Horizon 2020 Framework Programme

Grant Agreement: 732328 – FashionBrain

Document Information

Deliverable number: D3.1

Deliverable title: A set of crowdsourcing interfaces

Deliverable description:

Due date of deliverable: 31/12/17

Actual date of deliverable: Revised version submitted 20/04/18

Authors: Alessandro Checco

Project partners: University of Sheffield

Workpackage: WP3

Workpackage leader: Paul Clough

Dissemination Level: Public

Change Log

Version	Date	Status	Author (Partner)	Description/Approval Level
1	31/12/17	Final	University of Sheffield	Public

Table of Contents

1 Introduction

This document will provide a report on a set of Human Intelligence Task (HIT) designs, experimentally validated for object recognition in images, validation of named entity extraction, and image labeling, with particular focus on entity linking for images in fashion. This is the result of FashionBrain T3.1 “Tailored Crowdsourcing Tasks”, based on the investigation of the effect of task features (e.g., aesthetic, complexity, etc.) on the quality of the results collected from the crowd, and will be a building block for WP5 “Fashion Analysis in Social Media Streams”.

A crowdsourcing task is generally an automated task generated by a computer and published to a crowdsourcing platform via its Application Programming Interface (API). The crowd workers will perform the tasks at hand through a web interface and submit their result. An accurate design of crowdsourcing tasks allows for better quality and less expensive data collection. In this task we will investigate the effect of task features on the quality of the results collected from the crowd. We will focus in particular on creating user interfaces for the following tasks within the FashionBrain project: object recognition in images, validation of named entity extraction, and image labeling. We will show how the quality of a HIT is affected by the following aspects: task instructions, training, interface clarity, and overall task design. This is particularly important in situations where tasks are paid as workers will typically aim to minimise the time they spend on understanding what they are required to do.

The rest of the document is structured as follows: in Section 2, we describe in detail the crowdsourcing process. In Section 3, we present an extensive literature research, that will guide our task design approach. In Section 4, we present our analysis of work environments and the tools we developed to assist task design. In Section 5, we present our proposed solution for entity linking and images in fashion.

2 Crowdsourcing Process

Despite the fact that different types of tasks exist on crowdsourcing platforms, the process of implementing each task consists of basically the same stages. The mechanism of crowdsourcing works according to the following steps as shown in Figure 1: (1) Define the problem, (2) Collect data, (3) Design the task, (4) Launch the task online via a crowdsourcing platform, (5) Analyse the result, and, if the job accepted, (6) Send rewards to the workers.

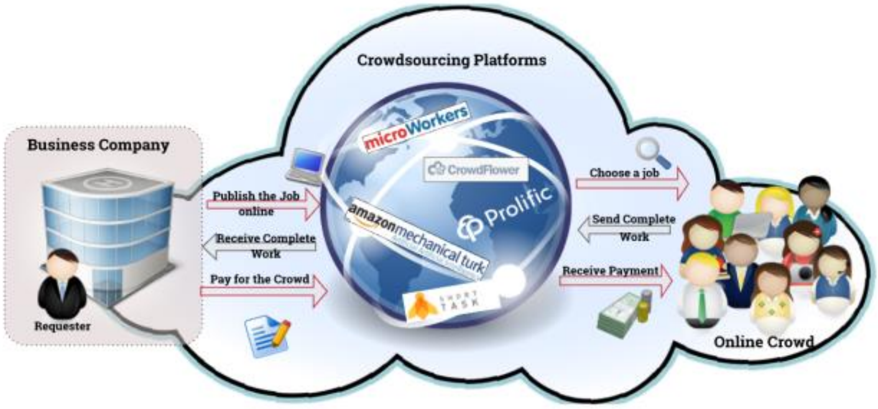

Fig. 1. Mechanism of crowdsourcing interaction with business companies.

The process of crowdsourcing could be analysed from three different perspectives: he worker, the requester and the task.

The worker registers on a platform and performs some unpaid tasks in order to become qualified in certain skills that might be required. On the platform, the workers will find a list of jobs available along with the specified reward for completing it accurately. The online crowd is invited to an open call for everyone who is interested in providing solutions or performing the tasks on behalf of the company, which will name a price for each task. In a particular situation, the crowd could be limited by the imposition of some constraints, such as needing certain experience in a given area [Brabham 2008].

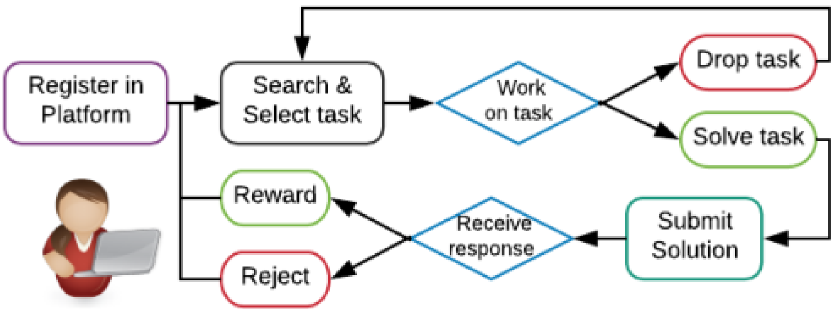

Fig. 2. Crowdsourcing process (UML activity diagram) from the worker perspective.

A number of recommendation systems appear to favour some workers for a specific task based on criteria, such as workers' history and their overall performance [Geiger and Schader 2014; Schnitzer et al. 2015; Yuen et al. 2015]. The worker will choose one of the listed tasks and attempt to complete it. They could decide at any point to leave the task or submit an answer if they succeed in completing it. The last stage of the worker process is receiving a response for the submitted job, either rejection or the pre-agreed reward (Figure 2).

The requester represents the company who will identify tasks or problems that need to be solved. The requester will gather the data and define the requirements, constraints, and output of the job and, for a long or complex task, the requester has to divide these task into smaller tasks (microtasks) which are released to the crowd online via one of the crowdsourcing platforms. For each task, the requester will determine a specific amount of time for the user to complete this task and submit it to the requester. When this time is up the requesters will analyse the quality of the received work and decide if the problem has been solved by the completed work or whether it should be rejected; based on this decision the worker will receive a response. This mechanism could vary from one requester to another (Figure 3).

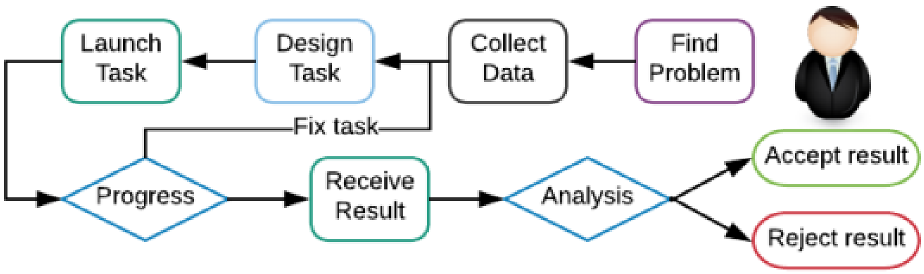

Fig. 3. Crowdsourcing process (UML activity diagram) from the requester perspective.

The task goes through three stages as shown in Figure 4. The first is the design process "off-line", where the task will be outlined using one of the predefined templates provided by the platform or designed from scratch [Luz 2015]. In this stage, the data or the input will be fed in, and the parameters of the task will be set. The measures for quality control will be implemented at this stage to guarantee efficient result and detect spam or malicious workers. Moreover, the long/complex task will be decomposed in simple/micro tasks. For example, identify the face of a specific person from a picture of a crowd in a football stadium will be a long task to perform by one worker. Such a task is able to be divided into micro-tasks by cropping the picture to small pieces and crowdsource each piece as a simple independent task to workers.

To correctly define task complexity we need to consider, as explained in [Finnerty 2013] and [Sweller 1988] task structure, task interdependence, task commitment, and cognitive load.

The second stage ("on-line" in Figure 4) is execution, where the microtasks appear online and become available to workers. The implementation of this process could be in *parallel* when a microtask does not depend on the result of another one. Alternatively, the microtasks could be implemented *sequentially* one after another, where the result of one task becomes the input for the next task. In this on-line stage, the task could be paused if it requires any modifications and then continues running again. In some situations, the task could be done by a number of different users, and each could be paid if they complete the task successfully. In other cases, such as logo design, the task could be completed in different ways depending on workers' understanding and creativity, and the payment would be given for the 'best' solution, as decided by the requester [Whitla 2009].

The last stage is when the requester receives the completed job and it reaches either a "finished" or "cancelled" state. In the case of reaching a finished state, the microtasks will go through the aggregation process where these small tasks will be merged together to form the final result of the job. A post-execution quality control methods will be used to identify the cheater workers based on their performance on the submitted job and if they meet the quality criteria that have been set up in the pre-execution quality control method.

