Analysis of User Proficiency and Involvement in a Citizen Science Project

Marinalva Dias Soares, Rafael Santos, Nandamudi L. Vijaykumar and Luciano Vieira Dutra
National Institute for Space Research (INPE)
São José dos Campos, Brazil
marinalva@dpi.inpe.br, rafael.santos@lac.inpe.br, vijay@lac.inpe.br, dutra@dpi.inpe.br

Abstract

Due to its popularity, ubiquity and relatively low cost of access, the Internet has become a major channel for interaction between people in different forms and for different purposes. One relatively new interaction paradigm is exemplified by citizen science, which allows scientists and common people to collaborate in different ways to solve a particular scientific problem. Several citizen science projects are already in execution, some of those being very successful both for the scientific purpose and in the sense of engaging the participants, showing its potential for science and education. In this paper is presented a citizen science project to voluntarily label imprecisely segmented images and show whether patterns and trends among the users can be identified through their proficiency and involvement with the project.

Keywords: *Citizen science, collective intelligence, user behavior, image segmentation, image labeling.*

1. Introduction

Citizen Science involves volunteers from the general public that act as participants or observers in some domain of science for data collection, classification or analysis [1]. This approach has been adopted in different science domains, including remote sensing that include observation, classification and analysis that are labor-intensive, time-consuming, costly and especially when the data collection or analysis is beyond the capacity of the core science team.

In remote sensing, one the most important activities is the identification of objects in the image, or scene interpretation. An important process in the identification of objects is the image segmentation. Image segmentation can be defined as a process that partitions a digital image into regions (usually polygons), so that elements belonging to each region (or polygon) are similar with respect to some properties [2].

The polygons obtained through segmentation must be labeled or identified usually associated with semantic information about them. For example, in urban scenes, example of labels may be roofs, trees, streets, pools, etc. The segmentation process may create a huge number of polygons and the image may not be properly partitioned or segmented due to the imperfection inherent of the segmentation algorithms.

When humans are involved in interpreting scenes, they use their experience, visual evidence, context of the scene, etc. to label each polygon. On the other hand, image processing systems for automatic scene identification, while possibly being faster, often cannot use that information in the same way that humans can.

A small region in an urban scene is shown in Figure 1; and its segmentation in Figure 2. In Figure 2 (this image was processed to improve its contrast for publication and to enhance the lines that define the polygons resulting from the segmentation) there are both oversegmented regions (perceptual objects divided into several regions) and undersegmented (regions which contains several different perceptual objects), which are practically unavoidable when using image segmentation algorithms.

Figure 1. Satellite Image of an Urban Scene

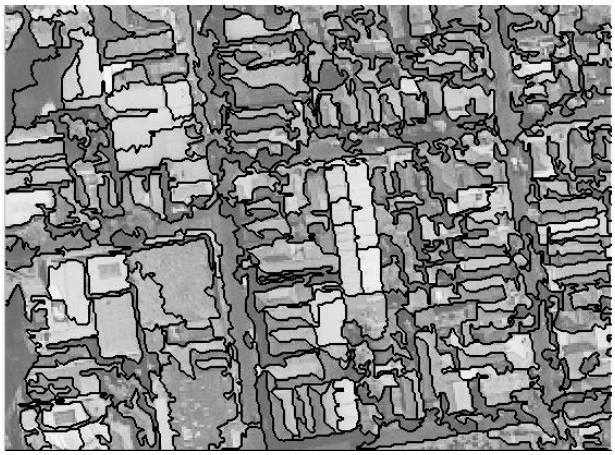

Figure 2. Segmentation of the Image in Figure 1

In Figure 3 there are four regions which were labeled by an expert and painted with dark gray (for ceramic-tiled roofs) and light gray (for trees) over the original

segmented image, which was also processed to improve its contrast for publication.

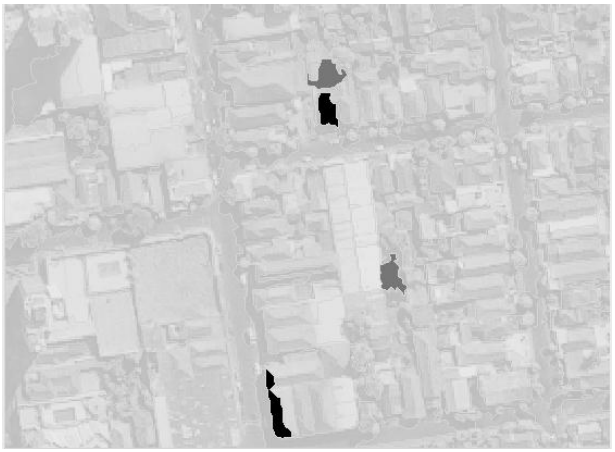

Figure 3. Manual Labeling of the Region

So, what would be the problem in using human expert or specialist to assign labels to the polygons as it is accepted as a fact that a human specialist can perform labeling based on some rules, samples or conditions that depend on a particular object, being potentially very effective at this task. The problem is that manual labeling is very time consuming as there are several polygons to be labeled. As it is a repetitive and a tedious task this process may lead to labeling errors. A possible solution is to hire several specialists which may check and corroborate others' results; however, a specialist's time is expensive, let alone several specialists.

One approach to overcome this is to make use of automatic labeling. However, in order to be successful, automatic labeling must have mechanisms to incorporate the knowledge human specialist use. This is definitely a difficult task, because it is an error-prone process since algorithms cannot faithfully reproduce the knowledge and experience from the specialists.

So what could be an alternative? A different approach, used in this research, to label objects is to use several different human agents, although they may not possess the same expertise as the specialists already mentioned. These agents could receive a very brief, superficial training and then are presented with different polygon labeling tasks. Thus, the entire polygon labeling task could be performed by several users, which could complement each other's opinions, hypothetically leading to good results (since they would use human knowledge and intelligence) without the expensive work of a single or several experts.

