This is the author's version of an article published in IEEE Transactions on Medical Imaging. The final authenticated version is available online at: https://doi.org/10.1109/TMI.2018.2883237.

Crowdsourcing of Histological Image Labeling and Object Delineation by Medical Students

Anne Grote*, Nadine S. Schaadt*, Germain Forestier, Cédric Wemmert, and Friedrich Feuerhake

***Abstract*—Crowdsourcing in pathology has been performed on tasks that are assumed to be manageable by nonexperts. Demand remains high for annotations of more complex elements in digital microscopic images, such as anatomical structures. Therefore, this work investigates conditions to enable crowdsourced annotations of high-level image objects, a complex task considered to require expert knowledge. 76 medical students without specific domain knowledge who voluntarily participated in three experiments solved two relevant annotation tasks on histopathological images: (1) Labeling of images showing tissue regions, and (2) delineation of morphologically defined image objects. We focus on methods to ensure sufficient annotation quality including several tests on the required number of participants and on the correlation of participants' performance between tasks. In a set up simulating annotation of images with limited ground truth, we validated the feasibility of a confidence score using full ground truth. For this, we computed a majority vote using weighting factors based on individual assessment of contributors against scattered gold standard annotated by pathologists. In conclusion, we provide guidance for task design and quality control to enable a crowdsourced approach to obtain accurate annotations required in the era of digital pathology.**

***Index Terms*—Crowdsourcing, human decision making, image classification, image delineation, digital pathology, annotation, confidence score, majority vote.**

I. Introduction

Crowdsourcing has been used for annotation of high-level objects in microscopic images, but mainly focusing on less complex tasks referred to as microtasks [1]. Successful examples include identification of cancer cells [2], [3], detection of nuclei [4], [5], scoring based on immunohistochemically stained images [2], [3], [6], and detection of *Plasmodium falciparum* in red blood cells for malaria diagnostics [7]. Other studies focus on the creation of training sets for convolutional neural networks for finding nuclei or mitoses in cancer [8], [9]. Crowdsourcing has also been applied in labeling of retinal images [10], text annotation in radiology reports [11], or for delineation of a single object per image [12].

The original idea of crowdsourcing started 1906 with estimating the weight of an ox by a crowd [13]. Since then, it has been shown that crowds may outperform individual experts [14], [15]. Currently, crowdsourcing is defined as a collaborative problem-solving activity, performed online, to work on a certain, well-defined, and simple task by an undefined and large group of contributors who can be quite heterogeneous regarding their knowledge about the problem [16], [17]. To design applicable tasks, it is recommended to implement simplicity, short duration, sufficient training phase, feedback, and reliability tests [18]. Many factors negatively influence the crowd's performance including insufficient experience, knowledge, and expertise of the participants or task difficulty [19]. Further sources for errors are handling of software, misunderstanding of tasks, motivation, intention to fail, and distraction. It has been also shown that volunteers are more reliable regarding quality than paid participants; however, their endurance to stay on the task is clearly lower [20]. In general, contributors' motivation is the most challenging aspect besides task design and quality control [1], [21].

Crowdsourcing benefits from combining multiple contributors and depends on the level of information to be collected. Evidence from other fields suggests that crowdsourcing can be expanded to complex tasks. Crowdsourcing in geosciences has been used to generate online maps, in general and for disaster management [22] as well as land cover and land use data from remotely sensed images [23], [24]. Another example for solving difficult tasks by crowdsourcing is sleep spindle detection from electroencephalographic data [25]. Crowdsourcing has also been included into gaming-like approaches for solving difficult multiple sequence alignments [26] or for predicting complex protein structures [27]. These promising examples indicate general feasibility of successfully solving macrotasks and should stimulate further research, given that approaches of subdividing macrotasks into less challenging microtasks are not always feasible [1]. This applies particularly for context-dependent microscopic image evaluation. To identify relevant morphological structures, interpretation of image objects and their surroundings is to some extent inevitable, as their spatial context can change their relevance dependent on the context like in other fields of automated image analysis [28].

The growing amount of whole slide images (WSIs) in the last decade increased the importance of automated workflows due to limited time and availability of pathologists. This includes machine learning tools for region of interest (ROI) detection that allow a reproducible, objective, and large-scale analysis [29]. Relevant examples are pathological conditions such as tumor regions or anatomical structures such as glands in colon or breast tissue [30]–[33]. Such workflows strongly rely on annotations; especially, deep learning needs huge sets of training data [34].

* Anne Grote and Nadine S. Schaadt contributed equally to this work.

A. Grote, N.S. Schaadt, F. Feuerhake are with the Institute for Pathology, Hannover Medical School, Hannover, Germany. G. Forestier is with the IRIMAS, University of Haute Alsace, Mulhouse, France. C. Wemmert is with the ICube, University of Strasbourg, Illkirch, France. F. Feuerhake is also with the Institute for Neuropathology, University Clinic Freiburg, Germany.

This work was performed in the framework of SYSIMIT (FKZ:01ZX1308A), ILUMINATE (FKZ:031 B0006C), and SYSMIFTA (FKZ:031L0085A) funded by German Ministry for Education and Research (BMBF).

In the current study, we address the demand for high quality annotations that can be used to develop automated ROI detection for example as training sets for machine learning. Building on own work that indicated general feasibility [35], we developed recommendations for application of crowdsourcing for two independent tasks (1) labeling and (2) delineation of complex anatomical structures in histopathological images in a real-world scenario, including task setup (complexity, number of classes) and use of scattered gold standard to measure the reliability of individuals and to provide a confidence score. Our goal was to test the feasibility, and to develop a workflow for quality assurance, for two complex tasks that both involve contextual interpretation of image content information. Being aware of this complexity, we decided to involve voluntary, highly motivated contributors who have some general domain knowledge (anatomy) but not yet the required expert knowledge (microscopic pathology), enabling them to interpret biologically relevant patterns beyond pure image content information.

II. Materials and Methods

We conducted three independent experiments with different crowds and slightly different tasks each time.

A. Crowd Composition

The crowd consisted of third-year medical students from Hannover Medical School, Germany without any experience in annotating histological slides. It included 76 students in total, of which 36 participated in the first experiment, 14 in the second experiment, and 26 in the third experiment. Each experiment consisted of 1–3 sessions on different days. An overview is shown in Table I. We used a username together with a password for each student as login for the tools in order to correlate their anonymous contributions on different tasks.

TABLE I
Overview over the Crowds Participating in three Independent Experiments.

Crowd	Size	Task	Session	Subsize
Experiment 1 (spring term 2016)	36	ROI labeling	1	9
			2	4
			3	4
		ROI delineation	1	10
			2	9
Experiment 2 (fall term 2016)	14	ROI labeling	1	4
		ROI delineation	1	12
			2	6
Experiment 3 (spring term 2017)	26	ROI labeling	1	23
			2	12
		ROI delineation	1	23
			2	11

B. Task Design and Images

Histological images were acquired as WSIs from sections either stained for hematoxylin and eosin stain (H&E) or immunohistochemical markers. ROIs were not specifically stained and in the case of immunohistochemistry were to be detected based on faint blue hematoxylin counterstain.

The use of tissue samples for digital pathology analyses was approved by the institutional review board of Hannover Medical School in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards (approval numbers 2968–2015, 2063–2013, 1121–2011, 1831–2013).

We asked the crowd to solve two different types of tasks:

1) Labeling of ROIs (microtask, single choice [1])
   Given a set of images, each showing a single candidate ROI (representing anatomical structures or pathological conditions), the participants should select one of several proposed categories to classify each image. All ROIs were highlighted by colored outlines. Only the area inside this outline was relevant for labeling and only a single object existed in the image. In real applications, this can be used for quality control of ROIs detected by some automated image analysis framework.
2) Delineation of ROIs (macrotask, single choice [1])
   Given an image showing a tissue region, the crowd should draw the outlines of all objects of some well-defined classes and mark the class names. Annotations like these can be used as training or test sets to develop image analysis tools which detect ROIs by automated segmentation.

