ICORD: Intelligent Collection of Redundant Data – A Dynamic System for Crowdsourcing Cell Segmentations Accurately and Efficiently

Mehrnoosh Sameki
Boston University
sameki@bu.edu

Danna Gurari
University of Texas at Austin
dgurari@cs.utexas.edu

Margrit Betke
Boston University
betke@bu.edu

Abstract

Segmentation is a fundamental step in analyzing biological structures in microscopy images. When state-of-the-art automated methods are found to produce inaccurate boundaries, interactive segmentation can be effective. Since the inclusion of domain experts is typically expensive and does not scale, crowdsourcing has been considered. Due to concerns about the quality of crowd work, quality control methods that rely on a fixed number of redundant annotations have been used. We here introduce a collection strategy that dynamically *assesses the quality of crowd work. We propose ICORD (Intelligent Collection Of Redundant annotation Data), a system that predicts the accuracy of a segmented region from analysis of (1) its geometric and intensity-based features and (2) the crowd worker's behavioral features. Based on this score, ICORD dynamically determines if the annotation accuracy is satisfactory or if a higher-quality annotation should be sought out in another round of crowdsourcing. We tested ICORD on phase contrast and fluorescence images of 270 cells. We compared the performance of ICORD and a popular baseline method for which we aggregated 1,350 crowd-drawn cell segmentations. Our results show that ICORD collects annotations both accurately and efficiently. Accuracy levels are within 3 percentage points of those of the baseline. More importantly, due to its dynamic nature, ICORD vastly outperforms the baseline method with respect to efficiency. ICORD only uses between 27% and 50% of the resources, i.e., collection time and cost, that the baseline method requires.*

1. Introduction

High-throughput microscopy technology enables researchers to produce large numbers of images of cells that must be segmented for further analysis [16]. Over the past decades, many automatic and interactive segmentation algorithms have been proposed (e.g., [5, 14, 20, 28, 29]). Finding a one-size-fits-all algorithm, however, that works well for segmenting cells with simple and complex boundaries, as shown in **Figure** 1, is a challenging task. An alternative option is to leverage crowdsourcing and design a "human-in-the-loop" solution. In this paper, we show how this option can be made scalable using computer vision and machine learning techniques.

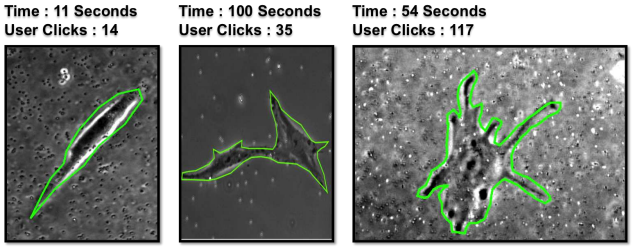

Figure 1. Given the outline of a cell in a microscopy image drawn by a crowd worker, how can we automatically determine the quality of this segmentation? Does, for example, the time a crowd worker takes to trace the outline of the cell correlate with segmentation quality? Or the number of mouse clicks the worker makes to create the cell boundary? Do automatically extracted shape and intensity features of the segmented region correlate with segmentation quality?

Researchers from communities like human computer interaction [10], computer vision [13, 18], computer graphics [2], multimedia [7], and bioinformatics [8] have proposed a variety of approaches for offloading labor-intensive image segmentation tasks to crowd workers. However, a clear message emerges from the literature: Crowd work is not reliable and thus needs redundant data collection (e.g., 5 workers per task [9]). In one study as much as 32% of annotations obtained from internet workers had to be discarded [2].

In this paper, we address the question: What automated mechanism should be applied to ensure the efficient collection of high-quality cell segmentations by the crowd? The question how to crowdsource high-quality segmentations has been asked for images of "everyday objects" such as birds or cars photographed with visible-light cameras [19]. Few works have considered the crowdsourcing of cells in microscopy imagery [9]. Consequently, little is known about how to specifically collect high-quality segmentations of cells from internet workers.

For general crowdsourcing tasks, i.e., not necessarily involving image analysis, a common approach to quality control is to aggregate a fixed number of annotations from multiple workers and then apply an aggregation policy. An example of a simple aggregation policy is majority voting [19]. Alternatively, the influence of different responses by crowd workers can be weighed by the level of expertise and interest of the worker [17]. Learning algorithms based on the Expectation-Maximization (EM) method [6] have been designed that evaluate the performance of multiple workers in the absence of ground truth by iteratively measuring the performance of annotators and using these measurements to refine estimates of the ground truth [15, 19, 24, 26, 27]. Task difficulty can also been estimated [27], as well as annotator bias [25]. Aggregation methods yield higher quality results yet introduce costs and delays that we can avoid with our proposed method, which dynamically determines how many annotations to aggregate.

Verification of the quality of crowd work by the crowd is an effective strategy for yielding higher quality work. As with aggregation approaches, this strategy also comes at the expense of additional costs and delays. According to this strategy, crowd workers are asked to vote whether submitted crowd work is sufficiently accurate. These votes are then used to decide whether to keep or discard crowd work [12, 22]. Our proposed method avoids the second verification round for annotation results that it deems sufficiently accurate. Only for crowd-submitted annotations that it flags to be potentially inaccurate, our method requires additional annotations from the crowd.