The use of common citizens who act as participants or observers in a domain of science is often called citizen science. Scientific research based on citizen science have been used, often with good results, in several different tasks that either could be performed by a lot of work by a human specialist or poorly by an automatic system. Citizen science is more often than not based on volunteer users – the motivation of the participants (often unpaid volunteers) has also been studied [3], [1].

In [3] the validity of using citizen science approach is discussed. It is pointed out that the information of these volunteer users is often better than the information coming from other groups being paid. The author comments that in a citizen science program, the same sort of standards (what is accepted and what is not accepted) must be applied to volunteers in terms of quality and types of information as if applied to paid specialists. Volunteers collaborate because they want to, not because it is considered as a paid job. Participation of volunteers in scientific research can bring benefits such as experience and increased knowledge about a particular topic [4].

Use of citizen science requires an infra-structure to control the distribution and integration of tasks. This infra-structure includes creating means and methods to present and collect information from volunteer collaborators. In order to use the data collected from those users one needs to assess the data quality, coherence, relevance, etc. Therefore, the analysis of the collected data may be more important than the data itself - for example, in an object labeling task one may not rely only on most of the opinions from the users, but on the past performance, reliability, inferred knowledge, etc. of those users. Modeling user knowledge is then of major importance when dealing with citizen science.

This paper presents a citizen-science project (volunteer labeling of imprecisely segmented image regions) and comments on the analysis that may elicit information about collective and individual user behavior.

The contributions of this paper include: demonstrate the feasibility of using citizen science to label polygons resulting from a segmentation process; evaluation of users' proficiency; and identification of patterns and trends between the users (e.g. identification of groups of users who tends to perform better with some labels than others);

The paper is organized as follows: Section 2 presents some applications of citizen science; Section 3 presents the methodology; Section 4 describes the experiment used in this research; Section 5 presents the analysis about the users' proficiency and Section 6 presents the final comments and future work.

2. Citizen Science: Concepts and Applications

2.1. Citizen Science

The term citizen science has been used to describe a range of ideas [5], including informal partnerships between scientists or scientific institutions and nonscientist volunteers who collaborate on specific projects. The volunteers (which may or not be paid for their work) provide important resources often their time, computational resources or specific (but not necessarily very specialized) knowledge.

Participation of volunteers in scientific research can bring benefits such as experience and increased

knowledge about a particular topic [4], therefore being also a tool to promote education and science awareness.

Citizen science is not a new concept: the American Association of Variable Star Observers [6] and the Audubon Society Christmas Bird Count [7] have both successfully partnered citizen scientists with professional scientists for more than a century, producing research results that could not otherwise have been achieved [1].

The availability of large scientific datasets through the Internet has allowed citizen science projects to engage volunteers in new ways, both allowing the use of those datasets to search for new information and allowing the collection of even more data. Some Internet-based projects that aims to promote public engagement with research, as well as with science in general include:

- Citizen Science Central (http://www.birds.cornell.edu/citscitoolkit): this site, created by the Cornell Lab of Ornithology, lists several other citizen-science based projects, a toolkit for the development of new projects and articles and other resources.
- SETI Search for Extraterrestrial Intelligence (http://setiathome.berkeley.edu): the goal of this project is to detect intelligent life outside the planet Earth. Users do not interact directly with the data, but they download an application which processes data collected from a central repository. The data processing is done while the computer is idle.
- Galaxy Zoo 2 (http://www.galaxyzoo.org/): the Galaxy Zoo 2 site invites users with Internet connection to classify galaxies accordingly to their shapes and to the Hubble classification scheme.
- Herbaria@home (http://herbariaunited.org/atHome): citizen science is used in this project to identify the undocumented plant samples that belong to museums and universities in the United Kingdom.
- Citizen Sky and the Mystery of epsilon Aurigae (http://www.citizensky.org): The main objective of this project is to understand a star that has been a mystery to scientists: epsilon Aurigae, located in the constellation Auriga, the charioteer.
- The ODP Open Dinosaur Project (http://opendino.wordpress.com) is a collaborative research effort, focused on developing a comprehensive database of limb bone measurements for some species of dinosaurs, in order to identify patterns of limb bone evolution. Collaborators are asked to submit measures of samples (from the literature or from direct measures from specimens) to the project.
- EpiCollect (http://www.spatialepidemiology.net/epicollect) this project collates data, such as spreading of a disease and occurrence of rare species using a mobile phones that feed a database.

2.2. Collective Intelligence

The increasing number of people contributing to the Internet, either deliberately or incidentally, has created huge sets of data that gives millions of potential insights into user experience, marketing, personal tastes, and human behavior in general [8]. From the users interactions with web applications, a large set of data that can be converted into intelligence can be acquired. From an application point of view, collective intelligence can be defined as the effective use of information provided by others to improve the application [9].

Collective intelligence has become increasingly popular, important and feasible. The methods for collective intelligence existed before the Internet [10]. However, the Internet provides a rapid way to collect information from thousands of people on the Web through interaction in general (e.g. on-line purchases, browsing, etc.). All these interactions between users and applications can be monitored and used to derive information.

Wikipedia (http://pt.wikipedia.org) is one of the most famous examples of collective intelligence. Wikipedia is an online encyclopedia which has been created from voluntary contributions from the users. It is possible for anyone to create or edit a topic on the encyclopedia, but since the entries are verified by a very small number of administrators, abuses and mistakes in general can be avoided.

As another example of collective intelligence, one can mention Google, the world most popular Internet search engine. It rates web pages based on number of links to those pages which is entirely based on the opinion of people who refer to that particular web page [10].

