Crowds for Clouds: Using an Internet Workforce to Interpret Satellite Images

Ling Yu[1], Sheryl Ball[2], Christine E. Blinn[3], Klaus Moeltner[1], Seth Peery[4], Valerie A. Thomas[3], and Randolph H. Wynne[3]

[1] Department of Agricultural and Applied Economics, Virginia Tech [2] Department of Economics, Virginia Tech

[3] Department of Forest Resources and Environmental Conservation, Virginia Tech [4] Virginia Tech Geospatial Information Sciences

Introduction

A chronologically ordered sequence of satellite images can be used to learn how natural features of the landscape change over time. For example, we can learn how forests react to human interventions or climate change.

Before these satellite images can be used for this purpose, they need to be examined for clouds and cloud shadow that may hide important features of the landscape and would lead to misinterpretation of forest conditions. Once clouds and their shadow have been identified, researchers can then look for other images that include the feature of interest, taken a bit earlier or later in time, to fill in the "missing information" for the original image. Therefore, the task of identifying clouds and their shadow is extremely important for the correct and efficient use of each image.

Computer algorithms are only imperfectly suited for this task. The aim of this project is to outsource the cloud interpretation task to a global internet community of "turkers" - workers recruited via amazon.com's online job market known as "Mechanical Turk" .

