Diversifying Demining
An Experimental Crowdsourcing Method for Optical Mine Detection

DAVID ANDERSSON

FOI
Swedish Defence Research Agency
Information Systems
P.O. Box 1165
SE-581 11 Linköping

Phone: +46 13 37 80 00
Fax: +46 13 37 81 00

www.foi.se

FOI-R--2619--SE
ISSN 1650-1942

Technical Report
October 2008

Information Systems

David Andersson

Diversifying Demining

An Experimental Crowdsourcing Method for Optical Mine Detection

Titel	Diversifiering av minröjning: En experimentell crowdsourcingmetod för optisk mindetektering
Title	Diversifying Demining: An Experimental Crowd-sourcing Method for Optical Mine Detection
Rapportnr/Report no	FOI-R--2619--SE
Rapporttyp Report Type	Teknisk rapport Technical report
Månad/Month	Oktober/October
Utgivningsår/Year	2008
Antal sidor/Pages	63 p
ISSN	ISSN 1650-1942
Kund/Customer	Försvarsmakten
Forskningsområde Programme area	4. Sensorer och signaturanpassning 4. Sensors and Low Observables
Delområde Subcategory	42 Sensorer 42 Above surface Surveillance, Target acquisition and Reconnaissance
Projektnr/Project no	E3084
Godkänd av/Approved by	Ola Kärvell

FOI, Totalförsvarets Forskningsinstitut
Avdelningen för Informationssystem
Box 1165
581 11 Linköping

FOI, Swedish Defence Research Agency
Information Systems
Box 1165
SE-581 11 Linköping

Sammanfattning

Detta examensarbete går igenom tanken bakom crowdsourcing och mångfaldens styrka tillämpad på optisk mindetektering. Tanken är att använda det mänskliga ögat och Internets skiftande och varierande arbetsstyrka som ett tillägg för att upptäcka minor tillsammans med dataalgoritmer.

Mångfaldsteorin i problemlösande diskuteras och speciellt ”Diversity Trumps Ability”-satsen och ”Diversity Prediction”-satsen och hur de ska genomföras för tillämpningar som kontrastigenkänning respektive ytreduktion.

Ett enkelt kontrastigenkänningsexperiment har genomförts för att jämföra resultaten mellan en lekmannagrupp och en expertgrupp. Grupperna tittar på delar av data från hyperspektrala bilder och klassifierar andel objekt eller minor och terrängtyp. På grund av lågt deltagande från expertgruppen och en felaktig experimentintroduktion ger inte experimentet några statistiskt signifikanta resultat, varför ingen slutsats dras.

Experimentförbättringar och framtida tillämpningar föreslås.

Nyckelord:

Crowdsourcing, mindetektion, minspaning, mångfald, bildanalys

Summary

This thesis explores the concepts of crowdsourcing and the ability of diversity, applied to optical mine detection. The idea is to use the human eye and wide and diverse workforce available on the Internet to detect mines, in addition to computer algorithms.

The theory of diversity in problem solving is discussed, especially the Diversity Trumps Ability Theorem and the Diversity Prediction Theorem, and how they should be carried out for possible applications such as contrast interpretation and area reduction respectively.

A simple contrast interpretation experiment is carried out comparing the results of a laymen crowd and one of experts, having the crowds examine extracts from hyperspectral images, classifying the amount of objects or mines and the type of terrain. Due to poor participation rate of the expert group, and an erroneous experiment introduction, the experiment does not yield any statistically significant results. Therefore, no conclusion is made.

Experiment improvements are proposed as well as possible future applications.

Keywords:

Crowdsourcing, diversity, mine detection, image analysis

Contents

1 Introduction **1**
- 1.1 Background . . . 1
- 1.2 Problem Description . . . 1
- 1.3 Goal . . . 2
- 1.4 Thesis Outline . . . 2

2 Theory **3**
- 2.1 Crowdsourcing . . . 3
  - 2.1.1 NASA Clickworkers . . . 4
  - 2.1.2 Amazon Mechanical Turk . . . 5
  - 2.1.3 The Steve Fossett Search . . . 6
  - 2.1.4 Peekaboom . . . 7
- 2.2 Diversity . . . 8
  - 2.2.1 Definitions for Diversity . . . 8
  - 2.2.2 Diversity Trumps Ability . . . 10
  - 2.2.3 Diversity Prediction . . . 11
  - 2.2.4 Predictive Markets . . . 12
- 2.3 Applied Theory . . . 13
  - 2.3.1 Possible Crowdsourcing Applications . . . 13
  - 2.3.2 Discussion of the Diversity Trumps Ability Conditions . . . 14
  - 2.3.3 Creating an Experiment . . . 15
  - 2.3.4 Implications . . . 16

3 Experiment **19**
- 3.1 Purpose . . . 19
- 3.2 Participants . . . 19
- 3.3 Data Set . . . 20
  - 3.3.1 Hyperspectral Measurement Data . . . 20
  - 3.3.2 Data Cells . . . 22
- 3.4 Implementation . . . 22
  - 3.4.1 Rewards and Cost . . . 22
  - 3.4.2 HIT-Builder . . . 24
  - 3.4.3 Reasons for Choosing the AMT/HIT-Builder . . . 24
  - 3.4.4 Experiment HITs . . . 25

3.4.5 Qualification Test . . . 28
3.5 Execution . . . 31
3.5.1 HIT Posting Problems . . . 31
3.5.2 Experiment Participants . . . 31
3.5.3 Expenses . . . 32
3.5.4 Feedback . . . 33
3.5.5 Cancellation of Experiment . . . 34

4 Analysis **35**
4.1 Descriptive Statistics . . . 35
4.1.1 Central Tendency . . . 35
4.1.2 Spread . . . 36
4.1.3 Observations . . . 36
4.1.4 Detection and False Alarm Rates . . . 37
4.2 Inferential Statistics . . . 38
4.2.1 Two-Tailed Correlation Tests . . . 38
4.2.2 Correlation . . . 39
4.2.3 Mean Value Test . . . 40

5 Conclusion **41**
5.1 Statistical Analysis . . . 41
5.2 Experiment Failure . . . 41
5.2.1 Method—Or the Lack of . . . 41
5.2.2 Erroneous Qualification Questions . . . 42
5.3 Lessons Learned . . . 42
5.4 Discussion . . . 42
5.4.1 Worker Verification . . . 42
5.5 Future Work . . . 43
5.5.1 Improvements for Future Experiment . . . 43
5.5.2 Diversity Trumps Ability Experiment . . . 43
5.5.3 Computer Algorithms Comparison . . . 44
5.5.4 Area Reduction Market . . . 44
5.6 Summary . . . 45

Bibliography **47**

A Theory **51**
A.1 Needs Assessment Excerpt . . . 51

B Experiment **53**
B.1 MATLAB Image Processing Code . . . 53
B.1.1 autocontrast.m . . . 53
B.1.2 labImadjust.m . . . 54
B.2 Qualification Test . . . 55
B.3 HIT-Builder Posting Errors . . . 57

Chapter 1

Introduction

1.1 Background

The Master's thesis *Diversifying Demining: An Experimental Crowdsourcing Method for Optical Mine Detection* is commissioned by the *Swedish Defence Research Agency* (FOI), Linköping, Sweden. It is a part of the *Multi Optical Mine Detection System* (MOMS) project, at the division of Sensor Systems. The MOMS project evaluates the possibility of an achievable and efficient electro-optical (EO) multi sensor system for mine detection.

In examining the previous internal reports [1], [2], and [3] of the MOMS project, it was noticed that there were yet no definite algorithms for extracting and fusing data of mines and unexploded ordnances (UXO) from different optical sources. However, it was also noticed that the human eye and brain could often easily detect deviations in the reports' measurement images, thus finding mines and UXOs.

1.2 Problem Description

The notion of using the human eye and humans—diversifying an otherwise expert-focused trade—is further explored and two main problems are defined:

1. Using crowdsourcing, is it possible to effectively *manually*[1] detect land mines and UXOs in optical image data used in the MOMS project?

2. Does a layman differ in recognizing mines and UXOs from experienced personnel from the demining community?

The answers to these questions will aid in determining if: *Could crowdsourcing become a useful tool in demining?*

[1] By a non-automated process involving people rather than algorithms and computers.

1.3 Goal

The goal of this thesis is to examine previous experiments made in similar image recognition tasks, to identify how randomly selected participants perform relative to expert participants, and to perform an open-call experiment with random and expert participants using MOMS data.

Experiences and lessons learned from both the theoretical study and the execution of the experiment will be discussed.

1.4 Thesis Outline

This paper is divided into 5 chapters. Chapter 2 surveys the theory of crowdsourcing and diversity. Furthermore, it takes into consideration how to apply this theory to an experiment in mine detection.

Chapter 3 explains the experiment that was carried out to find out how efficient laymen and experts are in detecting mines in optical data.

The statistical outcome of the experiment is presented in chapter 4.

In chapter 5, the experimental results are discussed, conclusions are made based on theory and experiment, and directions are given for future work.

Chapter 2

Theory

In demining, a lot of expertise and experience is required to spot, reduce, and clear mine contaminated areas. This is vital due to the sensitive nature of the profession and the imminent danger present in declaring an area clear for public access. Diversity, though, may induce new perspectives to demining, and mine detection in particular.

The MOMS project tries to optically detect mines in an automated environment. An added tool to that automation could be a semiautomated process of a diverse collection of solvers—adding a human eye and brain, skill, and perspective to the algorithms. As put by NASA in [4]:

> *[...] each and every human brain has image-processing abilities unrivaled by any supercomputer.*

Many solvers would be needed—a crowd—performing the task of browsing and classifying data for mine detection. A globalized internet makes those solvers available, and in the last two years, crowdsourcing has emerged as a new way to outsource and draw wisdom from crowds.

This chapter explains the term crowdsourcing, the theory behind crowdsourcing, the effectiveness of diversity and where it is effective, and how an experiment could be setup to test this in mine detection.

2.1 Crowdsourcing

In June of 2006, in the Wired Magazine article [5], journalist Jeff Howe coined the phrase *crowdsourcing*. Later on his blog [6][1], Howe further developed this term, defining it as follows:

Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.

[1] The blog is a part of the book he is writing on the subject, due in August 2008.

What differs crowdsourcing from outsourcing is the undefined group of people completing the job. A crowd of people may include both amateurs and professionals, thus a possible 'wisdom' of expertise and diversity, which is discussed in [7].

Just as outsourcing may imply both the practice of substituting cheap labour from other countries to cut costs at home, and the general meaning of using skills from outside when internal expertise is lacking, crowdsourcing may be used for both reasons.

2.1.1 NASA Clickworkers

In [4], NASA tested the "ability of pooled efforts"—crowdsourcing—to accomplish time-consuming tasks by letting so called *clickworkers* perform image-recognition tasks.

In the pilot study the clickworkers were exposed to images of the surface of Mars and its craters, the task was to detect and to draw circles on top of the craters. Figure 2.1 shows an example image. There was already a catalogue [8] to which the clickworkers' results were measured against. A training example with seven known craters was given to demonstrate how to classify and draw the crater circles.

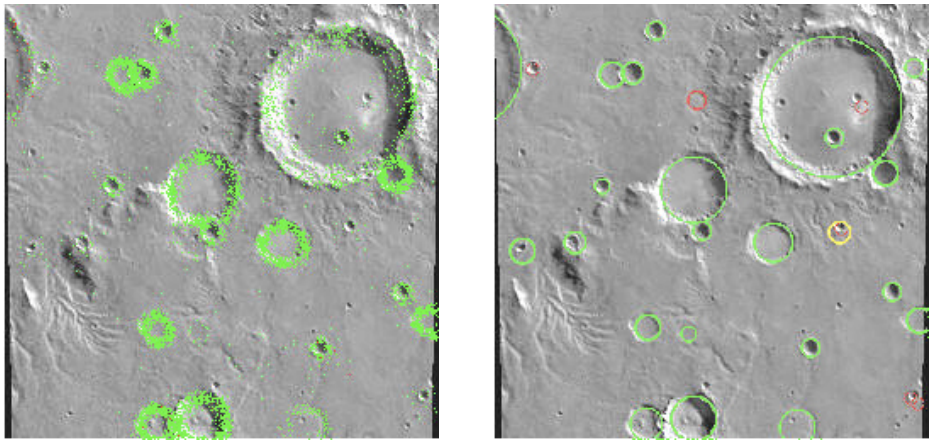


Figure 2.1. Left image: Clicks from 220 individuals. Right image: Consensus obtained by a weighted clustering of inputs. Courtesy of NASA Ames Research Center.