In [9] collective intelligence is classified in three categories: explicit, implicit and derived. In the case of explicit intelligence user directly provides the information/intelligence, for example, reviews, recommendations, ratings, voting, tags, bookmarks, user interaction, and user-generated content. In some cases, users may or not be linked to an application. Even then, they provide information that is considered as indirect such as posting messages on blogs, online communities and wikis. This is known as implicit intelligence. In this case, it is possible to mine data from external sources so that some value may be added to that particular application.

The third category of collective intelligence is known as derived intelligence. This is considered as a high-level intelligence. In this case, data mining techniques are extensively used to detect patterns based on the analysis performed on the data. Examples of this category include recommendation engines, use of predictive analysis for personalization, profile building, market segmentation, and web and text mining. The user's interaction with a citizen science project, when properly collected and measured, may serve to provide insights to the users'

decision processes and on the citizen science project itself, being therefore a way of derived intelligence.

2.3. Closing Remarks

Scientists and volunteers can work together to increase the potential of their research. The power of citizen science stems from the knowledge derived by the contributions of many people. The success of the projects that have used citizen science to collect scientific data has been the motivation for the work discussed in this paper to use volunteers to label unknown segments in a segmented image.

The use of citizen science is very important in the research described in this paper due to a significant number of data that can be collected from several volunteer users. The process of this collection can be characterized as a form of explicit intelligence and from then onwards it is possible to derive intelligence after analyzing the data using proper data mining artificial intelligence algorithms.

3. Methodology

In this paper a practical application is considered as a case study: labeling of segments or polygons extracted from high resolution satellite digital images of urban scenes. Figure 4 shows, at a glance, the processes considered for data preprocessing (step 1), acquisition (step 2) and analysis (step 3) required for the development of this work.

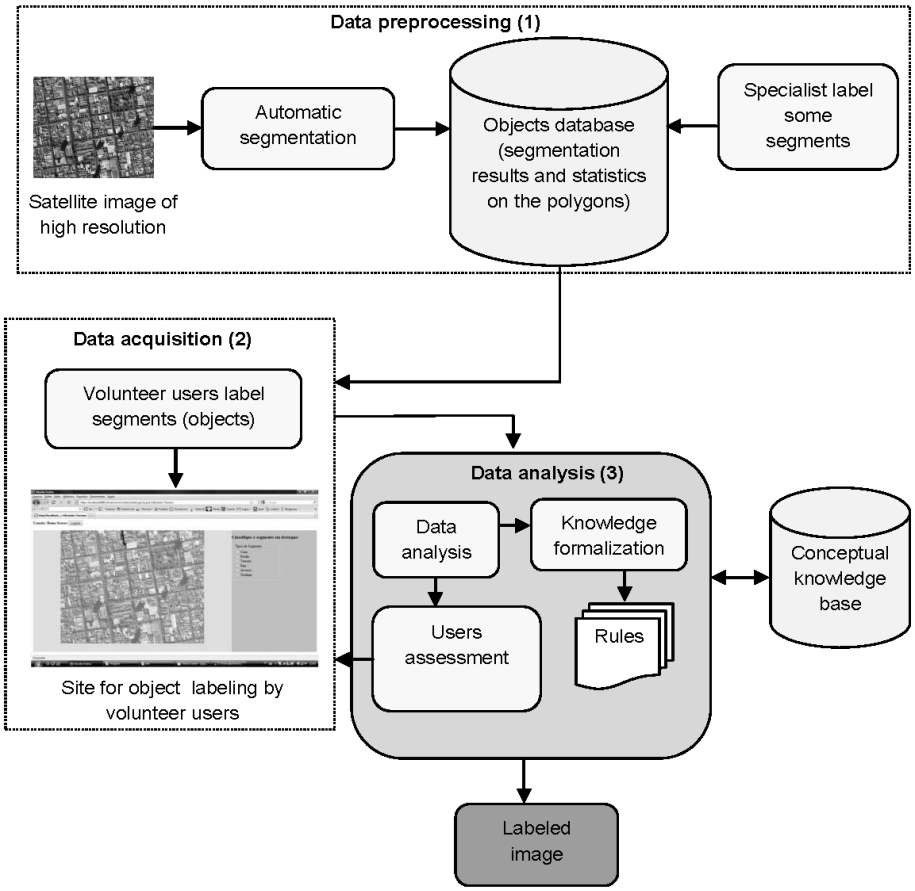

Figure 4. Data processing, acquisition and analysis tasks

3.1 Data Preprocessing

Before starting the process of collecting the user's knowledge for further analysis and application, a database of objects must be created and populated. Objects on this database and the database structure will depend, evidently, on the nature of the problem being considered. This subsection describes the steps used to create and populate a database with satellite image regions obtained by means of segmentation process (step 1 in Figure 4).

The required data preprocessing steps are:

- Selection of a high-resolution satellite digital image. It was decided to choose an urban scene consisting of different types of objects (buildings, occluded streets, shadows, etc.) and with several instances of such objects.
- Segmentation of this image using a traditional region growing algorithm [11]. Region growing is an image segmentation method that requires the selection of initial seed points. This algorithm examines neighboring pixels of initial seed points and determines whether the pixel neighbors should be added to the region in an iterative process. Some parameters of this algorithm control the number and size of the regions it creates and the similarity tolerance used to create those regions; and it is not trivial to determine the best parameters for a specific scene since the resulting segments are often evaluated subjectively by specialists.

  Due to the nature of the image and of the segmentation algorithm, is not expected to have clear, precise polygons as the segmentation result, which is not only acceptable but of interest for our research, which focus more on the user interaction than with the correct labeling for urban planning purposes. For the purposes stated in this work, was preferred have oversegmentation of the image (so an object may be broken into several small segments) than undersegmentation (which causes different targets on the image to be clumped into one single region). This is because it may be easier for the user to identify parts of a stated object (oversegmentation) than be in doubt when several different objects are present in a polygon (undersegmentation).