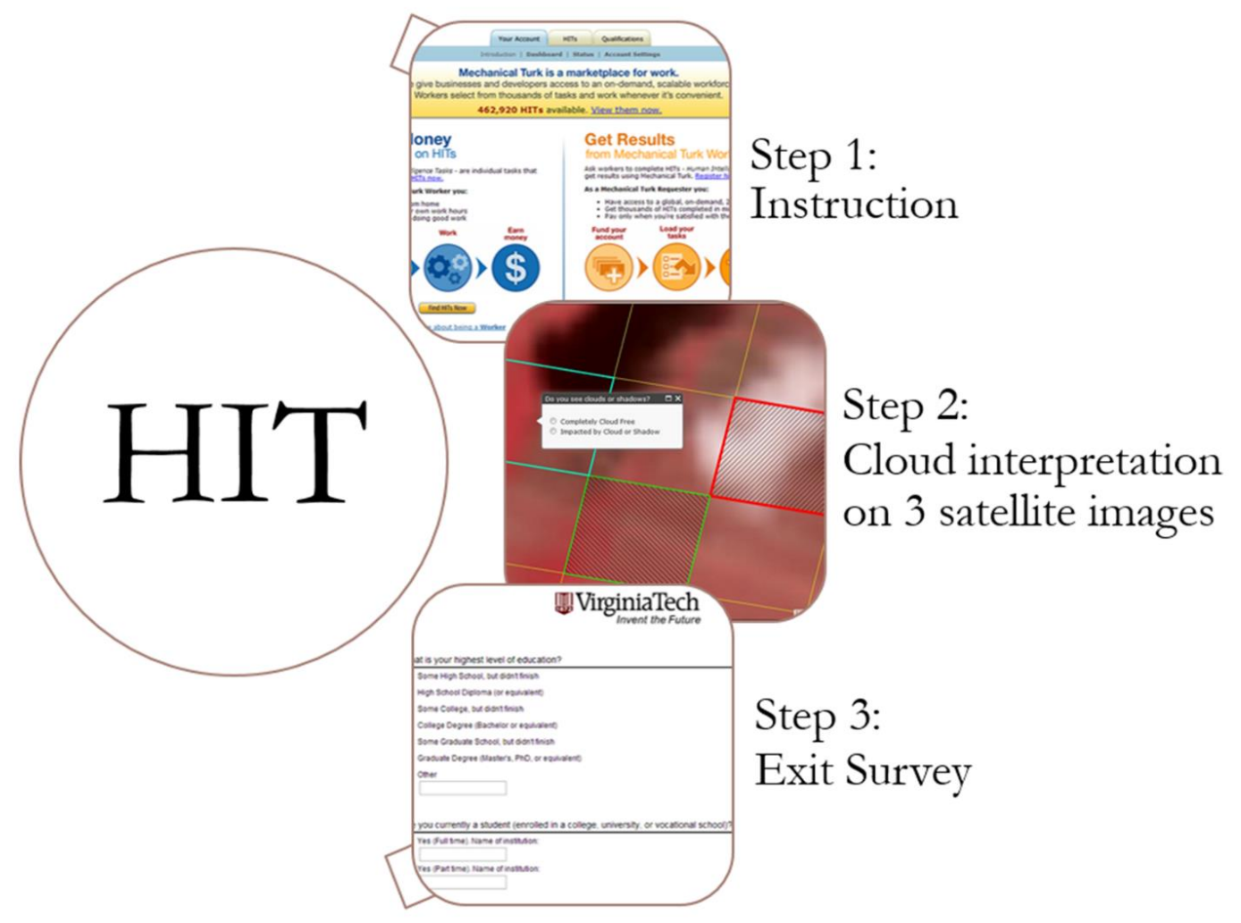
Aims of the Present Study

A labor experiment was implemented as an online Human Intelligence Task (HIT) at Amazon.com's Mechanical Turk job market platform. The experiment is designed to identify the separate effects of:

- wage ($0.5, $1.0, $1.5),
- complexity (easy, medium, hard) of satellite image, and
- learning/fatigue effects (the first, second, third image).

Methods and Experimental Design

The experiment was implemented in fall 2013 in three phases. Each phase offered a different wage, and provided six different versions, each with a different sequencing of the same three LANDSAT images. A training module was provided at the beginning of the HIT, and an exit survey collecting basic demographic information and feedback was given after the completion of the HIT.


HIT

Step 1: Instruction

Step 2: Cloud interpretation on 3 satellite images

Step 3: Exit Survey

Figure 1. The task for worker as one HIT on Mechanical Turk
One of 6 versions (EMH, EHM, MEH, MHE, HEM, HME) was randomly assigned.

Results

The accuracy of interpretation (based on expert benchmarks) for each image and for different wage levels are shown to the right. As can be seen from the figure, accuracy is primarily driven by image difficulty, and a "fatigue effect" that is independent of image complexity. In contrast, wage does not seem to affect workers' performance. We also find accuracy gains associated with U.S. workers (compared to international participants, who are primarily from India), and male interpreters.

Table 1. Regression results on accuracy percent per image

	(1)	(2)
constant	86.8396***	88.0775***
Phase II ($1 = 1)	-0.3999	
Phase III ($1.5 = 1)	-0.4129	
M (medium = 1)	-3.6636***	-3.7378***
H (hard = 1)	-13.2876***	-13.389***
second (second = 1)	-0.8955**	-0.602*
third (third = 1)	-0.5234	-0.1312
time on image (min)	-0.3262***	0.0636
time on training (min)	-0.1509***	
gender (male = 1)	1.4908***	
country (US = 1)	5.2842***	
college (at least college = 1)	-0.4175	
background (with background = 1)	1.1225*	
Individual fixed effects	No	Yes
R Squared	0.2019	0.1534
N	4713	4713

*** Indicates significance at the 1-percent level.
** Indicates significance at the 5-percent level.
* Indicates significance at the 10-percent level.