The purpose of the clickworker experiment was to answer the following questions: Are people interested in volunteering their free time for routine scientific work? Does the public have the training and motivation to produce accurate results in a scientifically important task?

They concluded yes in both cases.

Of 317 craters over 30 km in diameter that were contained in images assigned to five different clickworkers, 85% were found by at least two people. Of 86 faint craters, assigned to ten different clickworkers, 95% were found.

Their conclusion also suggests using crowdsourcing as to improve productive searching, using volunteers to prioritize results into a subset, which is then used, rather than an arbitrary sample.

2.1.2 Amazon Mechanical Turk

The *Amazon Mechanical Turk* (AMT) is a virtual marketplace for work and a typical example of a crowdsourcing implementation [9]. The work is divided into tasks, often small and quick, called *Human Intelligence Tasks* or *HITs*. The solvers of the tasks are called *workers*, and those publishing tasks to be completed by the workers are called *requesters*.

Sets of similar HITs are grouped in HIT groups; it is assumed that most HITs in a HIT group are the same type of HIT. A requester can make each worker take a *qualification test* to prove competent for the HIT. The requester may also reward the worker financially for submitting a valid HIT. AMT comes with an application programming interface (API).

Applications of Human Intelligence Tasks

AMT tasks are good for services like: transcription, translation, editing/proofreading, OCR/handwriting, content analysis, image recognition, blog content, filtering, and surveys. See table 2.1 for examples.

Table 2.1. HIT examples from [10]

.

Select the correct spelling for these search terms
Is this website suitable for a general audience?
Find the item number for the product in this image
Rate the search results for these keywords
Are these two products the same?
Choose the appropriate category for products
Categorize the tone of this article
Translate a paragraph from English to French

Demographics

A survey on the demographics of the AMT workers is available in [11]. Table 2.2 shows that workers are mainly from the U.S., India, the U.K., and Canada. The high American penetration is because of the AMT beta phase, requiring U.S. bank accounts to withdraw accumulated AMT rewards.

Slightly more are female than male, and most are the age of 21–30 and 31–40. Most have a Bachelor's degree, some have Master's or even PhD degrees.

Table 2.2. AMT worker demographics. Courtesy of P. Ipeirotis.

Country		**Gender**	
U.S.A.	76.25%	Female	58.19%
India	8.03%	Male	41.81%
U.K.	3.34%		
Canada	2.34%		
Philippines	1,34%		
Age		**Education**	
21–30	42,81%	High School	26,69%
31–40	33,68%	Bachelor's	53,04%
41–50	17,19%	Master's	15,20%
51–60	4,21%	PhD	5,07%
>60	2,11%		

2.1.3 The Steve Fossett Search

The aftermath of disappeared aviator Steve Fossett showed the infancy and inexperience of crowdsourcing as a tool.

In September 2007, Steve Fossett was reported lost in Nevada in [12]. Soon after his disappearance, Google and Amazon started an image-based search for Fossett, see [13]. Satellite data was divided into 85 m^2 sections, and Mechanical Turk workers were asked to flag images with "foreign objects" that might be a crash site or other evidence that should be examined more closely. See figure 2.2.

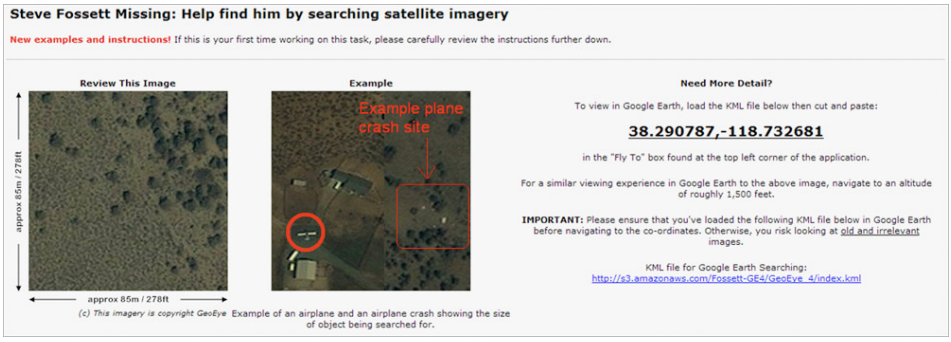


Figure 2.2. Screenshot of HIT from the Fossett search. The left image is the one to be examined, the right one is to get an idea of the scale. Courtesy of Amazon.

Other crashed planes were found, Fossett's own was not.[2] The discussions in

[2] After the presentation of this thesis, the wreckage of Fossett's plane has been found, reported in `http://www.iht.com/articles/2008/10/02/america/03fossett.php` on October 2nd, 2008.

[14] and the investigated search disclosed major problems with crowdsourcing. A number of factors left contributors conflicted and disappointed with the results:

Background Little understanding by contributors as to what they were actually looking for. Good background information was not given.

Interface Poor user interface. Users sought for the possibility to compare area images from dates prior to that of the disappearance.

Communication Lack of insight into what others had found. There was no information sharing as to what others had tagged or found.

2.1.4 Peekaboom

In 2006, von Ahn et al. constructed *Peekaboom* [15]—a game in which two players help out to classify and segment images.

The *peek* player is trying to guess what her partner is revealing by typing words that corresponds to what she sees. The *boom* player is revealing the most important parts of the image by clicking and piece-wise revealing the whole image, as to let the peeking come through as easily and quickly as possible. She then accepts the peekers guess when it is correct. (See figure 2.3.)

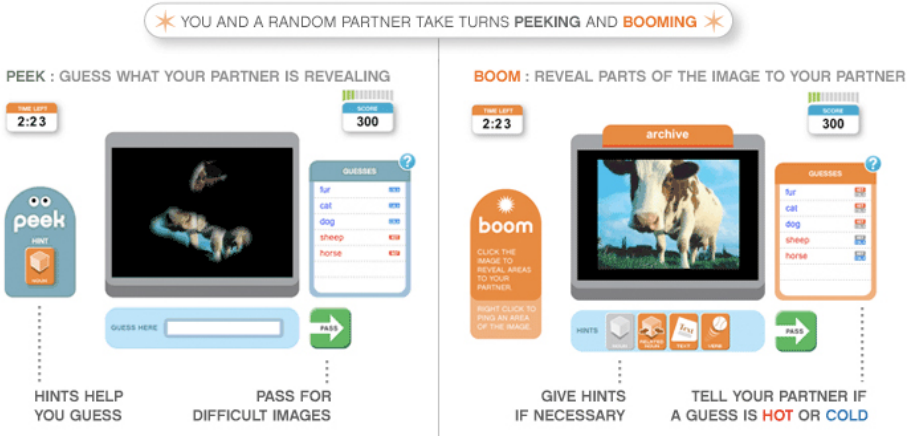


Figure 2.3. Peekaboom structure. Courtesy of Peekaboom.org.

All booming and peeking is stored, and a gamer does not know whether she is facing a real person or a stored booming/peeking. The system records all games and can use a bot to simulate both booming and peeking if there is an odd number of players online.

The Peekaboom team devised a simple algorithm to accurately calculate object bounding-boxes from user entries which would not only locate the object, but also show the extent of the object in the image.

2.2 Diversity

This section explains the differences between a diversity's ability to predict, what is popularily called the "wisdom of crowds" in [7], and a theorem saying that the results of a collection of random (smart) agents will trump the results of a collection of the best individual performers.

2.2.1 Definitions for Diversity

A distinction is made between *collections* and *groups*, they are defined as:

Collection Considered a collection is some number of individuals working on a problem sequentially or in parallel, but independently.

Group In contrast, a group consists of people who interact in space and time.

Lu Hong and Scott Page have proven in [16] and [17] that *under some conditions* diversity trumps ability; that is, the best individual performers' solutions of a problem, collectively, fare worse than a collection of random problem solvers.

A problem is defined as being either a difficult or easy problem; answering two plus two is an easy problem, building a dam is a difficult problem. Hong and Page argues that:

> *If the best problem solvers tend to think about a problem similarly, then it stands to reason that as a collection they may not be very effective and that random collections of intelligent agents may perform better owing to their diversity.* [16]

Distinguishing between *cognitive* (or functional) *diversity* and *identity diversity* is important; all Swedes do not view a problem the same way, just as a Swede and a Finn very well may. Identity-diverse people *may* induce a cognitive diversity, but it is not a given fact.

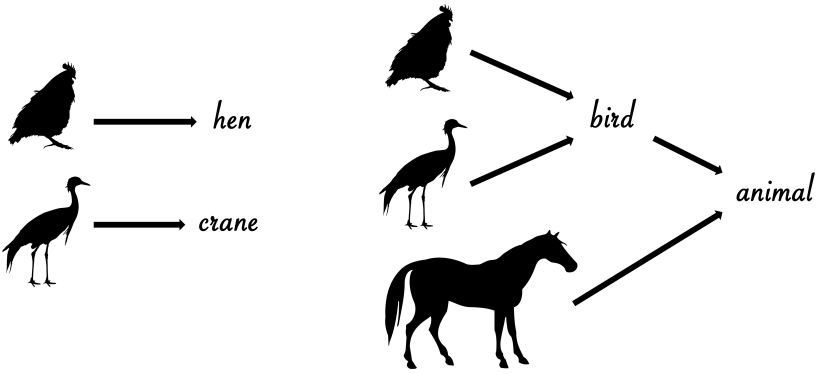
In his own book on diversity, see [18], Page explains the concept of *perspectives*, *heuristics*, and *interpretations*:

Perspective is a map from reality to an internal language such that each distinct object, situation, problem, or event gets mapped to a unique word.

Heuristic is a rule applied to an existing solution represented in a perspective that generates a new (and hopefully better) solution or a new set of possible solutions.

Interpretation is a map from objects, situations, problems, and events to words. In an interpretation, one word can represent many objects.

Perspectives are how a solver *perceives* a problem, heuristics are how he *solves* that problem. Polar and Cartesian coordinate systems are examples of perspectives, as heuristics can be examplified by either graphically using a coordinate system or using the quadratic formula to find roots to a quadratic equation.



(a) Perspectives uniquely map objects to distinct words.

(b) Interpretations group objects into words.

Figure 2.4. The difference between perspectives and interpretations. Horse image by mulp.com (Creative Commons Attribution license). Bird images by Vectoroom (GPL).

The perspective framework, however, assumes that there is a one-to-one mapping in our heads. Often this is not the case, there are not distinct words for everything, therefore interpretations are needed, see figure 2.4. Page illustrates:

> *In categorizing LEGO blocks, suppose that each block has a size and a shape and that no two blocks are identical. The encoding of blocks by their size and color creates a perspective (in the formal sense). Each block is uniquely defined. Interpretations put blocks in groups. One interpretation would be to categorize the blocks by color. Another interpretation would categorize them by size.*

Michael J. Mauboussin explains in [19] perspectives as "ways of representing situations and problems", interpretations as "ways of categorizing or partitioning perspectives", and heuristics as "ways of generating solutions to problems".

A General Mathematical Model

Here follows a more general mathematical model of the concepts of perspectives, heuristics, and interpretations, based on the foundation that gave Page and Hong the proofs of diversity trumping ability in their paper in [16].

A set of all objects X is assumed; it can be finite or infinite. Each problem solver is assumed to have an internal language Γ by which she perceives the objects at some neurological or metaphorical level. The perspective is the one-to-one representation of objects in the problem solver's internal language.

A **perspective** $P : R \to \Gamma$, where Γ is the internal language, and R is a subset of X.

The interpretation, however, is not one-to-one. The interpretation is more a coarse mapping from objects to the internal language, in which several objects may be assigned the same word of the internal language.

An **interpretation** $I : R \to \Gamma$, where Γ is the internal language, and R is a subset of X.

A problem solver's heuristic, denoted by H, is a mapping from elements of $P(R)$ in her internal language to subsets of $P(R)$. Given a $\gamma \in P(R)$, $H(\gamma) \subseteq P(R)$ is interpreted as the set of neighbouring objects in the internal representation of the problem solver that she would check to find an improvement. Let $S = P(R)$. Attention is restricted to a class of heuristics that consists of a collection of functions defined on S. For any $j = 1$, let $f_j : S \to S$ be a function. A heuristic is then defined as $H = \{f_1, \ldots, f_m\}$, where $H(\gamma) = \{f_1(\gamma), \ldots, f_m(\gamma)\}$.