- Creation of a generic database to hold the tables with data on the polygons, users and tasks. There are no special requirements on the database (e.g. it does not need to be a geographical or spatial database). Creation of a table in the database with all the polygons obtained in the segmentation step and their associated data: vertex positions; statistics on the pixels on that region (mean, variance, other measures); statistics on the polygon's shape (area, perimeter, etc.). This data is obtained from the segmentation algorithm itself.
- Creation of a hierarchical taxonomy for some classes (e.g. roofs and streets paved with different materials) presented to the users to identify. This taxonomy is very dependent on the task itself, and must be defined by a domain expert (e.g. one with expertise on urban planning, remote sensing image analysis, cartography, etc.; who is able to identify which targets are presented on the image). This taxonomy could be as detailed as possible, but when creating it one must consider that complex, multi-level hierarchical classes descriptions could lead to complex user interfaces, which could confuse or discourage users.
- Creation of a table in the database to collect the users' interaction with the data acquisition system. This database can be as simple as possible, but for the analysis there must be at least one table with the user identification (without personal data, but with information on the users' technical/educational background if possible) and the task record table (which records a decision by the user, i.e., the user's identification, the polygon he/she labeled, his/her choice for label and the timestamp for the inclusion of this data on the table).
- Partially populate the users' interaction table by preliminary labeling some of the polygons by a domain expert.

These initially labeled polygons could be considered as ground truth, i.e., labels that are considered to be correct; and can be used to assess answers from the non-expert users, effectively allowing the evaluation of the users' ability. The domain expert will label only some polygons for each class in the labeling task, so he/she will not need to dedicate a significant amount of time for this task.

3.2 Data Acquisition

Data acquisition (step 2 in Figure 4) is central to the idea presented in this paper: in this step data will be collected from the volunteer users that will be used to determine which labels will be used for the polygons and to model the users' knowledge and behavior.

In order to use citizen science, it is necessary to create means and methods to present and collect information from collaborators. There are many ways of to collect data from users, but as most of people have access to Internet, web based tools are presented as an efficient way to collect data from a large number of users in a short time.

So, there is a single task for the data acquisition: the creation and deployment of a web-based interface that presents the tasks for the users and collects the users' decisions. A web site is a natural choice for presenting a task to the user: it will be designed without any special software or hardware requirements; it uses software (internet browsers) which is already well-used and well-known. The basic mechanism of interaction is based on what is used by the Galaxy Zoo project [1], with some changes to allow collection of some metadata (e.g. timestamps of the users' interactions) and allowing users to skip decision tasks.

To perform the labeling, the user should access the website and login into it. For this, he/she must to register stating his/her name, a login and password. This registration is only done the first time the user accesses the site. This information is stored in a database for identification of the user at the time of data acquisition.

In this step, data acquisition, naturally there is a need to interact with the user. It is important that this step should be as much non-intrusive as possible: the volunteer will be presented with a polygon in the context of the image (see Figure 5) and a simple form with choices for the classes that can be used to label it (with the option to skip the labeling for that particular polygon).

Figure 5. Region to be labeled, highlighted and in context in the satellite image

The choices offer suggestions of labels: *roofs* (generic), *ceramic roofs*, *tin roofs*, *cement asbestos roofs*, *trees*, *streets*, *swimming pools*, *shadows*, *open fields*, *bare soil*, *water* and *mixed* targets. Any choice by the user will be stored and a new task (polygon) will be presented. Apart from the login into the site (to identify the user) no further interaction will be required.

3.3 Data Analysis

Data analysis is the most important task in this research, because it is in this stage that the quality of the collected data and user collaboration/behavior in the process of labeling are evaluated.

It is important to point out that the data collection and analysis are continuous processes. As long as there are more data it is possible to perform further, more detailed analysis, and the results of some of the analyses may be used to guide the data collection process.

Several different analysis tasks and scenarios are considered. Among those are:

- The most trivial analysis, which is highly related to the target application (in this case image region labeling), is the evaluation of users' opinions on the labeling of the polygons obtained from the image segmentation. Simple rules could be used such as labeling polygons only when enough opinions have been collected about it, and use a simple weighted majority rule for labeling it. Weights could be derived from an user reliability metric.
- Labeled polygons won't be shown to the users to avoid influencing their future decisions, but could be evaluated by a domain specialist to assess quality of labeling for further use.
- One of the most important analyses, specially considering that the users may be mostly volunteers without training, is the evaluation of the reliability of the user through a reliability metric, which could be calculated using the number of hits" of the user against his/her misses". For example, when the user is still under evaluation, he/she could be presented with some labeling tasks for which the expected classes are already known (being identified by the domain specialist beforehand). The evaluation will be taken into consideration for further analysis but will not provide any feedback to the user, in order influence him/her. User reliability could be reassessed over time, periodically or when statistics indicate that labeling errors are becoming more frequent. The user reliability metric could be calculated for each different label. Users with a record of reliable decisions for a particular label could be assigned more tasks, i.e. show similar polygons, related to that label (e.g. disambiguation tasks). Decisions by a particular user and label that are considered not reliable would imply in a smaller weight when deciding which label ought to be applied to which polygon. This reliability metric is essential to guide the task selector algorithm, which will determine which tasks each user should receive next.
- A measure of reliability could be calculated for each polygon when enough opinions were acquired about it, to determine whether it is an easy or difficult labeling task. Further data could be collected differently for easy or difficult polygons, and eventually difficult enough tasks could be sent to the domain expert for disambiguation. A simple measure could be the entropy of the users' classes for that polygon - if most users decide for a class this measure will be low, if several different opinions arise, the polygon could be tagged as difficult. Entropy is essentially a measure of the uncertainty or disorder associated with a random variable. Entropy of the users in this paper is the disagreement about the class for certain polygon. The higher the entropy, the higher the confusion about the polygon.
- Polygons for which labeling task were often skipped by users could also be labeled as difficult for further evaluation by the domain expert.