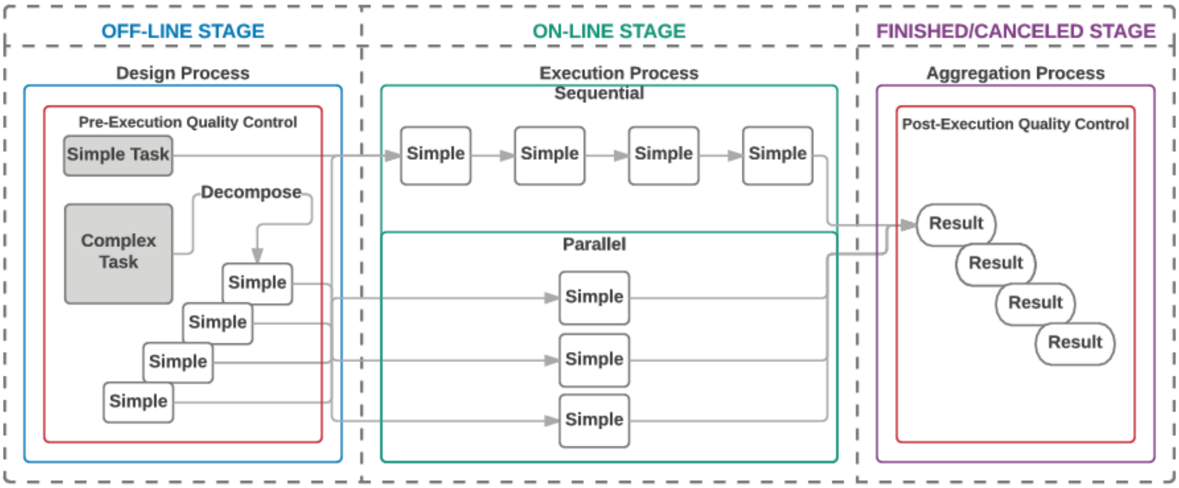

Fig. 4. Crowdsourcing task process.

3 Related Work

The aim of this section is to present a comprehensive survey on crowdsourcing task design, focusing on the technical factors that have a significant impact on the quality of the design. This is a novel endeavour, because even the most recent surveys on crowdsourcing, e.g. Chittilappilly et al. [2016] focus on different topics, such as types of incentives, task recommendation and quality control systems.

A number of surveys have been conducted in the field of crowdsourcing [Chittilappilly et al. 2016; Mao et al. 2015a; Pan and Blevis 2011; Xintong et al. 2014; Yuen et al. 2011]. A short survey by Pan and Blevis [2011] presented a literature review of crowdsourcing and interaction design among academic, business, and social domains. This study was the first step for providing some insights and recommendations for designing crowdsourcing tasks and highlighted some challenges in task creation within Human-Computer Interaction (HCI). Yuen et al. [2011] showed different classification of crowdsourcing systems based on their applications, algorithms, performances and datasets.

Xintong et al. [2014] presented the state of the art of using crowdsourcing in data mining. Mao et al. [2015a] conducted a survey on the use of crowdsourcing to support software engineering field.

Designing the task appropriately can lead to high-efficiency outcomes and a reduction in disagreements in the result [Garcia-Molina et al. 2016]. Catallo and Martinenghi [2017] define a taxonomy of designing crowdsourcing tasks based on four design dimensions inspired from the explicit control aspects of human computation mentioned in Law and Ahn [2011]. These dimensions are defined as *What* kind of task need to be solved, *Who* is going to solve it, *Why* the workers need to work on it, and *How* to process these tasks. This classification along with low levels components presented the main factors that involved in the process of designing crowdsourcing tasks.

Several studies considered that task design has a significant effect on the task outcomes. McDonnell et al. [2016] showed that designing the task in a way that reduces the cognitive load on workers significantly increases performances. Related to this, Yang et al. [2016] showed how task design properties are highly correlated with perceived task complexity.

Allahbakhsh et al. [2013] considered task design as one of the main dimensions that control the quality of the crowdsourcing system. Their proposed quality control approaches are the *design-time* approach, where the requester could use various techniques to control the quality of the task in the design stage; and the *run-time* approach, where requesters include some monitoring during the task running to prevent any mistakes or low-quality performance. These two approaches can help to control the quality of the result and can be applied separately or simultaneously on one task.

Moreover, the main aspect that should be considered is that the task will be done by a human not a machine, which is why the psychological aspects in designing the task should be analysed [Alonso 2013]. Deng et al. [2016] enumerates guidelines for workers, requesters, and platforms developers to enhance the services in crowdsourcing field. The study conducted survey instrument based on the worker's experiences and how they interact with the crowdsourcing system. The aim of this study is to enhance the workers position by providing governance mechanisms to ensure transparency and fairness in the work environment.

There are several factors that affect the task design: the length of the task, the nature of the required work (for example writing, classifying, or designing), the use of training questions and examples, and also the graphical user interface which often varies according to the complexity of the task. This section will present some of these studies and highlight a number of essential factors that have been discussed in the last few years.

Psychological Factors

Human factors, such as psychological traits, are one of the main aspects that influence performance. Many researchers have studied the influence of personality traits which can have a positive or negative effect on the accuracy of different task design [Harrison et al. 2013; Kazai et al. 2011]. For example, distinguishing different visual designs for a task could trigger different emotion in the workers leading to variation in the results.

Kazai et al. [2011] analysed workers' behaviours that lead to the classification of workers into five different types: *Diligent, Competent, Sloppy, Incompetent*, and *Spammers*. The authors tried to connect workers' characteristics and their personality traits with the accuracy and the average time for completing the tasks. The findings of this study showed that workers' behaviours have a significant effect on accuracy for labeling tasks. On the other hand, the average time the workers spent in solving the task did not have a positive effect on the accuracy level for the task results for *Sloppy, Incompetent*, and *Spammer* workers. For connecting workers behaviours with the personality traits, *Conscientiousness* workers achieved a high level of accuracy.

Morris et al. [2012] looked at priming effects in micro-task crowdsourcing environments. They showed that priming workers could increase performances in creative tasks. While they show that priming has positive effects, they also note that it should be unconsciously provided to workers and that it does not substitute training done by means of instructions and examples. Consistently, in Harrison et al. [2013], they used emotion priming in visual judgments tasks. They pointed out that while positive emotion has significant effects on the performance, negative emotions could also priming workers positively in some situations. Moreover, there are environmental priming that could affect workers differently and these beyond of control requesters.

André et al. [2014] looked at how groups of workers performed and showed that asking workers to contribute sequentially worked better than simultaneous collaboration in complex creative tasks. This finding proved the importance of crowdsourcing microtasks rather than sending it to a group of people to solve it together. The implications of this study point out that workers feel more secure when working independently. Several factors need to be investigated, such as the identity of the workers, the time of releasing the tasks online, and the nature of the tasks purposely motivated virtual teamwork among crowdsourcing systems.

Another study investigated intertask effects for image labelling; i.e. how workers are influenced by the type of task they have previously completed when working on a new task [Newell and Ruths 2016]. Moreover, the impact of any previous experience that the workers had and the rejection of a completed job have a significant effect on the workers' expectation for the upcoming work. McInnis et al. [2016] studied the impact of unfair job rejection on the workers and the subsequent risk. As a result of unfair rejections, workers tend to act safe and minimize the risk by accepting the same type of tasks or selecting a task from a limited number of requesters who have good repetition or previous successful business with them. This safe action could keep the workers safe from rejection, but it will also prevent workers from expanding their experience on the new type of tasks. Furthermore, the new requesters could face the risk of lack of turnout or have malicious workers only for their tasks.

Type of Tasks

An optimal task design for a crowdsourcing task might not translate well in a task of different nature. An analysis of the effect of different variables (e.g., interface, length, the number of items) on task completion performance have been shown not to generalise when the nature of the task is variable, e.g Marcus et al. [2012] has shown this effect for labeling (assign a label to an image) and counting (count the number of objects in an image) tasks.

Using closed questions, such as multiple choice answers, or using open question, such as providing a text box area for answers, has an influence on the workers performance. Some studies found that using predefined answers could save time and gain accurate answers [Jain et al. 2017]; whereas other studies found that these type of tasks could increase the number of malicious workers who complete all answers in the task quickly, in order to just gain rewards [Eickhoff and de Vries 2013; Gadiraju et al. 2015].

In addition, Moussawi and Koufaris [2013] highlighted that giving the workers some level of freedom in the way they perform and respond to the task leads to high motivation of the workers. Similarly, Eickhoff and de Vries [2013] state that the use of open questions could result in more creative answers and less cheating. In Eickhoff and de Vries' [2013] study, it appears that using questions that require a text answer to get feedback is very helpful. The author conducted a number of repetitive tasks that handle large datasets incorporating some factors that could enhance the overall result. The first factor was forming the question correctly. He recommends that the question should asked in a simple and straightforward way so that it could be consistently understood by all workers. For questions with labelling answers, he suggested replacing them with a numerical scale to prevent misunderstanding. Moreover, it is preferable to use a broad range of labels (with no more than 6-7 categories) to give the workers some level of flexibility in giving the right answer.

Task Graphical User Interface (GUI)

Since the task graphical interface is the main way for the workers to understand the job, it is fundamental to design adequate graphical user interfaces, that can help the workers understand the task requirements, the processes they need to follow, and the results expected from them. This process will have a strong impact on the worker's performance. Allahbakhsh et al. [2013] define one of the factors that affect the quality of the outcomes as the user interface, which is the graphical design of the task: they found that implementing a simple interface could help the workers complete the task in a short time and increase the accuracy of the completed job. Furthermore, the study by Jain et al. [2017] showed that writing long instructions providing a detailed description of the task, and using examples, will have a positive effect on the quality of the result, particularly for complex tasks.