Before the task started, the students were instructed by a pathologist for about five minutes regarding the specific tasks. This introduction represented an overview on the characteristics of the images and an explanation about the precise definitions of each class (terminology) used for the task, with representative example images. We excluded images where the classification would be ambiguous. ROI labeling was designed such that it could be completed in about 15 minutes and ROI delineation in about half an hour per image.

C. Setting and Tools

For the first experiment, the crowd had to be present in a computer room. The participants were fully concentrated and did no further activities in parallel. For this experiment, we used a Java-based GUI called c17 implemented by ourselves for ROI labeling and the commercial software Aperio ImageScope (Leica Microsystems, Wetzlar, Germany) for ROI delineation. Tool c17 displays the current image, a progress line, a score comparing labels to the ground truth (GT), and a radio button for each class (see Supp. Mat., Fig. 1). For ImageScope, we prepared a template with class names to ensure a common terminology for the crowd.

For application at large-scale and convenience for students, we decided to switch to web applications such that the tasks could be finished outside the classroom. For the second and third experiment, we used a php-based tool called c13 developed by ourselves for ROI labelingand the open-source tool Cytomine [36] running on an own server for ROI delineation. The labeling tools were both designed using a similar layout to allow comparability. c13 additionally splits the images into a training phase in which the participants receive feedback about the correct class directly after labeling, and a test phase afterwards in which GT remains hidden. The delineation tools

had some differences in handling. ImageScope ensures unique class labeling for each object, whereas Cytomine allows multiple classes for a certain object, which was here unfavorable. On the other hand, Cytomine avoids unclosed polygons and prevents accidental terminology changes by the users. We assume that handling of both tools is comparable. The GT by a pathologist was always done with the same tool as used by the crowd for the corresponding task.

D. Answer Aggregation

Crowdsourcing allows to combine several annotations, potentially increasing the overall accuracy under the assumption that individuals produce different misclassifications. Here, we consider two concepts to build a final crowdsourced annotation by joining individual statements:

- Majority vote (MV):
  A class is assigned to an image (ROI labeling) or a pixel in an image (ROI delineation) when the relative majority of individuals picked it. Images with equal votes remain unclassified.
- Weighted vote (WV):
  For each class, we sum up the training/reliability score (see Section II-G2) of all individuals who selected this class. The class with the highest sum is assigned to the corresponding image to be classified (ROI labeling) or to the corresponding pixel in an image (ROI delineation). Thus, high performers have stronger impact on the result.

E. Evaluation Scores

Based on full GT provided by experts, we used F_1 score to study crowd's performance:

$$F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} \tag{1}$$

as well as precision (positive predictive value, PPV)

$$PPV = \frac{\sum\limits_{i \in C} \frac{TP_i}{TP_i + FP_i}}{|C|} \tag{2}$$

and recall (sensitivity, true positive rate, TPR)

$$TPR = \frac{\sum\limits_{i \in C} \frac{TP_i}{TP_i + FN_i}}{|C|} \tag{3}$$

where C is the set of classes, TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

F. Correlations between Different Images and Tasks

In experiments where the same participants completed both ROI labeling and ROI delineation, took part in two sessions for the same task, or annotated several images for ROI delineation, Spearman's rank correlation coefficient ρ_s of their performance in both assignments was computed, in order to test whether the quality of participants' work is transferable between tasks and sessions. For ROI labeling, the accuracy was used as measure, for ROI delineation, the F_1 score was used. The correlation between two assignments was only assessed if at least eight participants completed both tasks.

G. Simulation of an Application Case

For the application case, it is important to reduce the weights of low quality annotations and to measure the quality of crowd annotations in order to ensure their usefulness. In our setting, classroom training and supervision mostly avoids problems with tools and tasks as well as distraction. The fact that all participants were intrinsically motivated reduces the risk for fake answers to fulfill the task in a short time period. To address error sources (insufficient background knowledge and experience), we analyzed the use of a qualification set to measure the reliability of individuals in an application setting. For this, a scattered gold standard quality assurance is included and compared to the full GT.

In the case of ROI labeling, images with known label including at least one example for each class were scattered during the test phase. In the case of ROI delineation, a pathologist can annotate a couple of objects of each class representing a small area in the images. Then, the performance quality was measured using these subsets.

1) Reliability Tests for Participants: In the following, we refer to (1) images given at the beginning of the session (labeling) and/or objects (delineation) used to instruct participants (feedback intermediately provided) as "training phase", (2) images/objects scattered during session used for quality assurance as "qualification phase", and (3) remaining images/objects as "test phase". We use answer aggregation weighted by probabilities indicating individuals' reliability, which was expected to be related to their performance during the training phase. We defined the reliability $r_{j,i}$ of a single participant j and a class i for ROI labeling as

$$r_{j,i} = a \cdot ACC_{i,\text{train}} + (1 - a) \cdot ACC_{i,\text{qualification}}, \tag{4}$$

where a is the weight for training phase and ACC_i the accuracy for class i.

$$ACC_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{5}$$

As the distribution of the classes is unknown for an application case, we decided to select the examples for qualification phase randomly, where each class has to occur at least once. We afterwards compared the reliability averaged over all classes of each individual with their overall performance to find a value for parameter a, such that the error $E = |ACC_{\text{test}} - r_j|$ is minimal. We repeated the random selection of the qualification set 50 times.

For ROI delineation, the random selection of objects used as qualification phase from an image was here repeated 50 times, as well. For one image, the number of selected objects varied between $|C|$ and $|C| + 10$ ($|C|$: number of classes of interest in the current image) in order to examine the variation of the reliability score depending on the sample size. Then, the reliability r_j of a participant j was calculated as the ratio of correctly classified area to the total area in the selected objects. To easily adopt the approach to real-world application without full GT, the selected regions include only objects that were part of the GT.

2) Confidence Score for Annotated Objects: For actual application, we propose to combine the individual results using a WV weighted by their reliabilities, i.e. for each class i, we added together the reliability $r_{j,i}$ of each contributor j that voted for i and then chose the class with the highest sum. Additionally, we measured a confidence score for each image in ROI labeling as given in equation (6),

$$c_o = \frac{1}{n} \cdot \sum_{j=1}^{n} \Big(r_{j,i} \cdot (vote(j) == i) - r_{j,i} \cdot (vote(j) \neq i) \Big) \quad (6)$$

where n is the size of the crowd. The annotation of images/objects o with small confidence score should be reevaluated by an expert, whereas we trust labelings with scores close to 1. For ROI delineation, the confidence score (7) for each pixel was computed as the normalized sum of all reliability scores r_j of all participants who voted for the class i of the pixel o.

$$c_o = \frac{1}{n} \cdot \sum_{j=1}^{n} r_j \cdot (vote(j) == i) \quad (7)$$

The confidence score was given as a heatmap image indicating areas of high and low confidence.

III. Results

The results of the three experiments (Table I) are presented for ROI labeling and ROI delineation in different tissue types and settings. Here, we focus on how to increase and measure reliability in application settings.

A. ROI labeling

1) Experiment Design and Crowd Performance: We studied the performance of the crowd that was asked to label images representing single objects based on a given terminology. For this, we consider six different experimental setups. The intention to change the settings was to study the influence of terminology and data set composition on the quality of the annotations. The variable components including the number of images, number of classes, tissue type, and staining are listed in Table II for each single experiment. The ROIs represent either preexisting/healthy or pathological structures (examples in Supp. Mat., Fig. 2–4) and differ in their complexity between the experiments. For example, experiment 1 focuses on anatomically well-defined classes (breast tissue), experiment 2 used an hierarchical order of the classes (breast tissue), and experiment 3 includes a class that consists of subcategories (renal tissue). The experiments comprise different sessions that differ in the number of classes and images.