A **heuristic** $H = \{f_1, \ldots, f_m\}$, where $f_j : S \to S$ for $j = 1$ to m.

2.2.2 Diversity Trumps Ability

There are four conditions that need to be fulfilled for diversity to trump ability, according to Page in [18].

1. **The Problem Is Difficult.** No individual problem solver always locates the global optimum.

   It is not an easy problem. 2 + 2 is an easy problem, saving the planet is not.

2. **The Calculus Condition.** The local optima of every problem solver can be written down in a list. In other words, all problem solvers are smart.

From this holds that the solvers can not be monkeys; to solve a calculus problem, solvers need to know calculus, to solve a statistics problem, solvers need to be statisticians.

3. **The Diversity Condition.** Any solution other than the global optimum is not a local optimum for some nonzero percentage of problem solvers.

   This states that there exists some problem solver who can find an improvement.

4. **Good-Sized Collections Drawn from Lots of Potential Problem Solvers.** The initial population of problem solvers must be large and the collections of problem solvers working together must contain more than a handful of problem solvers.

The experiment has to have many participants—several problem solvers.

Following these conditions, the *Diversity Trumps Ability Theorem* holds and is proven. Under some special conditions the same result will apply as well, but under these conditions, it always holds.

The Diversity Trumps Ability Theorem. Given conditions 1–4, a randomly selected collection of problem solvers outperforms a collection of the best individual solvers.

Keep in mind, the agents in a collection take advantage of previous solutions from previous agents in their respective collection, thus perfecting the result.

Hong and Page's models and theorem disregard incentives, communication, and learning. They also set aside two aspects of problem solving: they assume errorless communication and solvers assigning the same values to solutions.

2.2.3 Diversity Prediction

Page sets a framework for a logic of diversity with which the wisdom of crowds can be explained. First, the intuitive notion of a prediction is defined as a *predictive model*:

Predictive Model An interpretation together with a prediction for each set or category created by the interpretation.

Mauboussin [19] explains the predictive models as "ways of inferring cause and effect". To separate predictions from heuristics and explain them, examples serve well:

The phrase "It looks like rain" is a prediction, the predictor *thinks* rain will come. "It's raining, *let's run for cover*" is a heuristic, a solution to the problem of rain.

The Diversity Prediction Theorem. Collective Error = Average Individual Error – Prediction Diversity. Prediction diversity is defined as the prediction variance, i.e. the average squared distance from the individual predictions to the collective prediction.

The Diversity Prediction Theorem says that being different is equally important as being good. Diversity and ability contribute equally to the collective predictive performance. Mauboussin confirms the theorem in class experiments in [19].

This yields the following:

The Crowd Beats the Average Law. Given any collection of diverse predictive models, the collective prediction is more accurate than the average individual predictions. Collective Prediction Error $\leq$ Average Individual Error.

The crowd is at least as good as the average, that is, it is equally good or better.

Wisdom of Crowds

James Surowiecki in [7] popularly calls diversity prediction and its anecdotes the "wisdom of crowds", which he defines as diversity helping collections of people make accurate predictions.

Surowiecki has three conditions for "a collection of people to make accurate predictions":

1. Diverse predictive models.

2. People are independent (no influence on one another).

3. Decentralized prediction process (no communication with one another).

The collection is just a collection of independent individuals. They are smart in that they have experience in the area.

In the famous ox-weighing contest observed by Galton in [20] in which buyers collectively—albeit independent of each other—reached a very accurate prediction of the weight of an ox, all buyers had experience of oxes. On that topic they were 'smart'.

The Difference Between Diversity Prediction and the Diversity Trumps Ability Theorem

According to the Diversity Trumps Ability Theorem, a randomly chosen or diverse collection of less individual performance, will perform better than a collection of the top individual performers, an expert collection. The members of the expert collection have individually found the best solutions—local optima—to a problem. Because the experts at work think in similar ways, their collective performance will not be much better than their individual ones.

The diverse collection on the other hand has more diverse perspectives and tries things in new ways so their collective performance outdoes the expert collection. Both collections build upon the results of other members within that collection, thus reaching better collective results.

The Diversity Prediction Theorem tells how a crowd may predict or guess equally or better than individual experts. The difference from the Diversity Trumps Ability Theorem lies in that the crowd's members independently—and without using the solutions of other members—reach a better conclusion than an average of the top-guessers.

2.2.4 Predictive Markets

Predictive markets[3] work like stock markets. Predictors 'bet' on statements about the future by buying or selling. A stock with a market value of \$1 is considered to have a 100% fulfillment confidence, the market thinks the statement will occur. As with stock markets, bubbles may and do occur, but in general they predict well. [18]

Those people who believe their predictive models to be accurate can place larger bets and those who are unsure can bet less, thus weighing the predictions.

Predictive markets create incentives for less confident people to stay out, and for confident people to bet more. It also incites to be diverse, as well as being accurate. Because winnings are split in a market, predictors want to be alone with the accurate prediction, i.e. diverse from the other predictions.

The incentives should not disturb the solver, or make her prefer a certain solution over another. A good incentive would be to promote solvers to acquire diverse heuristics and perspectives, to learn—which markets in effect do.

[3] Or information, decision, or futures markets.

The Policy Analysis Market

On July 28th 2003, the *Policy Analysis Market* (PAM) was disclosed and highlighted in the U.S. Congress reports, see [21], a *Defense Advanced Research Projects Agency* (DARPA) project that would, through a commodity-style market—i.e. a predictive market—, trade with forecasts of Middle Eastern political events such as possible terror attacks, coups, and political murders. The PAM would be open to traders both inside and outside of the U.S. government and intelligence community, allowing the public to trade with their views and beliefs of coming events.

The very next day, in [22], Pentagon is reported having dropped PAM, after senators having critiqued the project as "ridiculous and grotesque" on the previous day's press conference.

In [23], it has been showed that media coverage became more favourable to the idea of a futures market as it became better informed. Conclusively it seems that:

> *[...] the results suggests that while uninformed opinion disliked PAM, informed opinion favored it.*

Therefore, an intuitive approach to counter public resiliance would be through an informative campaign on introduction.

2.3 Applied Theory

The theory above forms the basis for what can be done in applying theory to practical applications in demining.

2.3.1 Possible Crowdsourcing Applications

Connecting to theory, crowdsourcing and diversity can be used in *mine detection*—contrast interpretation[4]—and *area reduction*.

Mine Detection or **Contrast Interpretation** takes advantage of the human eye. Solvers examine contrast data of mine-contaminated areas and try to find and spot mines and UXOs in that data.

Area Reduction Predictors guessing whether or not an area is contaminated with mines, on the basis of information databases.

The clickworker conclusions from section 2.1.1 suggest that crowdsourcing could be used to filter out and find more difficult areas, which in turn could be examined by real experts. This applies to both mine detection and area reduction.

Contrast data—processed measurement images—from LIDAR[5], infrared or other types of cameras may somewhat easily be interpreted by humans, whereas it is harder for a computer algorithm[6] as can be found in [1], [2], [24], and [25].

[4] Not to be mistaken with the framework definition of interpretation in 2.2.1.
[5] **Li**ght **D**etection **a**nd **R**anging.
[6] Rather, it is harder to develop such an algorithm.

This type of imagery would be suitable for crowdsourcing mine detection, marking areas or bounding boxes where mines and UXOs are situated.

Area reduction could be a very good task for crowdsourcing prediction of contaminated areas. However, diversity prediction aggregating information from the wisdom of the crowd requires available databases with area-specific information from a wide range of sources: mine maps, photographs, discussions with locals, history, etc, for the crowd to have any data to make predictions from.

2.3.2 Discussion of the Diversity Trumps Ability Conditions

In section 2.2.2, it is stated that four conditions are needed for the Diversity Trumps Ability Theorem to apply. In mine detection and area reduction, the conditions apply under the following assumptions.

1. **The Problem Is Difficult.** Fulfilled.

There is no trivial solution in neither mine detection nor area reduction, there are answers, but it is impossible to know in advance; thus, the condition is satisfied.

2. **The Calculus Condition.** The smart solvers condition. Fulfilled.

In mine detection, solvers need to have the ability to analyze images, have knowledge of mine shapes, and preferably an understanding of the principles of mine laying. Following an introduction to mine shapes they are as valid as NASA's clickworkers are in crater marking.

In area reduction, predictors need access to the underlying data to be able to make predictions from it. Laymen cannot be considered to have experience of area reduction, but after a good introduction such as [26] (see Appendix A.1 for excerpt) and demining instruction material like [27], they would have enough knowledge to be considered 'smart'.

3. **The Diversity Condition.** Fulfilled. Recall how this condition demands that there always needs to exist another problem solver who can find an improvement.

In the case of mine detection, this may pose a problem. Say there exists an image in which, owing to the quality, it is impossible to find a mine. For the sake of the purpose, that mine has to be found. Because of the impossibility of finding it owing to the quality, it still meets the diversity condition, because the global optimum is that it is impossible to detect an invisible mine. Also, the fact that another solver always may interpret the image differently, fulfills the condition.

This also applies to area reduction. All new input is welcome, thus there is always an improvement.

4. **Good-Sized Collections Drawn from Lots of Potential Problem Solvers.** Fulfilled.

All Internet users can be considered a good-sized collection with lots of potential problem solvers in mine detection as well as in area reduction.

However, for diversity to trump ability, the solvers need to work together, or at least build upon each other's solutions.

2.3.3 Creating an Experiment

As a first step in seeing how crowdsourcing may be an added tool to demining and to achieve the goals of the thesis, an experiment will show if it is possible to detect mines in images, see how well laymen perform compared to experts—personnel from demining institutions[7]—and how the Diversity Prediction Theorem can be used. This is done in chapter 3.

As a future step, solvers should be classified in terms of high-ability solvers and diverse solvers, and let the experiment be made such that the two types of solvers can be put in collections and take advantage of their collection's results, to see the impact of the Diversity Trumps Ability Theorem.

This first experiment focus only on the actual mine detection performance of manual labour in optical data in a crowdsourced environment and gives indications for the Diversity Trumps Ability Theorem and the Diversity Prediction Theorem. It is composed in the way that current crowdsourcing is being used today. Not all theoretical aspects discussed above are covered by the experiment.

A crowd of 'smart' agents—Mechanical Turk workers—represent laymen. Despite the workers' undocumented experience of mine detection[8], they are well-versed and experienced in image recognition.

The strict purpose of the experiment is to find out how well mines actually can be detected in general in a crowdsourced environment, and if experts can be beaten by laymen in mine detection—to see whether or not laymen can contribute to demining.

In the experimental environment, all that is considered is that there are mines in the area, making as much as possible of the experiment to be about the actual image recognition part of mine detection.

Narrowing it down to only image recognition and contrast interpretation, all conditions are valid:

1. The problem is difficult.
2. The solvers are smart.
3. There is no 'right or wrong',[9] the solutions may always be improved.
4. The vast collection of Internet users is enough.

Going with the thoughts and successful experimental results of Hong and Page, the 'best' problem solvers—experienced deminers—would tend to think similarly about a problem, which as a collection is not very effective, therefore, a random collection of intelligent agents may perform better by bringing new perspectives and heuristics.

In the experiment, the experienced personnel collection is compared to that of the laymen workers, to indicate on how the Diversity Trumps Ability Theorem

[7] FOI and SWEDEC.

[8] That is known of. However, in a crowd, everything is possible.

[9] Of course, either there is or there is not a land mine in that specific area, but no one knows the correct answer until the mine as been detected in real life.

would do. Despite the experts not having means to cooperate, it is interesting to see how they fare compared to laymen, to see if there is a difference.

2.3.4 Implications

Letting the public take part in operations like demining has three main implications: *public opinion*, *cyberterrorism*, and *governmental resilience*.

Public Opinion and Credibility

Public opinion and public credibility may affect mine detection operations based on crowdsourced data. The public could be offended by the idea of having your next-door neighbour doing what has long been considered a highly skilled task done by professionals.

Just as in the study of PAM, though, an informed public may very well be positively inclined.

A land mine predictive market for area reduction may also be very controversial. In the long run, the market is 'gambling' on peoples lives, as it very possibly may be interpreted as. On the other hand, if it has documented good outcomes, it might be accepted by the public.

Cyberterrorism

A coordinated cyberterrorism hacker attack of automated or manual agents could quite easily organise an attack on mine classification, classifying hostile areas as safe.