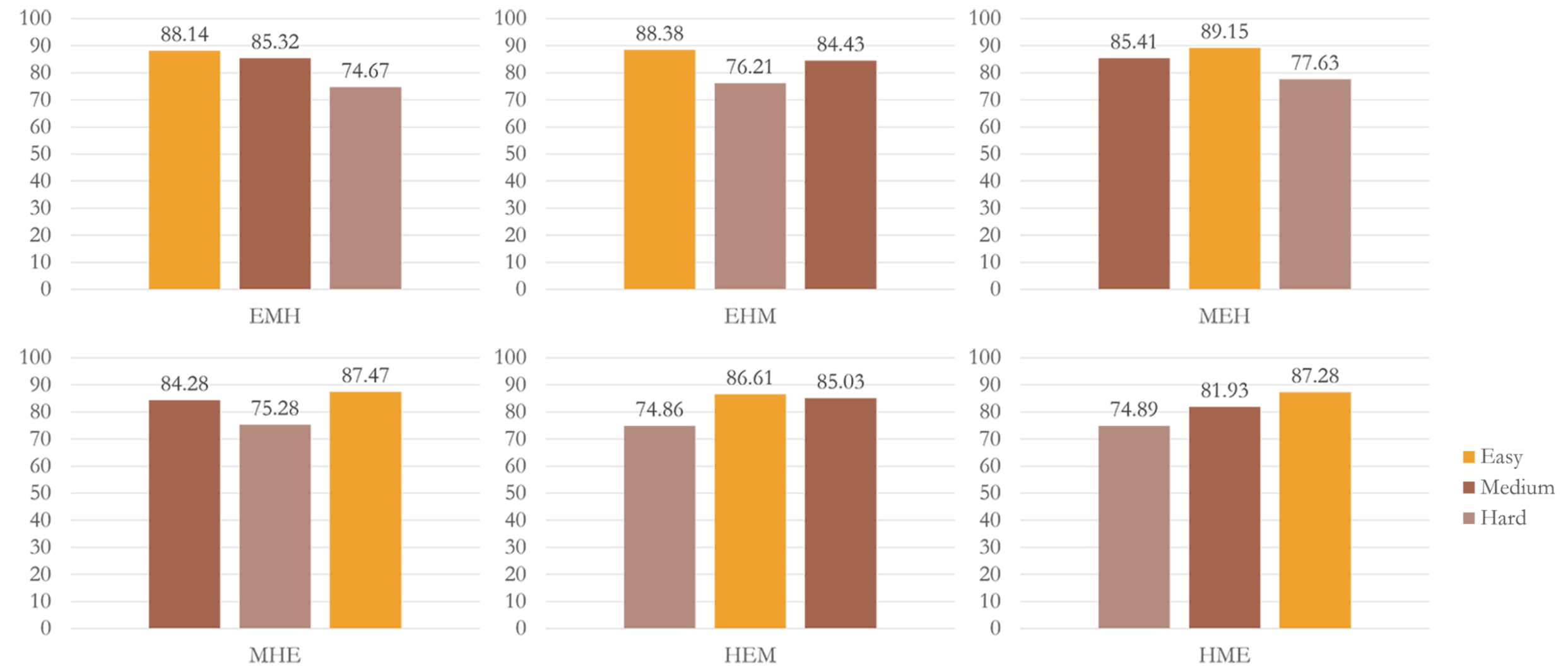


Figure 2. Accuracy percent for each image for each version for phase II
Accuracy is primarily driven by image difficulty.