Fig. 1 depicts that agreement between crowd participants increased with the quality of their annotations, but was not obviously linked to task complexity. In addition, we assessed the inter-annotator agreement between experts, confirming that the GT was not perfect, but disagreement was limited to an acceptable range ($\kappa > 0.6$) and clearly lower than between crowd participants (see Supp. Mat., Fig. 5). The experts differed almost exclusively for "pathological changed glomerula", a class with multiple characteristics (combination of subclasses), in experiment 3. The crowds F_1 score for this class was with 0.326 for session 1 and 0.453 for session 2 on average clearly lower than for other classes.

TABLE II
Data Sets and Terminology ($\mathfrak{T}$) used for ROI labeling

Experiment	Images train	Images test	Staining	Tissue	$\mathfrak{T}$
ex_1, se_1	95		Ki-67	breast	$\mathfrak{T}_1$
ex_1, se_2	120		Ki-67	breast	$\mathfrak{T}_2$
ex_1, se_3	107		Ki-67	breast	$\mathfrak{T}_2$
ex_2, se_1	25	125	ER	breast	$\mathfrak{T}_3$
ex_3, se_1	20	100	H&E, PAS	kidney	$\mathfrak{T}_4$
ex_3, se_2	35	140	H&E, PAS	kidney	$\mathfrak{T}_5$

$\mathfrak{T}$	Categories
$\mathfrak{T}_1$	"lobule", "duct", "no epithelial structure"
$\mathfrak{T}_2$	"lobule", "lobule with extralobular ducts", "duct", "no epithelial structure"
$\mathfrak{T}_3$	hierarchically ordered categories, starting with highest priority: (1) "technical artifact", (2) "invasive tumor", (3) "intraepithelial neoplasia", (4) "glandular epithelium", (5) "other anatomical structures"
$\mathfrak{T}_4$	"normal glomerulum", "pathologically changed glomerulum (not sclerotic)", "sclerotic glomerulum"
$\mathfrak{T}_5$	"normal glomerulum", "pathologically changed glomerulum (not sclerotic)", "sclerotic glomerulum", "no glomerulum"

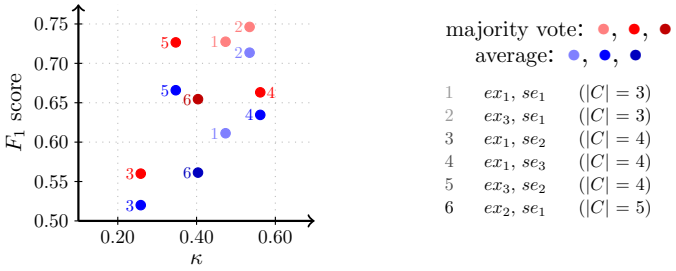

Fig. 1. **Overall F_1 scores in relation to Fleiss' κ for inter-annotator agreement (ROI labeling).** Shown are the average (blue) and majority vote (red) in the different experiments (ex_1, ex_2, ex_3) and sessions (se_1, se_2, se_3). The darker the color, the higher the level of complexity indicated by the number of classes $|C|$. The ids 1–6 are only used to label the dots.

2) Minimal Requirements for MV: Assuming that the agreement between the contributors increases with decreasing difficulty of the corresponding image, we built a MV, where a decision for a certain class is accepted if at least a minimum number of individuals l vote for this class. We calculated the MV for each minimum number between the number of votes required for a single majority and the crowd size n (i.e., $\forall\ l \in \{\lceil \frac{n}{|C|} \rceil, \ldots, n-1, n\}$ where $|C|$ is the number of classes). We then compared the number of unclassified images with the accuracy of classified images depending on the minimum number of identical answers required to accept a MV labeling.

This analysis gives insight into the use of crowdsourced image labeling in practice (e.g., large-scale quality control to exclude FPs of automatically detected ROIs). Crowdsourcing could reduce the need for a detailed review of images by pathologists.

In experiment 2, the requirement of at least three identical

labels resulted in a F_1 score of 0.753 for classified images (about 80% of all images). Accepting only labels where the full crowd was in agreement resulted in a F_1 score of 0.966 for classified images (about 50% of all images). In experiment 3, the crowd size allowed to fully investigate the influence of minimal requirements on the accuracy and the number of remaining, unclassified images. Obviously, the accuracy and the number of unclassified images increased with the required number of equal votes (Fig. 2).

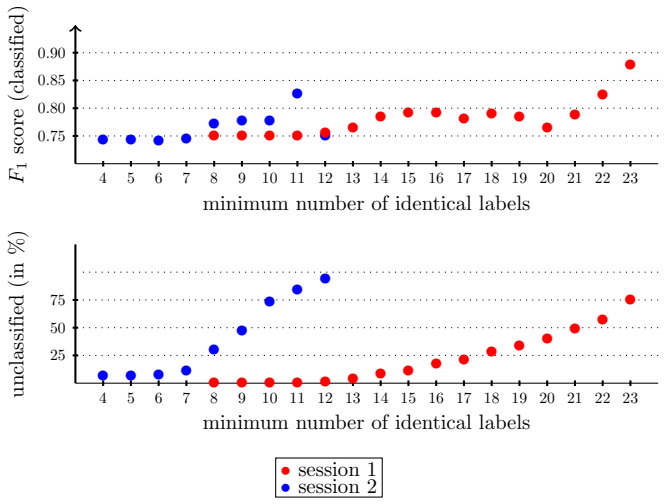

Fig. 2. **Experiment 3, ROI labeling.** F_1 score of classified images (top) in comparison to the number of unclassified images (bottom), where an image is assigned to a class, if a certain minimum number of individuals selected this class. To compute the score, only classified images were considered. Red dots correspond to session 1 (in total 100 images, a total crowd size of 23, and three classes; i.e., simple majority vote requires eight counts), blue dots to session 2 (in total 140 images, a total crowd size of 12, and four classes; i.e., simple majority vote requires four counts).

The first session included 23 participants and three classes, such that a majority requires at least eight votes for a certain class (equal to MV itself). Here, the number of unclassified images was zero and the F_1 score 0.746. In the second session (including 12 individuals), at least four identical labels represented the standard MV with a F_1 score of 0.743 for classified images (about 95% of all images).

For a real application, we envision an arrangement in which only 25% of the images need to be evaluated by an expert. This would sufficiently reduce the expert's evaluation time to justify the effort for crowdsourcing. In our current set-up, this would require 17–18 ($\sim$ 75% of crowd, session 1) and/or about eight ($\sim$ 67% of crowd, session 2) identical labels and would result in a MV F_1 score of classified images around 0.8.

3) Correlations: The correlation between both sessions was tested only in experiment 3 because the crowd size was not sufficient in the other experiments. Average F_1 scores of individuals participating in both sessions (9 participants) were measured, the corresponding correlation coefficient was 0.510. Supp. Mat., Fig. 6displays the F_1 score of individuals.

4) Confidence Score in a Simulated Application Case: The task included 20–25% gold standard images labeled by domain experts to measure the performance of individuals in order to weight the labeling. These images (qualification set) were randomly selected and scattered in the test phase. The whole process was repeated 50 times. The reliability of an individual was closest to its test accuracy with an error E of 0.036 in session 1 and 0.034 in session 2 on average for experiment 3 using a training weight of 35% and qualification weight of 65%. Based on this partition, Fig. 3, which illustrates the distribution of correctly classified images in the test phase compared to different confidence levels, shows that the accuracy increases with an increasing confidence score.

