Using Crowdsourcing for Scientific Analysis of Industrial Tomographic Images

CHEN CHEN, National University of Singapore
PAWEŁ W. WOŹNIAK, Chalmers University of Technology
ANDRZEJ ROMANOWSKI, Lodz University of Technology
MOHAMMAD OBAID, Chalmers University of Technology
TOMASZ JAWORSKI, JACEK KUCHARSKI, and KRZYSZTOF GRUDZIEŃ, Lodz University of Technology
SHENGDONG ZHAO, National University of Singapore
MORTEN FJELD, Chalmers University of Technology

In this article, we present a novel application domain for human computation, specifically for crowdsourcing, which can help in understanding particle-tracking problems. Through an interdisciplinary inquiry, we built a crowdsourcing system designed to detect tracer particles in industrial tomographic images, and applied it to the problem of bulk solid flow in silos. As images from silo-sensing systems cannot be adequately analyzed using the currently available computational methods, human intelligence is required. However, limited availability of experts, as well as their high cost, motivates employing additional nonexperts. We report on the results of a study that assesses the task completion time and accuracy of employing nonexpert workers to process large datasets of images in order to generate data for bulk flow research. We prove the feasibility of this approach by comparing results from a user study with data generated from a computational algorithm. The study shows that the crowd is more scalable and more economical than an automatic solution. The system can help analyze and understand the physics of flow phenomena to better inform the future design of silos, and is generalized enough to be applicable to other domains.

Categories and Subject Descriptors: H.5.m [**Information Interfaces and Presentation (e.g., HCI)**]: Miscellaneous

General Terms: Design, Algorithms, Human Factors

Additional Key Words and Phrases: Silo, crowdsourcing, particle tracking, tomography

---

Chen Chen, Paweł W. Woźniak, Shengdong Zhao, and Morten Fjeld would like to thank the Swedish Foundation for International Cooperation in Research and Higher Education (STINT, grant 2013-019). The research leading to these results has received funding from the People Programme (Marie Skłodowska-Curie Actions) of the European Union's Seventh Framework Programme (DIVA, REA grant agreement no. 290227) and the Sixth Framework Programme—Marie Curie Transfer of Knowledge Action (DENIDIA, contract No.: MTKD-CT-2006-039546). Paweł W. Woźniak thanks The Adlerbertska Research Foundation for its support for this research.

Authors' addresses: C. Chen and S. Zhao, National University of Singapore, School of Computing, 13 Computing Drive, Singapore 117417; emails: a0095639@nus.edu.sg, zhaosd@comp.nus.edu.sg; P. W. Woźniak and M. Fjeld (corresponding author), Department of Applied IT, Chalmers Univeristy of Technology, 41296 Gothenburg, Sweden; emails: {pawelw, fjeld}@chalmers.se; A. Romanowski, T. Jaworski, J. Kucharski, and K. Grudzień, Institute of Applied Computer Science, Lodz University of Technology, Łódź, Poland; emails: {androm, tjaworski, jkuchars, kgrudzi}@iis.p.lodz.pl; M. Obaid, KUAR, Koç University, Rumelifeneri Yolu 34450 Sarıyer İstanbul, Turkey; email: mobaid@ku.edu.tr.
Author's current address: M. Obaid, KUAR, Koç University, Istanbul, Turkey.

DOI: http://dx.doi.org/10.1145/2897370

ACM Reference Format:
Chen Chen, Paweł W. Woźniak, Andrzej Romanowski, Mohammad Obaid, Tomasz Jaworski, Jacek Kucharski, Krzysztof Grudzień, Shengdong Zhao, and Morten Fjeld. 2016. Using crowdsourcing for scientific analysis of industrial tomographic images. ACM Trans. Intell. Syst. Technol. 7, 4, Article 52 (July 2016), 25 pages.
DOI: http://dx.doi.org/10.1145/2897370

1. INTRODUCTION

Intelligent systems will be developed more quickly by first creating crowd-powered systems before gradually moving to fully automated solutions [Bigham et al. 2014]. Accordingly, pushing the scope of problems solvable with crowdsourcing implies (a) engaging with expert crowds, (b) embedding needed expertise in the tools that nonexpert crowds use, and (c) using a flexible combination of the two approaches. Inspired by previous work [Cooper et al. 2011; Lasecki et al. 2012], we know that crowdsourcing can be used to help analyze large amounts of scientific data instead of assigning these time-consuming tasks to experts with higher costs and limited availability. This kind of participatory science is a promising new application domain for crowdsourcing. A notable example is the Foldit[1] project, which has developed a game wherein nonexperts help scientists figure out how proteins fold, which is important for understanding how to target them with pharmaceuticals. Eiben et al. [2012] and several others have successfully carried out and published research using Foldit. In this article, we focus on how crowdsourcing may assist particle-tracking methods. These methods have been extensively studied in the domain of analyzing fluid flow, based on determining the velocity vectors of moving particles. Our case concerns the flow of dense bulk solids in closed, opaque containers—an industrial application for which flow can be monitored using X-ray imaging. We expect that our work can inform a larger spectrum of problems under investigation, specifically those using the measurement techniques of Particle-Tracking Velocimetry (PTV), Particle-Imaging Velocimetry (PIV), and process tomography [Grudzien et al. 2012].

Bulk solids have significant bearing on industry and society, in which accurate design guidelines for silos and process flow are crucial to efficiency, safety, and sustainability. Bulk solids are mainly foodstuffs, construction materials, and pharmaceuticals, while approximately 60% of industrial solid materials are stored, transported, or processed as bulk solids [Seville et al. 1997]. There are a number of problems with silo utilization, from uneven flow to blockage to silo collapse. These eventual disasters caused by uneven flow lead to major environmental and economic impacts due to the waste created. Bulk solids experts [Schulze 2008] state that approximately 1% of silos collapse, which is a serious problem considering their widespread use.

Despite many years of scientific work, the physics behind bulk solid flow still requires more understanding. New measurement techniques have been introduced and more computer-processing power is needed. These developments allow the detection of tracer particles, introduced to the mixture of bulk solid; by using tomographic imaging to monitor their trajectories, the general flow of all particles can be derived. The movement of these tracers can be captured while the solid flows out from the silo, allowing experts to study what aspects of a silo's geometry or material properties affect flow. However, the resulting 2D tomographic images are challenging to interpret due to their low quality, and the reconstruction of tracer-particle movements in the images is tricky and laborious since thousands of images must be interpreted. Grudzień and Hernandez De La Torre Gonzalez [2013] showed that algorithms, even when made with careful parametric choices, still do not produce sufficiently good results when

[1]http://fold.it/portal/info/about.

analyzing images. From that experiment, they found it challenging to achieve even 50% precision by algorithm. The reason may be that some particles appear deformed or blurred in the captured images because they are in motion, thus increasing automatic recognition difficulty. On the other hand, experts can determine movement patterns quite reliably by reading these images. However, such a solution is very tiring, and reading thousands of images does not, in itself, result in a better understanding of flow phenomena—trajectories must be further analyzed after being tracked. Since experts must spend time performing deeper analyses and research rather than recognizing particles, we propose that nonexperts could also perform this task. We conducted a user study crowdsourcing the image recognition tasks, resulting in around 70% precision—much better than the automatic algorithm results.

In the context of this article, experts are defined as researchers with experience and understanding of the mechanics behind bulk solids and industrial process tomography, with more than 4yr of experience. We have engaged and worked closely with a group of such experts, who provided us with the problem space along with related materials. We have developed the crowdsourcing system in response. Nonexperts are defined as laypeople who have some experience with scientific subject matter and can interact with images using a computer interface. Both experts and nonexperts worked with the crowdsourcing system in the same way, but the experts also verified the work of the nonexperts to see if the proposed system was sufficiently accurate for analysis. These experts thus both informed and participated in our study.

We used this system to investigate ways to organize expert and nonexpert workers with directive crowdsourcing [Bigham et al. 2014], an interface that guides human computation toward a specific goal. It assists in analyzing the tomographic images, resulting in detailed data about the physical properties of bulk solids, which can inform design improvements for silos. In the future, enough data gathered from this process has the potential to inform a fully automated system—for example, the data can be used as a training set for artificial intelligence–based pattern recognition algorithms. We investigate whether people are effective at analyzing both fast and slow flow speeds, and establish a system of redundancy that allows for more iterative rounds of the analysis procedure. Since missing data from each particle that goes unrecognized may increase the probability of a collapse, we want to find out how many iterative rounds will balance the results and the human resources.

Our work presents four contributions. First, we have identified a novel application domain for human computation, specifically within the study of crowdsourcing. It is a unique approach that helps to solve a problem to date not fully understood [Schulze 2008]. We suggest ways to engage with both expert and nonexpert workers to achieve effective human analysis of tomographic images. Second, we have realized a crowdsourcing system design and its application, and present early insights for strategies to combine image-analysis results. Third, we conducted a computational analysis that compared the crowdsourced results against an automatic algorithm. We demonstrated that using human computation in this case was more effective and efficient, validating the results and effectiveness of the crowdsourced solution. Last, we contribute insights from this study that investigate an example problem from the research domain. We support the second contribution by proposing a generalization of our system (Figure 1) that may be scalable to other domains for which crowdsourcing workflows may apply. We believe that the system has potential applications in a number of industrial and medical research applications, especially when we deal with a sequence of images. We also present strategies to organize workers for this specific task and examine how such a crowd behaves in the given setting, what factors affect their performance, and their subjective assessments of the task.

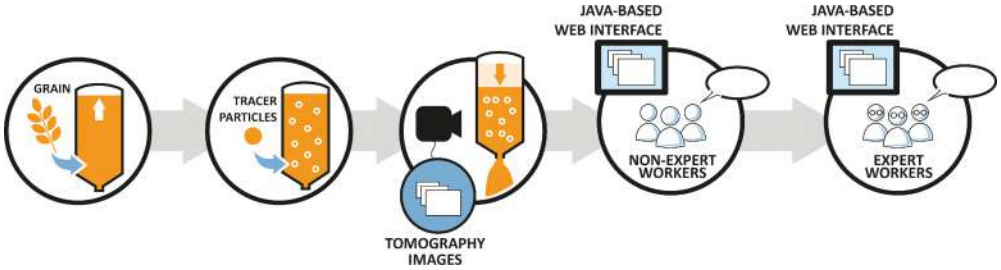

Fig. 1. Overall workflow, from left to right: Silo is filled with material mixed with tracer particles. Tomographic images are taken of the silo flow as it is emptied, images are sent to the web interface for analysis by nonexpert workers, then expert workers verify the analysis.