A study by Kim et al. [2015] used a crowdsourced task to match the appearance of the colour of some products on a website with the real colour of the same products. The lighting and the image quality that had been used in the task had a strong impact on the accuracy of the result. Other studies, such as Finnerty and Kucherbaev [2013] compared the outcomes of two tasks with simple and complex interfaces and the results proved that using a simple clear interface records higher result than using the same task contents, but with a complex interface (as defined in [Galitz 2007]).

Furthermore, a study by Alonso [2013] presented an interface design by following the guidelines of Nielsen [1993] to point out the basics of task design: write clear instructions, show examples, highlight and colour what is important and required for the job, which can reduce the effort to complete the task. Also, using a relevant, clear, and attractive title for the job will make it easy for workers to find it quickly when they are searching the platform for possible tasks to accept and complete.

McInnis et al. [2016] investigated a number of factors that lead to unfair rejection, such as insufficient task design, misleading instructions, technical errors, and requesters with poor knowledge. They concluded their study with a number of suggestions that could reduce the risk and enhance the connections between workers and requesters to achieve a better final outcome. One of these suggestions was to provide in the design of the task *an alarm* for a broken task, which notifies the requester of any error in of the task design during the work process.

Recently, Wu and Quinn [2017] outlined best practice guidelines for writing task instructions that could optimise the quality of the outcome. This study found that regardless of the facts that long and clear instructions will improve the result, workers tend to favour tasks with short guidelines and few lines of instructions. Therefore, the requesters should make a balance between presenting full instructions and defining attractive short steps which will be easy to read and deliver the full format of the task specification at the same time.

Gadiraju et al. [2017] also investigate the effect of the task clarity on the worker's performance. Through surveying workers regarding their previous tasks and how often they found it clear, feedback from workers pointed out that most of the features of an unclear task were because of the writing style and lack of detail in the presentation of the instructions. Also, they refer to the rare use of examples which make it less clear for them to understand the requirement of the job. The finding of this study showed that task clarity could be predicted and supervised via the proposed model and guide the requester in the task design. Further investigation could drown from this work to examine the relationship between task clarity and complexity and the effect of task clarity on workers dropout rates.

Training Questions

Training questions can be formed in a variety of ways which may be helpful for some tasks but not others. Several studies have looked at the training of workers before or whilst performing a particular task and a number of training techniques or methods were used. These can be summarised as follows:

1. *Control method:* does not have any training questions and the workers will read the instructions and start solving the task directly.
2. *Solution method:* adding a number of training tests before the real task questions without stating explicitly that first tasks are for training.
3. *Gold Standard:* the same setup as the solution method but after solving the first training tasks workers are shown the correct answers for the tasks and informed that they had been used for training purposes. Oleson et al. [2011] used this method in tasks as quality control mechanism rather than using it as a training method.

4. *Example method*: design task instructions to explain that workers will be shown some examples completed by an expert and that they are not allowed to start the task until the 30 second demonstration has been completed; this forces workers to read the examples and understand how they were solved. A recent study by Jain et al. [2017] and Wu and Quinn [2017] proved that using examples is crucial and plays a key factor in increasing the accuracy of results and the total agreement amongst workers. Similarly, Mitra et al. [2015], presented some examples for the workers followed by test questions to measure the improvement in their performance and to determine if they learned from the examples.
5. *Validation method:* in this method workers were shown two answers of other previous workers and asked to validate these answers by filling out some specific questions about them. Zhu et al. [2014] found that using the validation method in subjective tasks which required some creativity in devising the solution, was more effective than making the workers do more training tests.

Another study by Doroudi et al. [2016] presented five different ways of using training questions to improve the overall result of what they defined as a complex task. They used all five methods to find the most beneficial training method. The findings of this study reported that showing the workers expert examples increased the overall accuracy of the answers compared with using other methods. Moreover, using the validation method was the most effective way of training workers.

Length of the Microtask

The length of the crowdsourced task can be designed to vary in length. To maintain a balance between the length of the task and the desired quality of the outcomes, several solutions have been proposed in different studies. One of these solutions is to decompose the long task into shorter ones (microtasks), which corresponds with the main goal of crowdsourcing platforms - to keep the tasks simple.

The main goal of using a crowdsourcing platform is to break down a task into smaller tasks - as we mentioned previously - which can be solved by the crowd, achieving high-quality performance as well as saving time and money [Cheng et al. 2015; Kittur et al. 2011]. These microtasks should have a low level of complexity to achieve their purpose. Doroudi et al. [2016], defined the level of complexity for tasks: as a task which cannot be decomposed into micro-tasks and workers can use different mechanisms to perform such tasks. For complex task, a high level of accuracy is not achievable with low expertise workers.

Previous related work in the area of microtask crowdsourcing has looked at the effect on crowd performance of task granularity. For example, Cheng et al. [2015] showed that having shorter tasks lead to increased overall completion time but also to better quality contributions. Similarly, Allahbakhsh et al. [2013] discussed the granularity long tasks, which affects the quality of the outcomes. The final result of such a task is a combination of the results of a number of smaller or shorter tasks.

Another solution is to break the long task up with some activities to keep the worker interested in completing the task. Dai et al. [2015] proposed including some entertainment micro-tasks as a short break in performing a long task. They used the MTurk platform to design three different long tasks: (1) Classifying images, (2) Rating Wikipedia articles, and (3) Merging freebase entities. For each type of task, they inserted three different "micro-diversions": no diversion, a narrative webcomic story, and a dice game to keep workers on track and motivate them to continue working on the task. The findings of this study proved that using micro-diversions can significantly maintain workers' motivation to continue working on a long task as well as enhancing the speed of the answers. There are some variations in the findings depending on the task type and the micro-diversions combination. A complex cognitive task, such as rating a Wikipedia article, was performed more effectively using a diversions task. Moreover, the story acts better than a game diversion in speeding up workers' performance.

Moreover, Brambilla et al. [2015] propose prototyping methods for task design that will be implemented first in small datasets in order to gain better result for designing the same task for large datasets. This approach reports significant results for image relevance judgment tasks; further work could use the same strategies in other types of tasks.

Ordering Effects

In the process of implementing the task the ordering of data in the microtasks could lead to a variation on the overall result. The requester has the option to organise the data in the batch and present it in ascending order of difficulty that gradually prime the workers and improve their performance.

Cai et al. [2016] looked at how the sequence of writing tasks impacts crowd worker efficiency. They observed that by varying the order of task complexity and task type, workers' performance would vary thus indicating the potential to optimise worker efficiency by appropriately sorting tasks in a batch. Lasecki et al. [2015] looked at the effect of interruptions and of changing tasks type (i.e., context switch) on sequences of crowdsourcing tasks, showing how worker speed would significantly decrease in such situations.

Damessie and Culpepper [2016] investigated the impact on the *inter-rater agreement* of presenting documents in two different ways: (1) descending order of relevance, and (2) document identifier order where the documents vary depending on the topic. They designed a judgment task for 30 documents across *easy* and *hard* topics extracted from TREC collections and with a four-level relevance scale. The results showed that ordering by document identifier leads to a higher agreement in both easy and hard topics and a better result in term of identifying the relevant documents.

Recent study by Yang et al. [2016] proposed a high-dimensional regression model to measures the impact of task structural features on the complexity of the task and conversely using these features to predict the complexity and the tasks outcomes, showing that the semantic description and the visual appearance of the task are the most useful features to predict the complexity of the task and improve the quality of the output.

4 The Role of Microtask Work Environments

An aspect that has remained largely invisible in microtask crowdsourcing is that of *work environments*; defined as the hardware and software affordances at the disposal of crowd workers which are used to complete microtasks on crowdsourcing platforms. In [Gadiraju et al 2017], we reveal the significant role of work environments in the shaping of crowd work. First, through a pilot study surveying the good and bad experiences workers had with UI elements in crowd work, we revealed the typical issues workers face. Based on these findings, we then deployed over 100 distinct microtasks on CrowdFlower, addressing workers in India and USA in two identical batches. These tasks emulate the good and bad UI element designs that characterize crowdsourcing microtasks. We recorded hardware specifics such as CPU speed and device type, apart from software specifics including the browsers used to complete tasks, operating systems on the device, and other properties that define the work environments of crowd workers.

Our findings indicate that crowd workers are embedded in a variety of work environments which influence the quality of work produced. To confirm and validate our data-driven findings we then carried out semi-structured interviews with a sample of Indian and American crowd workers from this platform. Depending on the design of UI elements in microtasks, we found that some work environments support crowd workers more than others. Based on our overall findings resulting from all the three studies, we introduce ModOp, a tool that helps to design crowdsourcing microtasks that are suitable for diverse crowd work environments. We empirically show that the use of ModOp results in reducing the cognitive load of workers, thereby improving their user experience without affecting the accuracy or task completion time. An example of ModOp is shown in Figure 5.