Related work aim to automatically infer the quality of segmentations created by algorithms [4, 1]. For example, Kohlberger et al. [11] used nonlinear regression to predict segmentation error in CT images of the lung, liver, and other organs. Both works did not test their proposed regressors in a crowd setting. Our method is also based on a regression model. It differs from the above approaches by analyzing behavioral cues of the internet worker who creates the annotation.

The first behavioral cue that our prediction method considers is effort. When internet workers annotate the boundary of a cell with the software we provide, they select a series of points that the software connects sequentially with straight lines to create a closed polygon around the cell, as shown in **Figure** 1. The selection of each point is performed by the right-button click of the mouse. Our method trains workers by providing instructions on how to accurately draw the outline of example cells. However, they have to make choices themselves about how many points to click in order to accurately capture the details of a cell with complicated protrusions (or how few points are needed to define the outline of a round cell).

The second behavioral cue that our prediction method considers is annotation time. Inaccurate segmentations may occur when a crowd worker is uncertain how a cell should be separated from the background or from other cells. The worker's uncertainty may result in hesitation and slower annotation time.

For image segmentations performed by the crowd, existing literature reports the crowd worker's time and effort, which are the time a worker spends to draw a boundary [3, 23] and the number of clicks a worker makes to demarcate the boundary [2, 18, 21]. These works, however, did not investigate, as we do, if there is any correlation between these behavioral cues and the quality of segmentations. Our work complements existing efforts by demonstrating the value of predictive models for cell segmentation using both behavioral features and image features.

We first created training data by collecting segmentations from crowd workers, comparing them to expert-drawn segmentations, and computing a quality label for each crowd-drawn segmentation. We then developed a prediction model using both image features and behavioral cues to predict the quality of crowd work for microscopy images. We used our prediction model to propose ICORD (Intelligent Collection Of Redundant Annotation Data), an intelligent system that incorporates the predicted scores into a dynamic platform to detect whether a collected segmentation is sufficiently accurate to be used as a final result. If a segmentation is not deemed accurate, ICORD sends the image back to the crowdsourcing platform to collect an additional outline. We compared the performance of ICORD with that of two other baseline strategies. Our results show its effectiveness in terms of accuracy and, most significantly, efficiency.
In summary, our contributions are as follows:

- We propose a dynamic system for crowdsourcing redundant segmentation data, called ICORD.
- Our experiments involved five rounds of crowdsourcing which produced a total of 1,350 segmentations for 270 phase contrast and fluorescence images of cells.
- Our results demonstrate that analysis of behavioral cues of the crowd worker, augmented by analysis of image features, can be used to infer segmentation quality dynamically. ICORD predicts the accuracy of a crowd-drawn cell outline and determines the need to seek additional annotations.
- Comparisons with two baseline crowdsourcing strategies show that ICORD collects annotations from the crowd intelligently by effectively balancing annotation accuracy and collection efficiency.

2. Training ICORD

We first describe how we generated labeled training data for ICORD: We obtain the data by developing a crowdsourcing system and collecting redundant annotations of

cell boundaries, and the labels by computing a quality score for each annotation (Sec. 2.1). We then define the features that ICORD uses to predict the accuracy of a crowd-drawn cell segmentation in the absence of a quality label (Sec. 2.2). Finally, we describe the prediction model ICORD uses to learn the relationship between segmentation quality and extracted features (Sec. 2.3).

2.1. Training Data Generation

To capture a range of possible segmentation tasks and difficulty levels, we selected the training data for ICORD to involve crowd-drawn outlines of a variety of cells (smooth muscle, fibroblast, and melanoma) imaged with two modalities (phase contrast and fluorescence microscopy). The image data is described in Section 3. Each image contains one cell.

Annotation Tool. To collect crowd-drawn segmentations, we configured the freely-available source code for the online image annotation tool LabelMe [18] to run in the Amazon Mechanical Turk (AMT) Internet marketplace (**Figure 2**). Workers trace the boundary of a cell by clicking on points in the image. LabelMe connects consecutive points with straight lines. Workers complete the segmentation of the cell by clicking on the first point of the boundary to create a closed polygon. Workers have the option to delete and redraw the cell boundary, in case they made a mistake. To support the annotation effort, LabelMe automatically enlargens the display of an image to span the maximum possible width and/or height of the allotted space in the worker's browser window (while maintaining image resolution and proportions). We release our configuration of the LabelMe drawing environment for ATM with step-by-step instructions that explain how to set it up and connect it to AMT (`http://Anonymous`).

Annotation Instructions. Before a crowd worker on AMT could accept our posted Human Intelligence Tasks (HITs), he/she was shown our five-step set of instructions, in English, followed by pictures exemplifying accurate and inaccurate annotations to clarify the aim of the task (**Figure 3**).

Measuring the Quality of Crowd Annotations. Our system measures the quality of crowd segmentations by estimating the similarity of each crowd segmentation to a gold-standard segmentation provided by a domain expert. We use the Jaccard index to measure how closely two segmented regions resemble each other. The index computes the ratio of the number of pixels common to two segmented regions to the number of pixels in the union of both regions, i.e., $\frac{|A \cap B|}{|A \cup B|}$, where A represents the set of pixels in the crowd-segmented region and B represents the set of pixels in the expert-segmented region. Resulting scores range from 0 to 1 with larger values indicating greater similarity between the two regions.