2. RELATED WORK

Understanding the problem at hand and how to best approach it involved reviewing a variety of research from different fields. The starting point was the human skill of visual recognition, and how it can be employed in human computing to solve problems in which computers struggle. Next, we drew upon research on the concept of crowdsourcing and how it can be applied. Utilizing knowledge and experience from these fields, the next step was to research the specific problem of bulk solid flow in silos after consulting experts in process engineering and industry. We reviewed research on flow imaging and analysis as well as on detection of slow-moving phenomena to inform this study and its design.

2.1. Flow of Bulk Solids

Bulk solids are materials made up of solid particles, such as powders, granulates, and grains. They are generally fairly slow moving due to being stored in large volumes. The handling and transportation of bulk solids is important to society; thus, the problem-free operation of bulk transport and storage systems is crucial. When transported between places, bulk solids are stored in silos on trucks, and transport within a factory flows between large hoppers of various kinds. Solids must flow between all of these different containers in a safe and efficient way. In almost all cases, material is meant to flow from the top of the silo to the bottom; similar blockage problems occur across all methods. A common example of blockage is the formation of stagnant zones of nonflowing material on the interior walls of the silo (Figure 2(b), right), reducing throughput and storage capacity as well as allowing product such as food to spoil.

According to Schulze [2008], many silos and hoppers are not built or engineered with the physical properties of particular solids in mind. This results in several problems, such as flow obstruction, fouling, or unwanted separation. In the worst cases, silo damage and collapse occur, which can amount to disaster for large volumes (Figure 2(a)). These problems have a widespread effect across industries regardless of the volume of solid, as [Schulze 2008] notes that the mechanical processes of flow that affect bulk storage are basically the same, independent of scale.

Research into the design of silos and hoppers has not developed sufficient guidelines to address every possible operating condition. Combined with the wide variety of properties of bulk solids, monitoring and experimenting become important methods for understanding flow phenomena in order to determine the best possible silo design for efficient, trouble-free, and safe flow. Data from experiments, combined with proven design methods, can determine the most appropriate geometry for bulk solid storage and transportation [Schulze 2008].

(a) (b)

Fig. 2. (a) Silo collapse spilling several tons of barley over railway tracks. Photo used with permission. (b) Example diagram of two silos: On the left, we see a silo exhibiting a flow that is typically preferred. On the right is a silo exhibiting a more typical "funnel flow." The reason for the formation of the stagnant zones of bulk solid (right, dark orange areas) is poorly understood, and the processes causing it are thought to lead to further and more complex flow problems. This is caused by incompatibility between the design of the silo and the flow properties of the bulk solid, leading to spoilage, reduced efficiency, and mechanical stress on the silo. Appropriately designed silos would allow for a more efficient flow, resembling the example on the left.

Babout et al. [2013] developed a method for analyzing bulk solid flow properties in the laboratory using a small silo subjected to X-ray tomographic imaging. The silo is filled with a bulk solid as well as numerous tracer particles that are more absorbable to the X-rays, making them more visible on the resulting images. By monitoring the position of the tracer particles traveling over time through the silo, one can detect trends in the movement of the bulk solid and therefore its flow properties. The particle trajectories are most interesting to experts; deviations to the trajectories have the potential to inform the science of understanding bulk solid flow. The example images in Figure 3 show cross-sections of silos filled with grain, with large tracer particles mixed in to track the downward movement of the material. Since these properties are generally independent of scale [Schulze 2008], a small silo may provide enough data to inform the design of a full-size storage and transport system. However, as described in the introduction, the detection of the tracer particles in the images is difficult and time consuming. Albaraki and Antony [2014] examined how the internal angle of hoppers affects bulk flow. In their experimental studies, they used PIV. They suggested a PIV technique to study granular materials as a way of not having to use any tracer particles, thus avoiding tomographic imaging.

Our work is relevant for a number of possible particle-tracking techniques, as it is independent of acquisition and processing modality. We aim to deliver solutions useful for the investigation and understanding of flow in complex multiphase phenomena currently investigated using PTV, PIV [Chung et al. 2010], process tomography, or Multiple-Particle Tracking (MPT) [Meijering et al. 2012]. These are relevant for research in fields such as biology, medicine, biomedicine, engineering physics, and chemical engineering.

2.2. Slow-Moving Phenomena

Crowdsourcing data has benefited the understanding and monitoring of slow-moving processes. For example, Lu et al. [2014] showed how crowdsourcing may work in pace with landslide monitoring; similar works have followed in related fields [Meissen and Fuchs-Kittowski 2014]. In ecology, the employment of people to validate scientific data

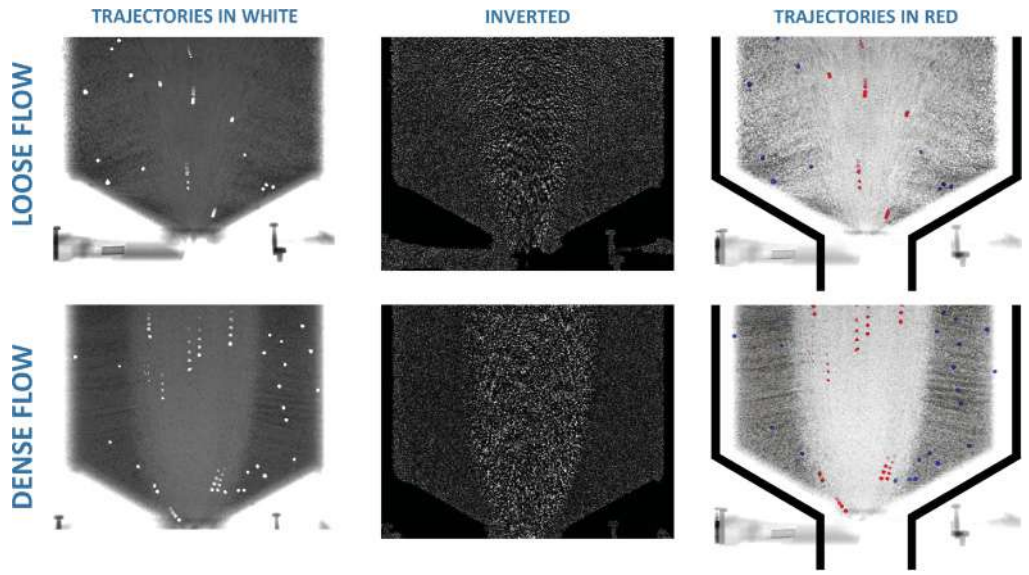

Fig. 3. Example tomographic images of trajectories in laboratory silos. Each horizontal series here—loose and dense—consists of the same respective images post-rendered in different ways to help the reader's understanding. The loosely packed particles that flow rapidly are difficult to track, while the dense particles have a higher contrast and are thus easier to track. Note the different patterns of stagnation on the inner walls of the silos. From left to right, the post-rendered images show: tracer particle trajectories shown in white; inverted image to reveal overall structure with stagnant zones; and a high-contrast image with illustrated silo outline, where tracer trajectories are shown in red and stagnant tracers are blue.

has also been examined [Bonter and Cooper 2012]. In FeederWatch, a continent-wide bird-monitoring program, a data-validation protocol was designed to increase researchers' and participants' confidence in the data being collected. In medical-image analysis, Eickhoff [2014] recently showed how "the comparably cheap results produced by crowdsourcing workers can serve to make experts more efficient and more effective at the same time." In short, crowdsourcing image analysis can allow medical experts more time to engage in activities such as treatment and research, rather than time-consuming repetitive analysis. Even with traffic monitoring, crowdsourcing has proven to be helpful in understanding the complexity of traffic congestion and related events [Artikis et al. 2014].

2.3. Crowdsourcing and Human Computation

Quinn and Bederson [2011] defined human computation partly as (a) computation problems that may one day be solvable by computers, and (b) computational processes that direct and assist humans to complete computational tasks. They emphasized the differences and commonalities between human computation and, for instance, crowdsourcing, social computing, and data mining. Crowdsourcing seeks to replace traditional or expert workers with workers from the general public recruited via open call [Quinn and Bederson 2011], thus can be used as a tool to employ nonexperts to provide computational skills to interpret data. Given the limitations of computers and algorithms, such an approach would offer a significant contribution to applications for which traditional data mining is insufficient. A crowdsourcing system may be expected to produce enough data to allow for better understanding of the processes involved and the creation of a fully automated system further down the road [Bigham et al. 2014].

In this context, Quinn and Bederson also provided a classification system for identifying different uses of human computation. In one of their examples classed as quality control—the Folditprotein folding game—they showed how players used a graphical interface to predict the folding of protein structures [Cooper et al. 2011]. In a second example—the ESP Game—human visual recognition skills were employed to validate answers provided by multiple contributors. In our own related but unique work, we developed a web interface leveraging people's innate visual recognition skills in order to enable a task that is not yet possible with computer algorithms. Like the Foldit game, our system also helps solve a real-world problem, for which detailed data is required to design appropriate solutions for bulk storage and transport [Schulze 2008].

In our work, following the classification developed by Quinn and Bederson [2011], we can describe our system as being about visual memory and sensitivity skills, supported by a system of redundancy—multiple answers for the same tasks are compared against each other, by both the machine and expert workers—weeding out mistakes and poor work. It could also be classed as an output agreement, in which completed tasks are accepted only if a group of users (in our case, experts) can agree on quality. The process order moves from the requester to the worker and then to the computer, which processes. There is a many-to-many task request cardinality, with several workers (either expert or nonexpert) working on each of several images [Quinn and Bederson 2011].