- With enough labels collected, one or more simple agents could be developed and applied to use the label on some well-known polygons (polygons which were unanimously or almost-unanimously labeled by enough users) to label polygons with similar spectral/shape statistical features. The performance (reliability) of these agents could be measured using their decisions compared against the users' decisions for further fine-tuning.
- Although this is not expected to happen, the original labeling of some polygons by the expert could be verified against a large majority of the users' decisions. Differences on the chosen labels could indicate either a mistake on the decision from the expert or a difficulty on labeling a particular polygon, with the expert possessing information that is not available or known by the user base.
- Eventually, with enough data collected from an user, it is possible to model the users' abilities (related to the task being performed) to assess whether he/she is performing better with time.

All those analysis tasks (and others that may be considered in the future) using several different classes of algorithms, ranging from using basic statistics to clustering and classification techniques to data mining techniques [12][13]. In particular, it is expected that algorithms and approaches used to mine and explore recommendation systems [14], [15] and user activity modeling [16] could be successfully applied to the data collected in this setup.

4. Experiment

This section presents the experiment which uses citizen science to identify objects resulting from segmentation process applied in a satellite image with urban scenes (image of Sao Jose dos Campos city, state of Sao Paulo, Brazil).

The image size is 900x900 pixels (Figure 6) and was segmented using the traditional region growing algorithm available in the Spring software [11]. Some parameters of the segmentation algorithm control the number and size of the regions it creates and the similarity tolerance used to create those regions. The parameter values for the similarity between pixels and minimum size of each segment was 20 and 50, respectively. The segmentation process generated 2430 polygons.

Figure 6. Satellite image chosen for this study (pixel size: $1m^2$)

A website was developed to enable the process of data collection in which volunteer users are presented with polygons and a list of options for labeling those polygons. This site is available at http://www.lac.inpe.br/UrbanZoo (Figure 7).

The following classes (labels) for identification of targets of interest was established: *generic roof*, *ceramic roof*, *tin roof*, *cement asbestos roof*, *tree*, *street*, *swimming pool*, *shadow*, *field* (any kind of vegetation other than trees), *bare soil*, *water* and *mixture* (for when the polygon is composed of different classes of objects). These classes are shown to the user along with the options *none of the above* (in case the user knows the polygon class, but its class is not shown in the list of options) and *unknown* (when the user does not know which is the correct class for the polygon).

Users label the polygons by accessing the UrbanZoo site (which requires registering for identification purposes) and selecting one of the classes for a particular polygon. This is repeated until the user decides to finish his/her interaction with the site. Each interaction (presentation of a polygon and recording of the choice of label by the user) is stored.

This experiment began on April 26, 2010. Polygons were presented as follows: in the first phase of this study, polygons were shown randomly for each user.

Figure 7. Web interface to record users' decisions

After the first week of running the site, 43 users had provided 3000 labels with few repetitions (very few objects were labeled more than once). At this stage we could not perform any kind of analysis due to small number of repetitions.

In the second phase users were presented with a list of 13 polygons which were specifically chosen and labeled by an expert user (directly on the database, instead of using the site). The volunteer users were not informed that the polygons were selected in a non-random way. This strategy serves two purposes: first, to obtain a significant number of votes from these known polygons which class was already known, so a measure of agreement with the expert user's opinion could be calculated; and second, to assess the reliability of each volunteer user (akin to a multiple-choice test). Two weeks after the new strategy of presenting objects to users, there were 6900 labels with a total of 56 users on the database. Some polygons were labeled up to 35 times and most of the polygons were labeled more than once.

In the third phase users were presented with a sequence of 20 polygons, of distinct classes. This phase has two goals: first, to ensure that the users label different classes of polygons for further analysis of their ability to recognize polygons of a given class; second, to verify whether non-randomness of the classes influences the user behavior. One week after the beginning of this phase there were 7530 labels with a total of 63 users.

In the fourth phase, users were presented with a sequence of 25 polygons in complex and simple shape. The purpose here was to try identify whether polygon with more irregular shape cause more difficulty in correctly identifying its class. Until this moment, there are 69 users and approximately 9500 labeled polygons.

5. Analysis of User Behavior

5.1. User Proficiency versus Involvement

This section examines whether there is any relationship between the proficiency of the users and their involvement with the project. Different systematic tests were applied to users.

The first test was applied in the second phase of the experiment (started on May 4 and completed on May 23); 20 of the participating users were asked to label 13 known polygons. These polygons were previously labeled by an expert and presented to the users with the same interface and without any hint that their labels were already known; so the volunteers did not know that they were being evaluated. Figure 8 shows the proficiency (measured as the rate of correct choice/decision of labels for those polygons) for users.

One can observe that 13 users (65%) correctly labeled more than 50% of the polygons, with only 3 users (15%) incorrectly labeled less than 20% of the polygons. Although the test was a simple one, the results are considered encouraging. This is a number considered good given the imprecise nature of data and untrained users.

Another analysis that can be done is the calculation and evaluation of a consensus for each polygon that was part of each test cited here, but this analysis is not within the scope of this paper. The consensus for each polygon that was part of the test can be seen in [17], which was conducted with the test still in progress. The analysis showed that there were more disagreements for polygons of the classes ceramic roof, field, tree, water and cement asbestos roof. During this experiment, one domain expert has visited the site and labeled 484 polygons, among which 438 were distinct. In [18] the entropy for each class which was both labeled by the expert at least once and labeled by the users at least five times was calculated based on the users' decisions.