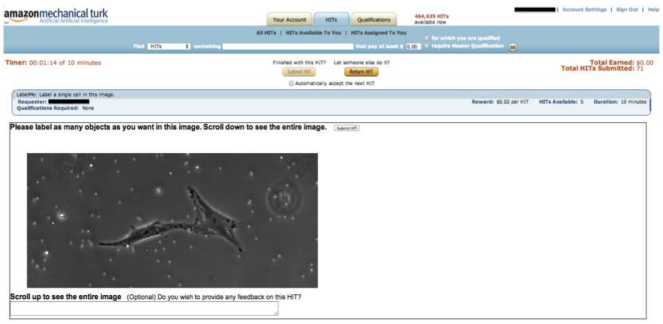

Figure 2. Our drawing interface enables crowd workers on Amazon Mechanical Turk to draw polygonal cell outlines with LabelMe [18].

Obtaining Training Data and Labels. Using our annotation tool and instructions, we collected five crowdsourced segmentations per cell image. Each segmentation represents a labeled training instance for our ICORD system. The training label is the quality score of the segmentation (i.e., the Jaccard score measuring its similarity to an expert-drawn segmentation).

2.2. Prediction Features

We propose three categories of features to describe crowdsourced cell segmentations: geometric features, intensity-based features, and the crowd worker's behavioral features. The first two categories of features were obtained by using the crowd-drawn boundaries of the cells and extracting features on the cell foreground and background. The third category was extracted from the post-task statistics that AMT provides. We standardized features by removing the mean and scaling to unit variance.

2.2.1 Geometric Image Features

For each cell image, our method uses the crowd-drawn boundary to mask out the background, i.e., the portion of the image which was not part of the cell, and extracts six geometric features:

Area. Number of pixels within the cell interior.

Convex Area. Number of pixels within the smallest convex polygon that contains the cell interior.

Perimeter. Number of pixels on the crowd-drawn boundary of the cell.

Euler Number. Number of annotated regions in the foreground minus the number of holes within these regions (should be 1 if worker annotates cell correctly).

Orientation. The angle between the x-axis and the major axis of the ellipse that has the same second-moments as the cell region.

Solidity. Number of pixels in the convex hull that are also in the cell interior, i.e., (area/convex area).

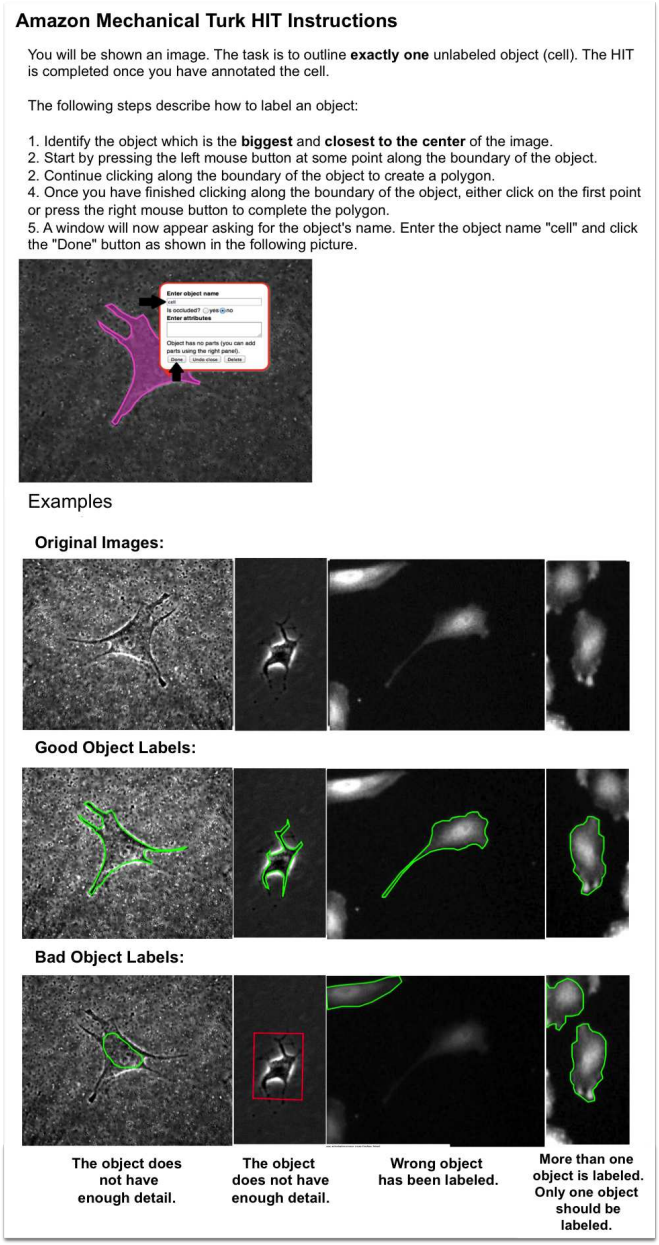

Figure 3. Crowdsourcing drawing instructions. The two examples on the left are cells in phase contrast microscopy images, the two on the right in fluorescence.