2.4. Recognizing Visual Changes

Though tricky, marking details and movement in the tomographic images of the silos is manageable enough for people to do. From our collaboration with tomographic imaging experts, we know that humans are equipped by nature to understand and remember the continuity between images. However, understanding changes and remembering continuity presents challenges that computer vision cannot currently solve. In a series of images, when a tracer particle appears distorted from one image to the next, a computer-vision system cannot maintain the continuity of the particle. According to the tomography experts, this is because the computer cannot know that the particle from before has distorted. It either registers it as a new particle or does not register it at all due to the fact that the particles, while they are physically spherical, may not appear spherical, a different shape than the computer was programmed to recognize. This would also be a problem if particles appear too close together. The human capacity for sensitivity can quickly understand the "story" of the particle's movement. Simple as that may seem, it involves the complex workings of the visual short-term memory system and the understanding that shapes can change. Humans have the ability to keep something in mind after it disappears, and to detect change and compare to the memory of the original, especially over the short term [Sligte et al. 2010]. While considerable research has been made over the years about how this process works in humans, current computer technology is unable to offer better performance than human computation. Simulating the neural networks and visual recognition of humans is in the experimental stages and currently not available for commercial use [Merolla et al. 2014]. However, gathering enough accurate data may help inform the design and training of an automated computer-vision system.

3. SYSTEM DESIGN

This section presents the tools developed for our experiment: the crowdsourcing system and the automatic algorithm for processing tomographic images. First, we introduce the crowdsourcing system.

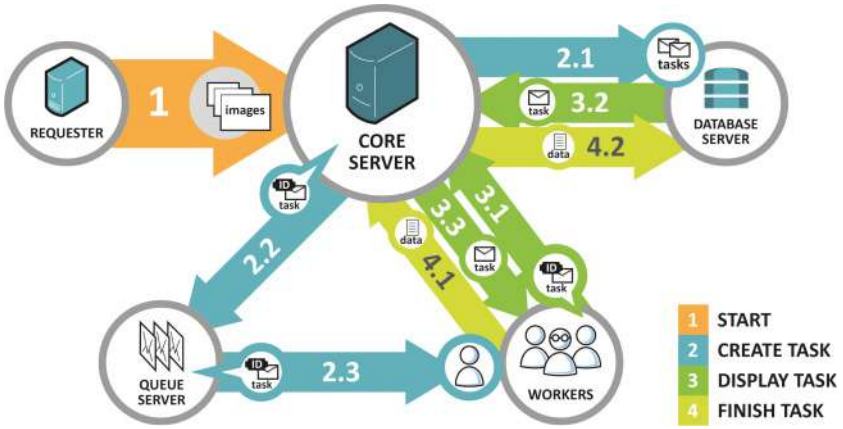

Fig. 4. The system structure implemented in the experiment (see Section 3 for details).

3.1. Crowdsourcing System

The crowdsourcing solution presented in this article was developed through a user-oriented design process in which four industrial process tomography experts were in constant contact with the design team. The inquiry began with semistructured interviews with two of the experts. The scientists explained the nature of their work and the variety of methods that they use to process tomographic images from silo models. Next, we proceeded to gathering requirements, scheduling four sessions with the experts. These interviews helped identify the target task and which datasets to use. Initial interface prototypes were discussed with the experts in two additional sessions.

We designed and implemented a crowdsourcing system[2] structure for analyzing images that operates in four stages (Figure 4). To start, the requester (whoever needs work done) uploads a series of images to a core server (1). The core creates one or more tasks for each received image and saves them in the database (2.1). At the same time, the core informs a queue server of these tasks by their tickets (2.2). The queue server then selects available workers, notifying them with a task ticket (2.3). After the worker receives the ticket, the worker requests the task from the core (3.1), and the core retrieves the task from the database (3.2), sending it to the worker (3.3). After finishing the job, the worker submits the results back to the core (4.1), which then updates the images database entry accordingly (4.2). Multiple workers will work on the same task, independently of each other. After all tasks for an image are finished, the results can be viewed by experts for verification.

This overall workflow is generalized enough to be scalable to other domains. With enough crowd work and expert verification, it is possible to produce a large amount of accurate data. As suggested by the experts, this data may be used to train and develop a system of computer vision to fully automate a task. This is in line with the statement of Bigham et al. [2014] about intelligent systems, as mentioned in the introduction.

3.1.1. Web Interface. Our interviews with the tomographic-imaging experts informed our decisions about what to include and pay attention to in the crowdsourcing interface. When marking tracer particles in each frame, a worker using the interface (Figure 5) *double clicks* to mark a new particle with a red circle, as shown in the figure, *drags* to

[2]The core server is built in the Spring Framework (version 3.2). We used Mongodb (version 2.4.9) as our database and developed the queue server with Redis (version 2.4.8) and Socket.io (version 0.9.11) to make task ticket queuing and to push new tickets to the workers. By using Mongodb, which is a NoSQL database,

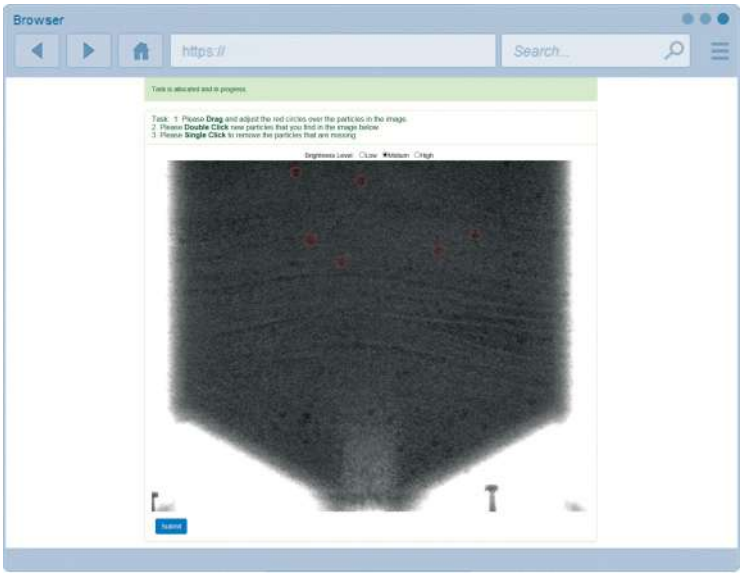

Fig. 5. An example of our web-based interface showing the particle marking and adjusting process.

adjust the position of a mark that has already been made, and *single clicks* to remove a mark. The worker then drags the marks to adjust their position to where the particle has been moved, and deletes marks for particles that have disappeared or left the silo. As the images are difficult and complex, workers are given the option to look at them using three brightness levels, changing visibility of different details in the images, as shown in Figure 3.

To better inform the interface, we conducted a pilot study with 5 nonexpert workers and asked them to rate the appearance of the marks after they finished their tasks. The results showed that all of them agreed that a circle was better, as compared to quadrant marks and crossed marks. The most common reason was that the particle was also a circle; thus, they could better see its distance to the edge of the circular mark. Four agreed that a red color was easy to recognize; they also suggested using a thicker circle to make it easier to see.

3.1.2. Task Design. In the experiment, *frames* are tomographic images uploaded from the requester server to the core server. A *data set* refers to a sequence of consecutive frames. *Tasks* are generated by the core server after it has received frames sent from the requester. Workers work on one task at a time, identifying and marking all the particles in the image shown to them in each task. The task design consists of three aspects: *interaction elements* in each task, *control flow* of all tasks, and *aggregation*.

Interactions define the operations that workers can perform when completing tasks. They are: *double click* to draw a circular mark around a tracer particle, *single click* to remove a mark, and *drag* to move marks. The system records all mark coordinates and submits them to the server for later processing.

Control flow defines how tasks are organized throughout the process. We define two tasks as being sequenced if they have dependency relationships and as being paralleled if they are independent. To start, the first task for all frames is sent to workers. In the

we can easily adjust the data structure to change the system to complete different tasks. By using Redis, we can customize the priority of the tasks and control the order of tasks sent to workers.

first round, workers will either see an image without marks (if no frame's first task has been completed) or an image with marks (from the nearest frame in the series that has had its first task completed). In the former, workers add marks to the images. In the latter, the marks will have a large offset in relation to the particles, since particles move (or disappear) between frames. Workers must thus adjust the accuracy of the marks. Among all tasks in the first round, the tasks are sequenced since the initial particles shown in a task are based on previously accomplished tasks.

After all first tasks are completed, the second tasks for all frames are sent. In this second round, workers will see the marked images from the first tasks of their respective frames. Most marks will be accurate, with only a tiny offset. Here, workers review marks from the previous round. Subsequent rounds repeat the second round over and over until all tasks are completed. In the second and subsequent rounds, tasks are sequenced among two rounds and paralleled within each round.

Aggregation is a step merging particles marked in the tasks to particles in the frames. We applied a machine algorithm to aggregate results from task n to $(n-1)$ of the same frame: (1) for each mark x in task n, find its closest mark y in task $(n-1)$; (2) if their distance from each other is smaller than a threshold (equal to the average radius of the biggest three visible particles), merge mark x and y to a midpoint; (3) if their distance is larger than the threshold, save mark x as a new mark of the frame. This approach will produce particle data in each frame and reflect the trajectories of the tracer particles in the dataset, as shown in Figure 9.

3.2. Algorithmic Approach for Automatic Tracer-Particle Detection

The image processing algorithm was based on the well-known Hough Transform (HT) [Hough 1962; Mukhopadhyay and Chaudhuri 2015], modified for circle detection [Yuen et al. 1990]—CHT, with a gradient-based detection kernel [Atherton and Kerbyson 1999]. During preliminary analysis, other HT-based methods such as RCD [Chen and Chung 2001] or RHT [Xu and Oja 1993; Xu et al. 1990], along with non-HT methods such as genetic algorithms [Ayala-Ramirez et al. 2006] were considered.