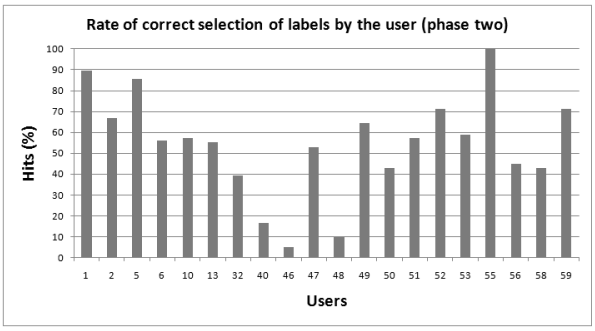

Figure 8. Rate of correct selection of labels by user on the second phase of the experiment (using known polygons)

In the second test of the experiment (applied during the third phase, started on May 24 and completed on June 16), 26 users were presented with a list of known polygons of classes field, tree, exposed soil, street and shadow. The polygons were presented sequentially according to the class, i.e, first all the polygons of the class field were presented, then all the polygons of the class tree and so on. As mentioned in Section 3, the objective of this test is to verify if the presentation of a sequence of polygons in the same class leads the user to commit fewer mistakes.

One can be note in Figure 9 that 12 users (46%) were able to correctly identify 50% or more of the polygons presented in this fashion, although a considerable number of volunteers (8 volunteer or 31%) correctly identified 20% or less than 20% of the polygons.

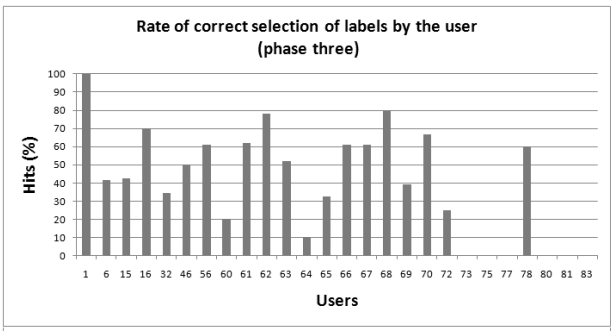

Figure 9. Rate of correct selection of labels by user on the third phase of the experiment (presenting a sequence of polygons of the same class)

The test two was performed after being identified, through analysis of the collected labels, some confusion among chosen labels for polygons of classes tree, field, shadow and exposed soil classes. This confusion is due to the fact that reflectance of the classes tree and shadow has a similar perceptual spectral value. The same confusion occurred among classes field and exposed soil.

On the other hand, polygons of the class street, when automatically segmented, often are undersegmented, containing small regions of other classes in it such as trees, field, etc. This might be the reason for the users to choose classes other than street. It is important to point out that the users were presented with the label option mixture that should be used in the case that some polygons present mixtures of classes due to the segmentation process. However, even though this option has been presented, it was observed that most of the volunteer users opted for a different class.

The third test was applied in the fourth phase of the experiment (started on June 1 and completed on June 16). In the fourth phase the users were presented with a list of 25 polygons where a majority of those have a complex shape, which presents some difficulty due to the discontinuity of the shape (see example in Figure 11). Figure 10 shows the correct labeling rate in this test and can be seen that just a few of the volunteers (8 users or 30%) correctly identified more than 50% of the presented polygons, and 13 users (48%) correctly identified 20% or less.

Although the shape of the polygon does not have any direct relationship with their class, it can be concluded that polygons with a complex shape cause more confusion, since they may be more difficult to identify or may confuse the volunteer who may be expecting in several cases a simple, regular man-made shape. For example, a roof usually has a rectangular shape, but that is not always correctly achieved by the segmentation process due to the imprecise nature of the algorithms and of the image itself. Figure 11 shows an example of polygon of ceramic roof class with a complex shape.

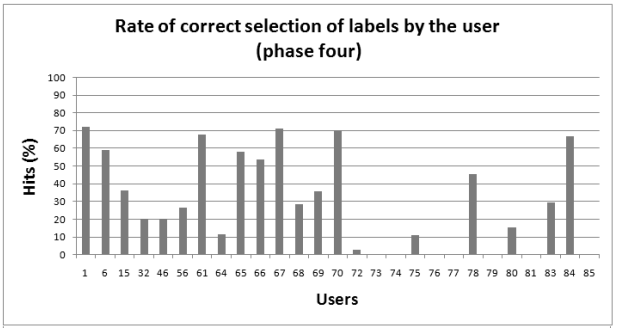

Figure 10. Rate of correct labeling by the users on the fourth phase of the experiment (presenting polygons with complex shapes)

Analyzing the graphs shown in Figures 8, 9 and 10 it may be noted that users 1, 6, 32, 46 and 56 had participated in all tests. Just for comparisons and further analysis, these results are summarized in Figure 12.

Figure 11. Polygon of the ceramic roof class generated by segmentation

Looking at the chart in Figure 12 and the chart of total of classifications by user in Figure 13 it may be noted that the number of classifications does not influence the proficiency of the user. That is, one can not say with certainty that the user becomes better by involving him/her in more labeling. In Figure 12, user 32 has a low rate of success. Looking at Figure 13, one can note that this user has a large number of labeling. One the other hand, users 1, 6 and 56 had a high hit rate and had a smaller number of interactions.

5.2. Entropy of Users' Decision per Class

In order to identify the classes that are more difficult for the users (i.e. the classes which are harder to label) the entropy of the users (users who participated in all tests) for each class was calculated. These values are summarized in Tables 1, 2 and 3, corresponding to the three tests. The entropy with value 0 means that the users hit all the polygons for that class. Low entropy means low confusion, and high entropy means high confusion.