2.2.2 Image Intensity Features

Our method computes eight features of each cell image that are based on analyzing its intensity:

Average Gray-scale Value of Drawn Cell Region.

Average Gray-scale Value of the Background.

Intensity Separability. The difference between the intensity averages of foreground and background (cell interior and exterior).

Average Contrast of Cell Pixels. The standard deviation (σ) of the intensity of the foreground pixels.

Intensity Smoothness of the Cell Region. $1 - 1/(1+\sigma^2)$.

Skewness of the Intensity Distribution. Third moment that specifies how asymmetric the intensity histogram is.

Uniformity Measure. Sum of the squared number of pixels p_i in each bin i of the intensity histogram of the cell region.

Entropy. $-\sum_i p_i \log p_i$.

2.2.3 Behavioral Features

We use the following three features to characterize the drawing behavior of a crowd worker:

Time per Task. Lapsed time for each completed HIT, from the time a worker clicks the "Accept HIT" button through the time the worker clicks the "Submit HIT" button.

Number of Clicks. Number of points that the worker clicks on the image to delineate the boundary of an object.

Average Time per Click. In order to normalize against the effect of boundary complexity (e.g., a circular cell versus a cell with many protrusions), we compute the average time per click (i.e., Time per Task / Number of Clicks).

2.3. Creating a Prediction Model

We next propose a framework to learn a model that predicts the quality of a given crowd segmentation based on the three categories of features described above. We chose a regression model to capture the continuous nature of our measure of segmentation quality. Specifically, we performed supervised learning by training a random forest regression model to determine whether the extracted features can be predictive of the quality of crowd-drawn annotations. This regression model exploits individual regression trees and grows a forest of many trees. Candidate splits at a node are chosen based on minimizing an impurity measure, the sum of the squared deviation from the leaf mean. The training procedure can be summarized as follows:

1. Draw N bootstrap samples from the training data randomly with replacement.

2. Grow a regression tree for each bootstrap sample.

3. At each node, randomly sample m predictors at random out of all M possible variables and choose the best split using only the selected predictors.

4. Aggregate the predictions of the N trees by averaging the responses of the trees and use this aggregate make a prediction for the new test data.

As described below, we used this procedure to create several regression models in order to evaluate the predictive power of various combinations of features. Our final system, ICORD, uses the best performing model.

3. Testing the Use of Prediction

We conducted three studies using the proposed prediction approach to answer: 1) Can we predict the quality of a given crowd-drawn cell segmentation? 2) Does accuracy improve if we train and test a prediction model on images from only one image modality? 3) If we train a model on

one image modality, can we use this model to predict quality for other image modalities?

3.1. Datasets

We used a freely available image library [9] that includes 151 phase contrast microscopy images showing rat and rabbit smooth muscle cells and mouse fibroblasts. The dataset also contains 119 fluorescence microscopy images of Lu melanoma cells and WM993 melanoma cells. The dataset consists of raw images and expert-drawn annotations to be used as pixel-level-accurate ground-truth segmentations.

3.2. Collection of Crowdsourced Segmentations

We recruited crowdsourced workers through AMT and accepted all workers who had previously completed 100 HITs and maintained at least a 92% approval rating. We paid workers \$0.02 upon completion of each object segmentation task and approved all submitted HITs. We allotted a maximum of ten minutes to complete the task. In total, 40 unique workers created our 1,350 collected segmentations (i.e., 5 crowd segmentations x 270 images).

3.3. Evaluation of Prediction Models

We analyzed the predictive power of our proposed regression models by comparing predicted and observed segmentation quality scores. Our measures for comparison are the Pearson's correlation coefficient r ($-1 \leq r \leq 1$), the coefficient of determination R^2 ($0 \leq R^2 \leq 1$) and the mean absolute error (MAE) between predicted and observed Jaccard index ($0 \leq \mathrm{MAE} \leq 1$).

3.4. Study 1: Which Feature Combination?

In this study, we used 5-fold cross-validation to train and test our regression model. Specifically, we randomly partitioned all 1,350 segmentations into 5 independent sets of equal size with all 5 segmentations of each unique image in the same fold. For each of 5 iterations, a different set was reserved as the test set and the combination of the remaining sets were the training set. We used the predictions for all crowd segmentations collected from the 5 partitions to evaluate the quality of the model.

We first examined whether the quality of crowd segmentations may be inferred based on crowd workers' number of clicks or time to annotate. Specifically, we trained two prediction models independently based on these two features using all 1,350 crowd segmentations from both images modalities. This study reveals a moderate correlation ($r = 0.52$) between human annotation behavior and segmentation quality when all three behavioral features are included (**Table** 1, row 4). The number of points a worker clicks to define a cell contour was the best predictor of segmentation quality among the behavioral features ($r = 0.59$ in **Table** 1, row 2).

We next trained three prediction models, the first using only the extracted geometric features from the foreground and background of the segmented regions, the second using only the intensity features, and the third using both (**Table** 1, rows 5–7). Our results illustrate that there is a strong correlation between geometric features and accuracy of segmentations ($r = 0.66$), between intensity features and accuracy ($r = 0.73$), and between both feature categories and accuracy ($r = 81$). This indicates that collected crowd annotations can be used as masks to extract static image features that are promising for predicting the quality of annotations.