Past work in the area [Grudzień and Hernandez De La Torre Gonzalez 2013] indicated that design of any algorithm addressing radiography particle tracking is highly constrained by the need to produce visually meaningful data for further analysis, the intricacies of the measured material (the X-ray absorption rates for sand are relatively very high), and the measurement setup. In an X-ray system for measuring the silo discharging process, in which the granular material is sand, X-ray radiation by flowing material is very high. Even accurate empirical selection of the X-ray system parameters (current and voltage settings on the X-ray tube, exposure time, magnification, and field of view) cannot ensure high signal-to-noise ratio (SNR) in radiographic images. Improvement of SNR could be easily achieved by, for example, increasing exposure time; however, such a solution increases blurring on the radiographic images as the result of gathering signals in time on the detector. The proposed methodology (based on HT) appears to be the optimal approach to detect particles given the characteristics of the data obtained. Other types of algorithms often require binarization during image processing. The low SNR level of the radiographic images leads us to the conclusion that any approach requiring binarization is simply too expensive in terms of intensity information loss. Figure 6 shows an example set of enlarged images of a single observed tracer particle (a metal ball, i.e., an X-ray absorbing material) and their contrast with the background.

The HT required significant modification to process the radiographic images to offer optimal methods for particle detection. We adapted the algorithm accommodating the additional information provided by the physical aspects of the registration method. First, the metal balls used in the experiment have equal radii. Any negligible size

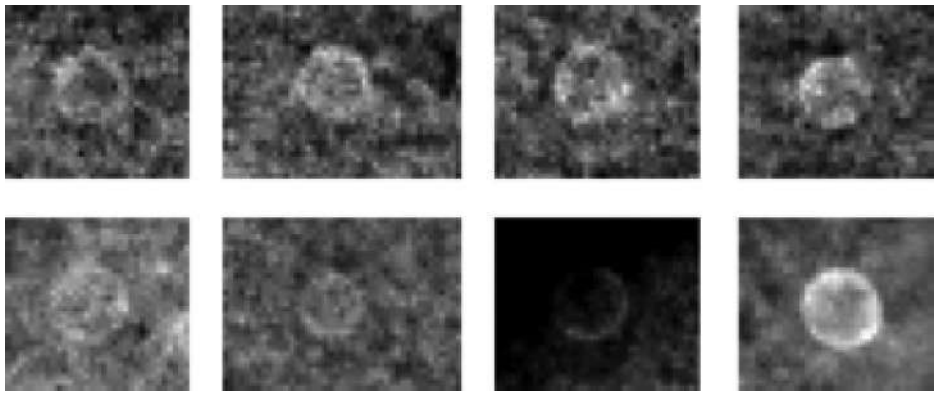

Fig. 6. Typical shapes of particles that can be found in the loose flow (top row) and dense flow (bottom row) datasets. The images were obtained during flow experiments. Images are enlarged with no pixel approximation method (or nearest neighbor). Note the low SNR of the images, which effectively limits the variety of applicable detection algorithms.

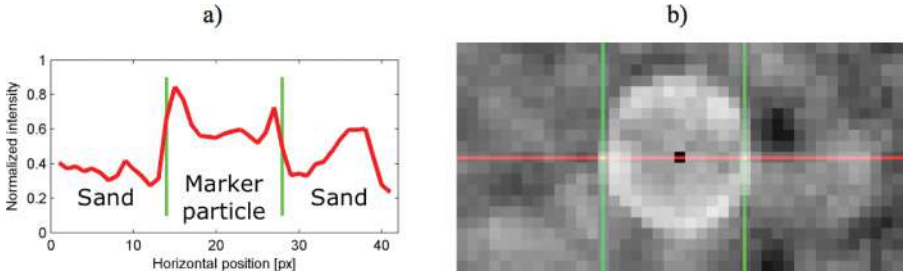

Fig. 7. An intensity profile (a) obtained along the red line, as shown in (b).

variation from an analysis point of view observed in registered images is caused only by perspective projection, in which the distance between particle and X-ray sensor plays an essential role. Second, particles appear not as a solid objects (circles) but rather as an intensity ring, in which the intensity is radius-dependent, as shown in Figure 7. This can be considered as registering pseudo-gradient images, since the X-ray intensity at the sensor is the inverse of absorption of the material along the emitter-sensor ray.

To account for these insights, we have chosen a method of size-invariant kernels for circle detection [Atherton and Kerbyson 1999], which allows configuring the degree of size invariance. This reduces the classical CHT parameter space to two dimensions and can be implemented as convolutional filtering. Another important aspect of the applied algorithm was the image preprocessing step. Each image fed to the HT algorithm was preprocessed in order to normalize differences in the insensitivity between the silo wall area and the center of the image. During the normalization procedure for each point in an image, a pair of values (minimum and maximum) was determined. These values were obtained by averaging a set of frames for a full silo (before adding tracer particles) and for an empty silo (after the discharging process). This allows us to reduce the intensity distribution in the X-ray images, along with the perspective projection distortion related to the size of the X-ray emitter, the silo, and the X-ray detector.

It is worth mentioning that the algorithm does not use temporal information about particle position in previous radiographic images. It may appear that considering the temporal aspects of particle position should produce higher algorithm performance; however, two physical aspects of the registration method limit the utility of such a

method. In fact, incorporating the temporal aspects of the data may cause errors in result interpretation. Due to low SNR in radiographic images and the fact that our analysis is limited to the 2D view (without 3D reconstruction) of granular material distribution in the silo, the particles may disappear from radiographic images at some time points. It is possible that, on the path between the X-ray source and detector, the changes in sand distribution and/or concentration cause high X-ray absorption, such that tracer particles would no longer be visible on the radiographic images (i.e., the SNR level would be too low for detection, but the physical metal balls would still be present in the sand mixture). Moreover, changes of particle position in 3D space cannot give us, for such a measurement methodology, any information about particle movements because the images are acquired in 2D. Consequently, our approach ignores the temporal aspect of the data in order to ensure the best possible precision.

4. METHODOLOGY

After the two systems were implemented, we proceeded to conduct their evaluation. Our evaluation strategy consisted of four activities. First, we conducted a study with nonexpert users in which they processed two different datasets at two different task-per-frame rates (i.e., the number of times that a single image is processed by the crowd). This enabled us to evaluate the feasibility of the technical solution, identify whether there are significant differences between the datasets in terms of human computation, and determine the adequate crowd effort to perform the task. Second, we conducted a study with expert users. In this activity, experts identified particles in two datasets to establish a near-to-ground truth for further comparisons. They also subjectively rated the results obtained from the crowd. Third, we compared the performance of the crowd system and the automatic algorithm using the near-to-ground truth produced by the experts. Last, we introduced a third dataset to test the scalability of both systems.

4.1. Crowd Study

We conducted an in-house experiment to evaluate the crowdsourcing system. In the experiment, we considered two factors. First, we specified two types of datasets—loose or dense—for which the material has different densities. The experts defined the two types as being more or less difficult to analyze, respectively. Second, we counted the number of tasks allocated and distributed per frame: three or six.

Thirty-two nonexpert participants from the university community (16 males and 16 females, all right-handed, age range from 22 to 44, $\mu = 29.5$, SD $= 5.2$) were recruited. Two of the expert participants who had previously informed our study were also recruited.

A between-subjects design was used with the group of nonexpert workers. We made a sequence of tasks with both the loose and dense datasets, and small and large numbers (three and six) of tasks per frame. It was counterbalanced, with four conditions, in the following order: 3-dense, 3-loose, 6-dense, and 6-loose. For example, the first group of nonexpert workers did tasks using a dense dataset and with three tasks generated for each frame. We had eight participants for each condition, who were tested in separate sessions each with different participants, making for a total of 32.

Previous work [Park et al. 2014] described alternative methods to organize training for crowdsourcing. At the beginning of the experiment, we gave participants a training session of ten tasks per participant, on average, to familiarize them with the interface and tasks. They were also provided instructions on paper with both text and images explaining what a particle is and how to use the system. The experiment leader went through all the instructions with the participants. After the training session, the participants knew how to interpret the images and process them efficiently.

In our design (excluding practice tasks) we had a total of two datasets—dense and loose—and two levels of task numbers—three and six—for each frame. There were 137 frames in each dataset, making for 548 frames in the study. There were 2,466 tasks in total. There were 51.375 tasks per person for the conditions 3-dense and 3-loose, and 102.750 tasks per person for conditions 6-dense and 6-loose.

We measured task completion time by average time taken for task completion (the interval between the worker receiving the task and pressing the "submit" button), accuracy of each frame (precision score and sensitivity score), and accuracy of identified particle trajectories (evaluated by experts). Since it is hard to objectively estimate the total number of tracer particles visible to the naked eye, an absolute ground truth is difficult to establish. However, we suggest employing a near-to-ground truth based on expert results of identifying the particles in the frame. After the experiment, participants were asked to grade their subjective experience on the NASA TLX scale.

4.2. Expert Study

In order to generate near-to-ground truth results for each dataset, we used results from four different tomography analyses. Two experts completed one task per frame of each dense-flow dataset, then evaluated the work of the nonexperts working on each loose-flow dataset. Two different experts did the same for each loose-flow dataset, then evaluated the nonexpert work on the dense-flow dataset. We generated only one task for each frame since the experts knew what to identify as particles.

The crowdsourcing system was used to gather data from experts. Before the experiment, experts performed the same training session as nonexperts had completed, with instructions and ten tasks per expert, which familiarized them with the system. In each session, an expert processed 137 tasks (excluding the training session). Afterwards, the experts filled out a questionnaire evaluating the quality of crowd work by judging the number and quality of the trajectories in the resulting images. Each dataset was processed by both experts. We asked them to identify the trajectories in each result image and to give a score between zero and ten to each image, zero being the worst and ten the best.

4.3. Comparison with the Automatic System

We now present how the automatic algorithm was applied and the methods used to compare the crowdsourcing system with the automated solution.

4.3.1. Apparatus. The algorithm for automatic localization of trace particles, based on HT as described in Section 3.2, was implemented in the MATLAB environment, as we found it suitable for fast algorithm prototyping. However, performance assessment is fairly coarse. For the MATLAB-based implementation, run on an Intel i3 3.3GHz, Windows 32b platform, the amount of time needed to analyze one image is about 1.6s. This implied that the algorithm was sufficient for a proof-of-concept experiment, especially taking into account the fact that on-line analysis is not required for bulk-flow process radiography.