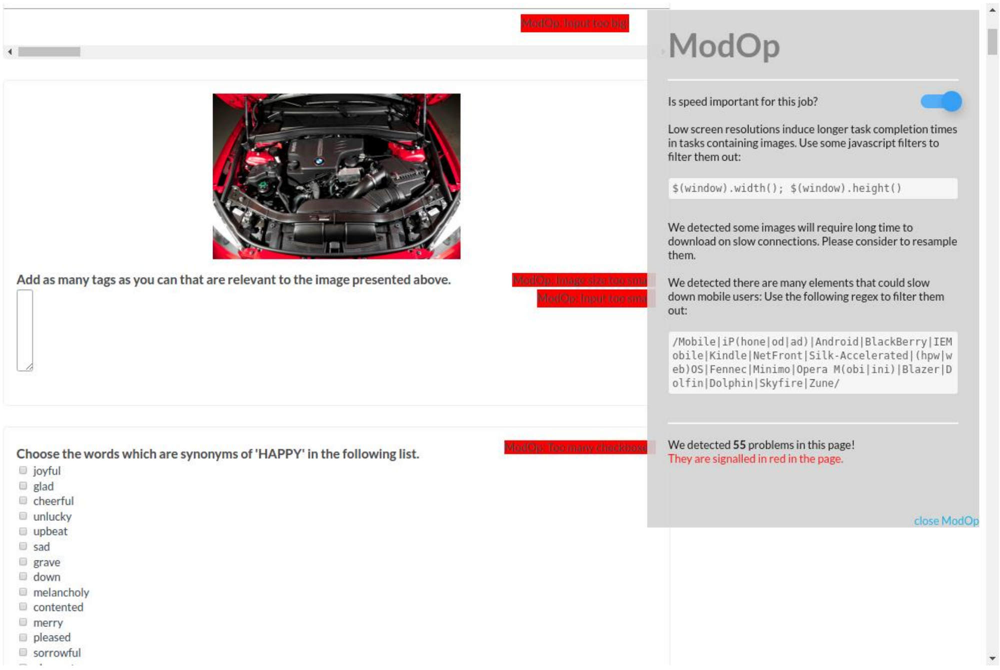

Fig. 5. Example of ModOp plugin in overlay over a task design.

5 Crowdsourcing Interfaces in Fashion

In WP3 we focus on entity linking: given a real world image we use crowdsourcing to link all the fashion products that are visible in the image to an existing catalog. In the optimal case the crowd should link the 'perfect match' (the exact same product) from a predefined set of products. In case the 'perfect match' is not present in the database or the crowd does not find the optimal solution the most similar product should be annotated.

Task Description

Product Linking Task:

The Product Linking Task has the following flow:

1. Each worker is presented with an independent task, no collaboration allowed (as suggested by André et al. [2014]).
2. Each independent task has one image that represents one or multiple products.
3. The worker has to signal (optionally) if this record has a data error (missing image etc) or if the images does not contain fashion products.
4. Each product can be considered as a microtask
    a. The worker is presented with an image patch containing a single product and a sublist (16) of suggested products out of the entire product db.
    b. The workers' task is to find the "perfect match" out of the entire product db.
    c. In case the sublist does not contain the "perfect match" the worker has the following options to modify the list: text search and "iterative image search"
    d. If no perfect match can be found, the worker should select the most similar product.
    e. Once a product is selected, the worker is signaling the quality of the match with a star rating (1 to 3).
5. After all microtask have been done, the worker clicks on the "save" button and the next task is presented to the worker.

The structure of the possible worker response (predefined options followed by text and image search) is carefully selected to minimise the worker load when a creative solution is not needed (Jain et al. [2017]) and providing a level of freedom when needed, increasing workers motivation (Moussawi and Koufaris [2013], Eickhoff and de Vries [2013]).

Task Instructions

Each worker receives a one hour training (via instant messaging) before the beginning of the job: this has been proven to have a positive effect on the quality of the result, in accordance to the study of Jain et al. [2017] for complex tasks. The training consisted of a multiple run of the task via examples (Alonso [2013]). Moreover, the worker is able to continuously communicate with the requester via instant messaging, receiving instantaneous feedback and clarification in case of doubts.

Task Graphical User Interface (GUI)

An overview of the Task GUI is shown in Figure 6:

- The task GUI consists of header (top part) and body (bottom part).
- The header presents the worker with an image and some additional information (e.g. who created the image in case the image comes from social media post, and some text description that accompanies the image).
- The body presents the worker with micro-tasks, i.e. with image patches for each product found in the header image.

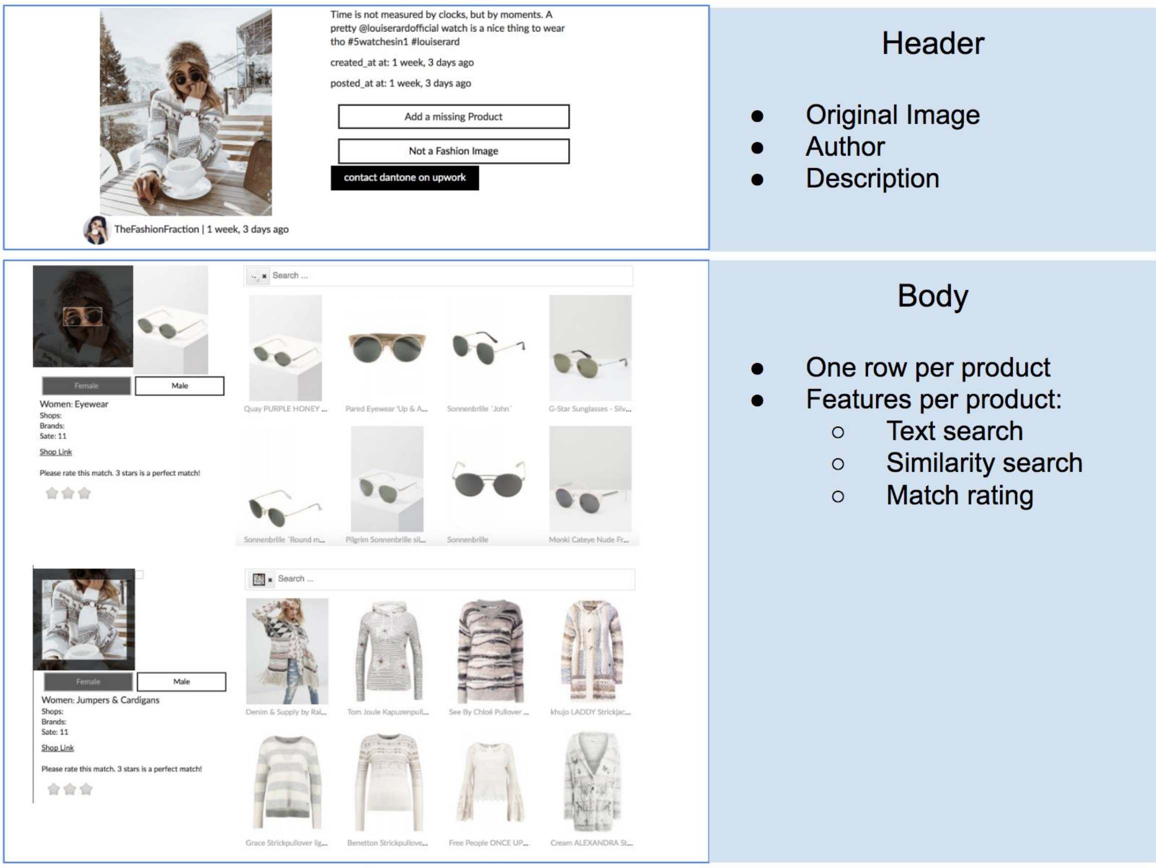

Fig. 6. Overview of Task Graphical User Interface.

An overview of the row structure is shown in Figure 7. Each row (micro-task) consists of:

- The image patch (**original image**) of a product
- **Info box** which presents the worker with information of gender and fashion category (e.g. jumpers, pants, sunglasses, etc.)
- An image of a **matched product** from the product database
- **Product suggestions** which allow the worker to select the "perfect match" product
- **Text search** field through which product suggestions list can be refined by fashion category, description, brand
- **Match rating** which the worker uses to rate how good the match is, ranging from 1 to 3 stars, higher is better (i.e. in case of the "perfect match" the worker would select 3 stars).

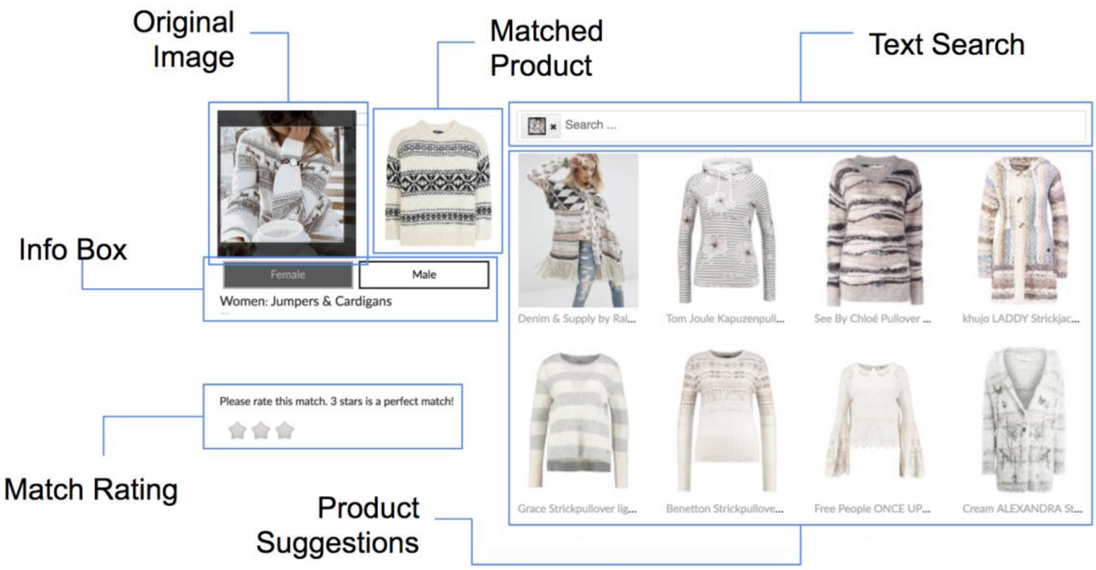

Fig. 7. Overview of the row structure.