The company *Subvert and Profit*, found at [28], allegedly, successfully uses crowdsourcing to raise site ranks on sites like Digg, StumbleUpon and YouTube[10], which themselves depend on user production. It is legal, however, Subvert and Profit users violate the Terms of Use of their accounts on the sites on which they are raising ranks.

One of the challenges of PAM was, according to DARPA [29], if "futures markets [can] be manipulated by adversaries". A reasonable question that was never put up to the test.

Governmental Resilience

According to Ralf Andrén of SWEDEC in [30], from a governmental side—i.e. the police—there is a built-in resilience against publicly sharing information:

> *From an information perspective, there are several information systems as the EOD IS. From a Swedish side and that of the manufacturers, it is preferable to keep such a system open to the ones that use it, to the ones that stand the 'question of responsibility'. The idea is that if you are well-enough trained and educated, then you are in the demining system—then you are not a bomb-maker.*

[10] `http://digg.com/, http://www.stumbleupon.com/, http://www.youtube.com/`

Andrén says that "the police is very negative towards having images [of land mines and UXO] floating around", adding:

> *'Bomb-making' in the loop, so to speak. Eventually, it is the police that will have to take care of [the bombs].*

Despite that, on request to get access to land mine databases for preparing an experiment, he added that the information that e.g. SWEDEC could supply with is "the same information that anyone would find on the Internet by searching for the different mine types".

A contingent crowdsourced mine detection system would, thus, have a lot of red tape to cross, before being realized.

Chapter 3

Experiment

An experiment was conducted using the HIT-Builder and the Amazon Mechanical Turk services. Laymen and expert participants were to watch and classify snippets of optical data of different areas and find contingent mines.

An erroneous qualification question was discovered in and the experiment was cancelled, read more in section 3.5.5, only to be resumed and finished later on.

In the following order the experiment is described: *purpose*, *participants*, *data set*, *implementation*, and *execution*.

3.1 Purpose

To hint on crowdsourcing's possibility in this area, the experiment wants to find out to what means it could be used. The scope of the thesis and time-frame-wise, it was considered suitable to create an experiment similar to those run at an existing service. For future work (see section 5.5) there is a wide array of options to be tried, such as the Peekaboom approach.

The experiment's purpose was aligned with that of the thesis, and the thesis purpose questions were rephrased to the following:

- Can humans detect land mines and UXO in optical data?
- Is an expert collection better than a laymen collection (i.e. the crowd) in examining that kind of imagery?

This was to shed light on some indications of diversity's ability in prediction.

3.2 Participants

In the search for Steve Fossett on the AMT there were five submissions used for each HIT, Hong and Page used 1,000 (mathematical) problem solvers in [16], and in [19] Mauboussin used 73 students for the Jelly Bean Experiment confirming the

diversity prediction theorem, which was compared to Galton's 19 guessers in an ox-weighing contest in [20].

No results were found that could reason for a specific number of experiment participants. Based on the financial situation (for details, see section 3.5.3), the loose ground of the experiment with theory rather than crowdsourcing culture, and the lack of reasoning against it, it was decided to use 100 participants of which 10 would be experts; by adding real experts it was possible to see if there existed a performance difference, and indications on whether diversity beat ability or not. Thus, two collections were predefined: a laymen and an expert collection.

3.3 Data Set

The experiment data was taken with Specim ImSpec hyperspectral camera at the SWEDEC test site in Eksjö, Sweden. Three surfaces had been measured at a road on the test site at different times a day:

- Right-hand side of the road bordered by moderately forested terrain
- The road
- Left-hand side of the road bordering with a grass field

The available data set, the *measurement data*, had minor quirks that had to be adjusted due to position mismatch between different measurements. Bear in mind, the measurement data was not at all adapted for an experiment of this sort. In total, 16 different measurements were used from three surfaces.

3.3.1 Hyperspectral Measurement Data

The measurement data was sampled at 150 wavelength bands, stretching from 253.47 nm to 1114.87 nm in intervals of ∼5.78 nm. The spatial resolution was 875 × 1600 pixels. To access the hyperspectral measurement data, the *Hyperspectral Imaging Toolbox* (HSI Toolbox), described in [24], was used in MATLAB. The HSI Toolbox is a collection of tools developed internally in various FOI projects.

Measurement Data Naming Convention

To ease data handling, a measurement data naming convention was used. Later this would be very helpful to name the data cells used in the experiment. A measurement data file was named: `eksjo_Location_Time` (`_Extra`). E.g. location 1 (right-hand surface of the road) at lunch time was named `eksjo_1_2`.

All measurements were made with mines placed on the surfaces, except for the *Background/0* measurements.

As seen in table 3.1, suffixes were added when several measurements were made at the same time or of the same location.

Table 3.1. Surfaces and measurements.

	Background (0)	Morning (1)	Noon (2)	Evening (3)
Surface 1	1_0	1_1	1_2	1_3
Surface 2	2_0	2a_1	2a_2	2a_3
		2b_1	2b_2	
			2c_2	
Surface 3	3_0	3_1	3_2	3_3
			3_0_2	

RGB Images

Three spectral ranges were extracted from the measurement data: *visible light* (RGB), *near infrared* (NIR), and *mid-infrared* (MIR), by combining three wavelengths into new *RBG images*.

Table 3.2. Spectral ranges used for RGB images.

	Red	Green	Blue
RGB	663 nm	560 nm	454 nm
NIR	768 nm	821 nm	899 nm
MIR	899 nm	1009 nm	1114 nm

Mismatch Problem in Measurement Data

Measurements of the same area made at different times of the day, differed in position in most images, see figure 3.1.

Figure 3.1. Two measurements from location 2 at different times. *From left:* Background measurement without mines (2_0), noon measurement with mines (2c_2), and the difference image of the two.

A rough but quick method to correct this was used: the extracted RGB images were spatially transformed in *Adobe Photoshop* using the *Warp Transform* function to match better. Key points such as landmarks, mines, and measurement equipment were used for reference in matching.

In an automated environment there should be strict requirements for the source data. If that cannot be fulfilled, mismatching images can be automatically stitched with other tools, such as the free open source software package *Panorama Tools*[1].

3.3.2 Data Cells

The measurement data was divided into 13×13 *data cells*. The cells were made to contain the corresponding area of the RGB images—i.e. visible light, near infrared, and mid-infrared. These data cells were to become the images to be examined in the experiment.

Image Processing

To improve detectability, three image processing operations were applied to each cell RGB image: RGB histogram equalization (function `histeq` in MATLAB), L*a*b* luminance channel histogram equalization, and auto contrast using contrast stretching.[2] A margin was added around the interested cell area to enhance understanding and orientation.

Cell Images

Finally, *cell images* (see figure 3.2) were made by composing the contents of the processed image data cells—i.e., a cell image contains nine images of the same area from the three spectral ranges with the three kinds of image processing applied to them. These were the final cell images to be examined in the experiment.

3.4 Implementation

The AMT was used in the experiment.

3.4.1 Rewards and Cost

For most HITs, the worker is rewarded. Rewards span in the range of USD \$0 to USD \$16 per HIT, most common being a reward about USD \$0.02 per HIT. By community practice, an hourly rate of USD \$2.50 for work in a HIT group is considered fair [31].

By the AMT, the Requester is charged a minimum of USD \$0.005 per HIT, or 10% of the reward.

[1] `http://www.panotools.org/`

[2] See appendices B.1.1 and B.1.2 for code.

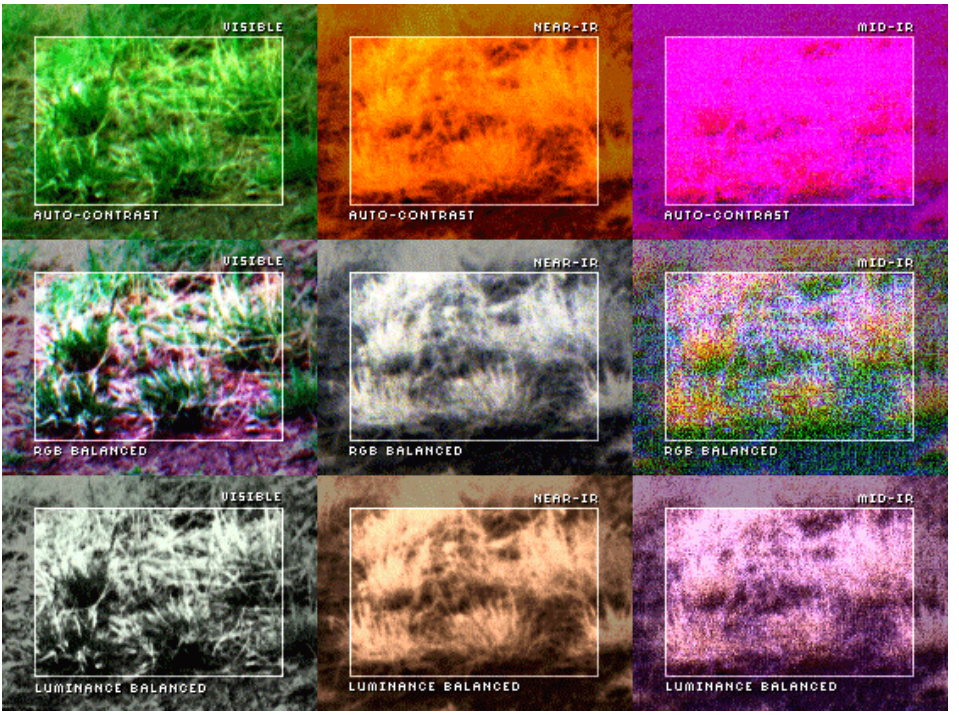


Figure 3.2. Cell image layout. Cell images in three spectral ranges and with the following image processing applied: Auto contrast using contrast stretching (`AUTO-CONTRAST`), RGB histogram equalization (`RGB BALANCED`), and L*a*b* luminance channel histogram equalization (`LUMINANCE BALANCED`). For illustratory reasons, the margin has been faded.

3.4.2 HIT-Builder

The AMT is still in beta and allows only U.S. citizens to become requesters. The *HIT-Builder* service, found at [32], allows non-U.S. citizens to setup a requester account on its service, which further sets it up on the AMT.

David Pfeiffer of HIT-Builder was very helpful in the development of the experiment HITs. Due to the fact that this is a scientific experiment, Pfeiffer did not charge any extra fees other than what the original AMT service charges, on the condition that he could mention the experiment in marketing purposes.

HIT-Builder has an easy-to-use web interface for uploading data sheets, creating HITs and HIT groups, and exporting and handling HIT results. The service was chosen to save time and effort in finding a U.S. citizen to sponsor with her U.S. bank account, having to write a form using the AMT API, and thanks to a good HIT result export feature.

Unfortunately, the HIT-Builder server could sometimes be very slow, and bugs were experienced when using other browsers than *Microsoft Internet Explorer*. The limited interface for editing HIT questions did complicate the development of the HIT questions and qualification.

When posting HITs, not all HITs were posted due to an unknown error. Also, the posting occupied the browser for more than one hour's time! See section 3.5.1 for more on this.

In hindsight, the time difference between using the HIT-Builder and programming towards the AMT API may have been insignificant. At this point it is uncertain if better results could have been derived in a similar time frame.

3.4.3 Reasons for Choosing the AMT/HIT-Builder

An alternative approach would have been to write a new image-recognition interface, as in the NASA Clickworkers project. Considering the time frame of the thesis, there were several key factors that made a combination of the Mechanical Turk and HIT-Builder beneficial:

- An environment, i.e. the AMT service, in which there is an active user base which also has experience of similar tasks. The high activity of the workers gave 'guaranteed' high response rates, and saved time in finding them. These were a suitable crowd or collection of laymen.

- Putting the experiment on a random server would have demanded advertising and attracting users. Further, there is an unofficial worker forum called *Turker Nation* [33], an active community in which the members discuss and comment on HITs, requesters and related topics. Turker Nation was used for worker feedback in preparation of and during the time of the experiment.

- The working reward and payment procedure included in the AMT.

- HIT-Builder's practical interface for handling operations such as uploading of data sheets, managing users, and exporting results.

3.4.4 Experiment HITs

The *experiment HITs* consisted of a cell image with the three band images processed in three ways, and four questions. The worker was supposed to watch the cell image for mines, UXO or odd objects, count the objects, and classify the area. It was also possible to flag the HIT as difficult and comment it. See figure 3.3 for a screenshot from the Mechanical Turk implementation.

Experiment HIT Questions

Each experiment HIT had four questions; two of them, objects and classification—marked with asterisks*—required answers to submit the HIT.