The algorithm used the same input data as the crowdsourcing system—a set of subsequent TIFF images. It also provided the same output—sets of particle coordinates for each of the input frames. The full compliance with respect to format allows for fast comparisons. We used the distance between tracer particles identified by the crowd, the experts, and the algorithm to conduct the comparison.

4.3.2. Comparison Method. In order to further assess the feasibility of the crowdsourcing system, we conducted a comparative analysis. The results obtained by the crowd in various combinations were compared with the results obtained by the proposed

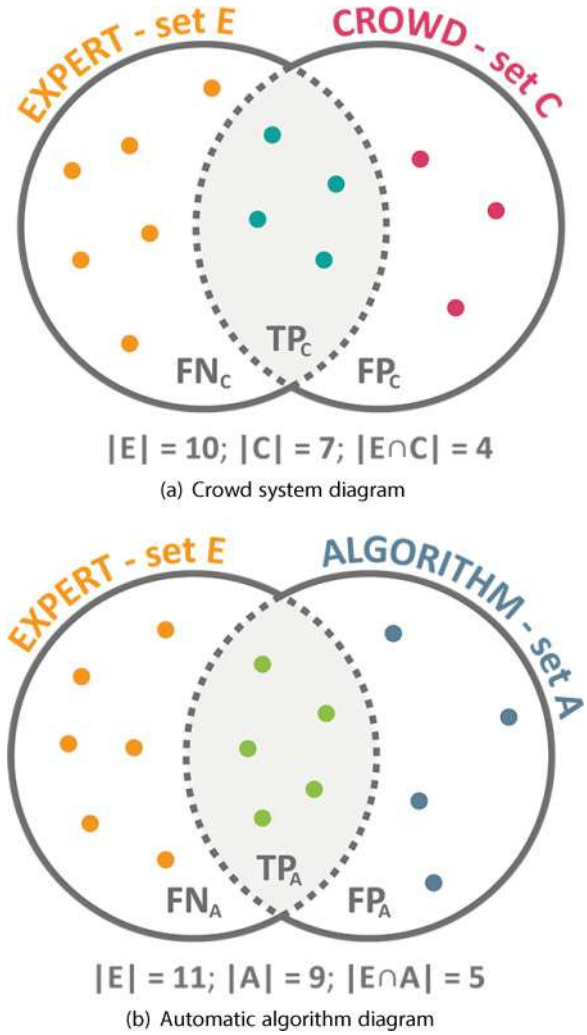

Fig. 8. Diagrams illustrating the method used for assessing the quality of the crowd system (a) and the algorithm (b): For the first diagram (a), orange particles are those identified by the expert only ($FN_C = 6$), dark-green particles were identified by the crowd and confirmed by the expert ($TP_C = 4$), and red particles were detected by the crowd only ($FP_C = 3$). For the second diagram (b), orange particles are those identified by an expert only ($FN_A = 6$), green particles are those identified by the algorithm and confirmed by the expert ($TP_A = 5$), and blue particles were detected by the algorithm only ($FP_A = 4$).

algorithm. The quality assessments for both classification approaches were based on the data obtained by the experts.

The performance of the crowdsourcing system and the algorithm were assessed using the methodology shown in Figure 8—a confusion matrix. Classification results obtained from the crowd and the algorithm were referred to the expert results (the assumed ground truth) and presented in terms of *sensitivity*, *precision*, and *accuracy*. A full confusion matrix cannot be computed, as identification of objects that are not particles is not the case in the presented research (there is no true negative cases).

Assuming that an incomplete confusion matrix for each of the frames of the measurement is computed for both the algorithm and the crowd, the following measures can be obtained (C denotes the crowdsourcing system and A denotes the algorithm):

—TP_C (or TP_A): true positives—the total number of particles detected by the crowd (TP_C) or algorithm (TP_A) and confirmed by the expert;
—FP_C (or FP_A): false positives—the total number of particles detected by crowd (FP_C) or algorithm (FP_A) and not confirmed by the expert;
—FN_C (or FN_A): false negatives—the total number of particles detected by the expert, but not the crowd (FN_C) or the algorithm (FN_A).

A full confusion matrix cannot be computed, as the experts do not identify objects that are not particles. The relationship between TP, FP, and FN and the total count of particles P (their population) is defined as follows:

$$P_i = TP_i + FP_i + FN_i. \tag{1}$$

The index $i = A, C$, where A stands for the algorithm and C for the crowd (e.g., P_A is the total number of unique particles detected in the expert-algorithm comparison).

The ground truth value (the total number of particles found by an expert in a given dataset) is defined as follows:

$$G = TP_A + FN_A = TP_C + FN_C. \tag{2}$$

Consequently, the following algorithm quality indicators can be defined for the incomplete confusion matrix and these measures. The classifier sensitivity (algorithm or crowdsourcing) $SENS$—also known as recall—defines the probability of the classifier detecting a particle confirmed by the expert.

$$SENS_i = \frac{TP_i}{TP_i + FN_i} = \frac{TP_i}{P_i} \tag{3}$$

The classifier precision $PREC$ is defined as the probability of a particle detected by the classifier being a particle also identified by the expert.

$$PREC_i = \frac{TP_i}{TP_i + FP_i} = \frac{TP_i}{P_i} \tag{4}$$

Finally, the classifier accuracy ACC is defined as the probability of correctly detecting a particle, and can be used as an overall measure of the performance of this method.

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

A particle is treated as having been correctly detected by measuring the Euclidean distance between the detected points. For a given particle x detected by the classifier, if a particle y detected by the expert exists so that $d(x, y) < D$, then the particle is treated as detected. D is defined as the average radius of the three largest particles obtained manually from the data set.

4.4. Evaluating the Scalability of the Solution

In order to test the scalability of the solutions, we introduced an additional dataset called 3-sandpaper (using images from a different silo type—a silo with a sandpaper wall, and generating 3 tasks per frame) and processed it both using our crowdsourcing system (with $n = 8$ participants) and the algorithm. A single expert was also asked to process the dataset. We reapplied the comparison method presented earlier to see how the two systems were affected by the dataset change.

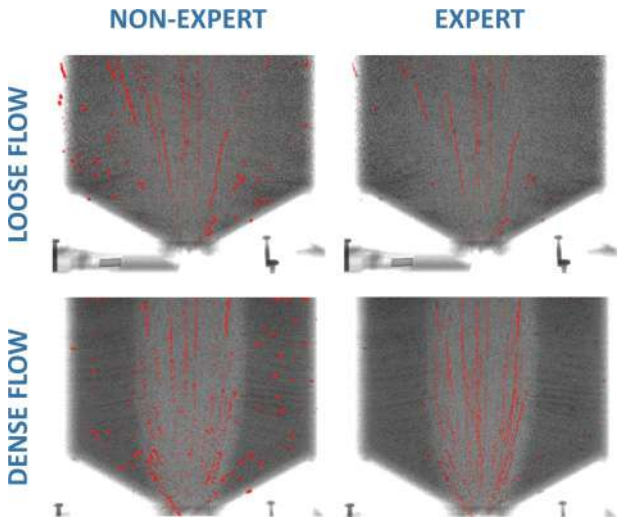

Fig. 9. Results from experiment comparing accuracy of trajectories between expert and nonexpert workers for both loose and dense flows, without eliminating particles that are identified by only one worker.

5. RESULTS

This section shows the results of the four phases of the evaluation. First, we present the results of the evaluation of the crowdsourcing system. Next, we present how experts assessed the results produced by the crowd. Finally, we present the comparison between the automatic and crowdsourcing methods.

5.1. Crowd Study Results

For each condition, we could generate result images as seen in Figure 9. We also ran a two-way between-subject ANOVA for each dependent variable and calculated statistics on the frame level. This section presents the effects of varying the task-per-frame rates and the dataset on completion time and performance.

5.1.1. Task Completion Time. We measured the average time for processing each task; the average time for each frame is the sum of the average time of all of its tasks. We recorded the timestamp for when a worker received the task, and the timestamp for when the worker pressed the "submit" button. The interval was the time for processing this task.

We found a significant main effect of the dataset type on time $F_{1,544} = 9.04$, $p < .01$. Pairwise t-test (with Bonferroni correction) shows that processing each task in the loose dataset ($\mu = 67.89$s, SD $= 22.69$s) is significantly faster ($p < .01$) than processing each task in the dense dataset ($\mu = 74.12$s, SD $= 26.20$s). This suggests that nonexpert workers can process images in the loose dataset faster than processing images in the dense dataset.

There was also a significant main effect on the number of tasks for each frame $F_{1,544} = 10.63$, $p < .01 (p = .001)$. Pairwise t-test shows that, if we generate six tasks for each frame ($\mu = 67.62$s, SD $= 22.26$s), work is significantly faster ($p < .01$) than if we generate three tasks for each frame ($\mu = 74.38$s, SD $= 26.50$s).

5.1.2. Performance. In order to establish the performance of the system, we consider the precision score and the sensitivity score for each frame (Figure 10). Precision and sensitivity are defined as in Equations (4) and (3).