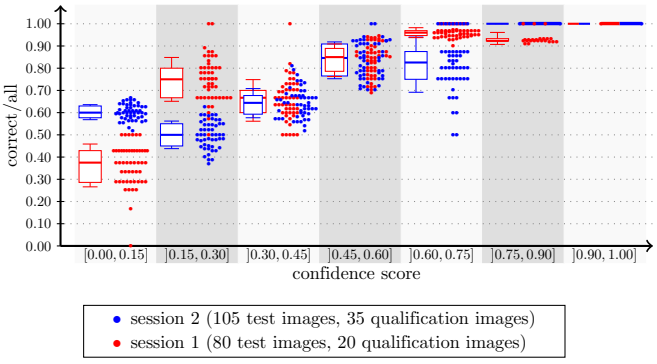

Fig. 3. Confidence score (ROI labeling); experiment 3, compared to the relative number of correctly labeled test images (correct/all, i.e., number of correctly labeled test images divided by the number of all test images; qualification images not included) based on majority vote weighted by individual reliability scores. Displayed are distribution, interquartile range, median, and standard deviation over 50 randomly selected qualification sets.

B. ROI delineation

1) Experiment Design and Crowd Performance: In experiments 1 and 3, participants delineated classes in renal tissue, and classes in breast tumor tissue were delineated in experiment 2. Table III gives an overview over all considered images and the used classes. $I_{ex_3,se_1,G_1,2}$ refers to the second image of session 1 of the third experiment, G denotes the number of a participant group.

We considered different sizes of images (2,735x2,735–26,986x21,487 pixels), number of classes (2–8), and object characteristics (e.g., well-defined anatomical structures, fuzzy tumors) to analyze the relation between task design and crowd performance. The size of each image and the number of included objects (Supp. Mat., Table 2–4) and the images with reference annotations (Supp. Mat., Fig. 9–11) are provided in Supp. Mat.

Overall results for MV and average F_1 score (Fig. 4) mostly increase with an increasing inter-annotator agreement measured by Fleiss' κ (Spearman's rank correlation coefficient between κ and MV: 0.393) and show the general feasibility of the tasks, especially when considering the fact that annotations provided by two experts also slightly varied at the border and in identification of objects (Supp. Mat., Fig. 12). In contrast to the crowd, the experts never disagreed in the class of a delineated object. Images showing the difference between MV and reference annotations are shown in Supp. Mat., Fig. 13.

TABLE III
DATA SETS AND TERMINOLOGY ($\mathfrak{T}$) USED FOR ROI DELINEATION

Image	Participants	Staining	Tissue	$\mathfrak{T}$
$I_{ex_1,se_1,1}$	10	CD3/CD20	kidney	$\mathfrak{T}_1$
$I_{ex_1,se_2,1}$	9	CD3/CD34	kidney	$\mathfrak{T}_2$
$I_{ex_2,se_1,1}$	9	ER	breast	$\mathfrak{T}_3$
$I_{ex_2,se_1,2}$	12	CD8	breast	$\mathfrak{T}_3$
$I_{ex_2,se_2,1}$	6	ER	breast	$\mathfrak{T}_3$
$I_{ex_2,se_2,2}$	5	CD8	breast	$\mathfrak{T}_3$
$I_{ex_3,se_1,1}$	22	CD68	kidney	$\mathfrak{T}_4$
$I_{ex_3,se_1,G_1,2}$	8	CD68	kidney	$\mathfrak{T}_4$
$I_{ex_3,se_1,G_2,2}$	10	CD68	kidney	$\mathfrak{T}_4$
$I_{ex_3,se_2,1}$	11	CD68	kidney	$\mathfrak{T}_4$
$I_{ex_3,se_2,G_1,2}$	5	CD68	kidney	$\mathfrak{T}_4$
$I_{ex_3,se_2,G_1,3}$	5	H&E	kidney	$\mathfrak{T}_4$
$I_{ex_3,se_2,G_2,2}$	6	CD68	kidney	$\mathfrak{T}_4$
$I_{ex_3,se_2,G_2,3}$	6	H&E	kidney	$\mathfrak{T}_4$

$\mathfrak{T}$	Categories
$\mathfrak{T}_1$	"normal glomerulum", "tubulus"
$\mathfrak{T}_2$	"normal glomerulum", "tubulus", "artery", "dysfunctional glomerulum", "sclerotic glomerulum", "muscle", "collageneous tissue/septae"
$\mathfrak{T}_3$	"duct", "lobule", "non-malignant precursor lesion", "invasive tumor", "intraepithelial neoplasia", "technical artifact", "large blood vessel", "other anatomical structures"
$\mathfrak{T}_4$	"glomerulum", "artery", "muscle"

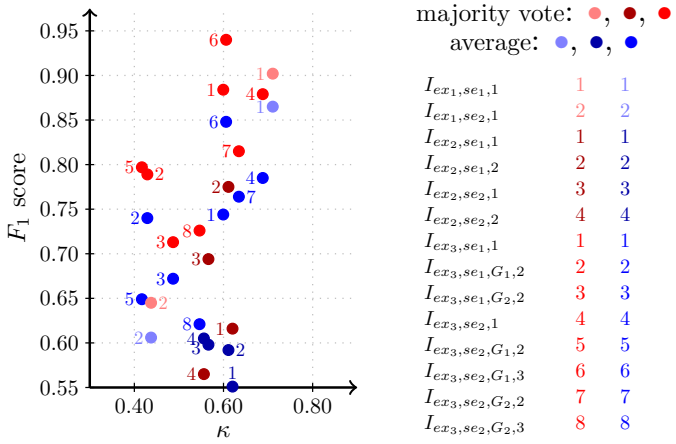

Fig. 4. **Overall F_1 scores in relation to Fleiss' κ for inter-annotator agreement (ROI delineation).** Shown are the average (blue) and majority vote (red) for different images. Color intensities refer to different experiments. The ids 1–8 are only used to label the dots.

2) Subcrowd Size and Robustness: In order to analyze how the number of participants influences the result of the MV image, we examined subgroups of participants.

Since the total number of possible combinations n_c can be very large even for moderately large crowd sizes n, we restricted the number of combinations by randomly selecting a subset of all possible participant combinations. Thus, for each subgroup size k from 1 (individual notations) up to $n-2$, a randomly selected set of participant combinations was examined. The number of selected participant combinations was set to $n_c(k=2)$, with an upper limit of 45. The case $n-1$ was not considered because it is very similar to the whole crowd.

For each participant combination, a MV image is created from the annotated images of the corresponding participants. Since, particularly for small k, there can be large regions without clear majority, a class was only assigned to a pixel if there is a clear relative majority for this class. If two or more classes at the top share the same number of votes, the corresponding pixels were labeled "ambiguous".

F_1 scores were computed for all MV images and the average F_1 score for each subgroup size was determined. The "ambiguous" class was not considered for F_1 score computation; instead, the average area of "ambiguous" regions was reported for each subgroup size.

To study the number of required contributors to obtain crowdsourcing results of acceptable quality for object delineation, we considered the results for all images with annotations from at least nine participants (seven images in total).

Fig. 5 shows average F_1 scores of MVs for the analyzed group sizes as well as the percentage of area labeled as "ambiguous" for each image. The F_1 score improved compared to the average individual F_1 score starting from a group size of three and did not improve much further beyond a group size of seven. In four images (Fig. 5A,D–F), the F_1 score stabilized early: The standard deviation as a measure of variations between different subgroups decreased markedly with growing number of contributors, down to a value below 0.02 at a group size of about eight participants (three for Fig. 5A). In the other three images (Fig. 5B,C,E), the standard deviation remained higher, although it also showed a decreasing trend. In two of these images, there were relatively large areas misclassified by several participants. $I_{ex_3,se_1,G_2,2}$ (Fig. 5F) showed only minor improvements of the F_1 score with growing group size. This indicated that there may be situations where only better training but not increased crowd size could improve results.

