Phase I Grand Challenges Explorations Financial Report

Organization Name: Universidad Nacional Autónoma de México
Project Title: Software to identify and quantify pathogenic helminth eggs
Report Type: (Final)
Reporting Period *From* **(01/11/2011)**
To **(12/06/2013)**

Total Expenses by Type	PHASE I PROJECT: ACTUAL EXPENSES (in US dollars)	
Personnel	$6,142.31 (6.14%)	
Equipment	$4,765.71	$23,404.39 (23.40%)
	$292.23	
	$2,063.26	
	$948.26	
	$292.23	
	$2,020.56	
	$284.11	
	$2,225.93	
	$1,601.69	
	$4,285.31	
	$2,809.79	
	$1,815.31	
Travel	$0.00	
Consultants	$0.00	
Supplies	$43,019.12 (43.02%)	
Subcontracts	$27,434.18 (27.43%)	
Project Expenses Subtotal	$100,000.00	
Other sources of project support	$100 gift (a genera of HE from Brazil) $4,000 (lab work) $11,015.34 (equipment and supplies) $5,000 (patent) $5,000 (lab work)	
Total Project Cost	$125,115.34	

Phase I Scientific Final Report

UNAM TEAM

Software to identify and quantify pathogenic helminth eggs

The analytical technique used to identify and quantify helminth eggs, besides its time consuming nature, involves the critical step of identification of the eggs. This is the major source of uncertainty in quantifying results, because microbiologists are not capable of performing it without a long training programme. This project is focused on creating software that will automatically perform helminth egg identification and quantification, reducing the consumption of materials and expenditure on highly trained personnel while obtaining reliable results.

Activities

Setting a workstation

The first task to begin the project was to put together a microscopic image processing workstation, which was set to acquire digital images of different helminth egg species. This was possible once the grant funds were available.

Generating helminth eggs image database

With the workstation ready, several helminth eggs images were acquired including approximately 23 genera and 36 representative species from different wastewater and sludge samples. This led to conform a digital database of up to 720 helminth eggs images. An optical microscope Carl Zeiss AxioLab A1 and two digital photomicrography cameras were utilized to reach this objective. From this compilation, a few species were formerly selected to study their particular characteristics (mainly size, shape and texture). This was made by using very detailed and clear images, but also pictures where the target structures were surrounded by an artifact saturated environment, reproducing the most common difficulty levels on the examination of sample. The species utilized for this objective were selected, based on their medical importance and worldwide ubiquity: *Ascaris lumbricoides* as fertile and unfertile*, Trichuris trichiura, Toxocara canis* and *Taenia saginata.*

Developing identification software

The next step was the development of the software. Given the complexity of analyzing biological images, to obtain the most suitable system to identify helminth eggs, a comparative study of the available recognition protocols and image processing techniques was performed.

A specific image processing protocol for the helminth eggs was developed. This algorithm had two different parts: the first one consisted in detecting and labeling all of the objects in an image using a median filter and carrying out an image treatment including adaptive histogram equalization, edge detection and watershed. The second part of the algorithm performed the classification of each detected object. This classification is the resultant of measuring shape properties (area, perimeter and eccentricity) and texture properties (energy, mean gray level, contrast, correlation and homogeneity) for each studied object. 360 images regarding the selected species were used as training pictures to establish their range of values for each classification property. One class should be assigned to each helminth species and another one should be assigned to objects that appear in the images but are not helminth eggs, like detritus, bubbles, pollen particles, etc.

By using such training data, it was possible to identify a new object using a nearest neighbor classifier based on Mahalanobis distance, which is a function that weighs the similarities

between an object in the sample that is intended to be identified, and the reference images of helminth ova from the database.
With this algorithm an average sensibility of 0.85 was obtained among the four genera tested. The sensibility was the measure of correct identification of each species, regarding the number of true positive and false negative identifications. The fertile *Ascaris* recognition was the one that yielded the lowest sensibility of 0.66, and the identification of unfertile *Ascaris* had a sensibility of 0.86, as was also similar for *Toxocara* and *Trichuris*. *Taenia* yielded 0.80 of sensibility and the correct identification of objects different than helminth eggs, was 0.88 effective.
At this point it was considered that the system needed to be improved and it was also necessary to increase the number of involved species for identification (*Hymenolepis nana, Hymenolepis diminuta* and *Schistosoma mansoni*). The additional genera were selected, in the case of *Hymenolepis*, because of the difficulty level of identification and, in the case of *Schistosoma*, because it is a widespread genus especially in Asia, Africa and South America.
To accomplish the goal, new image processing tools and changes were applied to the software development. The first modification was a new image filter algorithm. The median filter that was used in the previous results had problems preserving the borders of the eggs, while the anisotropic diffusion filter allowed establishing more clearly the border of each identified object while smoothing the rest of the image.
Thus, software changes were made and the system was subject to more and deeper validation tests, including three different water quality conditions.