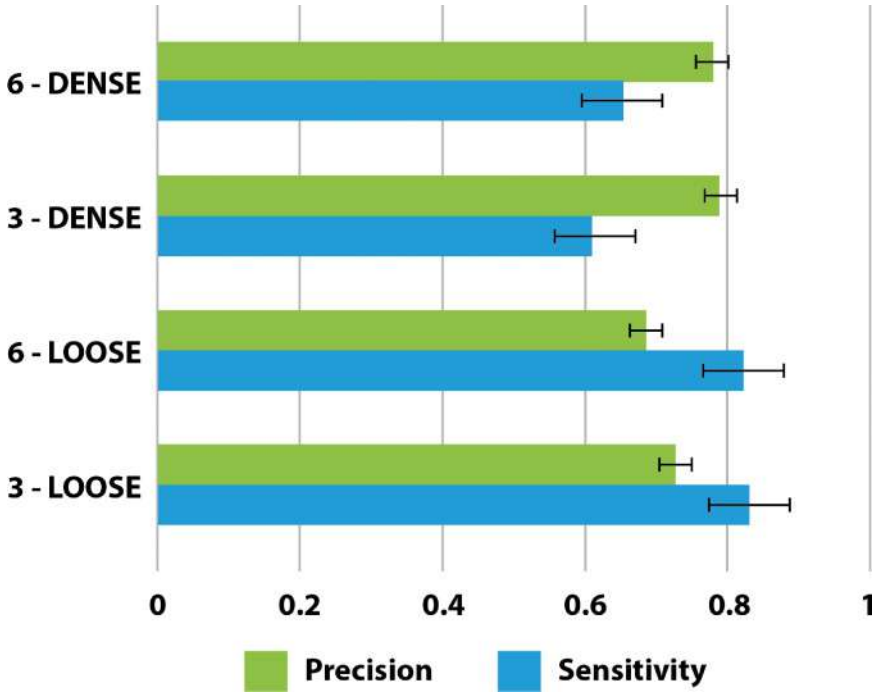

Fig. 10. Precision and sensitivity for the four experiment conditions. Note that, for the dense dataset, an increase in the number of tasks per frame yields better sensitivity, while the effect was not observed for the loose dataset. Error bars are .95 confidence intervals. The results on the graph are provided without eliminating particles that were identified by only one worker.

We assumed that expert results were near-to-ground truth results. The precision score of expert results should thus be approximately equal to 1.0, which means that every particle identified by the expert is correct. The sensitivity score should also be approximately equal to 1.0, which means that the experts identified all particles on each frame.

We measured precision and sensitivity scores for each frame completed by nonexpert workers by comparing its result to the corresponding frame result from the experts. Since the system has a feature to eliminate particle results that are identified by only one worker, we calculated the scores separately based on whether we would eliminate those results.

Without eliminating particles that are identified by only one worker, particle results will include every particle identified by the nonexpert workers. In this situation, there was a significant main effect of the dataset type on the precision score $F_{1,544} = 92.43$, $p < .001$ and on the sensitivity score $F_{1,544} = 746.98$, $p < .001$. Pairwise t-test shows that the results of the dense dataset ($\mu = .783$, SD $= .082$) are significantly more precise ($p < .001$) than results of the loose datasets ($\mu = .706$, SD $= .106$). However, the sensitivity of the dense datasets ($\mu = .630$, SD $= .089$) is significantly lower ($p < .001$) than that of the loose datasets ($\mu = .825$, SD $= .080$). Dense flow has a higher precision score than loose flow, meaning that workers are less likely to identify wrong particles in dense than in loose flow. However, dense flow has lower sensitivity scores, meaning that there are also more particles that are difficult for workers to identify.

There was also a significant main effect of task number for each frame on precision scores $F_{1,544} = 9.97$, $p < .01$ and on sensitivity scores $F_{1,544} = 6.95$, $p < .01$. Pairwise

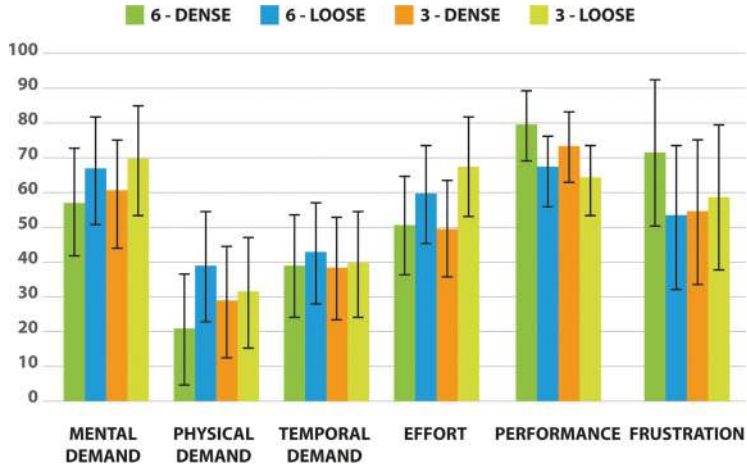

Fig. 11. Average NASA TLX scores for 32 participants, split into 4 sets (dataset type—number of tasks per frame). Error bars are .95 confidence intervals.

t-test shows that, if we generate 3 tasks for each frame, the precision score ($\mu = .757$, SD = .104) will be higher, but the sensitivity score ($\mu = .719$, SD = .141) will be lower than if we generate 6 tasks, for which the precision score is .732 (SD = .099) and the sensitivity score is .737 (SD = .115). With increasing task numbers for each frame and thus increased review rounds, the increases in sensitivity and precision scores are predictable. With more review rounds, more correct particles will be identified, while, of course, more wrong particles will also be identified. This result meets our expectation.

Finally, we found an interaction between dataset type and task number for the sensitivity value $F_{1,544} = 11.37$, $p < .01$. For the dense dataset, when the number of tasks for each frame is increased, the sensitivity value also increases from .609 to .652. However, for the loose dataset, with an increase in the number of tasks for each frame, the sensitivity value does not seem to change too much: from .828 to .823. We can see that the optimal number of tasks for one frame will be different based on different data sets. For the dense data sets, we can generate more tasks for one frame to get a better sensitivity, for loose data sets, three tasks are enough.

When eliminating particles that are identified by only one worker, we still found a significant main effect of data set type on precision scores $F_{1,544} = 63.17$, $p < .001$ and sensitivity scores $F_{1,544} = 661.56$, $p < .001$ and a main effect of task number of each frame on precision scores $F_{1,544} = 34.79$, $p < .001$ and on sensitivity scores $F_{1,544} = 14.94$, $p < .001$. Moreover, for both independent variables, if we eliminate the particles that are identified by only one worker, we can find an increase in precision but decrease in sensitivity. This is because, in the eliminated particles, there are both particles incorrectly identified by only one person and correct particles that are identified by only one person.

5.1.3. NASA TLX Results. All 32 participants completed the NASA TLX questionnaire [Hart and Staveland 1988] after the study (see Figure 11). The participants reported that the task was moderately mentally demanding (mean score of 64.00). Physical demand and time pressure were not perceived as problems. The participants reported that they performed the task correctly (mean performance score observed was 71.25). The task also produced a moderate level of frustration in the participants (mean frustration score was 60.00).

Table I. Expert Evaluation of Datasets

	A			B		
	Expert-loose	3-loose	6-loose	Expert-dense	3-dense	6-dense
No. of trajectories	14	13,5	14	14	12,5	12,5
Score (out of 10)	9	8,5	9	9,5	8,5	9

Note: In (A), experts working on the dense dataset evaluated the results of the loose dataset; in (B), experts working on the loose dataset evaluated the results of the dense dataset.

In order to investigate whether the number of tasks or dataset type were affected by subjective measures as reported by the workers, we performed 12 Wilcoxon Signed-Rank tests with Bonferroni correction. The only significant result was obtained for dataset type and perceived performance ($p < .05$).

5.2. Expert Results and Feedback

There were two experts. The first expert worked on the dense dataset and was evaluated by the second expert. The second expert worked on the loose dataset and was evaluated by the first expert. Afterwards, these two experts switched roles, and the same procedure was repeated. Also, each expert evaluated all results for both loose and dense datasets done by crowd workers. From Table I, we can see that the experts can find almost all trajectories from nonexpert results, with a difference of one to three trajectories, and that both experts gave high scores for nonexpert results. The expert evaluating the loose dataset even gave higher scores to the nonexperts' work. Finally, both experts said that the crowdsourcing results would help them very much with their work.

5.3. Comparison Results

Nine datasets were obtained to obtain the result of the comparison, products of comparing the observations by the two experts, the crowd, and the algorithm. We compared results for different material densities, different task-per-frame ratios, silo types, and the way that the dataset was presented to the experts. The following datasets are presented (X denotes that a dataset was available for experts A and B):

—3-dense-X and 6-dense-X: Image set from a smooth-wall silo filled with high-density material analyzed by expert A or B, the algorithm, and the crowd with low (3-*) and high (6-*) task-per-frame ratios[3].
—3-loose-X and 6-loose-X: Image set from a smooth-wall silo filled with low-density material analyzed by expert A or B, the algorithm, and the crowd with low (3-*) and high (6-*) task-per-frame ratios.
—3-sandpaper-B: Image set from a sandpaper-wall silo filled with low-density material analyzed by expert B, the algorithm, and the crowd with a low task-per-frame ratio. This dataset was used for additional scalability testing.

Table II presents the results of the comparison; relative quality measures for particle detection for the algorithm and the crowd were obtained using Equations (3) to (5). It can be observed that the accuracy index for the crowd outperforms the automatic algorithm in all cases. This is even true for an additional dataset (3-sandpaper-B), which was not available at the design stage for both the crowdsourcing system and the automatic solution.

An additional frame-by-frame analysis of the number of particles detected shows that, while the ground truth expert result shows a number of particles with low variability (e.g., for 6-dense-A, the experts found $\mu = 53.59$ particles per frame with $\sigma = 3.83$),

[3]We use the star (*) notation here as a wildcard, as in file names—*-B denotes all image sets with names ending with B.

Table II. Comparison of Particle Detection Performance Between Our Crowdsourcing System and Our Automatic Algorithms

Dataset	$PREC_C$	$PREC_A$	$SENS_C$	$SENS_A$	ACC_C	ACC_A
3-dense-A	0.6157	0.2732	0.7841	0.7946	0.5264	0.2551
3-dense-B	0.9553	0.5261	0.4865	0.6581	0.4757	0.4132
6-dense-A	0.6537	0.2732	0.7741	0.7946	0.5489	0.2551
6-dense-B	0.9151	0.5261	0,5011	0.6581	0.4789	0.4132
3-loose-A	0.7833	0,4386	0.7387	0.4130	0.6133	0.2702
3-loose-B	0.8324	0.4437	0.7021	0.3736	0.6151	0.2544
6-loose-A	0.7795	0.4395	0.7122	0.4129	0.5928	0.2705
6-loose-B	0.8251	0.4440	0.6750	0.3736	0.5905	0.2545
3-sandpaper-B	0.6262	0.4578	0.2565	0.1759	0.2225	0.1456

Note: The two datasets from the crowd study are used along with an additional dataset that features a different experiment configuration. Results without eliminating particles that were identified by only one worker were used for the crowdsourcing system.