The average area of the "ambiguous" regions always has a peak at the subgroup size of two, because with two participants there is a relatively high probability that they do not agree on the classification. For the same reason, there is often a small drop in the F_1 score at even numbers of participants. The amount of "ambiguous" area depended on the amount of tissue which contains relevant objects, on the number of classes, and on the characteristics of classes.

3) Correlations: The correlation of the performance of individual participants on different images was analyzed for those cases where at least eight participants took part in two sessions or completed several images given in one session. The F_1 score was used in the comparisons. Fig. 6 illustrates the F_1 scores of the compared images and the corresponding correlation coefficient. Supp. Mat., Table 5 shows the examined image pairs with average, minimum and maximum absolute differences, and the average signed differences.

The correlation coefficients for experiment 3 show moderate correlations, the correlation for the examined image pair in experiment 2 was weak. The average signed difference showed a slight tendency for worse results in the second image, but in comparison to the absolute difference it can be seen that some participants had better results and some had worse results in the second image. An exception was the first session of experiment 3, where most participants had worse results on the second image. In this case, the first image was simpler to annotate than both second images. This analysis indicates

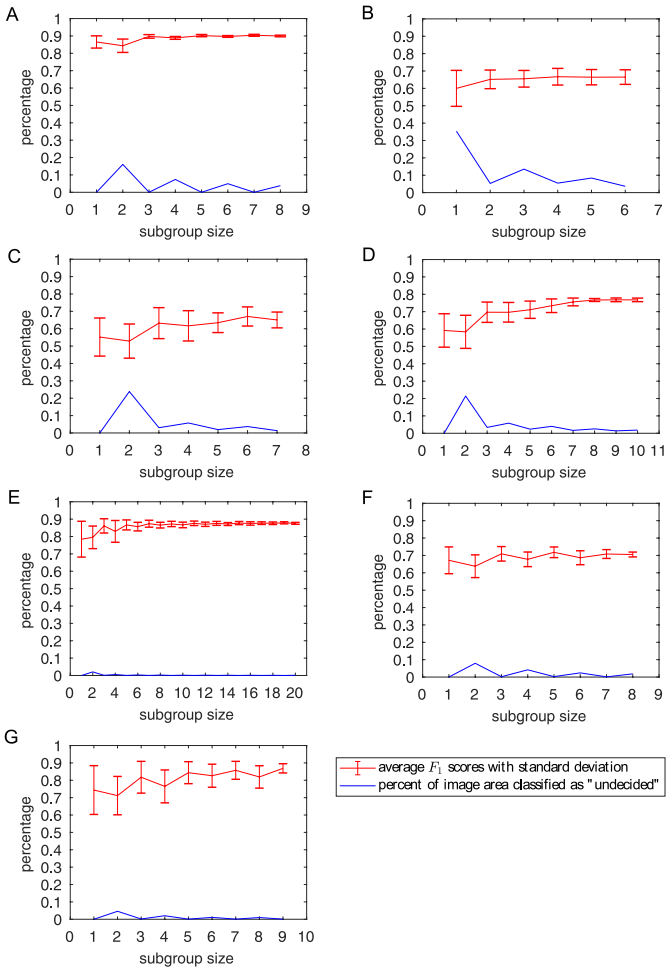

Fig. 5. **Subgroup analysis for experiment 1, ROI delineation.** Red: average F_1 scores with standard deviation, blue: percent of image area classified as "ambiguous". A: $I_{ex_1,se_1,1}$. B: $I_{ex_1,se_2,1}$. C: $I_{ex_2,se_1,1}$. D: $I_{ex_2,se_1,2}$. E: $I_{ex_3,se_1,1}$. F: $I_{ex_3,se_1,G_2,2}$. G: $I_{ex_3,se_2,1}$.

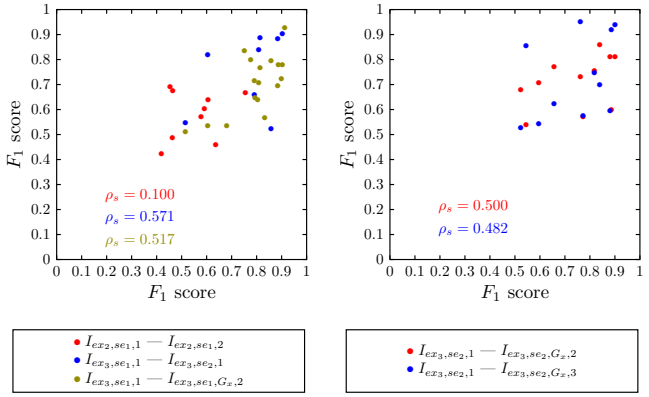

Fig. 6. **Scatterplot of individuals' F_1 scores for different images (ROI delineation)**. Spearman's rank correlation coefficient given as ρ_s. G_x: result data combined from G_1 and G_2.

that the performance of individual participants is influenced by image content.

4) Confidence Score in a Simulated Application Case: Accuracy scores were computed for randomly selected samples in varying sample sizes for $I_{ex_3,se_1,1}$ in experiment 3. Average standard deviations and ranges of the reliability score for all participants over 50 runs are shown in Supp. Mat., Table 6. The variability of the reliability score expectedly decreased with increasing number of GT samples. As a compromise between practical applicability and accuracy, we chose the number of samples to be at least five and at least one from each class. The number of five samples was in the range of 10–30% of total sample count for most images used in this study. This number of samples was used in the following experiments.

The GT sample selection and WV for 50 runs was performed for each image. F_1 scores were computed for the WV images. Detailed results for all images, with a comparison to the unweighted MV, are listed in Supp. Mat., Table 7. The weighted average F_1 score was better than the unweighted F_1 score for half of the examined images. This is due to the fact that the reliability score is a measurement of recall (see Section II-G1) and the WV will therefore tend to produce annotations with higher recall but potentially lower precision. Recall for WV was in fact higher than for unweighted MV for all images.

For experiment 3, with several comparable images per session, we used the reliability scores from the first image of the session for the WV from the second session. Table V shows the results for the second and third images of the sessions. In four of the six images, the F_1 score from the transferred reliability scores was lower than the F_1 score from the reliability scores of the same image. This indicates that it is desirable to have GT objects in each image, if possible.

Using the WV, a confidence image was produced that facilitates checking for false annotations. Fig. 7 (top) shows an example of a confidence image together with the comparison to GT. Regions where the WV differed from the GT mostly had lower confidence values, indicated as darker colors in the confidence image. Fig. 7 (bottom) shows average confidence scores for annotated objects from example WV images (remaining images in Supp. Mat., Fig. 15, 16), for different percentages of agreement with the reference image.

The confidence scores were averaged over all objects per image. The confidence scores for objects that did not agree with the reference were typically low, which means that these objects can be found efficiently using this confidence score image. Confidence scores for correct objects were typically higher than those for incorrect objects, while the absolute value depends on the nature of the task. This is illustrated by the fact that in experiment 3, where the given classes were relatively simple and few, the confidence scores were higher than in experiment 2 with its higher class complexity and number of classes. Checking for false annotations should therefore focus on relative confidence score values between regions in an image.

C. Correlations between both Tasks

Experiment 3, where the crowd size was large enough, was used to measure the correlation between the F_1 scores for ROI labeling and ROI delineation of the individuals. In session 1, 22 participants performed both tasks. In session

TABLE IV
TRANSFERABILITY OF WEIGHTS BETWEEN DIFFERENT IMAGES IN A SESSION (ROI DELINEATION).