Test one covered the classes water, field, cement asbestos roof, ceramic roof, street, swimming pool, bare soil (or exposed soil) and tree. Test two covered the classes field, street, bare soil, tree and shadow. Test three covered the classes field, cement asbestos roof, ceramic roof, street, tree and shadow.

Looking at Table 1, one can observe the confusion made by users. In this test one, most of the users, except users 32 and 46, presented entropy 0 or low entropy for classes water, field, street, swimming pool, bare soil and trees. Users 32, 46 and 56 presented high confusion for class ceramic roof when compared with other classes. But user 32, unlike others users, presented high entropy for class tree.

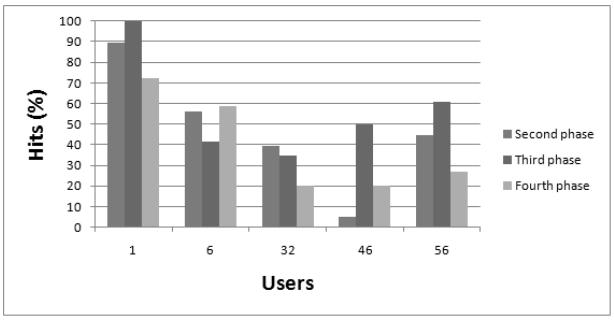

Figure 12. Rate of correct labeling by users which participated in all tests

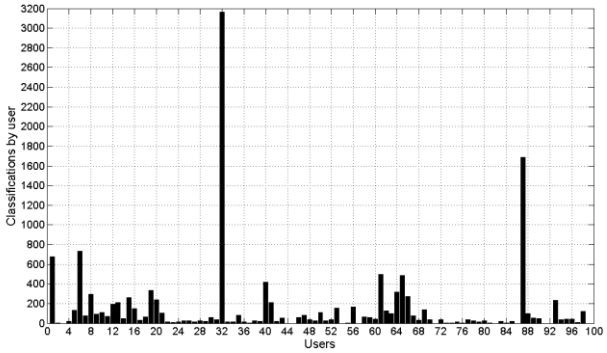

Figure 13. Total of classifications by user

Table 1. Users' entropy for each class in phase two (test one). CAR: Cement Asbestos Roof; CR: Ceramic Roof; BS: Bare Soil

User	Water	Field	CAR	CR	Street	Pool	BS	Tree
1	0	0.81	0	0.92	0	0	0.97	0
6	0	0	0	0.92	0	0	0	0
32	1.24	1.22	1.3	2.6	0	0.68	0.59	2.67
46	1.52	1.58	1.58	3.17	1.92	1.58	1	0
56	0	0	1	2.58	2	1	0	0

Looking at Table 2, one can observe that user 1 hit all the classes. User 6 presented high entropy for class tree, unlike that presented in test 1. User 32 kept the entropy 0 for street class, but had an increase in the entropy for classes field and tree.

User 46 did not label polygons of the classes street, bare soil and shadow in this test, but had a decrease in the entropy for class field, when compared with test one, and had confusion for class tree. User 56 presented confusion only for class tree.

Table 2. Entropy of users for each class phase three (test two)

User	Field	Street	Bare soil	Tree	Shadow
1	0	0	0	0	0
6	0	0.92	0	3.93	0
32	2.84	0	2	3.68	2
46	0	-	-	1	-
56	0	0	0	1	0

Looking at Table 3, user 1 presented entropy 0 for almost all the classes, presenting low confusion for classes field and cement asbestos roof. User 6 had a decrease in the entropy for street class, when compared with test two, and had an increase in the entropy for classes field, cement asbestos roof and ceramic roof when compared with test one. User 32 had a decrease in the entropy for classes cement asbestos roof and shadow and had an increase for street class. User 46, in this test, had labeled only polygons of the classes field, ceramic roof and tree. For class field, this user kept the same entropy as in test one. For class ceramic roof this user had a decrease in the entropy when compared with test one and had an increase in the entropy for class tree when compared with test one and test two. User 56 had a decrease in the entropy for class ceramic roof when compared with test one, had the same entropy for class tree in test two and three and presented entropy 0 for the other classes.

Table 3. Users' entropy for each class in phase four (test three). CAR: Cement Asbestos Roof; CR: Ceramic Roof

User	Field	CAR	CR	Street	Tree	Shadow
1	0.81	0.92	0	0	0	0
6	1	1	1.92	0	0	0
32	4	0	4.18	2.58	1.5	0
46	1.58	-	1.58	-	2	-
56	0	0	1.58	0	1	0

Classes such as field, ceramic roof and trees have lead users to a greater confusion. There is a reason for this. Class tree has an approximate spectral value as water, and as there are few polygons representing water, this confusion may have been enhanced with respect to trees. Water and shadow lead to less confusion. This might have occurred because there are only few polygons for these classes. The class ceramic roof and bare soil have an approximate spectral value, and this may have caused confusion. The swimming pool class has a distinct spectral reflectance leading to less confusion. It is important to reflect and accept that the labels may have to be renamed so that the confusion in labeling is less. As an example, the class field that had caused high confusion can be mentioned. It was assumed that this class is a field just with plants and grass but has neither trees nor bare soil. Probably the change of the label in the options might lead to less confusion and this will be evaluated in a future experiment.

6. Final Comments and Future Work

Considering the imprecise nature of polygons and untrained users, the users' proficiency in tests can be considered good. This reinforces our beliefs on the benefits of using citizen science for labeling of imprecisely segmented images and encourages the continuity of this work. The results are acceptable from the scientific point of view and it is possible to acquire enough data for a timely usage. The results and methodology also indicates that it is possible to use data collected from interactions of volunteers of citizen science projects and tasks assigned to them to extract information about the users' behaviors, both individually and collectively.