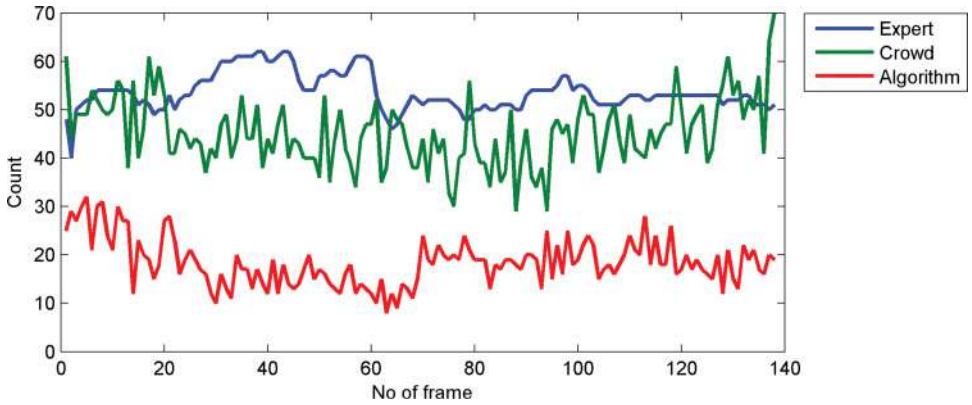

Fig. 12. The number of particles detected in each frame of the sample by the expert (blue), the crowd (green) and the algorithm (red) for the 6-dense-B dataset.

both the crowdsourcing system ($\mu = 45.25$, $\sigma = 7.22$) and the algorithm ($\mu = 18.15$, $\sigma = 4.89$) produced more variable results. This is illustrated in Figure 12.

6. DISCUSSION

We interpret the results of our experiment in terms of implications for crowdsourcing design. We also address the advantages of a crowdsourcing solution over an automatic system. Finally, we discuss future research based on our work.

6.1. Implications for Crowdsourcing

Task completion time measurements showed that the datasets perceived as easy by the experts received less attention from the nonexpert workers. This may be explained by the fact that the experts' perception of the images is different than the nonexpert workers, due to the nature of the experts' work. This result suggests an implication for the design of future systems, for which the images should be examined by nonexperts in prestudies to assess their complexity. A future complexity assessment, in turn, would be useful to determine the number of times that each image needs to be processed. While we observed a positive impact of increasing the number of tasks per frame on

the output quality for the dense datasets, such an effect was not observed for the loose datasets. This implies that, in future crowdsourcing system design, the difficulty of the datasets needs to be assessed and the parameters of the crowdsourcing method needs to be adjusted accordingly. When we combine this result with the subjective expert assessments of the crowdsourced results, one may wonder what could be the lowest number of tasks per frame that produce an acceptable result. This may be a key question for minimizing implementation costs if future systems are to be scaled up and eventually become training sets for AI-based solutions. Furthermore, one must not forget that, while an increased task-per-frame ratio may yield more accurate results and shorter per-frame processing times, it will still generate longer total processing times and, consequently, higher costs.

For the task of detecting tracer particles in tomographic images, we have shown that using crowdsourcing yields accurate results. We have gained novel insights about how to build a crowdsourcing system and how to organize the crowd for this specific task. In terms of their accuracy and sensitivity, we have learned how a nonexpert crowd behaves when using the proposed system and, with the given results, consulted the same experts as in the beginning of our investigation. They explained that the larger precision score for the dense dataset could be explained by the fact that the contrast between the funnel zone and the stagnant zones facilitates particle detection within the funnel. This may make nonexpert workers focus on the funnel zone, and thus miss particles in the stagnant zones. The loose dataset instead produced more uniform images, which led to them investigating the entire image, allowing for better sensitivity. The scores may also be affected by the fact that the number of particles between datasets is not equal. The experts noted that the number of particle markers in a given sample at a given time interval cannot be controlled, as the particles are combined with sand in an industrial mix.

The NASA TLX questionnaires completed by study participants indicated that the task was cognitively demanding. We believe that this is acceptable for such a proof-of-concept inquiry as presented in this article, but future solutions should seek to alleviate this problem. One solution could be to reduce the task-per-worker load and to make workers process a lower number of frames at a time. This, however, may result in lower accuracy, as workers will be less familiar with datasets and will not be able to capitalize on the visual memory capabilities unique to human computation. The nonexpert's perceived performance measure correlated with the dataset type. This is an interesting result, as it was not reflected in task completion time measures. It is possible that different varieties of images produce a subjective impression of complexity on workers, which may not be reflected in objective performance.

We observed that the potential for using crowdsourcing for processing flow images was confirmed by the experts, with one even ranking the crowdsourced result as being more accurate than the expert's own work. This shows that experts are eager to use the help of the crowd and are open to new solutions. It also confirms that we managed to find a problem space in which other methods have been exhausted, and that a proper crowdsourcing system can produce tangible solutions. Furthermore, the participants' assessment of their own performance (as measured by the NASA TLX) corresponds to the expert assessment. In general, these results positively support using a workflow such as our own for the utilization of human computation in scientific-image processing. It may be generalized and scale to other domains for which a crowdsourcing workflow may apply.

6.2. Advantages of Crowdsourcing Over Automation

The results indicate that the crowd offers superior particle-detection quality compared to the algorithm. This can be observed in the difference between the absolute true

positive measures (TP_C and TP_A) and by the indices relative to the total population count (ACC in Table II). The crowd exhibits better performance in terms of precision ($PREC$) in all cases. This indicates that the crowd detects a larger fraction of true particles than the algorithm. With respect to sensitivity ($SENS$), the algorithm and the crowd produced comparable performance in the case of *-dense-A, which cannot be observed in *-loose-*. This difference can be attributed to the fact that the experts suggest that the quality of X-ray imaging is better for dense flow. A comparable level of $SENS_C$ and $SENS_A$ ($SENS_C$ and $SENS_A \geq 0.75$) indicated that both the crowd and the algorithm identified the majority of the particles observed by the expert.

The results for the sandpaper dataset indicate that the automatic algorithm is inferior to the crowdsourcing system in terms of scalability. We believe that the automatic system would require additional effort and modification to achieve a performance level similar to that achieved for the loose and dense datasets. This points to the conclusion that a crowdsourcing system may be preferred when there is high variability of the measurement images and/or the images come from many sources.

While we believe that improvements can be made to the algorithm, we have shown that it represents the current state of the art in the application domain. The algorithm is refined, having taken an estimated 100 hours of work from a computer science professional (PhD). Furthermore, the very nature of the experimental setup in industrial tomography implies that the algorithm would need to be modified each time a new experiment is conducted. Different experimental configurations produce differing values of contrast; thus, a possible automatic solution would require significant refinement. From an economical standpoint, the cost of development is thus significantly higher than the cost of processing the dataset with the crowdsourcing system. For example, for the 3-loose-B condition, the total cost can be estimated at $3\frac{tasks}{frame} \times 137\,frames \times \frac{\$0.033}{task} = \$13.7$[4]. Furthermore, while the crowdsourcing system is scalable and designed to accept particle detection tasks in silos of many shapes and configurations, an automatic algorithm will need to be redesigned for each experiment. This process would also most likely include manual data preprocessing. Crowdsourcing work can begin immediately after the dataset is obtained, while the algorithm needs to be designed in advance and the employment of an image-processing expert must be scheduled. Taking these factors into account, there is a strong economic argument for applying crowdsourcing in this case.

6.3. Future Work

Looking at the results of our work, we find it justified to ask whether we can train people sufficiently so that the difference between expert work and nonexpert work will be negligible. If this would be possible, nonexpert workers could complete parts, or even all, of the workload carried out by bulk-solids experts in monitoring flow.

Worker motivation is another topic affecting our system's workflow. For how long are workers willing to take part, and for what kind of compensation? Can they maintain long-term motivation? How can we design a system like the one proposed here in such a way that workers stay motivated? In earlier studies, motivational effects include gamification, competitive standing, nonmonetary points and rewards, reputation, community, and monetary compensation. In our studies, subjects received a small gift as compensation for their work. However, we also believe that alternative forms of motivation could work, having observed that many participants were keenly interested to know more about the actual problem being solved.

[4]This assumes an average Amazon Mechanical Turk wage of USD~~D~~$2.00 per hour, according to Ross et al. [2010], and a task length of 60s.

We see two directions for conducting further system development. First, we see crowdsourcing as a potential way of producing extensive training sets for machine-learning systems. This can eventually lead to designing more effective algorithmic automatic systems.

Second, we see that there is a potential to use a crowd to verify and cross-check false-positive and true-negative particles initially identified by an automated solution. However, an extensive study is needed in order to determine whether the task of first verifying the particles detected by the algorithm and then looking for the ones missed by the algorithm would still be an effective and feasible way to tackle the problem. It may happen that such a complicated task would be easier for the crowd workers, but it may cause more cognitive load.

By applying our workflow, researchers will be able to effectively gather learning sets for enhanced AI-based solutions. Using this approach, artificial neural networks or fuzzy systems can be developed for automatic processing of tomographic images, especially for recognition of their important patterns. Moreover, data-mining methods applied to the gathered datasets, generated both by experts and nonexperts, can help in a thorough understanding of human perception of these kinds of images, leading to a possible enhancement of algorithmic procedures such as better image intensity conditioning or coping with high noise (low SNR).

Systems such as the one that we have proposed may be implemented in the future to learn how human users can pick out relevant details in images for which previous machine learning has failed. This would be in line with the development envisioned by Bigham et al. [2014]. The problem studied here might even be analyzed without any need for experts. If that should be the case, how many nonexperts compared to experts would be needed to average out errors caused by the nonexperts? If the number of nonexperts were too high, would it then be beneficial for society as a whole to use crowdsourcing? These particular questions require human insight.