Image	Participants	r_{avg}	mean F_1 transferred	mean F_1 same	unweighted F_1
$I_{ex_3,se_1,G_1,2}$	8	0.408	0.750	0.786	0.789
$I_{ex_3,se_1,G_2,2}$	10	0.457	0.683	0.736	0.713
$I_{ex_3,se_2,G_1,2}$	5	0.654	0.775	0.780	0.797
$I_{ex_3,se_2,G_1,3}$	5	0.915	0.938	0.740	0.940
$I_{ex_3,se_2,G_2,2}$	6	0.626	0.865	0.846	0.815
$I_{ex_3,se_2,G_2,3}$	6	0.613	0.640	0.744	0.726

Results of weighted majority vote for second and/or third image in sessions of experiment 3. r_{avg}: average participants' reliability score; mean F_1 transferred: average F_1 score from weighted majority vote with weights from first image in session; mean F_1 same: average F_1 score from weighted majority vote with weights from the same image.

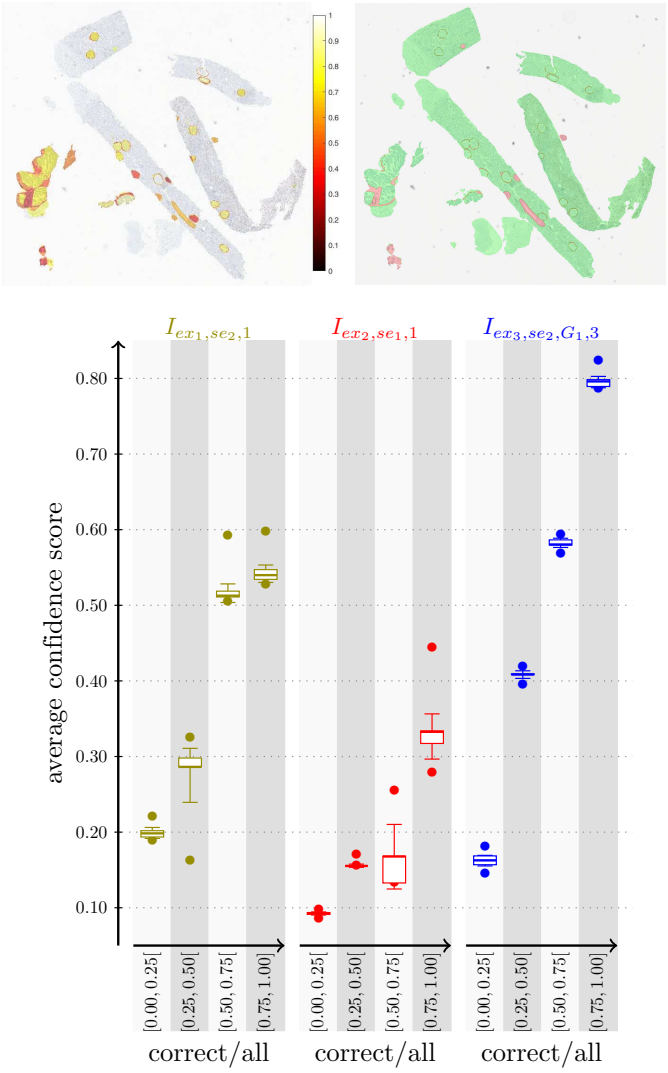

Fig. 7. **Confidence (ROI delineation).** Top: Confidence image from weighted vote for $I_{ex_3,se_1,G_1,2}$. Left: overlay of confidence image with original image (bright colors: high confidence; dark colors: low confidence). Right: difference of majority vote to ground truth (green: agreement to ground truth, red: different from ground truth). Bottom: Confidence scores for different percentages of agreement (correct/all) for objects in the images. Scores were computed from weighted majority vote images and averaged over all objects in an image. Shown are interquartile range, median, and standard deviation for 50 runs.

2, 11 participants performed both tasks. For both sessions, the results for second or third images were combined due to grouping participants in ROI delineation. The correlation coefficients given in Fig. 8 indicate that the correlation between labeling and delineation quality was relatively weak, especially in session 2 (values close to zero). The F_1 scores of ROI delineation vary more than those of ROI labeling.

These results suggest that the individual performance cannot be projected from one task to another, indicating that participants' contributions should be weighted for each task separately. However, it is possible to transfer quality estimates of contributors between sessions of the same task.

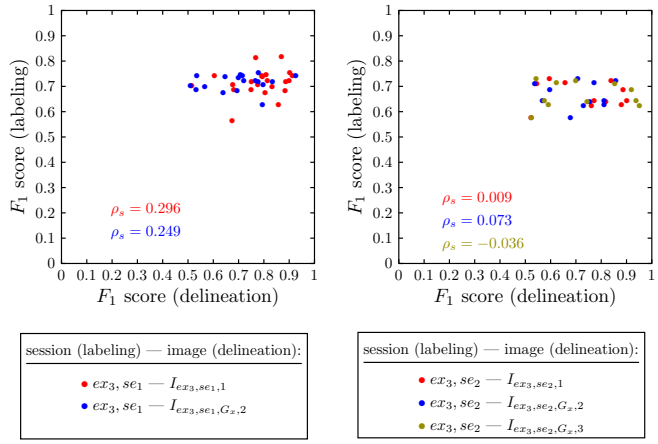

Fig. 8. **Scatterplot of individuals' F_1 scores for ROI labeling and ROI delineation** in experiment 3, session 1 (left) and session 2 (right). Spearman's rank correlation coefficient given as ρ_s. G_x: result data combined from G_1 and G_2.

IV. DISCUSSION

This study analyzed the potential of a relatively small "educated" crowd, as opposed to the common practice of large crowds of contributors. Students with comparable motivation and knowledge at the same stage of the medical school curriculum can be regarded as "semi-experts". We asked whether this setting would be suitable for labeling and delineation of histopathological images, a task of considerable complexity that is usually assigned to experts. Our tests cover a variety of image objects that are for example relevant for cancer or transplantation research that includes ROI detection. We therefore discuss strengths and limitations of this approach for machine learning applications.

A. Crowd Performance and Limitations

Although it is recommended that crowdsourcing tasks should be kept as simple as possible, our findings are in line with the notion that handling of more complex tasks is feasible in more controlled settings of collective problem solving [18]. A previous study showed that Crowdsourcing of image annotation tasks was generally feasible with variations corresponding with task complexity, even the concept of an hierarchically ordered terminology (ROI labeling) resulted in reliable annotations. Nevertheless, attention should be paid to object characteristics and number of classes. Multiple shape- or texture-related features and semi-quantitative characteristics like size or cell density, may mark limits of feasibility for crowdsourced image labeling [35].

For ROI delineation, the experiments covered a broad range of class complexity and image size. Classes with relatively constant size and appearance (such as "glomerulum") were more suitable than more variable classes with overlapping characteristics. A comparison between the three experiments shows a relation between performance and number of classes (quality from experiment 2, which featured a number of diverse classes, was relatively low compared to the setups with fewer classes). Additionally, our results suggest a higher recall for smaller images (experiment 1). In particular, good performance can be achieved by limiting the task to a low number of classes for annotations on images of limited size. In accordance with [37], the error rate increases with increasing complexity.

We paid particular attention to the problem of image objects with poorly defined levels of feature variability, such as "invasive tumor". Experiment 2 confirmed that outlining the invasive edge of tumors was in fact more difficult than anatomically defined structures (e.g., differentiated epithelial cells or basal membranes). However, GT in digital pathology is made by individual manual drawing and represents an approximation instead of the entire truth. In line with this, the class "invasive tumor" almost inevitably showed some difference to GT in the border region. Interestingly, the outline based on MV sometimes seemed even more exact than the GT on visual inspection. We conclude that the interpretation of quality measures of classes with irregular borders like "invasive tumor" should take into account that lower quality values do not necessarily mean low crowd performance. In contrary, the observation that MV-based outlines sometimes seemed to be a little closer to the tumor edge than the GT seems to confirm [13], claiming that crowds can outperform experts in particular settings. However, it is difficult to quantify this effect and it remains to be investigated whether this trend holds true in larger series. This would be a promising field where crowdsourcing could substantially contribute to more reliable annotations.

