytology diagnosis via digital image analysis. *Analytical and Quantitative Cytology and Histology*, 15(6):396–404, 1993.