An overview of the iterative visual search is shown in Figure 8. The iterative visual consists of the following steps:

- The worker is first presented with product suggestions for the **original image** (top part).
- Then the worker can select a similarly looking product from the product suggestions list.
- After that, product suggestion list is refreshed with products similar to the **matched product** (bottom part), thus allowing user to move in space of similar fashion product until he/she gets to the most similar one.

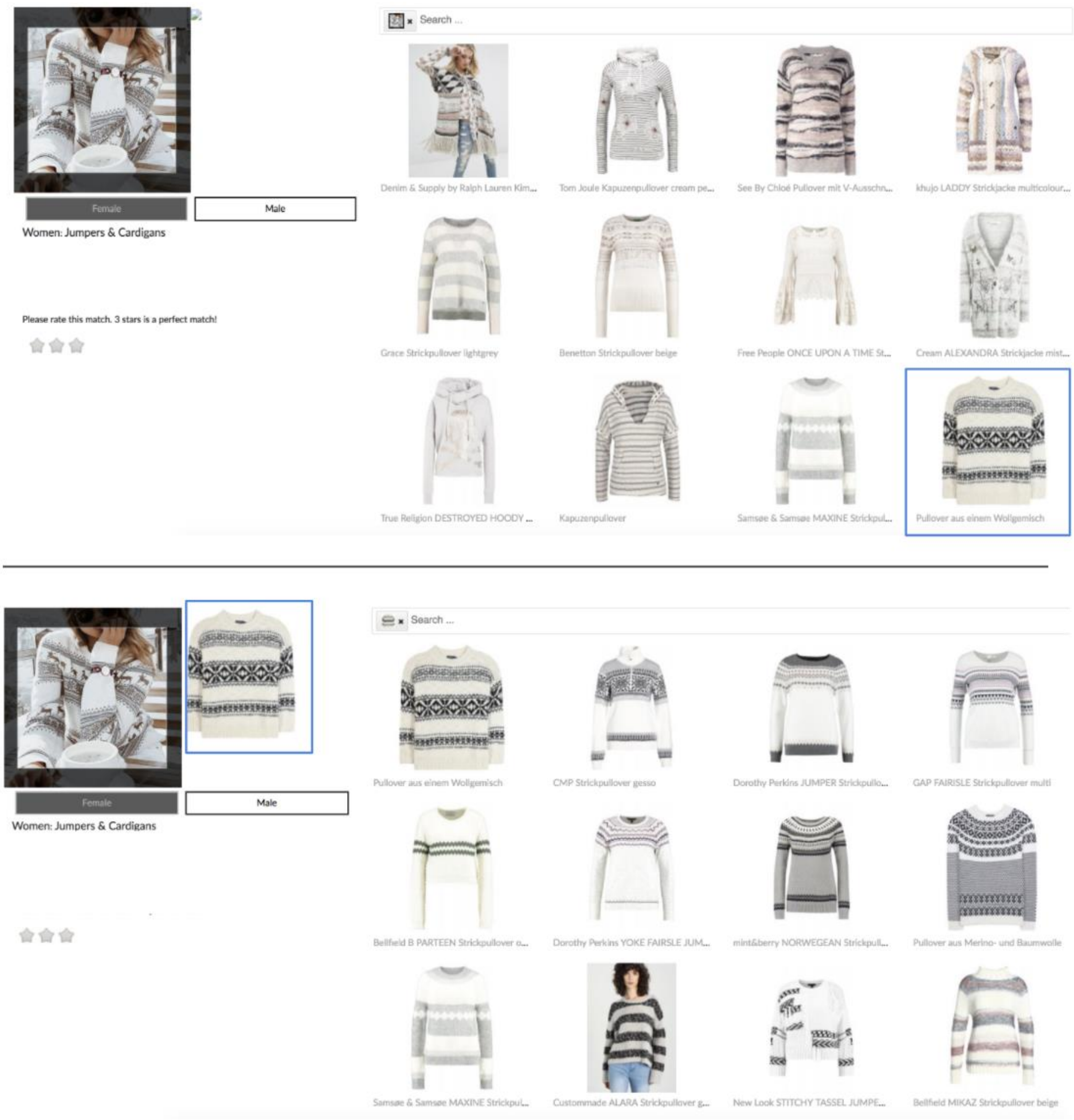

Fig. 8. Overview of the iterative visual search.

An overview of the final rating is shown in Figure 9: in the end of the task the worker rates the match, e.g. in the example shown in the figure the jumpers are quite similar but not a “perfect match”, so the worker should rate it with 2 stars.

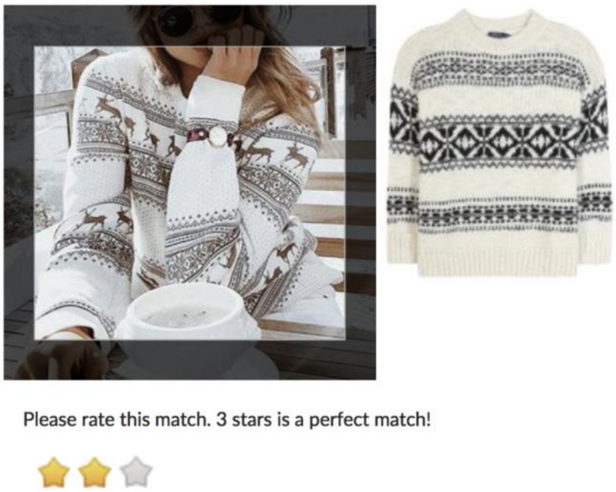

Figure 9. Overview of final rating system.

Strategies on GUI design

The following guidelines has been used to maximise the quality:

- We focused on designing simple and clear interfaces (Finnerty and Kucherbaev [2013], Nielsen [1993], Gadiraju et al. [2017]).
- We minimise the risk of unfair rejection and worker dissatisfaction by providing continuous feedback with clear explanation of potential mistakes (McInnis et al. [2016]).

Training

- The training of the workers is important for the success of the overall task. The workers have to be highly skilled for this task, because they need to have a deep fashion knowledge in order to point out the difference between two different products.
- The learning curve is steep. We noticed that trained and skilled workers have a much higher throughput (up to 3-4 times) with significant higher quality. Speed and quality increase once they understand the performance of the ‘iterative image search” and once the worker get a good feeling for the product catalog.
- The training of the worker is done using video conference and screen sharing.
- The trainer starts with 5 standard task, so that the worker gets familiar with the task.
- The next step is that the worker performance 2-3 task and the trainer helps in that process.
- Once the worker is familiar with the standard task, the trainer shows some borderline cases.
- After the first day the trainer reviews random task and gives direct feedback. The same process is performed several times throughout the month in order to ensure the quality of the worker.

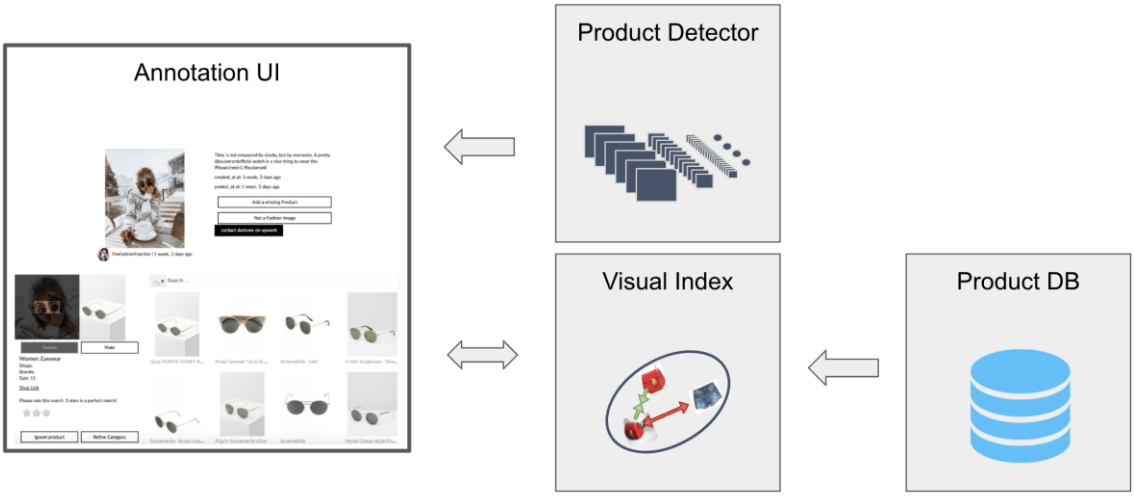

Fig. 8. Task Infrastructure. In order to provide the 'suggestions' and the 'visual similar' product feature several components needed to be implemented. The product detector runs over the image and localizes if a product is visible. The Visual Index performs the 'similarity search' for a given image patch, this means it finds the most similar product out of our product database.