Another aspect that may influence the level of difficulty in our setting was the individual staining of the images. The annotation of ROIs will often be based on faint blue counterstain or normal H&E, and not on specific staining of tissue components of interest. An exception in our setting was a marker for blood vessels (experiment 1). In our case where only large blood vessels should be delineated, it was probably this staining that lead to a number of false positives. Similarly, a staining for ER, which can be positive or negative in tumors and variable in normal epithelium can have an effect on ROI detection. In two images of experiment 2, ER prominently marked the tumor area, which could, in these images, have guided the delineation. We found that both ER-stained images had precision and recall values of over 0.9 for the tumor class, whereas in the other two images either precision or recall was lower than 0.9. We conclude that use of immunohistochemical stainings did not seem to have a negative influence.

We recorded cases where the MV overruled single better contributions. In general, this occurs in crowds with clearly detectable differences between low and high performing contributors, as observed in our experiments. A higher weight for better performers may reduce this risk (therefore, weighted MVs are often considered [38], [39]). As this assumes that the quality of individual contributors remains constant during larger time spans, we analyzed the correlations of contributors between sessions. Our analysis confirmed that a transfer of individual estimates for the same task is feasible. This allows the application of a confidence score for weighting the contributions according to capability, which strengthened the input of high-performing contributors and ensured reliability of the results.

B. Recommendations for Application

1) Task design and teaching
   Our experience showed that the layout of the tasks (terminology, ROI complexity) has a strong impact on the crowd performance, consistent with [12]. If object characteristics are complex, it may be advisable to keep the number of classes low. The terminology should be well-defined, categorical, and easy to explain. Our observations confirmed the importance of a teaching session with direct interaction between the crowd and the instructors to explain terminology, give technical support, and to avoid misunderstanding. This is in agreement with published data reporting that face-to-face or video-based teaching improved the result compared to written illustrated descriptions [40].
2) Crowd composition
   In contrast to the usual large and heterogeneous crowds [16], [17], our crowd was relatively small but with homogeneous background knowledge. The general possibility of working with small crowds is supported by published data [4], [9], [11]. Our subset analysis of ROI delineation suggests that results often do not improve much further when the crowd exceeds 7–8 participants. Most objects were clearly delineated by a MV from this crowd size, and objects that were very difficult to identify remained unstable anyway. Further studies are required to investigate whether those "difficult" images may still benefit from a larger crowd.
3) Tool design
   Difficulties in tool handling were manageable, and more

common in the technically more complex ROI delineation. In our experience, Cytomine [36] seemed to be more suitable for crowdsourcing than ImageScope because of automatic saving, and the possibility to restrict user actions. However, users have to actively avoid double assignment of a class to a single ROI in Cytomine. In our case, this occurred very rarely. In contrast, some participants did not follow the instructions to produce closed lines leading to incomplete objects in ImageScope.

4) Adjustments to individual applications
Our approach provides several variable elements that can be adjusted to the requirements of future applications. For a set-up like ROI labeling, we suggest to give the contributors the option to label images "not classifiable", e.g. for images which contain two or more objects inside the outlines.
To avoid the disadvantage of MV losing information about certainty (relative frequency of the majority class) [41], the minimal requirement for agreement for MV can be adjusted, resulting in a trade-off between accuracy and the number of unclassified images. This reduces the pathologist's time for review of unclassified images at acceptable reliability of the crowdsourced labeling.
5) Quality control
In order to provide a way to control quality with limited GT, we tested reliability measures that were based on scattered gold standard. The introduced confidence score clearly stated the quality of an annotation for both tasks. Since the reliability measure in the case of ROI delineation effectively measures recall, WV images tended to have higher recall than unweighted ones, sometimes at the cost of precision. For a scenario where crowdsourced annotation data is intended to reduce the workload for subsequent quality control, this is advantageous as it is often easier to reject FPs than to identify FNs. The confidence score image provided useful guidance to quickly find potentially wrong labels. Therefore, we strongly recommend to provide a confidence score based on the generally suggested reliability tests [18]. We used a percentage of about 10–35% control instances, compared to published examples with 0.1% (however, in a huge data set) [23], [24] or 20% [1], [7] of the data set.

C. Perspective: Application for Complex Annotations

Published results suggest that crowds are able to classify nuclei [2], [5], [6], including a study where 28 participants achieved high correlation with GT in the detection of positively stained cells [4]. Our study expands the scope of crowdsourcing in pathology, using more complex terminologies in two demanding tasks that address a high current need for WSI annotations. The presented concept of human decision making differs from classical crowdsourcing projects in the composition of the crowd (education, size) and use of classroom teaching. We hypothesize that the performance of the crowd may be related to this specific setting, and that a substantial component of success is the medical background knowledge of the participants. To test this hypothesis, it is necessary to evaluate the quality of a heterogeneous crowd. We anticipate a notably higher need for crowd teaching and training and we expect further need for adjustments of task design. Complex tasks in a civil engineering context have been evaluated with heterogeneous crowds [42], where it was found that with increasing task complexity more communication with participants was needed.

For application in development of automated ROI detection, the labeling approach produces a high accuracy for classified images and a reduced remaining data set for which expert annotations are still required. For delineation, the heatmap of the confidence score provides guidance for the pathologist who can review slides focusing on areas with low confidence score, which is more efficient compared to full expert annotations.

Acknowledgment

The authors thank all contributing students. Further, the authors thank Dr. Maja Temerinac-Ott, University of Strasbourg for participating in data selection for experiment 3.

References

[1] G. Li, J. Wang, Y. Zheng, and M. Franklin, "Crowdsourced data management: a survey," IEEE T Knowl. Data En., vol. 28, pp. 2296–2319, 2016.
[2] F. J. C. dos Reis, S. Lynn, H. R. Ali, D. Eccles, A. Hanby, E. Provenzano, C. Caldas, W. J. Howat, L.-A. McDuffus, B. Liu et al., "Crowdsourcing the general public for large scale molecular pathology studies in cancer," EBioMedicine, vol. 2, pp. 681–689, 2015.
[3] J. Lawson, R. J. Robinson-Vyas, J. P. McQuillan, A. Paterson, S. Christie, M. Kidza-Griffiths, L.-A. McDuffus, K. A. Moutasim, E. C. Shaw, A. E. Kiltie et al., "Crowdsourcing for translational research: analysis of biomarker expression using cancer microarrays," Brit J Cancer, vol. 116, pp. 237–245, 2017.
[4] V. Della Mea, E. Maddalena, S. Mizzaro, P. Machin, and C. A. Beltrami, "Preliminary results from a crowdsourcing experiment in immunohistochemistry," Diagnostic pathol., vol. 9, p. S6, 2014.
[5] H. Irshad, L. Montaser-Kouhsari, G. Waltz, O. Bucur, N. Jones, F. Dong, N. Knoblauch, and A. Beck, "Crowdsourcing image annotation for nucleus detection and segmentationin computational pathology: Evaluating experts, automated methods, and the crowd," in Pac Sym Biocomp, 2015, pp. 294–305.
[6] H. Irshad, E.-Y. Oh, D. Schmolze, L. M. Quintana, L. Collins, R. M. Tamimi, and A. H. Beck, "Crowdsourcing scoring of immunohistochemistry images: Evaluating performance of the crowd and an automated computational method," Scient. Rep., vol. 7, 2017.
[7] S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen, and A. Ozcan, "Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study," PloS ONE, vol. 7, p. e37245, 2012.
[8] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images," IEEE T Med. Imag., vol. 35, pp. 1313–1321, 2016.
[9] E. Kim, S. Mente, A. Keenan, and V. Gehlot, "Digital pathology annotation data for improved deep neural network classification," in SPIE Med. Imag., 2017, pp. 101 380D–101 380D.
[10] D. Mitry, K. Zutis, B. Dhillon, T. Peto, S. Hayat, K.-T. Khaw, J. E. Morgan, W. Moncur, E. Trucco, and P. J. Foster, "The accuracy and reliability of crowdsource annotations of digital retinal images," Trans. Vis. Sci. Technol., vol. 5, pp. 6–6, 2016.
[11] A. Cocos, T. Qian, C. Callison-Burch, and A. J. Masino, "Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation," J Biomed Informatics, 2017.
[12] D. Gurari, M. Sameki, and M. Betke, "Investigating the influence of data familiarity to improve the design of a crowdsourcing image annotation system," AAAI Conf. HCOMP, pp. 59–68, 2016.