Table 1. Evaluation of the predictive power of 8 combinations of features. Predicted and measured quality scores (Jaccard overlap index) are compared with the correlation coefficient (r), the coefficient of determination (R^2), and the mean absolute error (MAE). The regression model trained on all features is most predictive.

Regression model based on	r	R^2	MAE
1. Time	0.52	0.29	0.06
2. Number of clicks	0.59	0.42	0.05
3. Time per click	0.53	0.24	0.06
4. Only behavioral features	0.52	0.26	0.06
5. Only geometric features	0.66	0.44	0.05
6. Only intensity features	0.73	0.52	0.05
7. All image features	0.81	0.29	0.06
8. All features	**0.83**	**0.69**	**0.04**

Finally, we considered all three groups of features (geometry and intensity of the cell region and behavioral clues) to train a prediction model. The correlation coefficient improved to be the top predictor of annotation quality, $r = 0.83$ (**Table** 1, row 8). Similarly, the coefficient of determination R^2 was the highest for this regressor and the mean absolute error the lowest.

Overall, the results of study 1 demonstrate that image and behavioral features can be combined to train an accurate model, and this model can be used to predict the quality of crowdsourced annotations in the absence of ground truth.

3.5. Study 2: Per Modality Evaluation

In the previous study, we trained our prediction models with all the cell images in our library, irrespective of the imaging modality. In this study, we trained prediction models separately for phase contrast and fluorescence microscopy images. This study was motivated by the facts that an end user of ICORD would typically only have data collected by one modality and the visual appearance of cells in fluorescence versus phase contrast datasets differs. Phase contrast images show cells with more complicated boundaries than the fluorescence images, e.g., **Figure 3** bottom. Cell boundary protrusions, e.g., lamellipodia or filopodia, seen in the phase contrast images, are difficult to trace. Moreover, we observed a large range of gray-scale values within the area of the cells in phase contrast images, while cells in fluorescence images have gray-scale values more distinguishable from the background.

We split the data of each modality into training and test-

Table 2. Evaluation of the predictive power of models trained and tested separately on phase contrast (PC) and fluorescence (Fl) images. Training involves all image and behavioral features.

Train/ Test	r	R^2	MAE
1. PC / PC	0.76	0.58	0.05
2. Fl / Fl	0.90	0.81	0.02
3. PC / Fl	0.54	-0.68	0.12
4. Fl / PC	0.44	-0.10	0.09

ing data using 5-fold cross validation and trained a random forest regression model for each modality using both image and behavioral features. For the first model, we had 755 (151×5) crowd-drawn boundaries of cells in phase contrast microscopy images; for the second model, 595 (119×5) in fluorescence images. When we analyzed the predictive power of the two regression models by comparing predicted and observed segmentation quality scores, we found high correlations (**Table** 2, rows 1 and 2) for both. As we had anticipated, it is beneficial to train and test on the same modality, particularly for cells imaged by fluorescence microscopy ($r = 0.9$). Training and testing on phase contrast images yielded a correlation coefficient of $r = 0.76$ (**Table** 2, row 2), slightly lower than the coefficient of $r = 0.83$ we measured for regression model that was trained and tested by all images in our database (**Table** 1, row 8).

3.6. Study 3: Cross Modality Evaluation

In our third study, we investigated the effect of cross modality training and testing on prediction results. We first trained a regression model using all features on 755 phase contrast images, and then tested the model on 595 fluorescence images using the correlation coefficient r, coefficient of determination R^2, and the mean absolute error to compare predicted and observed accuracy scores. We repeated the experiment with training on fluorescence images and testing on phase contrast images.

Despite moderately strong correlation coefficients for both cases (0.54 and 0.44), the R^2 values were negative (arbitrarily worse), illustrating that the data is not able to fit the model accurately (**Table** 2, rows 3 and 4). These results suggest that there is a limited power in using one image modality for training and another dataset with different image modality for testing.

4. ICORD: When to Collect Redundant Data? Automatically Balancing Efficiency and Accuracy

As we described in Section 1, the collection of redundant annotations is widely recommended for crowdsourcing because the response of a single crowd worker is not deemed trustworthy [19, 24]. Quality control mechanisms that aggregate annotations from multiple workers can yield higher quality results, yet introduce additional costs and delays. Often ad hoc decisions are made by the designer of the crowdsourcing system about how to balance collection efficiency and annotation accuracy. For example, the designer may decide that 5 crowd workers are needed to ensure a sufficient accuracy level, but more than 5 would result in a collection effort that is too inefficient and costly [9].

We argue that ad hoc decisions on how to set up crowdsourcing of cell annotations does not scale to the large datasets produced by modern microscopy technology. An automated process is needed that trades off annotation collection efficiency and accuracy. We here propose ICORD, a system that collects annotations by the crowd by effectively balancing annotation accuracy and collection efficiency. ICORD predicts the quality of crowd work and decides when to collect additional data or when to stop and trust the results of crowd workers:

ICORD Process for Cell Segmentation:

Input: Raw images of cells, quality threshold τ, number of rounds N.