5 Conclusion

We have reviewed the literature on task design within crowdsourcing with an emphasis on user interface design and worker training. We then present interface designs for the problem of entity linking and images in fashion. The resulting process is complex, and requires a different approach of the traditional crowdsourcing methods: in particular we employed a continuous two-way feedback system with the workers, and an extensive initial training of one hour per worker, which have been proven being very effective in obtaining the expertise required for such a complex task. Many of the results in the literature regarding task clarity, micro-task segmentation and interface design have been confirmed by our experiments.

Bibliography

Salman Ahmad, Alexis Battle, Zahan Malkani, and Sepander Kamvar. 2011. The Jabberwocky Programming Environment for Structured Social Computing. *Architecture* (2011), 53. https://doi.org/10.1145/2047196.2047203

Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing* 17, 2 (3 2013), 76--81. https://doi.org/10.1109/MIC.2013.20

Omar Alonso. 2013. Implementing crowdsourcing-based relevance experimentation: An industrial perspective. *Information Retrieval* 16, 2 (2013), 101--120. https://doi.org/10.1007/s10791-012-9204-1

Vamshi Ambati, Stephan Vogel, and Jg Carbonell. 2011. Towards Task Recommendation in Micro-Task Markets. *Human Computation (2011)*, 1--4. http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewFile/4005/4264

Bertalan K. Andrásfalvy, Mark A. Smith, Thilo Borchardt, Rolf Sprengel, and Jeffrey C. Magee. 2003. Impaired regulation of synaptic strength in hippocampal neurons from GluR1-deficient mice. *The Journal of physiology* 552, Pt 1 (10 2003), 35--45. https://doi.org/10.1113/jphysiol.2003.045575

Paul André, Robert E Kraut, and Aniket Kittur. 2014. Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 139--148. https://doi.org/10.1145/2556288.2557158

Sepehr Assadi, Justin Hsu, and Shahin Jabbari. 2015. Online Assignment of Heterogeneous Tasks in Crowdsourcing Markets. *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing (2015)*, 12--21. http://www.seas.upenn.edu/~sassadi/stuff/papers/oacm.pdf

Yoram Bachrach, Thore Graepel, Tom Minka, and John Guiver. 2012. How To Grade a Test Without Knowing the Answers --- A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. *Proceedings of the 29th International Conference on Machine Learning (ICML-12) (2012)*, 1183--1190. https://arxiv.org/pdf/1206.6386.pdf

Daniel W Barowy, Emery D Berger, and Andrew Mcgregor. 2012. AUTOMAN: A Platform for Integrating Human-Based and Digital Computation. *ACM SIGPLAN Notices* 47, 10 (2012), 639--654. https://doi.org/10.1145/2384616.2384663

Wei Bi, Liwei Wang, James T Kwok, Zhuowen Tu, Hong Kong, United States, and United States. 2014. Learning to Predict from Crowdsourced Data. *Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference*, UAI 2014 (2014), 82--91. https://ftp.cse.ust.hk/~jamesk/papers/uai14.pdf

Rubi Boim, Ohad Greenshpan, Tova Milo, Slava Novgorodov, Neoklis Polyzotis, and Wang Chiew Tan. 2012. Asking the right questions in crowd data sourcing. *In Proceedings - International Conference on Data Engineering*. 1261--1264. https://doi.org/10.1109/ICDE.2012.122

Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Andrea Mauri, and Riccardo Volonterio. 2014. Pattern-based specification of crowdsourcing applications. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8541 (2014), 218--235. https://doi.org/10.1007/978-3-319-08245-5_13

Daren C. Brabham. 2008. Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies* 14, 1 (2008), 75--90. https://doi.org/10.1177/1354856507084420

Daren C. Brabham. 2013. Crowdsourcing. *Mit Press*. 176 pages.

Jonathan Bragg, Mausam, Daniel S Weld, Jonathan Bragg Mausam, and Daniel S Weld. 2013. Crowdsourcing Multi-Label Classification for Taxonomy Creation. *Proceedings of the First Conference on Human Computation and Crowdsourcing* (2013), 25--33. https://doi.org/10.1016/j.jneumeth.2015.02.025

Marco Brambilla, Stefano Ceri, Andrea Mauri, and Riccardo Volonterio. 2015. An Explorative Approach for Crowdsourcing Tasks Design. *In Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*. 1125--1130. https://doi.org/10.1145/2740908.2743972

Alice M Brawley and Cynthia L S Pury. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54 (2016), 531--546. https://doi.org/10.1016/j.chb.2015.08.031

Carrie J Cai, Shamsi T Iqbal, and Jaime Teevan. 2016. Chain Reactions: The Impact of Order on Microtask Chains. *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3143--3154.

Nicholas G Carr. 2010. The Shallows: What the Internet Is Doing to Our Brains. W. W. Norton.

Ilio Catallo and Davide Martinenghi. 2017. The Dimensions of Crowdsourcing Task Design. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10360 LNCS. 394âĂŞ402. https://doi.org/10.1007/978-3-319-60131-1

Kam Tong Chan, Irwin King, and Man-Ching Yuen. 2009. Mathematical modeling of social games. *In Computational Science and Engineering, 2009. CSE'09. International Conference on*, Vol. 4. IEEE, 1205--1210.

Chen Chen and Shengdong Zhao. 2017. ReTool: Interactive Microtask and Workflow Design through Demonstration. *In Chi.* 3551--3556. https://doi.org/10.1145/3025453.3025969

Jenny J Chen, Natala J Menezes, Adam D Bradley, and T A North. 2011. Opportunities for Crowdsourcing Research on Amazon Mechanical Turk. *Human Factors* 5, 3 (2011), 3. https://doi.org/10.1145/1357054.1357127

Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. 2015. Break it down: A comparison of macro-and microtasks. *In CHI '15*. ACM, 4061--4064.

Anand Inasu Chittilappilly, Lei Chen, and Sihem Amer-Yahia. 2016. A Survey of General-Purpose Crowdsourcing Techniques. *IEEE Transactions on Knowledge and Data Engineering* 28, 9 (9 2016), 2246--2266. https://doi.org/10.1109/TKDE.2016.2555805

J R Corney, C Torres-Sánchez, P Jagadeesan, A Lynn, and W Regli. 2009. Outsourcing labour to the cloud. *International Journal of Innovation and Sustainable Development* 4, 4 (2009), 294--313. https://doi.org/10.1504/IJISD.2009.033083

Matthew J C Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8, 3 (2013). https://doi.org/10.1371/journal.pone.0057410

Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 2015. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. *In Crowd Work and Crowd Process*. Vancouver. https://doi.org/10.1145/2675133.2675260

Tadele T Damessie and J Shane Culpepper. 2016. The Effect of Document Order and Topic Difficulty on Assessor Agreement. *In ICTIR* 6. 2--5. https://doi.org/10.1145/2970398.2970431

A P Dawid and A M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society Series C Applied Statistics* 28, 1 (1979), 20--28. https://doi.org/10.2307/2346806

R Dawson and S Bynghall. 2011. Getting Results from Crowds: The Definitive Guide to Using Crowdsourcing to Grow Your Business. *Advanced Human Technologies.*

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. *Proceedings of the 21st international conference on World Wide Web - WWW '12* (2012), 469--478. https://doi.org/10.1145/2187836.2187900

Xuefei Deng, K. D. Joshi, and Robert D. Galliers. 2016. The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to the Less Powerful Through Value Sensitive Design. *MIS Quarterly* Vol. 40, X (2016), 1--24.

Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. *The Dynamics of Micro-Task Crowdsourcing.* In Proceedings of the 24th International Conference on World Wide Web - WWW '15 (WWW '15). ACM Press, New York, New York, USA, 238--247. https://doi.org/10.1145/2736277.2741685

Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. 2011. Crowdsourcing Systems on the World-Wide Web. *Commun.* ACM 54, 4 (4 2011), 86--96. https://doi.org/10.1145/1924421.1924442

Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* - CHI '16. San Jose, CA, USA, 2623--2634. https://doi.org/10.1145/2858036.2858268

Carsten Eickhoff and Arjen P. de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16, 2 (2013), 121--137. https://doi.org/10.1007/s10791-011-9181-9

Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara. 2012. Towards an Integrated Crowdsourcing Definition. *J. Inf. Sci.* 38, 2 (4 2012), 189--200. https://doi.org/10.1177/0165551512437638

Ailbhe Finnerty and Pavel Kucherbaev. 2013. Keep it simple: Reward and task design in crowdsourcing. *In Proceedings of the CHI Italy 2013*. Trento, Italy, 2--5. https://doi.org/10.1145/2499149.2499168

Michael J Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. 2011. CrowdDB: answering queries with crowdsourcing. *In Proceedings of the 2011ACM SIGMOD International Conference on Management of data*. 61--72. https://doi.org/10.1145/1989323.1989331

Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. "Modus operandi of crowd workers: The invisible role of microtask work environments." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, no. 3 (2017): 49.

Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. *IEEE Intelligent Systems* 30, 4 (2015), 81--85. https://doi.org/10.1109/MIS.2015.66

Ujwal Gadiraju, Yang Jie, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. *Proceedings of HT (2017)*, 5--14.

Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A Taxonomy of Microtasks on the Web. *In Proceedings of the 25th ACM Conference on Hypertext and Social Media*. 218--223. https://doi.org/10.1145/2631775.2631819

Galitz, W. O. (2007). The essential guide to user interface design: an introduction to GUI design principles and techniques. John Wiley & Sons.