Since the collected data is not associated with users' personal data (e.g. names, professions, degrees, etc.) it was not possible infer relations between the users' personal or professional characteristics and their performance, which would be very interesting and could open further possibilities for customization of tasks for some users.

The next steps in this research are to study whether it is possible to:

- Label the yet-unlabeled polygons on a segmented image, using proper measures of reliability. Complement knowledge that can be partially extracted from a domain specialist (which, for reasons stated before, cannot do the whole labeling task alone) and eventually use this knowledge in similar labeling tasks e.g. statistics for a particular label will be most reliable when there are many different labeled polygons.
- Identify and possibly relabel polygons that may received an incorrect label by the domain specialist (e.g. based on a large number of different opinions from the users).
- Model the knowledge of the users, assessing his/her performance in general and related to specific classes (users may perform differently depending on the types of image regions presented to him/her) for further analysis and usage.

Many questions may be answered when a large enough amount of data is collected. Some them are:

- Can patterns and trends between the users be identified (e.g. identification of groups of users who tends to perform better with some labels than others)?

- Is it possible to evaluate temporal changes on users' performance?
- Is it possible to model the knowledge of a specific group of users (e.g. the most precise or reliable) to get information to try and automate the labeling task?

It is expected that more questions will appear as more and more data is collected and the basic analysis is performed.

7. References

[1] M. J. Raddick, G. Bracey, K. Carney, G. Gyuk, K. Borne, J. Wallin, and S. Jacoby, "Citizen Science: Status and Research Directions for the Coming Decade", in astro2010: The Astronomy and Astrophysics Decadal Survey, ser. ArXiv Astrophysics e-prints, vol. 2010, 2009, pp. 46P–+.

[2] C. Jung, "Unsupervised multiscale segmentation of color images", *Pattern Recognition Letters*, vol. 28, no. 4, pp. 523–533, 2007.

[3] S. Droedge, "Just because you paid them does not mean their data are better", http://www.birds.cornell.edu/citscitoolkit/conference/proceeding-pdfs/Droege%202007%20CS%20Conference.pdf, March 2010.

[4] D. Brossard, B. Lewenstein, and R. Bonney, "Scientific knowledge and attitude change: The impact of a citizen science project", *International Journal of Science Education*, vol. 27, no. 9, pp. 1099–1121, 2005.

[5] C. Willyard, "Using Citizens in Science Research", http://www.earthmagazine.org/earth/article/212-7d9-4-1e, April 2009.

[6] T. R. Williams, "Reconsidering the History of the AAVSO", *Journal of the American Association of Variable Star Observers*, vol. 29, no. 2, pp. 132–147, 2001.

[7] T. Root, *Atlas of wintering North American birds: An analysis of Christmas bird count data*. University Of Chicago Press, 1998.

[8] T. Sagaran, *Programming Collective Intelligence – Building Smart Web 2.0 Applications*. O'Reilly Media, Inc., 2007.

[9] S. Alag, *Collective Intelligence in Action*. Manning, 2009.

[10] B. Liu and P. S. Yu, *The Top Ten Algorithms in Data Mining*. Taylor & Francis, 2009, ch. PageRank, pp. 73–100.

[11] G. Câmara, R. C. M. Souza, U. M. Freitas, and J. Garrido, "SPRING: Integrating remote sensing and GIS by object-oriented data modelling", *Computers & Graphics*, vol. 20, no. 3, pp. 395–403, 1996.

[12] R. Santos, "Conceitos de Mineração de Dados na Web", in XV Simpósio Brasileiro de Sistemas Multimídia e Web, VI Simpósio Brasileiro de Sistemas Colaborativos – Anais, M. M. Teixeira, C. A. C. Teixeira, F. A. M. Trinta, and P. P. M. Farias, Eds., 2009, pp. 81–124.

[13] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Application)*. Springer, 2007.

[14] S. S. Anand and B. Mobasher, "Contextual recommendation", in From Web to Social Web: Discovering and Deploying User and Content Profiles – Workshop on Web Mining, WebMine 2006, Berlin, Germany, September 18, 2006, Revised Selected and Invited Papers (LNAI 4737), B. Berendt, A. Hotho, D. Mladeni˘c, and G. Semeraro, Eds., 2007, pp. 142–160.

[15] B. Mobasher, X. Jin, and Y. Zhou, "Semantically enhanced collaborative filtering on the web", in Web Mining: From Web to Semantic Web – First European Web Mining Forum, EWMF 2003, Cavtat-Dubrovnik, Croatia, September 22, 2003, Invited and Selected Revised Papers (LNAI 3209), B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, and G. Stumme, Eds., 2004, pp. 57–76.

[16] E. Fras-Martnez and V. Karamcheti, "A customizable behavior model for temporal prediction of web user sequences", in WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profiles, 4th International Workshop, Edmonton, Canada, July 23, 2002, Revised Papers (LNAI 2703), O. R. Zaane, J. Srivastava, M. Spiliopoulou, and B. Masand, Eds., 2003, pp. 66–85.

[17] M. D. Soares, R. Santos, N. Vijaykumar, and L. V. Dutra, "Citizen Science-based Labeling of Imprecisely Segmented Images: Case Study and Preliminary Results," in Simpósio Brasileiro de Sistemas Colaborativos, IEEE Computer Society, 2010, v. I., pp. 87-94.

[18] M. D. Soares, R. Santos, N. Vijaykumar, and L. Dutra, "Analysis of User Behavior and Difficulty in Labeling Polygons of a Segmented Image in a Citizen Science Project". To be published in Journal of Computer Science, (Impresso), 2011.