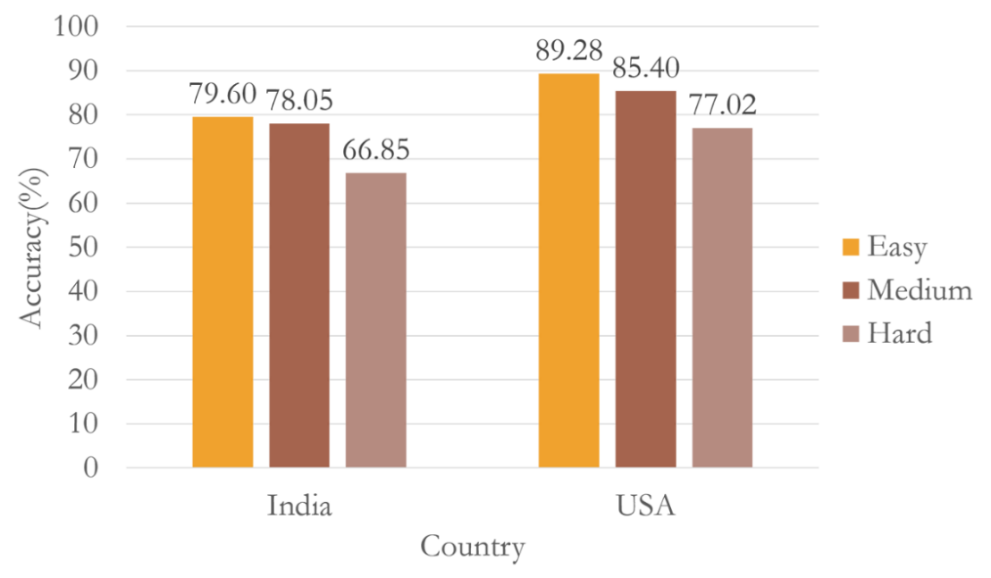
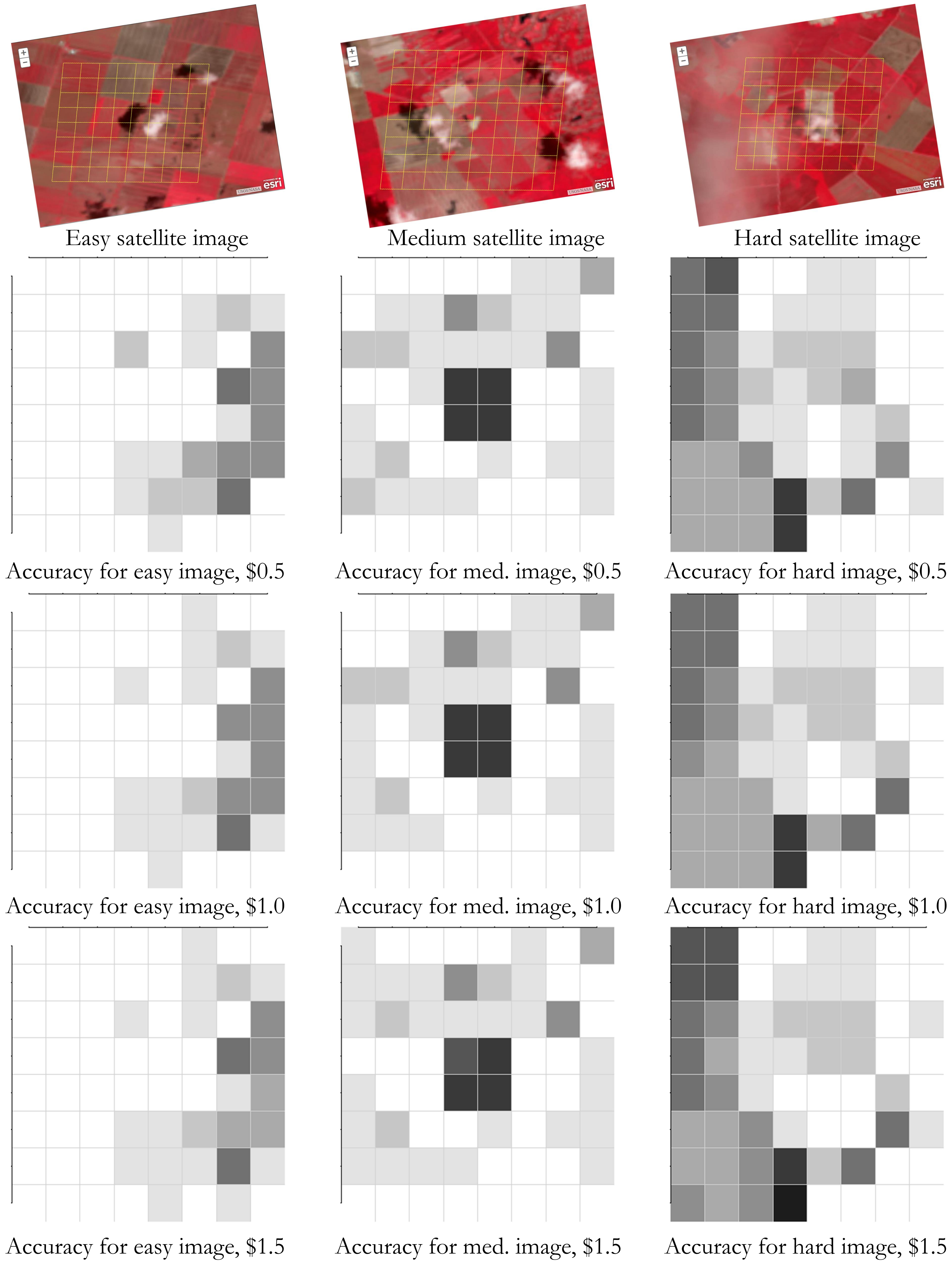


Figure 3. Accuracy percent for each image for US and India workers for phase II
Higher accuracy for US workers compared to India workers on average.


Easy satellite image

Medium satellite image

Hard satellite image

Accuracy for easy image, $0.5

Accuracy for med. image, $0.5

Accuracy for hard image, $0.5

Accuracy for easy image, $1.0

Accuracy for med. image, $1.0

Accuracy for hard image, $1.0

Accuracy for easy image, $1.5

Accuracy for med. image, $1.5

Accuracy for hard image, $1.5

Figure 4. Accuracy percent per tile for each image for each phase
A darker shading implies a higher degree of incorrectness. Accuracy is primarily driven by image difficulty, however, wage does not seem to affect workers' performance.

Discussion and Future Actions

Based on these findings and qualitative feedback by participants, we are now refining the training module and the cloud interpretation interface. We will also examine the comparative performance of computer algorithms and human workers, and possibly develop a hybrid approach that caters to the relative strengths of machine and human interpretation.