1. A single round of crowdsourcing is performed on all cell images. One segmentation is obtained per cell.
2. Crowd segmentations are converted to binary masks, and image and behavioral features are extracted.
3. The prediction model receives the feature vectors and evaluates the quality of each segmentation.
4. For each cell: If the predicted score is higher than threshold τ, the system accepts the annotation (step 7). Otherwise, the annotation is flagged as inaccurate (step 6).
6. Repeat until all cell segmentations are predicted to be accurate or N crowdsourcing rounds have been performed:
   6.1 A new round of crowdsourcing is performed on the cell images with annotations flagged as inaccurate.
   6.2 Steps 2.-4. are applied to the current segmentation.
7. For any cells still predicted to have inaccurate segmentations, the segmentation among the N collected is chosen that has highest predicted quality.

Output: Cell annotations and their predicted quality scores.

An example annotation collection process for a phase contrast image with the ICORD system with a threshold of $\tau = 0.75$ is shown in **Figure 4**. Here, after $N = 5$ rounds of collecting crowdsourced segmentations, the prediction score exceeds τ and the process stops. In another example, shown in **Figure 5**, ICORD deemed the cell segmentation obtained after the second round sufficiently accurate. By not requiring a fixed number of crowdsourcing rounds for each image, ICORD prevents collecting unnecessary data. We tested ICORD with various threshold values and up to $N = 5$ rounds of crowdsourcing.

Our experimentation showed that the performance of ICORD is sensitive to the cutoff threshold τ. If the selected threshold is too low, very few instances are selected and

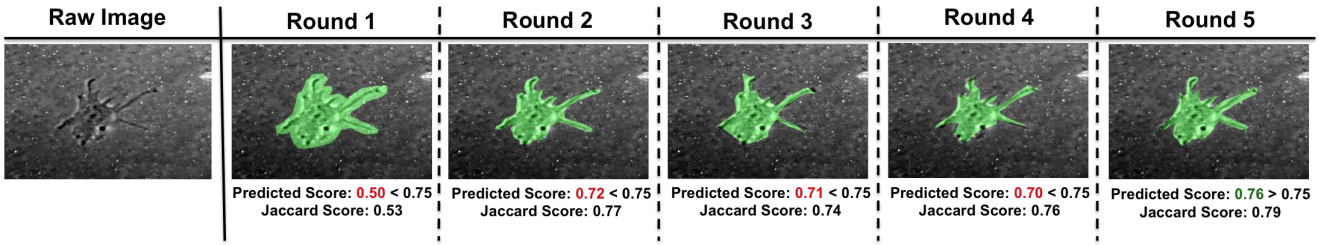

Figure 4. An example processed by ICORD: A phase contrast image of a cell and its segmentations, produced by crowd workers in 5 rounds. In rounds 1–4, the prediction model flagged the segmentations as not sufficiently accurate (quality score below threshold $\tau = 0.75$). In round 5, ICORD predicts that the shown segmentation is accurate (score > 0.75) and terminates the processing on this cell. For each round, the Jaccard scores measuring the overlap between expert-drawn and crowd-worker-drawn regions are also displayed (observed and predicted scores only differ by 6 or fewer percentage points).

sent to the next crowdsourcing round. If the threshold is too high, almost all of the instances will be sent to the next round, making the prediction model almost superfluous. A reasonable choice for a threshold that can be computed automatically is taking the average predicted score of the annotations obtained in the first round.

We compared the performance of ICORD to two baseline collection processes. The first is called Perfect Oracle Baseline and uses the same algorithm as ICORD except for step 6.1, where a new round of crowdsourcing is performed on the cell images with annotations that are flagged to be inaccurate. In ICORD, the inaccuracy flag is based on the prediction model. In the Perfect Oracle Baseline, however, the inaccuracy flag is based on ground-truth knowledge. The reason we designed the baseline so that a perfect oracle provides the inaccuracy score is that the performance of the ICORD framework can be tested irrespective of the "false negative detection rate" of the prediction model, i.e., the ability of the regressor to flag inaccurate outlines.

The second baseline process, called Fusion Baseline, involves combining multiple crowd annotations as is standard practice in crowdsourcing (e.g. [9]).

Fusion Baseline for Cell Segmentation:

Input: Raw images of cells, number of rounds N, aggregation number M.

1. A single round of crowdsourcing is performed on all cell images. One segmentation is obtained per cell.

2. Repeat for N crowdsourcing rounds:

2.1 A new round of crowdsourcing is performed on all cell images.

2.2 Any existing segmentations are combined with the most recently collected segmentation as follows: If a pixel is labeled as part of the cell for at least M segmentations, it is assigned to be in the combined new segmentation.

Output: Cell annotations.

The accuracy scores averaged for all fluorescence and all phase contrast images, respectively, are shown for ICORD and the two baseline processes per crowdsourcing round in **Figure 6**. Two thresholds were selected automatically per imaging modality. The average predicted score of the first batch of annotations was taken as one threshold (0.82 for fluorescence and 0.75 for phase contrast), and this score minus 0.05 as the other threshold.

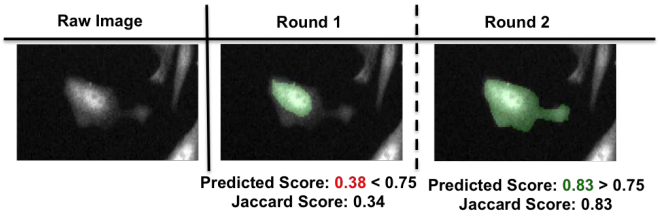

Figure 5. An example processed by ICORD involving a cell on a fluorescence microscopy image. ICORD detects in the second round that the outline is sufficiently accurate to be considered a final product ($\tau = 0.83$).