Objects* The first question *Objects** was required, and answered how many objects that where present in the cell area. For cells with four or more, the 4+ option was suggested.

Classification* Also the second question *Classification** was required, and asked the worker to classify the HIT area.

Difficult If the HIT was not clearly classifiable, there were unidentifiable objects present, or the HIT for some other reason contained difficulties, the worker could flag the HIT as *difficult*.

Comments The worker was urged to *comment* on HIT's flagged as difficult, or if for some reason he wanted to give information on how he thought with her classification.

Classification

Each experiment HIT could be classified as *field*, *grove*, *road*, *above horizon*, or *undistinguishable or unknown*.

Field Used for open fields.

Grove Used for forested terrain and areas that are not open fields.

Road Used for cells where a road is predominant.

Above Horizon Used for cells in which it was clear that they showed sky or other aerial objects.

Undistinguishable or Unknown Used for undistinguishable or unknown areas.

The classifications were chosen on the basis of the available measurement data and their content.

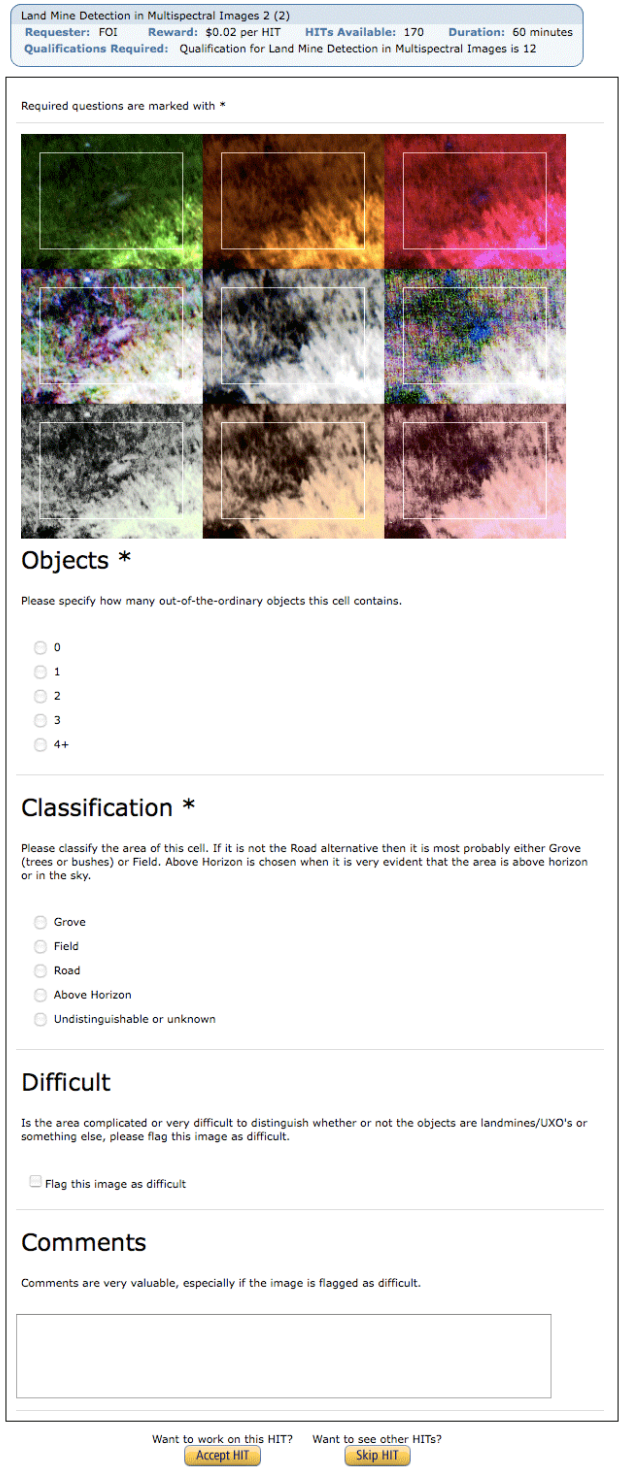


Figure 3.3. Experiment HIT screen.

Submission and Approval of HITs

Approval of a submitted HIT automatically pays out the HIT reward to the worker. AMT holds a possibility to manually, or automatically approve submitted HITs. Manual approval would have been very arduous, and unnecessary, because one of the purposes of the experiment was to see whether a manual optical approach was possible or not in detecting land mines and UXOs.

An automatic approval of the experiment HITs was set to 5 days unless it was explicitly rejected before the end of that time limit.

Time

On average, it was estimated that submitting a cell required 20–30 seconds of browsing the image data, clicking the radio buttons, and occasionally adding a comment.

Number of HITs

There were supposed to be 100 participants of which 10 were decided to be experts, personnel from FOI and SWEDEC. The number of HITs was based on the USD $384 available on the HIT-Builder account (see section 3.5.3 for expenses in detail), and these 100 participants.

The available funds did not cover use of all cells from all surfaces. To increase detectability, it was considered the best approach to use as many sources of information—i.e., the `Time` variable from the measurement data—as possible for each cell area, and rather reduce the number of areas.

Surfaces 1, 2, and 3 have 4, 7, and 5 measurements respectively.

Table 3.3. HIT amount calculation.

Surfaces	$S = S_1 + S_2 + S_3 = 4 + 7 + 5 = 16$
Available Money	$M = \$384$
Workers/HIT	$W = 100$
Cost/HIT	$C_1 = \$0.02 + \text{Fee}_{AMT} = \$0.022$
	$C_2 = \text{Minimum Fee}_{AMT} = \$0.005$
Number of HITs	$N = \dfrac{M}{(0.9 * C_1 + 0.1 * C_2) * W} \approx 189.16$
HITs per surface	$\lfloor N \rfloor / S = 11.81 \approx 11$

As seen in table 3.3, 11 HITs per surface were be used. A possibility was to add $\lfloor N \rfloor - 11 * 16 = 13$ HITs to the entire test. However, this was not considered

necessary to the results.

HIT Selection

From each surface, 11 cells were chosen, which can be seen in figure 3.4. The criteria was that they should include: mines, grove, field, road, or difficult objects.

3.4.5 Qualification Test

Each worker was obliged to complete a qualification test (attached in appendix B.2). The qualification test consisted of an introductory text with information about the experiment, examples of land mines, and twelve qualification questions. All HITs were linked to this qualification, making it impossible to perform HITs without having passed the test; to pass, the worker had to get all twelve questions right, receiving a value of 12, which was the qualifiying value for the HIT group.

The obligatory qualification made it possible to consider the solvers, i.e. workers, as smart, as discussed in section 2.2.2.

Purpose of the Qualification Test

The qualification test had three main purposes: to educate workers about demining and show typical land mines, prevent spam-like HIT submissions, and instruct workers on how to classify HITs.

Experiment Background

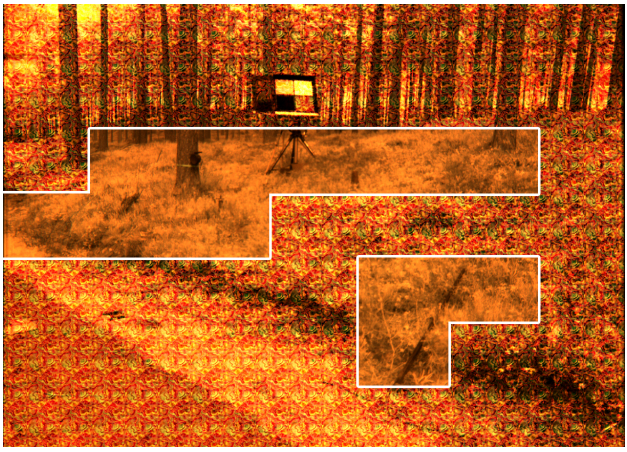
The first part introduced the worker to what the experience was about, how it was part of an M.Sc. thesis project at a military research institution, and asking the workers to pay attention to the fact that the results of the thesis can be used both for tactical and humanitarian demining purposes. It also laid out the purpose of the experiment and the questions to be answered.

Images and Site of Measurements

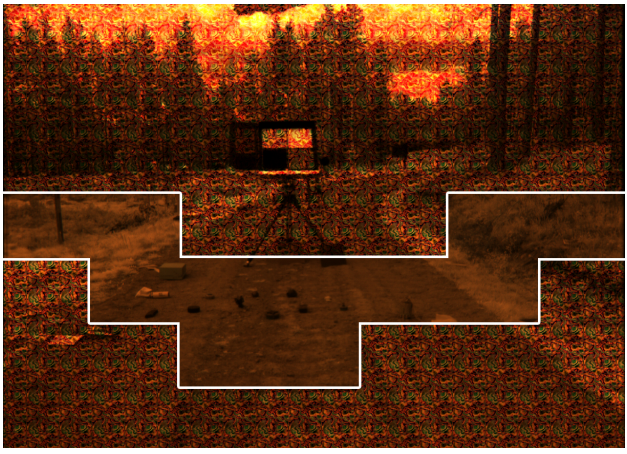
Information was given on the content and type of images to be examined, and where the measurements had been made. Also a panorama of three RGB images were shown to give a sense of orientation.

Cells

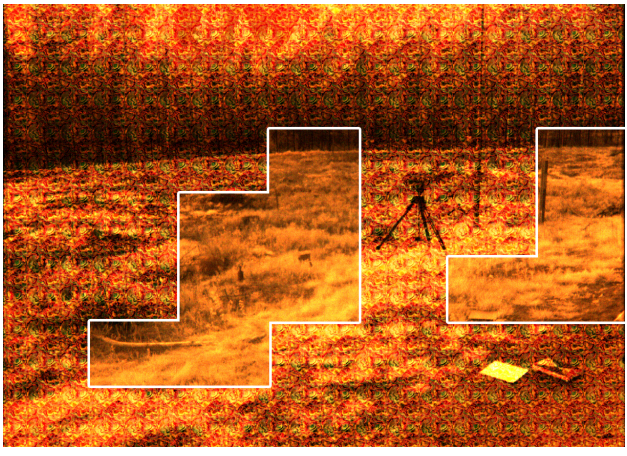
The content of each cell image was explained as well from what bands the images were extracted, as to what processing methods had been applied. Also, the example cell image shown previously (figure 3.2) was shown, explaining which area of the cell had what methods applied to it, and from which spectral band it had been derived.



(a) Surface 1.



(b) Surface 2.



(c) Surface 3.

Figure 3.4. Chosen cells from the three surfaces.

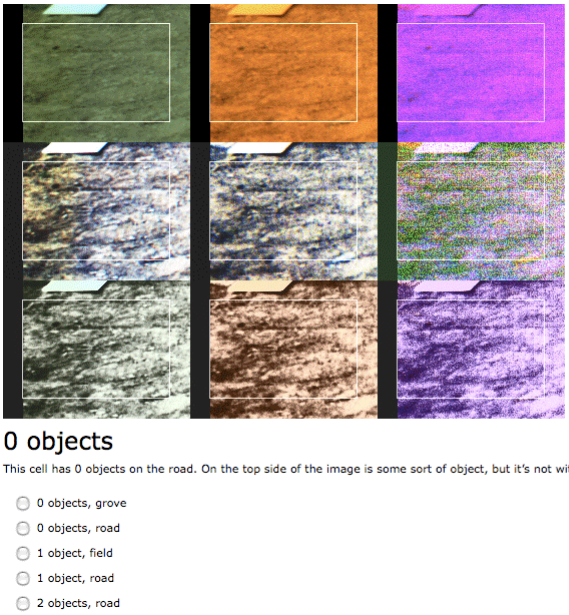


Figure 3.5. Question 2 from the qualification test.

Land Mines

Example images of 21 different AP and AT mines were shown to give the workers a sense of the general shape and form of land mines. The assortment was made from the selection of mines in [2].

Qualification Instruction

The workers were instructed that the twelve following qualification questions were taken from the real data set used in the HITs, and how to proceed in answering them. A mistake was made in not having experienced personnel examine these qualification questions before being put into the instruction—a mistake that would lead to cancelling the experiment (discussed in section 3.5.5). Lastly, they were thanked for their participation.

Qualification Questions

Each qualification question examplified a typical or difficult case, e.g. as in figure 3.5 where there is a visible object, but is located in the outer margins of the HIT, thus not being part of the HIT area.

Five radio button answers, of which one was correct, were given with descriptions on how to classify the HIT. By giving a description in text on how to classify the HIT, with answers requiring reading that description, it was possible to effi-

ciently prevent spam users and at the same time educate and instruct the worker how to approach such cases.