7. CONCLUSION

In this study, we contributed the identification of a novel application domain for human computation, more specifically, for crowdsourcing—a system designed to carry out a study, providing insights on a problem from the research domain, in our case, industrial tomography. We developed a workflow to support the study allowing for the utilization of human computation in scientific-image processing, generalized enough to be scalable to other domains (Figure 1). We carried out a user study using this crowdsourcing system and found that task completion time and task completion accuracy were positively validated by expert review.

Furthermore, by developing and comparing against an automatic algorithm, we have validated the results and the effectiveness of a crowdsourced solution against other possible methods. We found that the crowdsourced approach (a) delivered results close to the near-ground truth (as provided by experts); (b) performed better than the algorithm quantitatively and in terms of not requiring adjustment to new images; and (c) is more economical in terms of both development cost and time saved for expert researchers to focus on the next stages of their work. Compared to the algorithm, the crowd offered superior particle-detection quality.

These results positively support using a workflow such as the one that we have proposed for using the crowd to analyze scientific images. The system is scalable and can handle detection tasks for silos of many shapes and configurations, whereas an algorithm would have to be redesigned in each case. Our system may also be generalized and scale to other domains for which such a workflow may apply.

We believe that our crowdsourcing model can be applied to image processing and pattern recognition in cases in which standard algorithms are not sufficient. Research

into the detection of tracer particles in tomographic images to date has been largely dominated by algorithmic and experimental work seeking to find computational solutions. But a classical approach does not necessarily lead to the best results or the best outcomes. Clearly, the process introduced in this article is simply one of many possible improvements, and we expect that our process will be further improved. Further exploratory work is needed to understand how to organize the work for expert and nonexpert workers, as well as how this work meets the modality and pace of the task presented. Recent work clearly indicates that crowdsourcing will change the very nature of work [Bernstein et al. 2015] and adapting it to industrial needs emerges as a relevant challenge for computing. We hope that our work constitutes a step in that direction.

ACKNOWLEDGMENTS

Thank you to Barrie Sutcliffe for his editorial work. Special thanks to Eric Marie and Jérome Adrien from INSA Lyon for their help during the measurement campaign. We thank Adviye Ayça Ünlüer Çimen for illustrations and artwork created by her and used in Figures 1, 2b, 3, 4, 5, 7a, 8, 9, 10, and 11. We are grateful to Mattias Mellquist, Moss, Norway who kindly allowed us to use the photo shown in Figure 2a.

REFERENCES

Saeed Albaraki and S. Joseph Antony. 2014. How does internal angle of hoppers affect granular flow? Experimental studies using digital particle image velocimetry. *Powder Technology* 268, 0, 253–260. DOI:http://dx.doi.org/10.1016/j.powtec.2014.08.027

Alexander Artikis, Matthias Weidlich, Francois Schnitzler, Ioannis Boutsis, Thomas Liebig, Nico Piatkowski, Christian Bockermann, Katharina Morik, Vana Kalogeraki, Jakub Marecek, and others. 2014. Heterogeneous stream processing and crowdsourcing for urban traffic management. In *EDBT*. 712–723.

Tim J. Atherton and Darren J. Kerbyson. 1999. Size invariant circle detection. *Image and Vision Computing* 17, 11, 795–803. DOI:http://dx.doi.org/10.1016/S0262-8856(98)00160-7

Victor Ayala-Ramirez, Carlos H. Garcia-Capulin, Arturo Perez-Garcia, and Raul E. Sanchez-Yanez. 2006. Circle detection on images using genetic algorithms. *Pattern Recognition Letters* 27, 6, 652–657. DOI:http://dx.doi.org/10.1016/j.patrec.2005.10.003 cited By 0.

Laurent Babout, Krzysztof Grudzien, Eric Maire, and Philip J. Withers. 2013. Influence of wall roughness and packing density on stagnant zone formation during funnel flow discharge from a silo: An X-ray imaging study. *Chemical Engineering Science* 97, 0, 210–224. DOI:http://dx.doi.org/10.1016/j.ces.2013.04.026

Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: A word processor with a crowd inside. *Communications of the ACM* 58, 8, 85–94. DOI:http://dx.doi.org/10.1145/2791285

Jeffrey P. Bigham, Michael S. Bernstein, and Eytan Adar. 2014. Human–computer interaction and collective intelligence. In *The Collective Intelligence Handbook*, Thomas Malone and Michael Bernstein (Eds.). MIT Press, Cambridge, MA.

David N. Bonter and Caren B. Cooper. 2012. Data validation in citizen science: A case study from project feederwatch. *Frontiers in Ecology and the Environment* 10, 6, 305–307.

Teh-Chuan Chen and Kuo-Liang Chung. 2001. An efficient randomized algorithm for detecting circles. *Computer Vision and Image Understanding* 83, 2, 172–191. DOI:http://dx.doi.org/10.1006/cviu.2001.0923

Y. C. Chung, S. S. Hsiau, H. H. Liao, and J. Y. Ooi. 2010. An improved PTV technique to evaluate the velocity field of non-spherical particles. *Powder Technology* 202, 13, 151–161. DOI:http://dx.doi.org/10.1016/j.powtec.2010.04.032

Seth Cooper, Firas Khatib, Ilya Makedon, Hao Lu, Janos Barbero, David Baker, James Fogarty, Zoran Popović, and others. 2011. Analysis of social gameplay macros in the Foldit cookbook. In *Proceedings of the 6th International Conference on Foundations of Digital Games*. ACM, 9–14.

Christopher B. Eiben, Justin B. Siegel, Jacob B. Bale, Seth Cooper, Firas Khatib, Betty W. Shen, Foldit Players, Barry L. Stoddard, Zoran Popovic, and David Baker. 2012. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature Biotechnology* 30, 2, 190–192.

Carsten Eickhoff. 2014. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the 1st International Workshop on Gamification for Information Retrieval*. ACM, 53–56.

Krzysztof Grudzien, Zbigniew Chaniecki, Andrzej Romanowski, Maciej Niedostatkiewicz, and Dominik Sankowski. 2012. ECT image analysis methods for shear zone measurements during silo discharging process. *Chinese Journal of Chemical Engineering* 20, 2, 337–345.

Krzysztof Grudzień and Manuel Hernandez De La Torre Gonzalez. 2013. Detection of tracer particles in tomography images for analysis of gravitational flow in silo. *Image Processing and Communications* 18, 2–3, 11–22.

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology* 52, 139–183.

Paul V. C. Hough. 1962. Method and means for recognizing complex patterns. Retrieved March 10, 2016 from http://www.google.com/patents/US3069654 US Patent 3,069,654.

Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST'12)*. ACM, New York, NY, 23–34. DOI:http://dx.doi.org/10.1145/2380116.2380122

Ping Lu, Alexander Daehne, Julien Travelletti, Nicola Casagli, Alessandro Corsini, and Jean-Philippe Malet. 2014. Innovative techniques for the detection and characterization of the kinematics of slow-moving landslides. In *Mountain Risks: From Prediction to Management and Governance*. Springer, 31–56.

Erik Meijering, Oleh Dzyubachyk, Ihor Smal, and others. 2012. Methods for cell and particle tracking. *Methods in Enzymology* 504, 9, 183–200.

Ulrich Meissen and Frank Fuchs-Kittowski. 2014. Crowdsourcing in early warning systems. In *7th International Congress on Environmental Modelling and Software*. iEMSs.

Paul A. Merolla, John V. Arthur, Rodrigo Alvarez-Icaza, Andrew S. Cassidy, Jun Sawada, Filipp Akopyan, Bryan L. Jackson, and others. 2014. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345, 6197.

Priyanka Mukhopadhyay and Bidyut B. Chaudhuri. 2015. A survey of Hough transform. *Pattern Recognition* 48, 3, 993–1010. DOI:http://dx.doi.org/10.1016/j.patcog.2014.08.027

Sunghyun Park, Philippa Shoemark, and Louis-Philippe Morency. 2014. Toward crowdsourcing micro-level behavior annotations: The challenges of interface, training, and generalization. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*. ACM, 37–46.

Alexander J. Quinn and Benjamin B. Bederson. 2011. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1403–1412.

Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems (CHI EA'10)*. ACM, New York, NY, 2863–2872. DOI:http://dx.doi.org/10.1145/1753846.1753873

Dietmar Schulze. 2008. *Powders and Bulk Solids: Behavior, Characterization, Storage and Flow*. Springer-Verlag, Berlin, GmbH & Co. K. 512 pages.

Jonathan Seville, Uur Tzn, and Roland Clift. 1997. Storage and discharge of particulate bulk solids. In *Processing of Particulate Solids*. Powder Technology Series, Vol. 9. Springer, The Netherlands, 298–367. DOI:http://dx.doi.org/10.1007/978-94-009-1459-9_8

Ilja G. Sligte, Annelinde R. E. Vandenbroucke, H. Steven Scholte, and Victor Lamme. 2010. Detailed sensory memory, sloppy working memory. *Frontiers in Psychology* 1, 175. DOI:http://dx.doi.org/10.3389/fpsyg.2010.00175

Lei Xu and Erkki Oja. 1993. Randomized Hough transform (RHT): Basic mechanisms, algorithms, and computational complexities. *CVGIP: Image Understanding* 57, 2, 131–154. DOI:http://dx.doi.org/10.1006/ciun.1993.1009

Lei Xu, Erkki Oja, and Pekka Kultanen. 1990. A new curve detection method: Randomized Hough transform (RHT). *Pattern Recognition Letters* 11, 5, 331–338. DOI:http://dx.doi.org/10.1016/0167-8655(90)90042-Z

H. K. Yuen, John Princen, John Illingworth, and Josef Kittler. 1990. Comparative study of Hough transform methods for circle finding. *Image Vision Computing* 8, 1, 71–77. DOI:http://dx.doi.org/10.1016/0262-8856(90)90059-E

Received February 2015; revised August 2015; accepted November 2015