As can be observed in **Figure 6**, sending the suspicious instances to a second round of crowdsourcing increases the average Jaccard score of the new set for all of the cases. This pattern was observed for both phase contrast and fluorescence image sets and with all tested thresholds. For instance, crowdsourcing the predicted inaccurate annotations of fluorescence increased the average Jaccard accuracy from 0.79 to 0.83 for threshold of 0.85 images (**Figure 6**(a), group 2). A third round of collecting redundant data is helpful in 3 of the shown 4 groups of experiments, slightly increasing the average segmentation accuracy. The last two rounds do not help improve the average Jaccard score much.

Comparing the results of ICORD to those of the Fusion method demonstrates that for both modalities and thresholds accuracy levels are similar. ICORD outperforms the Fusion method in the second round with regard to accuracy by up to 5 percentage points; the Fusion method outperforms ICORD in later rounds by up to 3 percentage points.

More importantly, with regards to efficiency, ICORD vastly outperforms the Fusion method because it requires significantly fewer collections of annotations. Using a lower threshold, the fusion method requires 3.7 times more collections than ICORD for fluorescence imaging and 3.0

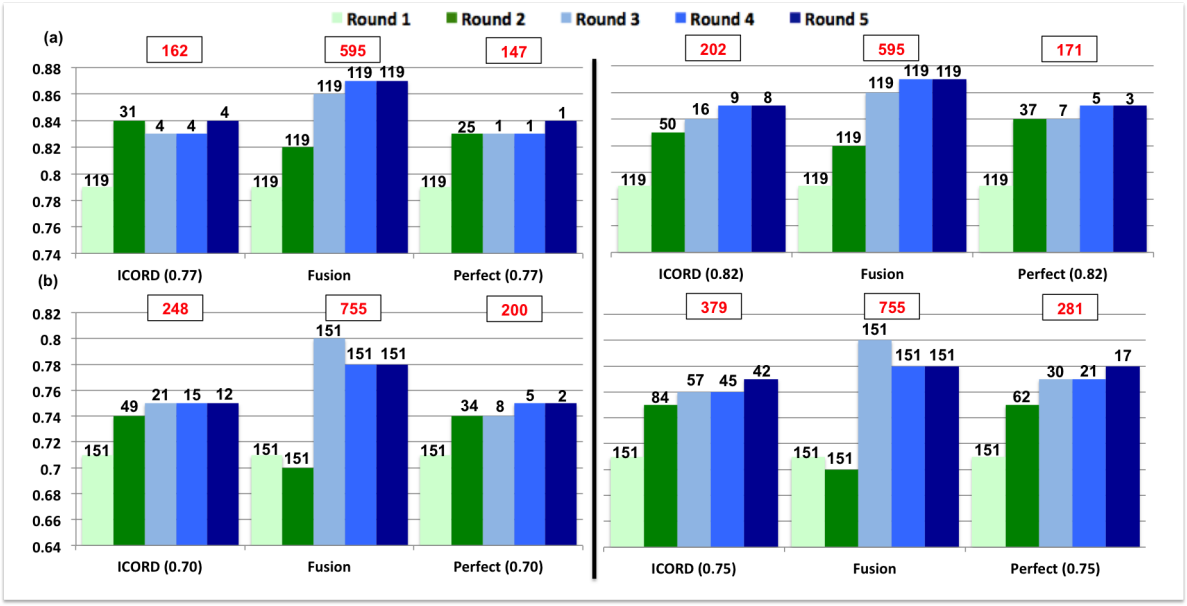

Figure 6. Average quality scores and the number of annotations collected per round of crowdsourcing for ICORD and the Fusion and Perfect Oracle benchmarks for (a) Fluorescence images with thresholds $\tau = 0.82$, and 0.85, and (b) Phase contrast images with $\tau = 0.75$, and 0.80. The number of collected annotations is the number shown above each bar. The total number of collected annotations for all five round is shown in red. ICORD requires the collection of significantly fewer annotations than Fusion.

times more for phase contrast. The savings are slightly smaller with a stricter threshold (3.0 and 2.0). This means that ICORD only uses between 27% and 50% of the resources, i.e., collection time and cost, that the Fusion method requires.

Comparing the results of ICORD to the results of the Perfect Oracle method shows the effects of a perfect prediction method incorporated into ICORD. It yields equal accuracy levels except in one case (0.1 difference) and even fewer collection costs (15 and 31 fewer collections of annotations on fluorescence images and 48 and 98 on phase contrast). (*Average* quality scores can be lower if crowdworkers happen to outline cells more accurately in an ICORD experiment than in a Perfect Oracle experiment, see supplemental materials.) The comparison shows the potential for improvement of the performance of ICORD if a prediction method was incorporated that had a lower rate of predicting inaccuracies when the cell outline indeed matches the ground truth well. Interesting future work would be to evaluate other machine learning methods that could substitute the random forest regression approach we selected here. Another interesting question is if other computer-vision approaches to characterize the image features of the segmented cell region could improve the automated assessment of the quality of these segmentations.