3.5 Execution

The experiment was conducted in the period between 2008-05-02–2008-05-16 with laymen, and 2008-06-02–2008-06-19 and 2008-06-23–2008-07-06 with experts, being cancelled and resumed, see section 3.5.5. 184 laymen AMT workers participated, and four of FOI personnel.

3.5.1 HIT Posting Problems

Using HIT-Builder to post HITs was a slow process. Posting $176 * 90 = 15840$ laymen HITs seemed to be done in a purely sequential mode requiring the client's browser. The browser was occupied for more than an hour's time, and it was not known what was actually happening.

Many hours later returning to the browser, the HIT-Builder system concluded several errors and only 170 out of 176 HITs were posted, see appendix B.3. It seems as the connection was lost to the AMT server, and that probably affected the consequent HITs. Errors say there are not available funds, which there were, at least at the time of the posting.

3.5.2 Experiment Participants

Two collections of participants were needed: laymen and experts. Laymen were AMT workers, and experts scientists and personnel of FOI and SWEDEC.

Laymen Reward

After HIT approval, a reward was given of USD $0.02 per HIT to a layman, giving an approximate hourly rate of USD $2.40 per HIT if a HIT took 30 seconds on average; in the AMT environment considered a fair reward.

Demographics

It was not possible to find any demographics of the laymen participants performing the experiment.

Experts

No reward was given to the expert collection, however the hours of the FOI personnel put in to completing the experiment, were charged on the MOMS project.

A call for participation composed of an introduction to the thesis and the experiment, followed by instructions on how to login with one of the 17 predefined accounts, complete the qualification, and start submitting HITs.

The call for participation was e-mailed by Jörgen Ahlberg to 10 FOI staff members, and by David Andersson to 4 members of SWEDEC and to 3 members of the *Croatian Mine Action Centre* (CROMAC).

3.5.3 Expenses

Explicit expenses of the experiment were: PayPal transaction fee to transfer money to the HIT-Builder account, rewards, and AMT fees. Implicit were the time cost for personnel to participate in the experiment. HIT-Builder did not charge the usual fee thanks to an agreement (appendix **??**) that was struck between David Andersson and David Pfeiffer .

PayPal

Two PayPal transactions were made to DPA Software's PayPal account. The transfer to the account incurred a 3.975% fee, in total USD $16.40 (of $400 and $10). Due to a misunderstanding of the PayPal account transfer fee, the extra USD $10 had to be transfered to cover for the expert HITs, which could not be posted because of an AMT assignment liability policy.

HIT-Builder and AMT Rewards and Fees

Usually, HIT-Builder charges the requester a fee on larger runs of HITs like this experiment, but an agreement was struck with Dave Pfeiffer that in exchange for the right to co-release a press release about the project (with customer approval rights before release), DPA Software was to provide the HIT-Builder account for no charge (see appendix **??**).

Money once transfered to the HIT-Builder account were non-refundable. Because both laymen and expert HITs were running at the same time, it was not possible to take advantage of contingent inaccessible funds to use for other HITs.

There should have been enough money in the HIT-Builder account to post the USD $\$0.005 * 176 * 10 \approx \$8.80$ worth of expert HITs[3]. For some reason, though, the account balance was only USD $1.60 instead of the expected $47.40[4]—no HITs could be posted due to a lack of available funds on the account. An extra USD $10 had to be transfered.

$$\begin{aligned}\text{Available Credits} &= \text{Initial Credits} - \text{Liable Credits} \\ &= \$384.00 - \left(170\ \text{HITs}_{posted} * 90\ \text{Workers} * \$0.022\ \text{Fee}_{AMT}\right) \\ &= \$384.00 - \$336.60 \\ &= \$47.40\end{aligned}$$

In e-mails [34], Pfeiffer never did respond properly as to why there were no credits for yet another batch of HITs, only to conclude the following:

[3] These were later extended to $10 + 4 + 3 = 17$ experts.

[4] As seen in the equation, only 170 out of 176 were actually posted because of the error discussed in section 3.5.1.

> *The Current number of assignments pending approval*[5] *is likely the problem. Maybe one of the HIT posting error setup HITs that have not been approved or rejected.*

There was an early run of HITs that contained mistakes that were discovered after some hours. Some HITs had already been approved by the time of cancellation and in cancelling all the HITs, something probably occurred with the counting mechanism of the HIT-Builder service.

Personnel

The MOMS project was charged for each hour spent on the experiment. It was approximated that each expert were to spend 2-3 hours on the experiment.

Total Costs

In table 3.4, the three HIT batches' costs are shown.

Table 3.4. Total HIT costs.

	HITs	Cost	Subtotal
Laymen	7156	$0.022	$157.4
Batch 1 Expert HITs	280	$0.005	$1.4
Batch 2 Expert HITs	0	$0.0	$0.0
		Total Sum	$158.8

As specified in the contract, all transfered funds are non-refundable. Thus, the experiment cost USD $410.00, despite an account balance of USD $196.00 at the end of the final expert HIT batch.

3.5.4 Feedback

Laymen 184 AMT workers participated, of which 6.9% (11) completed all tasks.

Experts 4 unique experts did some parts of the experiment, and none completed all tasks.

More experts would have been wished for, however the lesson was learned that time is the issue in getting people to voluntarily do experiments, not compensation.

After cancellation, the experiment was halted for some time. During that time, the SWEDEC personnel were not able to perform the experiment; once it was online again, all personnel were prevented to do the test due to different circumstances. CROMAC did not respond to the participation call.

[5] 1206 at the time.

Comments

Comments were gathered from e-mails to the requester account, from the HITs themselves, and from the Turker Nation forum.

3.5.5 Cancellation of Experiment

Unfortunately, in a very late stage of the experiment execution on 2008-06-19, an erroneous qualification question was discovered by SWEDEC personnel.

The first qualification HIT question included a mine in the processed cell image, despite the qualification text telling the participant to classify the HIT as road without any objects in it! The question wrongly instructed participants to classify mine HITs as non-mine HITs.

Not only did it wrongly instruct how to classify the question, nothing else but choosing the erroneous classification answer was possible.

For this reason, the results could not be considered reliant.

At this stage had already some 184 AMT workers participated and 4 from FOI and SWEDEC. In the scope of the thesis, there was no more time for yet another experiment.

The conclusion was made to continue the experiment and do statistical analysis, with reservation to only let those conclusions be hints of crowdsourcing and diversity's use in demining. Unfortunately, no one else participated.

Chapter 4

Analysis

The analysis was to be made in two parts: the *descriptive statistics* describing and summarizing the data, and the *inferential statistics* generalizing and making predictions about the future.

Thanks to the HIT-Builder and AMT services, the data was well-prepared, accurate and ready for analysis without major data preparation needed. The terms *objects*, *classification*, and *difficulty* are used in lieu of the answers of counted objects, classification, and difficulty respectively.

Due to the cancellation of the experiment and what gave rise to it, there is no real data to be analyzed, albeit the considered methods are discussed in this chapter and what was acquired is analyzed only to give hints—and nothing but hints—to the result.

4.1 Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. The descriptive statistics are simply describing what the data shows.

4.1.1 Central Tendency

Both the count of objects and classification answers are described using a histogram frequency distribution.

The count-of-objects central position is measured using the *mean*. The classification answers (field, grove, road, above horizon, or undistinguishable or unknown) are of a nominal data type for which the concept of mean is meaningless (no pun intended), thus the *mode* is the correct measure.

4.1.2 Spread

As spread measure, the objects' variance is used.

4.1.3 Observations

Comments

Comments could be considered a sign of seriousness. Of 184 laymen workers, 29.9% (55) did not comment on even a single HIT. 26.1% (48) did comment on ten or more HITs.

The comments were often highly informative as the random sample in the list suggests:

- *(0,25) something*
- *2nd obj outside cell, bottom l corner*
- *3 objects clearly are manmade (sign in background, pole in forground, object on lower left) but field appears to be filled with stumps making the possibility of contrasting other manmade objects very difficult*
- *A tulip at (80,70) and half a menorah at (30,30) or am I seeing things that aren't there?*
- *Could be Measuring Equipment*
- *Definitely two objects in the lower right corner, might be another one a little bit to the right from the center, and possibly one in the bushes.*
- *annular object at (20,35) looks like a washer*
- *i do not see other obj's , but i would be concerned about the parallel line shadows - not something you normally see in nature*
- *i think it's a dirt road, and there is a second object that is either a mine or a rock*
- *might just be a branch, but appears to be a long metal tube*
- *not clear*
- *not very confident on this one. images are very grainy*
- *very busy, def 4 w/stump? on R*
- *two mines or explosive objects - one in top left, one in top right.*

Valuable comments were also given as to how the experiment could be improved. A random sample of quoted comments and extracted suggestions:

- *In a sharp situation, the assessment be totally totally different because of the consequences in missing an object.*

- Difficult to see what the objects or the HIT are. A call for better image resolution.
- *Too small images on screen.*
- Rearrange the HIT layout so everything fits in one screen to prevent scrolling.
- Keyboard shortcuts for faster tagging.
- No size reference in the images.
- *It would also be helpful if there was a help page so that Mturk workers could go to the help section for any clarifications, examples.*[1]
- *You know, scale also matters. There's some kind of round object, probably just a pebble, and if I knew the scale of the photo, for example if I knew that all photos were taken with the same lens and not some wide angle and some zoomed in, then I would definitely say it's just a pebble. But if the shot was taken from, say 25 meters or a wider angle lens, then maybe it's an M1 AP DV 59 or an M14. Still, I think it' just a pebble. You might want to revise your instructions though, to say something about the lens/distance the photos were taken.*

Of four participating experts only two commented with one and two comments respectively.

4.1.4 Detection and False Alarm Rates

The probabilities—rather, the frequencies—for a given outcome where classified as seen in table 4.1. Remember, it is only known which areas *might* contain mines, and which do not contain any at all.

Table 4.1. Description of frequency rates.

$$p_{ij} = P(X = i|Y = j) \text{ for } i, j \in 0, 1$$

	p_{00}	*No mine found where non-existent*
Miss Rate	p_{01}	*No mine found despite possibly existent*
False Alarm Rate	p_{10}	*Mine found despite non-existent*
Detection Rate	p_{11}	*Mine found where possibly existent*

The results are seen in table 4.2.

[1] A copy of the qualification test was uploaded and noticed to the worker in question and posted on the Turker Nation forum.

Table 4.2. Laymen and expert frequency rates.

		Laymen	**Experts**
	p_{00}	0.3442	0.2817
Miss Rate	p_{01}	0.3079	0.2823
False Alarm Rate	p_{10}	0.6558	0.7183
Detection Rate	p_{11}	0.6921	0.7177

4.2 Inferential Statistics

Inferential statistics try to reach conclusions beyond the immediate data alone. For example, they may be used to make judgements of the probability that an observed difference between collections is a dependable one, or that one might have happened by chance in a study.

This section would use the statistics of *Pearson's correlation test* and the *two-tailed t-test* to make inferences from the experiment data to more general condictions—if possible. They are linked to the two questions posed in the experiment purpose[2].

4.2.1 Two-Tailed Correlation Tests

The correlation tests show how strong the correlation of two variables is—a linear correlation computed as in seen in table 4.3.

Using a two-tailed t-test, it is tested whether the correlation is statistically significant, that is, whether it can be concluded that the correlation is not chance. The MATLAB functions `corrcoef` and `ttest2` from the Statistics Toolbox were used.

Table 4.3. Correlation calculation with Pearson's product moment coefficient (PPMC)

Number of pairs of scores	N
First scores	x
Second scores	y

Correlation (PPMC)
$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

[2] Can humans detect land mines and UXO in optical data? Is an expert collection better than a laymen collection (i.e. the crowd) in examining that kind of imagery?

The two main questions posed in the experiment purpose are: *Can humans detect land mines and UXO in optical data?* and *Is an expert collection better than a laymen collection (i.e. the crowd) in examining that kind of imagery?*

Level of Significance

The rules-of-thumb level of significance of 5% is not enough. In a HIT batch with 30 submissions, that would be one and a half submission which is considered too much from such few samples. Therefore, a lower level of 1% is chosen.

Correlation Is Not Causation

If there is a significant correlation, it is vital to remember that it does not imply causation—causality. One of five situations can then be true:

1. There is a direct cause and effect relationship
2. There is a reverse cause and effect relationship
3. The relationship may be caused by a third variable
4. The relationship may be caused by complex interactions of several variables
5. The relationship may be coincidental

The correlation does not say anything about what is happening, only that the observations correlate.