Hector Garcia-Molina, Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios. 2016. Challenges in Data Crowdsourcing. I*EEE Transactions on Knowledge and Data Engineering* 28, 4 (4 2016), 901--911. https://doi.org/10.1109/TKDE.2016.2518669

David Geiger, Michael Rosemann, and Erwin Fielt. 2011. Crowdsourcing information systems: a systems theory perspective. *In Proceedings of the 22nd Australasian Conference on Information Systems (ACIS 2011).*

David Geiger and Martin Schader. 2014. Personalized task recommendation in crowdsourcing information systems - Current state of the art. *Decision Support Systems* 65, C (2014), 3--16. https://doi.org/10.1016/j.dss.2014.05.007

Buddhadeb Halder. 2014. Evolution of crowdsourcing: potential data protection, privacy and security concerns under the new media age. *Revista Democracia Digital e Governo Eletrônico* 1, 10 (2014), 377--393.

Lane Harrison, Drew Skau, Steven Franconeri, Aidong Lu, and Remco Chang. 2013. Influencing visual judgment through affective priming. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. 2949--2958. https://doi.org/10.1145/2470654.2481410

Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive Task Assignment for Crowdsourced Classification. *Proceedings of the 30th International Conference on Machine Learning* (ICML-13) 28 (2013), 534--542. http://yiling.seas.harvard.edu/sc2013/Ho.pdf

Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119--131. https://doi.org/10.1007/s40708-016-0042-6

Chi Hong. 2017. Generative Models for Learning from Crowds. (2017). https://arxiv.org/pdf/1706.03930.pdf

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. *Naacl-Hlt '13* 3, June (2013), 1120--1130. http://www.aclweb.org/anthology/N13-1132

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine* 14, 6 (2006), 1--4. https://doi.org/10.1086/599595

Jeff Howe. 2008. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business (1 ed.). *Random House Business Books*, New York, NY, USA.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. *In 13th International Semantic Web Conference (ISCW2014)*. 486--504. https://doi.org/10.1007/978-3-319-11915-1_31

Panos Ipeirotis. 2010. Demographics of Mechanical Turk. *Working Paper CeDER-10-01* August (2010). https://doi.org/10.2139/ssrn.1585030

Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. *In Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*. 64. https://doi.org/10.1145/1837885.1837906

Lily C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. *In Proceedings of the SIGCHI Conference on CHI17*. Paris, 611--620. https://doi.org/10.1145/2470654.2470742

Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment* 10, 7 (2017), 829--840.

Jing Wang, Siamak Faridani, and Panagiotis G. Ipeirotis. 2011. Estimating the Completion Time of Crowdsourced Tasks Using Survival Analysis Models. *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)* 1 (2011), 31- -34. http://ai2-s2-pdfs.s3.amazonaws.com/4744/4975fe62b6bd24f2fbc6d712cf3c23636fff.pdf

Hiroshi Kajino and Hisashi Kashima. 2011. A Convex Formulation of Learning from Crowds. 111, 275 (2011), 231--236. https://doi.org/10.1527/tjsai.27.133

David R Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative Learning for Reliable Crowdsourcing Systems Accessed Iterative Learning for Reliable Crowdsourcing Systems. *Advances in Neural Information Processing Systems* 24 (NIPS 2011) (2011), 1--9. https://web.engr.illinois.edu/~swoh/paper_ crowdsourcing_nips.pdf

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. *In Proceedings of the 2011 ACM international conference on Information and knowledge management.* 1941--1944. https://doi.org/10.1145/2063576.2063860

Faiza Khan Khattak and Ansaf Salleb-Aouissi. 2011. Quality Control of Crowd Labeling through Expert Evaluation. *In NIPS*. https://people.cs.umass.edu/~wallach/workshops/nips2011css/papers/Khattak.pdf

Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian Classifier Combination. *In Proceedings of the International Conference on Artifical Intelligence and Statistics (AISTATS)*. 619--627. http://proceedings.mlr.press/v22/kim12/kim12.pdf

Jaejeung Kim, Sergey Leksikov, Punyotai Thamjamrassri, Uichin Lee, and Hyeon-Jeong Suk. 2015. CrowdColor: Crowdsourcing Color Perceptions Using Mobile Devices. *In Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '15*. Copenhagen, Denmark, 478--483. https://doi.org/10.1145/2785830.2785887

Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. 2012. CrowdWeaver. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12 (2012),* 1033. https://doi.org/10.1145/2145204.2145357

Aniket Kittur, Boris Smus, and Robert Kraut. 2011. CrowdForge Crowdsourcing Complex Work. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11 (2011)*, 1801. https://doi.org/10.1145/1979742.1979902

Anand P. Kulkarni, Matthew Can, and Bjoern Hartmann. 2011. Turkomatic. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11 (2011)*, 2053. https://doi.org/10.1145/1979742.1979865

Walter S Lasecki, Jeffrey M Rzeszotarski, Adam Marcus, and Jeffrey P Bigham. 2015. The Effects of Sequence and Delay on Crowd Work. *In CHI '15, Vol. 1.* ACM, New York, NY, USA, 1375--1378. https://doi.org/10.1145/2702123.2702594

Edith Law and Luis von Ahn. 2011. *Human Computation.* Vol. 5. 1--121 pages. https://doi.org/10.2200/S00371ED1V01Y201107AIM013

Edith Law and Luis Von Ahn. 2009. Input-agreement: a new mechanism for collecting data using human computation games. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 1197--1206.

Edith Law, Luis von Ahn, and Luis Von Ahn. 2005. Human Computation. *Ph.D. Dissertation.* https://doi.org/10.2200/S00371ED1V01Y201107AIM013

Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. 2009. Evaluation of Algorithms Using Games: The Case of Music Tagging. *In ISMIR*. 387--392.

Hongwei Li, Bo Zhao, and Ariel Fuxman. 2014. The wisdom of minority: discovering and targeting the right group of workers for crowdsourcing. *Proceedings of the 23rd International Conference on World Wide Web (2014)*, 165--175. https://doi.org/10.1145/2566486.2568033

Joseph C. R. Licklider. 1960. Man-computer symbiosis. I*RE transactions on human factors in electronics HFE-1*, 1 (1960), 4--11. https://doi.org/10.1109/THFE2.1960.4503259

Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49, 2 (4 2017), 433--442. https://doi.org/10.3758/s13428-016-0727-z

Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2010. TurKit : Human Computation Algorithms on MTurk. *In Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 57--66. https://doi.org/10.1145/1866029.1866040

Qiang Liu, Jian Peng, and Alexander Ihler. 2012b. Variational Inference for Crowdsourcing. *Nips (2012)*, 701--709. http://papers.nips.cc/paper/4627-variational-inference-for-crowdsourcing.pdf

Xuan Liu, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, and Meihui Zhang. 2012a. CDAS: A Crowdsourcing Data Analytics System. (2012). https://doi.org/10.14778/2336664.2336676

Nuno Luz, Nuno Silva, and Paulo Novais. 2015. A survey of task-oriented crowdsourcing. *Artificial Intelligence Review* 44, 2 (2015), 187--213. https://doi.org/10.1007/s10462-014-9423-5

Ke Mao, Licia Capra, Mark Harman, and Yue Jia. 2015a. UCL DEPARTMENT OF COMPUTER SCIENCE A Survey of the Use of Crowdsourcing in Software Engineering. (2015), 1--36. https://doi.org/10.1016/j.jss.2016.09.015

Ke Mao, Ye Yang, Qing Wang, Yue Jia, and Mark Harman. 2015b. Developer recommendation for crowdsourced software development tasks. *In Proceedings - 9th IEEE International Symposium on Service-Oriented System Engineering*, IEEE SOSE 2015, Vol. 30. 347--356. https://doi.org/10.1109/SOSE.2015.46

Adam Marcus, David Karger, Samuel Madden, Robert Miller, and Sewoong Oh. 2012. Counting with the crowd. *Proceedings of the VLDB Endowment* , 6, 2 (2012), 109--120. https://doi.org/10.14778/2535568.2448944

Winter Mason and Duncan J Watts. 2009. Financial incentives and the performance of crowds. *Proceedings of the ACM SIGKDD Workshop on Human Computation* 11, 2 (2009), 77--85. https://doi.org/10.1145/1809400.1809422

Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. *Hcomp (2016)*, 139--148. https://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/viewFile/14043/13641

Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16 (2016)*, 2271--2282. https: //doi.org/10.1145/2858036.2858539

Patrick Minder and Abraham Bernstein. 2012. CrowdLang: programming human computation systems. (2012).

Tanushree Mitra, C J Hutto, and Eric Gilbert. 2015. Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems 1 (2015)*, 1345--1354. https://doi.org/10.1145/2702123.2702553

Robert R. Morris, Mira Dontcheva, and Elizabeth M. Gerber. 2012. Priming for better performance in microtask crowdsourcing environments. *IEEE Internet Computing* 16, 5 (2012), 13--19. https://doi.org/10.1109/MIC.2012.68

Sara Moussawi and Marios Koufaris. 2013. The Crowd on the Assembly Line: Designing Tasks for a Better Crowdsourcing Experience. In ICIS. 1--17.

Nakatsu, R. T., Grossman, E. B., & Iacovou, C. L. (2014). A taxonomy of crowdsourcing based on task complexity. Journal of Information Science, 40(6), 823-834.