5. Conclusions

State-of-the-art crowdsourcing techniques rely on ad hoc decisions about the fixed number of redundant annotations to be collected and aggregated. The more annotations are collected, the higher is the likelihood for accuracy, but the more costly the collection process becomes. The generalizability and scalability of determining (manually) how to balance accuracy and efficiency on a case-by-case basis are questionable. An automated decision process is needed that is scalable to the large datasets produced by modern microscopy technology. In this paper, we proposed such an automated process. ICORD dynamically determines during the collection process how many annotations should be collected. ICORD uses image feature analysis and random forest regression to automatically interpret the quality of cell annotations. ICORD decides when to collect additional data or when to stop. It intelligently balances annotation accuracy and collection efficiency.

We collected a total of 1,350 crowd-drawn segmentations for 270 cell images. To the best of our knowledge, we made a novel contribution by studying the correlation between worker's behavioral cues and the quality of their cell segmentations. Our idea to integrate automatically-extracted behavioral and image features to infer annotation accuracy is also new. Lastly, we propose a new crowdsourcing methodology for annotating images that use dynamic decisions about which images to re-annotate. Our experiments revealed that our strategy is highly effective for dynamically assessing cell segmentation quality.

6. Acknowledgments

The authors gratefully acknowledge funding from the National Science Foundation (IIS-1421943).

References

[1] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Proceedings of the Conferece on Computer Vision and Pattern Recognition (CVPR), 8 pages*, 2014.

[2] S. Bell, P. Upchurch, N. Snavely, and K. Bala. OPENSURFACES: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (TOG)*, 32(4):111, 2013.

[3] A. Carlier, V. Charvillat, A. Salvador, X. Giro-i Nieto, and O. Marques. Click'n'Cut: Crowdsourced interactive segmentation with object candidates. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 53–56. ACM, 2014.

[4] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *Proceedings of the Conferece on Computer Vision and Pattern Recognition (CVPR), 8 pages*, 2010.

[5] D. R. Chittajallu, S. Florian, R. H. Kohler, Y. Iwamoto, J. D. Orth, R. Weissleder, G. Danuser, and T. J. Mitchison. In vivo cell-cycle profiling in xenograft tumors by quantitative intravital microscopy. *Nature Methods*, 12(6):577–585, 2015.

[6] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

[7] L. Galli, P. Fraternali, D. Martinenghi, M. Tagliasacchi, and J. Novak. A draw-and-guess game to segment images. In *2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT), and 2012 ASE/IEEE International Conference on Social Computing (SocialCom), Amsterdam, Netherlands*, pages 914–917, 2012.

[8] B. M. Good and A. I. Su. Crowdsourcing for bioinformatics. In *Bioinformatics*, volume 29, pages 1925–1933, 2013.

[9] D. Gurari, D. Theriault, M. Sameki, B. Isenberg, T. A. Pham, A. Purwada, P. Solski, M. Walker, C. Zhang, J. Y. Wong, and M. Betke. How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms. *IEEE Winter Conference on Applications in Computer Vision (WACV)*, 2015.

[10] K. Hara, V. Le, and J. Froehlich. Combining crowdsourcing and Google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 631–640. ACM, 2013.

[11] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady. Evaluating segmentation error without ground truth. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 528–536, 2012.

[12] L. I. Kuncheva, C. J. Whitaker, and C. A. Shipp. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, 6(1):22–31, Apr. 2003.

[13] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

[14] J. Pan, T. Kanade, and M. Chen. Heterogeneous conditional random field: Realizing joint detection and segmentation of cell regions in microscopic images. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2947, 2010.

[15] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 889–896, New York, NY, USA, 2009. ACM.

[16] J. Rittscher. Characterization of biological processes through automated image analysis. *Annual Review of Biomedical Engineering*, 12:315–344, 2010.

[17] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In *ACM CHI '10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872, 2010.

[18] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1–3):157–173, 2008.

[19] A. Sheshadri and M. Lease. SQUARE: A benchmark for research on computing crowd consensus. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 156–164, 2013.

[20] Y. Song, W. Cai, H. Huang, Y. Wang, D. D. Feng, and M. Chen. Region-based progressive localization of cell nuclei in microscopic images with data adaptive modeling. *BMC Bioinformatics*, 14:173–188, 2013.

[21] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *IEEE Comptuer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08)*, pages 1–8, June 2008.

[22] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[23] S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 2262–2269, June 2009.

[24] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[25] P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS) 23*, pages 2424–2432. Curran Associates, Inc., 2010.

[26] P. Welinder and P. Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),*, pages 25–32, June 2010.

[27] J. Whitehill, T. Wu, J. Bergsma, J. R. Movellan, and L. R. P. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS) 22*, pages 2035–2043. Curran Associates, Inc., 2009.

[28] Z. Yin, H. Su, E. K. M. Li, and H. Li. Cell-sensitive phase contrast microscopy imaging by multiple exposures. *Medical Image Analysis*, 25:111–121, 2015.

[29] X. Zhang, F. Xing, H. Su, L. Yang, and S. Zhang. High-throughput histopathological image analysis via robust cell segmentation and hashing. *Medical Image Analysis*, 16:306–315, 2015.