4.2.2 Correlation

The data set lacks known positions of mines and UXOs. However, there is data from mine-laid areas and clear areas,[3] it is known which areas that *do not* have mines or UXOs.

Thus, it is possible to know in which HITs mines have been 'wrongly' detected. Though, it is not possible to know if mine-tagged HITs truly do have mines in them, because it is only known that the entire area contains mines, but not whether that specific HIT does.

Despite the incoherence in position discussed in section 3.3.1, it is interesting to see how well participants actually could detect mines, and answer to the first question in the experiment purpose.

The user entries of objects are denoted x (there is or is not a mine), and y denotes if the corresponding cell may contain a mine or not. The hypothesis testing is done by testing the null hypothesis $H_0 : x$ and y are *uncorrelated*, against $H_1 : x$ and y are *correlated*.

The correlations and resulting hypotheses using a two-tailed t-test at a significance level of 1% are shown in table 4.4.

[3] Background: `0`, and mine-laid: `1`, `2`, `3`

Table 4.4. Correlations

	Laymen	**Experts**
Correlation r	0.033	-0.0005863
Hypothesis, 1%	H_1	H_0

4.2.3 Mean Value Test

Denote x as user entries from HITs without mine and y as user entries from HITs with contingent mines. Testing $H_0 : x$ and y have the same mean, against $H_1 : x$ and y do not have the same mean using a t-test at a significance level of 1% yields table 4.5.

Table 4.5. Mean value test

	Laymen	**Experts**
Hypothesis, 1%	H_1	H_1

Chapter 5

Conclusion

As no valid experiment was completed, there are no real results to conclude about. The chapter discusses the statistical analysis, the failure of the experiment and what led to the erroneous qualification question, lessons learned throughout the work of the thesis, and future work in the area.

5.1 Statistical Analysis

It is interesting to notice the slightly higher frequency of both false alarm rates p_{10} and detection rates p_{11} of the experts relative to the laymen. Also, the expert miss rate p_{01} was lower than that of laymen. This could be a sign of an increased carefulness among the experts.

The only correlation that was significant was that of the expert answers on the object classification. However, the resulting $r = -0.0005863$ is practically zero, which means no correlation.

5.2 Experiment Failure

The experiment had to be cancelled because of an erroneous qualification question. This could be attributed to the fact that no method was used, which led to no evaluation being set up, but essentially it was the fault of the writer of the thesis.

5.2.1 Method—Or the Lack of

The whole thesis did *not* follow any method. Therefore, no techniques were followed or specific requirements met.

Hypotheses were proposed as experiment purposes as to what to be obtained by carrying out the experiment, and they *were* repeatable to dependably predict future results and set up as to confirm those purposes. Also, the design of the experiment was made to show the statistical significance of dependancies and correlations—with consideration to the data set at hand.

However, there was no scientific methodology that would protect from what eventually did occur, an unverified experiment. There should have been some sort of verification procedure which would have *professionally* evaluated the actual qualification.

5.2.2 Erroneous Qualification Questions

During the development of the qualification test, the focus was to find an average amount of questions which would be sufficient as to educate the participants on how to submit the HITs.

The example HITs used for the qualification were chosen at random without reasoning. To have a professional evaluation of the qualification questions was lost in the process of choosing suitable questions. That said, all qualification questions were not assessed by anyone. The error is attributed solely to the writer of the thesis.

Once the qualification test was completed in early April, experienced personnel of FOI and SWEDEC should have been brought in to verify the chosen HITs and the corresponding classifications.

5.3 Lessons Learned

HIT-Builder vs AMT API The HIT-Builder service suffers from problems, especially in browsers other than Microsoft Internet Explorer. To increase adaptability to the experiment, consider writing directly to the API.

Time Time is a very essential part of work, and for people it is of low priority to voluntarily take time to perform thesis experiment.

Methodology Use a scientific methodology to aid and give structure to exploring new areas of science.

5.4 Discussion

The section discusses discoveries made during the work of the thesis.

5.4.1 Worker Verification

It is not possible to know what drives the worker to complete the HITs. If it were for just the reward, it is possible that HITs she classes HITs on random. However, the high rate and quality of the comments suggest otherwise.

Verification of each worker by checking random submission samples before approving submissions could prevent unearnest workers.

5.5 Future Work

Apart from the improvements and lessons learned above together with a successful experiment, there is yet a lot more interesting to be made in the novel field of crowdsourcing in demining.

5.5.1 Improvements for Future Experiment

For future experiments similar to this, the following improvements are recommended:

Qualification Evaluation Have experienced and professional personnel choose qualification questions.

Interface Improve the interface to minimize scrolling and add keyboard shortcuts for quicker submission procedure.

Implementation Use a Flash or Java based implementation where the user can pinpoint positions for enhanced accuracy of object positions. Consider programming an independent implementation. Using the AMT service alone could be a way to handle transactions costs.

Unrewarded Experiment NASA Clickworkers were not rewarded, try to find workers either from the AMT or a stand-alone site to perform experiment without reward.

Positioned Measurements Have all measurement data come from a known position in space in relation to the surface measured, and provide a size reference object, e.g. the Coke can.

5.5.2 Diversity Trumps Ability Experiment

Having solvers cooperate within their collection would pave the way for an experiment on the Diversity Trumps Ability Theorem. Following the conditions in section 2.2.2 and adding the possibility to take advantage of other's solutions, both mine detection and area reduction could be tested on the theorem.

Solution-Building Challenge

The challenge lies in how to make internet users build on each others' solutions. A problem is also how to categorize them into groups based on performance—what performance?

Mine Detection

A mine detection experiment which tests the problem-solving theory of the Diversity Trumps Ability Theorem, in which each solver may watch the results of others in her collection, and iteratively build on that result. A more extensive qualification test where 'bad' solvers are disqualified, to acquire a smart crowd.

With properly processed images in which the problem is about valuing and interpreting contrast in images and object positions, as often is the problem in the MOMS results, totally random agents without demining experience would be very suitable.

Showing plots of objects others have made in the same collection that of the agent, and having the agent evaluate those plots, as the possibility to add her own.

Area Reduction

In an experiment in area reduction, though, all conditions of the Diversity Trumps Ability Theorem apply except for the second, the calculus condition, or the smart solvers condition. For the agents to be able to be considered smart, they need access to all the information.

It is a huge undertaking to formalize information aggregation out in the field. However, the possibilities are huge were the information available. Not only an entire new (huge) source of workforce, but a diverse one, would be made available.

5.5.3 Computer Algorithms Comparison

An algorithm could be considered an individual performer with limited heuristics and perspectives. Once MOMS and other projects have come so far as there are implemented mine detection algorithms, they should be compared with the collaboration of people. Page and Hong put it like this in [16]:

> *Computers and people differ in their abilities to exploit diverse perspectives and diverse heuristics. Computers can iteratively apply multiple heuristics with awesome speed, but they have a difficult time communicating across perspectives. Humans apply heuristics rather slowly, but can switch perspectives quickly and can communicate across diverse perspectives.*

5.5.4 Area Reduction Market

As seen in section 2.2.4 on predictive markets and the Policy Analysis Market, predictive markets have been thought to have an impact in prediction of futures and they are known to aggregate information.

The *Area Reduction Market* would trade with regions and zones possibly contaminated with mines, letting the 'smart' inhabitants of a mine-afflicted country or region contribute with their piece of wisdom to the market. The market would be accessed the line of work of mine action such as mine action centres, and through internet cafés and hot spots with computer access such as libraries.

The predictors own stories, experiences and knowledge of the area combined with open mine maps, databases, wikis, and forums, would aggregate information into the market.

5.6 Summary

In spite of the disappointing results due to the circumstances of experiment failure and low expert participation, and not having answers to the questions posed in thesis problem description; the thesis in itself holds important insights for further studies in the area of diversifying demining and using crowdsourcing.

The diversity of people, working parallell to computer algorithms, would bring the best of man-machine cooperation; the thesis shows the direction as to how to achieve this with a firm theoretical base.

It is proposed to further evolve information databases, which in turn can be used by crowdsourced area reduction markets. Also, an improved experiment would finally give the answers to whether or not the crowd beneficially can contribute to mine detection.

Bibliography

[1] Anna Linderhed, Sten Nyberg, Stefan Sjökvist, and Magnus Uppsäll. Optical Methods for Detection of Minefields. Technical Report FOI-R--1331--SE, Swedish Defence Research Agency (FOI), Sep 2004.

[2] S. Sjökvist, S. Abrahamson, P. Andersson, T. Chevalier, G. Forssell, C. Grönvall, H. Larsson, D. Letalick, A. Linderhed, D. Menning, S. Nyberg, I. Renhorn, M. Severin, O. Steinvall, G. Tolt, and M. Uppsäll. MOMS Multi Optical Mine Detection System - Initial Report. Technical Report FOI-R--1721--SE, Swedish Defence Research Agency (FOI), Sep 2005.

[3] D. Letalick, P. Andersson, T. Chevalier, C. Grönwall, A. Linderhed, H. Larsson, D. Menning, C. Nelsson, P. Nilsson, S. Nyberg, S. Sjökvist, O. Steinvall, G. Tolt, and M. Uppsäll. MOMS - Progress report 2006. Technical Report FOI-R--2147--SE, Swedish Defence Research Agency (FOI), Dec 2006.

[4] B. Kanefsky, N.G. Barlow, and V.C. Gulick. Can Distributed Volunteers Accomplish Massive Data Analysis Tasks? *Proceedings of the Lunar and Planetary Science XXXII*, 1272, 2001.

[5] J. Howe. The Rise of Crowdsourcing. *Wired Magazine*, 14(6), 2006.

[6] Jeff Howe. Crowdsourcing: Tracking the Rise of the Amateur. `http://crowdsourcing.com/`.

[7] James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Anchor Books, 2005.

[8] N.G. Barlow. Abstract 1475. *LPSC XXXI*, 2000.

[9] Amazon. `http://www.mturk.com/mturk/welcome`.

[10] Amazon. `http://www.mturk.com/mturk/welcome?variant=worker`.

[11] Panagiotis G. Ipeirotis. Unpublished Mechanical Turk survey. `http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html`.

[12] Steve Friess. Renowned Aviator Is Missing in Nevada. `http://www.nytimes.com/2007/09/05/us/05aviator.html`, Sep 5, 2007.

[13] Mike. Amazon Web Services Blog: Help Find Steve Fossett. `http://aws.typepad.com/aws/2007/09/help-find-steve.html`, Sep 8, 2007.

[14] Steve Friess. Online Fossett Searchers Ask: Was It Worth It? `http://www.wired.com/techbiz/it/news/2007/11/fossett_search`, Nov 6, 2007.

[15] Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 55–64, New York, NY, USA, 2006. ACM Press.

[16] Lu Hong and Scott E. Page. Diversity and Optimality. Unpublished, May 22, 2002.

[17] Lu Hong, Scott E. Page, and William J. Baumol. Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389, 2004.

[18] Scott E. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, 2007.

[19] Michael J. Mauboussin. Explaining the Wisdom of Crowds: Applying the Logic of Diversity. *Mauboussin on Strategy*, March 20, 2007.

[20] F. Galton. Vox populi. *Nature*, 75(1949):7, 1907.

[21] BBC News. Pentagon Axes Online Terror Bets. `http://news.bbc.co.uk/2/hi/americas/3106559.stm`, Jul 29, 2003.

[22] Noah Shachtman. The Case for Terrorism Futures. `http://www.wired.com/politics/law/news/2003/07/59818`, Jul 30, 2003.

[23] R. Hanson. The Informed Press Favored the Policy Analysis Market. *George Mason University*, 15, 2005.

[24] J. Ahlberg. A Matlab Toolbox for Analysis of Multi/Hyperspectral Imagery. Technical Report FOI-R--1962--SE, Swedish Defence Research Agency (FOI), March 2006.

[25] Daniel Westberg. A Sensor Fusion Method for Detection of Surface Laid Land Mines. Master's thesis, Department of Electrical Engineering, Linköpings universitet, 2007.

[26] Eric M. Filippino. *A Guide to Socio-Economic Approaches to Mine Action Planning and Management*. Geneva International Center for Humanitarian Demining, November 2004.

[27] Andy Smith. *Mined Area Indicators – Angola*.

[28] Subvert and Profit. http://www.subvertandprofit.com.

[29] Paul Courson and Steve Turnham. Amid furor, Pentagon kills terrorism futures market. http://www.cnn.com/2003/ALLPOLITICS/07/29/terror.market/, Jul 30, 2003.