[13] F. Galton, "Vox populi (the wisdom of crowds)," Nature, vol. 75, pp. 450–451, 1907.

[14] J. Surowiecki, The wisdom of crowds. Anchor, 2005.

[15] H. Hussain, P. Nirmala, and M. Swathy, "Mitogame: Gamification method for detecting mitosis from histopathological images using crowdsourcing," IRJET, vol. 4, 2017.

[16] J. Howe, "The rise of crowdsourcing," Wired magazine, vol. 14, pp. 1–4, 2006.

[17] E. Estellés-Arolas and F. González-Ladrón-De-Guevara, "Towards an integrated crowdsourcing definition," J Info. Science, vol. 38, pp. 189–200, 2012.

[18] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best practices and recommendations for crowdsourced qoe-lessons learned from the qualinet task force crowdsourcing," QUALINET, 2014.

[19] S. Liu, F. Xia, J. Zhang, L. Wang, and L. Wang, "How crowdsourcing risks affect performance: an exploratory model," Management Decision, vol. 54, pp. 2235–2255, 2016.

[20] J. Redi and I. Povoa, "Crowdsourcing for rating image aesthetic appeal: Better a paid or a volunteer crowd?" in Proc. 2014 Internat. ACM Work. Crowdsourcing Multimedia. ACM, 2014, pp. 25–30.

[21] A. Chittilappilly, L. Chen, and S. Amer-Yahia, "A survey of general-purpose crowdsourcing techniques," IEEE T Knowl. Data En., vol. 28, pp. 2246–2266, 2016.

[22] P. Boccardo and P. Pasquali, "Web mapping services in a crowdsource environment for disaster management: State-of-the-art and further development," Internat. Archives Photogrammetry, Remote Sensing Spatial Info. Sciences, vol. 39, pp. 543–548, 2012.

[23] S. Fritz, L. See, C. Perger, I. McCallum, C. Schill, D. Schepaschenko, M. Duerauer, M. Karner, C. Dresel, J.-C. Laso-Bayas, M. Lesiv, I. Moorthy, C. F. Salk, O. Danylo, T. Sturn, F. Albrecht, L. You, F. Kraxner, and M. Obersteiner, "A global dataset of crowdsourced land cover and land use reference data," Nature Scient. Data, vol. 4, p. 170075, 2017.

[24] L. See, A. Comber, C. Salk, S. Fritz, M. van der Velde, C. Perger, C. Schill, I. McCallum, F. Kraxner, and M. Obersteiner, "Comparing the quality of crowdsourced data contributed by expert and non-experts," PLOS ONE, vol. 8, p. e69958.

[25] S. C. Warby, S. L. Wendt, P. Welinder, E. G. Munk, O. Carrillo, H. B. Sorensen, P. Jennum, P. E. Peppard, P. Perona, and E. Mignot, "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods," Nature meth., vol. 11, pp. 385–392, 2014.

[26] A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, L. Sarmenta, M. Blanchette, J. Waldispühl et al., "Phylo: a citizen science approach for improving multiple sequence alignment," PloS ONE, vol. 7, p. e31362, 2012.

[27] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović et al., "Predicting protein structures with a multiplayer online game," Nature, vol. 466, pp. 756–760, 2010.

[28] M. Grega, A. Matiolański, P. Guzik, and M. Leszczuk, "Automated detection of firearms and knives in a cctv image," Sensors, vol. 16, p. 47, 2016.

[29] G. Kayser and K. Kayser, "Quantitative pathology in virtual microscopy: history, applications, perspectives," Acta histochemica, vol. 115, pp. 527–532, 2013.

[30] A. Grote, M. Abbas, N. Linder, H. Kreipe, J. Lundin, and F. Feuerhake, "Exploring the spatial dimension of estrogen and progesterone signaling: detection of nuclear labeling in lobular epithelial cells in normal mammary glands adjacent to breast cancer." Diagnostic Pathol, vol. 9, 2014.

[31] P. Kainz, M. Pfeiffer, and M. Urschler, "Semantic segmentation of colon glands with deep convolutional neural networks and total variation segmentation," arXiv:1511.06919, 2015.

[32] G. Apou, N. S. Schaadt, B. Naegel, G. Forestier, R. Schönmeyer, F. Feuerhake, C. Wemmert, and A. Grote, "Detection of lobular structures in normal breast tissue," Computers Biol. Med., vol. 74, pp. 91–102, 2016.

[33] B. E. Bejnordi, M. Balkenhol, G. Litjens, R. Holland, P. Bult, N. Karssemeijer, and J. A. van der Laak, "Automated detection of dcis in whole-slide h&e stained breast histopathology images," IEEE T Med. Imag., vol. 35, pp. 2141–2150, 2016.

[34] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," IEEE T Pattern Anal., vol. 35, no. 8, pp. 1798–1828, 2013.

[35] N. S. Schaadt, A. Grote, G. Forestier, C. Wemmert, and F. Feuerhake, "Role of task complexity and training in crowdsourced image annotation," LNCS, 2018.

[36] R. Marée, L. Rollus, B. Stévens, R. Hoyoux, G. Louppe, R. Vandaele, J.-M. Begon, P. Kainz, P. Geurts, and L. Wehenkel, "Collaborative analysis of multi-gigapixel imaging data using cytomine," Bioinformatics, vol. 32, pp. 1395–1401, 2016.

[37] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in data crowdsourcing," IEEE T Knowl. Data En., vol. 28, pp. 901–911, 2016.

[38] A. Kurve, D. Miller, and G. Kesidis, "Multicategory crowdsourcing accounting for variable task difficulty, worker skill, and worker intention," IEEE T Knowl. Data En., vol. 27, pp. 794–809, 2015.

[39] C.-J. Ho, S. Jabbari, and J. W. Vaughan, "Adaptive task assignment for crowdsourced classification," in Internat. Conf. Mach. Learn., 2013, pp. 534–542.

[40] J. Starr, C. M. Schweik, N. Bush, L. Fletcher, J. Finn, J. Fish, and C. T. Bargeron, "Lights, camera citizen science: assessing the effectiveness of smartphone-based video training in invasive plant identification," PLOS ONE, vol. 9, p. e111433, 2014.

[41] V. Sheng, J. Zhang, B. Gu, and X. Wu, "Majority voting and pairing with multiple noisy labeling," IEEE T Knowl. Data En., vol. PP, p. 14, 2017.

[42] M. Staffelbach, P. Sempolinski, T. Kijewski-Correa, D. Thain, D. Wei, A. Kareem, and G. Madey, "Lessons learned from crowdsourcing complex engineering tasks," PLOS ONE, vol. 10, p. e0134978, 2015.