Edward Newell and Derek Ruths. 2016. How One Microtask Affects Another. *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3155--3166. https://doi.org/10.1145/2858036.2858490

Jakob Nielsen. 1993. Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Besmira Nushi, Adish Singla, Anja Gruenheid, Erfan Zamanian, Andreas Krause, and Donald Kossmann. 2015. Crowd Access Path Optimization: Diversity Matters. *Proceedings, The Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15)* AUGUST (2015), 130--139. http://arxiv.org/abs/1508.01951

David Oleson, Alexander Sorokin, Greg Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. *Human Computation: Papers from the 2011 AAAI Workshop (2011)*, 43--48.

Yue Pan and Eli Blevis. 2011. A survey of crowdsourcing as a means of collaboration and the implications of crowdsourcing for interaction design. *In Proceedings of the 2011 International Conference on Collaboration Technologies and Systems, CTS 2011*. IEEE, 397--403. https://doi.org/10.1109/CTS.2011.5928716

Natalie Parde and Rodney Nielsen. 2017. Finding Patterns in Noisy Crowds: Regression-based Annotation Aggregation for Crowdsourced Data. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017), 1908--1913. http://www.aclweb.org/anthology/D17-1204

Eyal Peer, Sonam Samat, Laura Brandimarte, and Alessandro Acquisti. 2016. Beyond the Turk : Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology 70, January (2016)*, 153--163. https://doi.org/10.1016/j.jesp.2017.01.006

Alexander J Quinn and Benjamin B Bederson. 2009. A taxonomy of distributed human computation. Human-Computer Interaction Lab Tech Report, University of Maryland (2009).

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 (CHI '11)*. ACM, New York, NY, USA, 1403--1412. https://doi.org/10.1145/1978942.1979148

Alexander J Quinn and Benjamin B Bederson. 2014. AskSheet : Efficient Human Computation for Decision Making with Spreadsheets. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14 (2014)*, 1456--1466. https://doi.org/10.1145/2531602.2531728

Alexander J. Quinn, Benjamin B. Bederson, Tom Yeh, and Jimmy Lin. 2010. CrowdFlow: Integrating Machine Learning with Mechanical Turk for Speed-Cost-Quality Flexibility. *Human Computer Interaction Lab, 2010-05 (2010)*, 1--8.

Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* Vol. 8181 LNCS. 1--15. https://doi.org/10.1007/978-3-642-41154-0_1

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, Linda Moy, and Linda Moy@nyumc Org. 2010. Learning From Crowds. *Journal of Machine Learning Research 11 (2010)*, 1297--1322. https://doi.org/10.2139/ssrn.936771

Vikas C Raykar, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised learning from multiple experts. *In Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. 1--8. https://doi.org/10.1145/1553374.1553488

Filipe Rodrigues, Francisco C Pereira, and Bernardete Ribeiro. 2014. Gaussian Process Classification and Active Learning with Multiple Annotators. *Proceedings of the 31st International Conference on Machine Learning 32 (2014)*, 433--441. http://proceedings.mlr.press/v32/rodrigues14.pdf

Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10 March 2017 (2010)*, 2863. https://doi.org/10.1145/1753846.1753873

Anne C Rouse. 2010. A preliminary taxonomy of crowdsourcing. *ACIS 2010 Proceedings 76 (2010).*

Cristina Sarasua, Elena Simperl, and Natalya F Noy. 2012. CROWD MAP: Crowdsourcing Ontology Alignment with Microtasks. The Semantic Web. ISWC 2012 *Lecture Notes in Computer Science (2012)*, 525--541.

Daniel Schall, Hong-Linh Truong, and Schahram Dustdar. 2011. The human-provided services framework. *In Socially Enhanced Services Computing*. Springer, 1--15.

Eric Schenk and Claude Guittard. 2011. Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics* 0, 1 (2011), 93--107. https://doi.org/10.3917/jie.007.0093

Steffen Schnitzer, Christoph Rensing, and Sebastian Schmidt. 2015. Demands on task recommendation in crowdsourcing platforms - the workers perspective. *In workshop of Crowdsourcing and human computation for recommender systems (CrowdRec2015)*, ACM RecSys 2015, September (2015), 1--7.

Victor S Sheng. 2017. Label Aggregation for Crowdsourcing with Bi-Layer Clustering. 1 (2017), 921--924. https://doi.org/10.1145/3077136.3080679

James Surowiecki. 2004. The Wisdom of Crowds. How Collective Wisdom Shapes Business Economies Societies and Nations New York Doubleday (2004), 296. https://doi.org/10.3174/ajnr.A3417

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), 257-285.

Nguyen Thanh Tam, Huynh Huu Viet, Nguyen Quoc Viet Hung, MatthiasWeidlich, Hongzhi Yin, and Xiaofang Zhou. 2017. Multi-Label Answer Aggregation for Crowdsourcing. *ACM Comput. Surv. Article 1828* (2017), 53--59. https://doi.org/10.1145/1235

Stefano Tranquillini, Florian Daniel, Pavel Kucherbaev, and Fabio Casati. 2015. Modeling, Enacting, and Integrating Custom Crowdsourcing Processes. *ACM Trans. Web* 9, 2 (2015), 7:1--7:43. https://doi.org/10.1145/2746353

Alan M Turing. 1950. Computing machinery and intelligence. *Mind 59*, 236 (1950), 433--460.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based bayesian aggregation models for crowdsourcing. *In Proceedings of the 23rd international conference on World wide web - WWW '14*. 155--164. https://doi.org/10.1145/2566486.2567989

Matteo Venanzi, John Guiver, Pushmeet Kohli, and Nicholas R. Nick Jennings. 2016. Time-Sensitive Bayesian information aggregation for crowdsourcing systems. *Journal of Artificial Intelligence Research 56 (2016)*, 517--545. http://www.jair.org/media/5175/live-5175-9434-jair.pdf

Xuan Wei, Daniel Dajun Zeng, and Junming Yin. 2017. Multi-Label Annotation Aggregation in Crowdsourcing. (2017). http://arxiv.org/abs/1706.06120

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. 2010. The Multidimensional Wisdom of Crowds. *Most 6 (2010)*, 1--9. https://doi.org/10.1.1.231.1538

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Advances in Neural Information Processing Systems 22, 1 (2009)*, 1--9.

Paul Whitla. 2009. Crowdsourcing and Its Application in Marketing Activities. *Contemporary Management Research 5, 1 (3 2009)*, 15--28.

Meng-han Wu and Alexander J Quinn. 2017. Confusing the Crowd : Task Instruction Quality on Amazon Mechanical Turk. *In The Fifth AAAI Conference on Human Computation and Crowdsourcing*. 206--215.

Guo Xintong, Wang Hongzhi, Yangqiu Song, and Gao Hong. 2014. Brief survey of crowdsourcing for data mining. *Expert Systems with Applications 41, 17 (2014)*, 7987--7994. https://doi.org/10.1016/j.eswa.2014.06.044

Yan Yan, Romer Rosales, Glenn Fung, and Mark Schmidt. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. New York 9 (2010), 932--939. http://proceedings.mlr.press/v9/yan10a/yan10a.pdfhttp://jmlr.csail.mit.edu/proceedings/papers/v9/yan10a/yan10a.pdf

Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling Task Complexity in Crowdsourcing. October (2016), 249--258. https://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/viewFile/14039/13653

Yang Yang, Bin B Zhu, Rui Guo, Linjun Yang, Shipeng Li, and Nenghai Yu. 2008. A comprehensive human computation framework: with application to image labeling. *In Proceedings of the 16th ACM international conference on Multimedia*. ACM, 479--488.

MC Yuen, Irwin King, and KS Leung. 2012. Task recommendation in crowdsourcing systems. http://dl.acm.org/citation.cfm?id=2442661

Man-Ching Yuen, Ling-Jyh Chen, and Irwin King. 2009. A survey of human computation systems. *In Computational Science and Engineering, 2009. CSE'09. International Conference on, Vol. 4.* IEEE, 723--728.

Man Ching Yuen, Irwin King, and Kwong Sak Leung. 2011. A survey of crowdsourcing systems. Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE *International Conference on Social Computing, PASSAT/SocialCom 2011 (2011)*, 766--773. https://doi.org/10.1109/PASSAT/SocialCom.2011.36

Man Ching Yuen, Irwin King, and Kwong Sak Leung. 2015. TaskRec: A Task Recommendation Framework in Crowdsourcing Systems. *Neural Processing Letters* 41, 2 (2015), 223--238. https://doi.org/10.1007/s11063-014-9343-z

Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng. 2015. QASCA. I*n Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*. 1031--1046. https://doi.org/10.1145/2723372.2749430

Dengyong Zhou, John Platt, Sumit Basu, and Yi Mao. 2012. Learning from the wisdom of crowds by minimax entropy. *Advances in Neural Information Processing Systems* 25 (2012), 2204--2212. http://papers.nips.cc/paper/4490-learning-from-the-wisdom-of-crowds-by-minimax-entropy.pdf

Haiyi Zhu, Steven P Dow, Robert E Kraut, and Aniket Kittur. 2014. Reviewing versus Doing: Learning and Performance in Crowd Assessment. *In Proceedings Conference on Computer Supported Cooperative Work and Social Computing (CSCW '14)*. 1445--1455. https://doi.org/10.1145/2531602.2531718