[30] David Andersson. Interview with Ralf Andrén, SWEDEC, Nov 12, 2007.

[31] E-mail conversation with Dave Pfeiffer, HIT-Builder, Oct 31, 2007.

[32] DPA Software. HIT-Builder for Amazon's Mechanical Turk - Home. http://hit-builder.com/.

[33] Turker Nation. Turker Nation - Home. http://turkers.proboards80.com/index.cgi.

[34] E-mail conversation with Dave Pfeiffer, HIT-Builder, May 8, 2008.

Appendix A

Theory

A.1 Needs Assessment Excerpt

This excerpt is based on a list on needs assessment from the report *A Guide to Socio-Economic Approaches to Mine Action Planning and Management* [26] by the Geneva International Center for Humanitarian Demining.

Geographic

1. What is/was the pattern of current and former conflict?
2. Where are the mine- and battlefields?
3. What is the pattern of roads and bridges, and electrical and other utilities?
4. Where are health/education facilities and administrative centres?
5. What is the range of soil types and vegetal cover and climate zones and where are they located?

Demographic

1. What is the spatial distribution of the settled population?
2. What are the numbers and likely movements of refugees and internally displaced persons?
3. What are the numbers and migration patterns of nomadic groups?

Public Health

1. How many mine incidents are there and how many civilians have been affected (broken down by age, sex, position in household, occupation/ livelihood)?
2. What are the main reasons for risk-taking (e.g. ignorance, recklessness, economic or other survival pressures)?

3. What is the capacity of public heath facilities for treatment and rehabilitation?
4. How many victims are reaching treatment centres?

Economic

1. What is the level and structure (sectoral, geographic, public-private, market-subsistence) of economic activity?
2. What are the principal and secondary sources of livelihood in contaminated communities?
3. What is the extent of commercial activity and dependence of affected populations on factor (supplies, labour, credit) and product markets?
4. What are the types of land, resources, and infrastructure affected by mines and UXO?
5. What is the degree of inequality and pattern of poverty?
6. Where are critical natural resources located?

Appendix B

Experiment

B.1 MATLAB Image Processing Code

B.1.1 autocontrast.m

Automatically adjusts contrast of images to optimum level.

```
function img2 = autocontrast(img)
[m1 n1 r1]=size(img);
img2=double(img);
%---calculation of vmin and vmax---
for k=1:r1
    arr=sort(reshape(img2(:,:,k),m1*n1,1));
    vmin(k)=arr(ceil(0.008*m1*n1));
    vmax(k)=arr(ceil(0.992*m1*n1));
end
%---
if r1==3
    v_min=rgb2ntsc(vmin);
    v_max=rgb2ntsc(vmax);
else
    v_min=vmin;
    v_max=vmax;
end
%---
for i=1:m1
    for j=1:n1
        for k=1:r1
            img2(i,j,k)= 255*(img2(i,j,k)-v_min(1))/...
                         (v_max(1)-v_min(1));
        end
    end
end
```

```
%---
img2=uint8(img2);
img2=double(img2);
img2=img2./255;
```

B.1.2 labImadjust.m

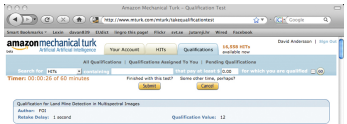
From MathWorks website[1]. Converts to L*a*b* colour space and does a histogram equalization on the light channel.

```
function shadow_histeq = labImadjust(input)
shadow = input;
srgb2lab = makecform('srgb2lab');
lab2srgb = makecform('lab2srgb');
shadow_lab = applycform(shadow, srgb2lab); % convert to L*a*b*

% the values of luminosity can span a range from 0 to 100; scale
% them to [0 1] range (appropriate for MATLAB intensity images
% of class double) before applying the contrast enhancement
% techniques
max_luminosity = 100;
L = shadow_lab(:,:,1)/max_luminosity;

% replace the luminosity layer with the processed data and then
% convert the image back to the RGB colour space

%% Histeq
shadow_histeq = shadow_lab;
shadow_histeq(:,:,1) = histeq(L)*max_luminosity;
shadow_histeq = applycform(shadow_histeq, lab2srgb);
```

[1] http://www.mathworks.com/products/image/demos.html?file=/products/demos/shipping/images/ipexcontrast.html

B.2 Qualification Test

The qualification test laymen and experts had to complete before browsing HITs.



Please read the background and instructions carefully.

Experiment

This scientific experiment is part of an M.Sc. thesis 'Crowdsourcing in Demining' at FOI (The Swedish Defense Research Agency). The thesis is undertaken by David Andersson, M.Sc. student in Media Technology and Engineering at the Faculty of Technology at Linköpings universitet, Sweden.

The thesis files under the Multi Optical Mine Detection System project (MOMS) at FOI. Please be aware this being a military research project for both tactical (saving conflict) and humanitarian (post-conflict) demining.

The purpose of the experiment is to examine and give indications of the following:

- Can humans detect land mines and Unexploded Ordnances (UXO) in optical data?
- Can they do that better than current computer algorithms applied to the same data?
- Is an expert group better than a layman group (e.g. the crowd, turkers) in examining this kind of imagery.

These public HITs will be published once again as closed HITs open only to an expert group consisting of personnel of SWEDEC, FOI, and other demining instances.

Images

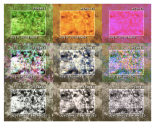
The images are all experimental. They are taken from a hyperspectral camera – Taking photos in 150 bands rather than 3 (red, green blue) – at a test site in Eksjö, Sweden, at the SWEDEC (Swedish EOD and Demining Centre) facility.

There are images taken from the same area, at different times of the day. There are also images from the same area, with no land mines or UXO.

The perspective is a couple of metres up on a tripod. The different views are all taken from the same tripod position.

Cells

In each HIT a cell is shown: each image column represents the visible band (RGB), the near infrared band (NIR), and the mid-infrared band (MIR). Each row of a column has had different image processing methods applied to it: Auto-contrast, RGB balance, and Luminance balance.



All images in a cell are from the same area, just in different spectral ranges and with different image processing algorithms applied. The different processing methods and spectral bands are supposed to improve detection in that very area.

Because of the perspective, some cells might show the sky. Further down we'll examine how to classify those cells and others. There might also be metal-related equipment in the area, please read further on how classify those cases.

Land mines

Please carefully examine this list of various land mines to get a hunch of their general shapes and forms.

Anti-Personnel (AP) Mines

PMA-1

PMA-1A, PMA-3, PMA-3

VS-50

OZM-72

MI4

PFM-1

M18A1

MON-100

PROM-1

PMR-3A

MI AP DV 59

Type-72

Anti-Tank (AT) Mines

TM-62M

FFV 013

AT-2

TMM-1

TMA-1

TMA-5

TMRP-6

Strvmina S

BLU-92B

Qualification Instruction

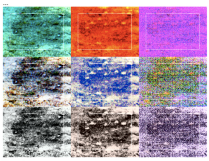
The following classification questions and cells are actually some examples from the actual HITs in the dataset. Please examine them carefully, read the small text with instructions for classification, and classify by using the checkboxes. Either 2 or 3 boxes are required to be checked at each cell.

Before starting with the Qualification, we'd like to thank you for your help in participating in this experiment.

David Andersson
April 29th, 2008

Qualification

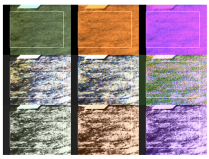
...



Road

This is a road with no objects in it.

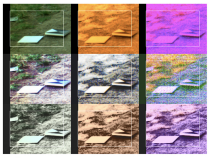
- 0 objects, road
- 1 object, field
- 1 object, road
- 0 objects, field
- 0 objects, grove



0 objects

This cell has 0 objects on the road. On the top side of the image is some sort of object, but it's not within the area of this cell.

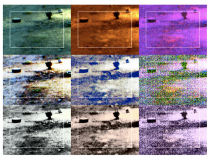
- 0 objects, grove
- 0 objects, road
- 1 object, field
- 1 object, road
- 2 objects, road



3 objects

Even though only approximately half of the cell's area is covered by road, it should be classified as road. Two objects are clearly present in the cell, there is also some sort of object in the top left area of the cell.

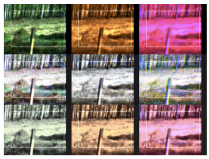
- 1 object, field
- 1 object, road
- 2 objects, field
- 3 objects, road
- 2 objects, road



3 or 4 or 5 Objects?

3 objects clearly lie on the road. In the upper right part of the cell looks there seems to be 2 objects located, of which one in the very corner. Should be marked as 4+ objects and flagged as difficult.

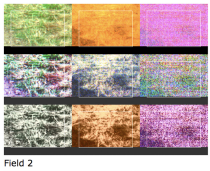
- 3 objects, road
- 3 objects, road, difficult
- 4+ objects, field, difficult
- 4+ objects, road, difficult
- 4+ objects, road



Field

This area clearly is not road, thus a field. Because of the view there seems to be a stump of wood, or it could be an object, and should therefore be marked as difficult and as one object.

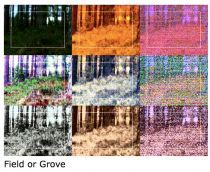
- 0 objects, field
- 0 objects, field, difficult
- 1 object, field, difficult
- 1 object, field
- 1 object, grove, difficult



Field 2

This is a patch of grass (field) which is not situated on the road. No discernible objects are present nor is it a difficult area.

- 0 objects, field
- 0 objects, field, difficult
- 0 objects, grove
- 0 objects, grove, difficult
- 0 objects, road

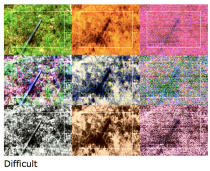


Field or Grove

This is clearly not a part of the road. Low bushes or grass imply a field, but might as well be classified as grove because of the stems of trees present.

Classify as grove without any objects.

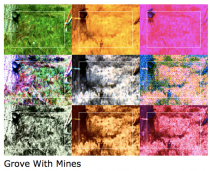
- 0 objects, field
- 0 objects, field, difficult
- 0 objects, grove
- 0 objects, grove, difficult
- 0 objects, road



Difficult

An object is lying in a field. It might be a tree branch or something else, flag it as difficult and as one object in a field.

- 1 object, field
- 1 object, field, difficult
- 1 object, grove
- 1 object, grove, difficult
- 1 object, road

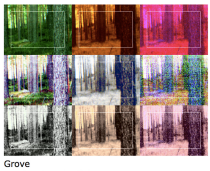


Grove With Mines

Clear part of a tree, defining the cell as grove. One mine is attached to the stem of the tree. There is also a mine in the grass in the bottom of the cell area.

Classify as 2 objects and grove.

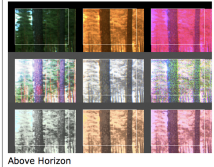
- 0 objects, field
- 0 objects, grove
- 1 object, field
- 1 object, grove
- 2 objects, grove



Grove

A grove of trees without any discernible objects.

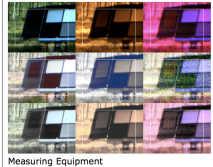
- 0 objects, above horizon
- 0 objects, field
- 0 objects, grove
- 0 objects, road
- 0 objects, field, difficult



Above Horizon

The visible stems and treetops tell us this is not ground, thus above horizon. No objects are seen.

- 0 objects, above horizon
- 0 objects, field
- 0 objects, grove
- 0 objects, road
- 1 object, above horizon



Measuring Equipment

All the cells are taken from hyperspectral measurement images from a test site. In the text were also other measurements made. Therefore there is often equipment present. Mark as one object and flag it as difficult.

- 0 objects
- 0 objects, difficult
- 1 object
- 1 object, difficult
- 2 objects, difficult

Finished with this test? Some other time, perhaps?
Submit Cancel

B.3 HIT-Builder Posting Errors

The following errors occured in the posting of the laymen HITs in the HIT-Builder service:

"2b_1___6_4 - 119" - Sorry! you are not able to connect Amazon server

"3_0_2___6_3 - 172" - This Requester has insufficient funds in their account to complete this transaction.

"3_0___6_3 - 173" - This Requester has insufficient funds in their account to complete this transaction.

"3_1___6_3 - 174" - This Requester has insufficient funds in their account to complete this transaction.

"3_2___6_3 - 175" - This Requester has insufficient funds in their account to complete this transaction.

"3_3___6_3 - 176" - This Requester has insufficient funds in their account to complete this transaction.