System improvement

One of the changes made to the system was the utilization of an anisotropic filter instead of the median filter used in the earlier version. The figure 1 presents, so far, the sequence algorithm describing the main steps performed by the system.

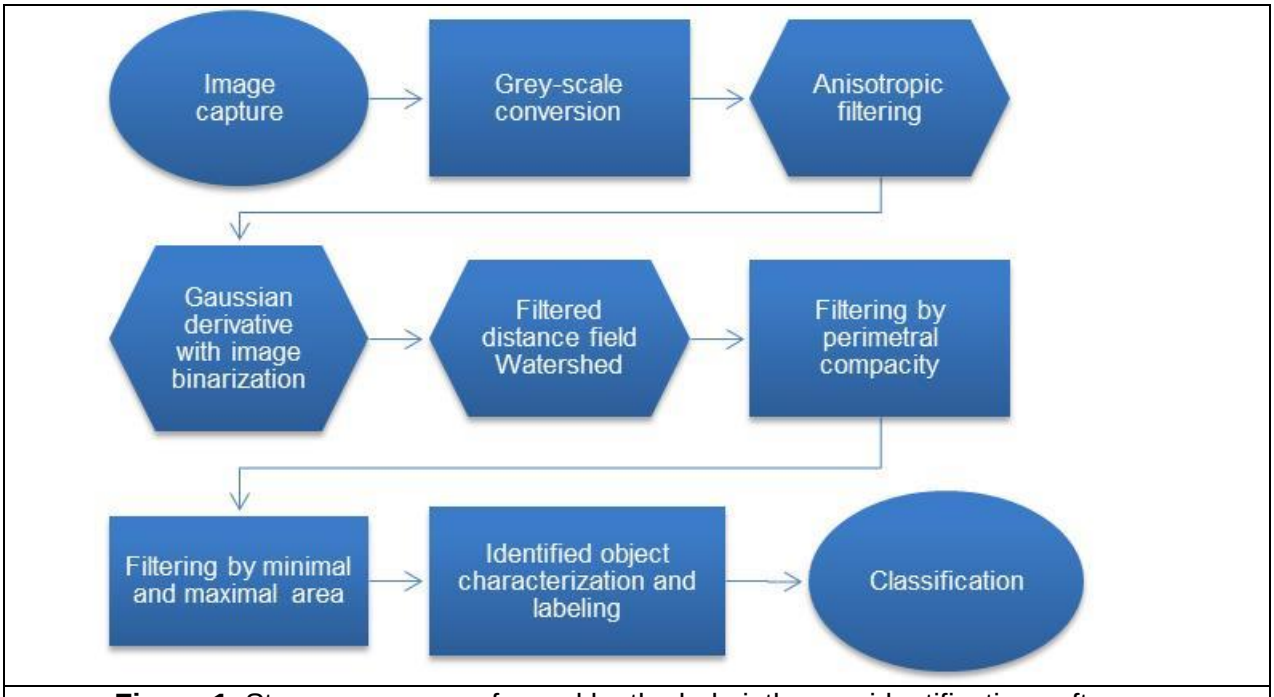

Figure 1. Step sequence performed by the helminth eggs identification software.

For the last modifications of the system, a step of segmentation with grey-scale profiles based on vectors was put in place but, to obtain better results, it was eliminated from the protocol, and

new changes were performed seeking superior identification performance. At this point a series of filters were added to the image processing software. In summary, the final version of the system for this first phase of the project, includes a Gaussian derivative protocol that is useful to better identify the borders of the object, yielding at the end a binarized image, which afterwards is filtered using a distance field watershed procedure, that allows the extraction of the borders of the objects by an improved segmentation of the identified objects. The perimetral compactness is evaluated using the shape of the object and the refilling of its spaces, and after that, a filtering regarding the minimal and maximal area is applied.
In figure 2 the complete image processing is shown for a *Schistosoma* egg structure, as an example of what the system can do so far.

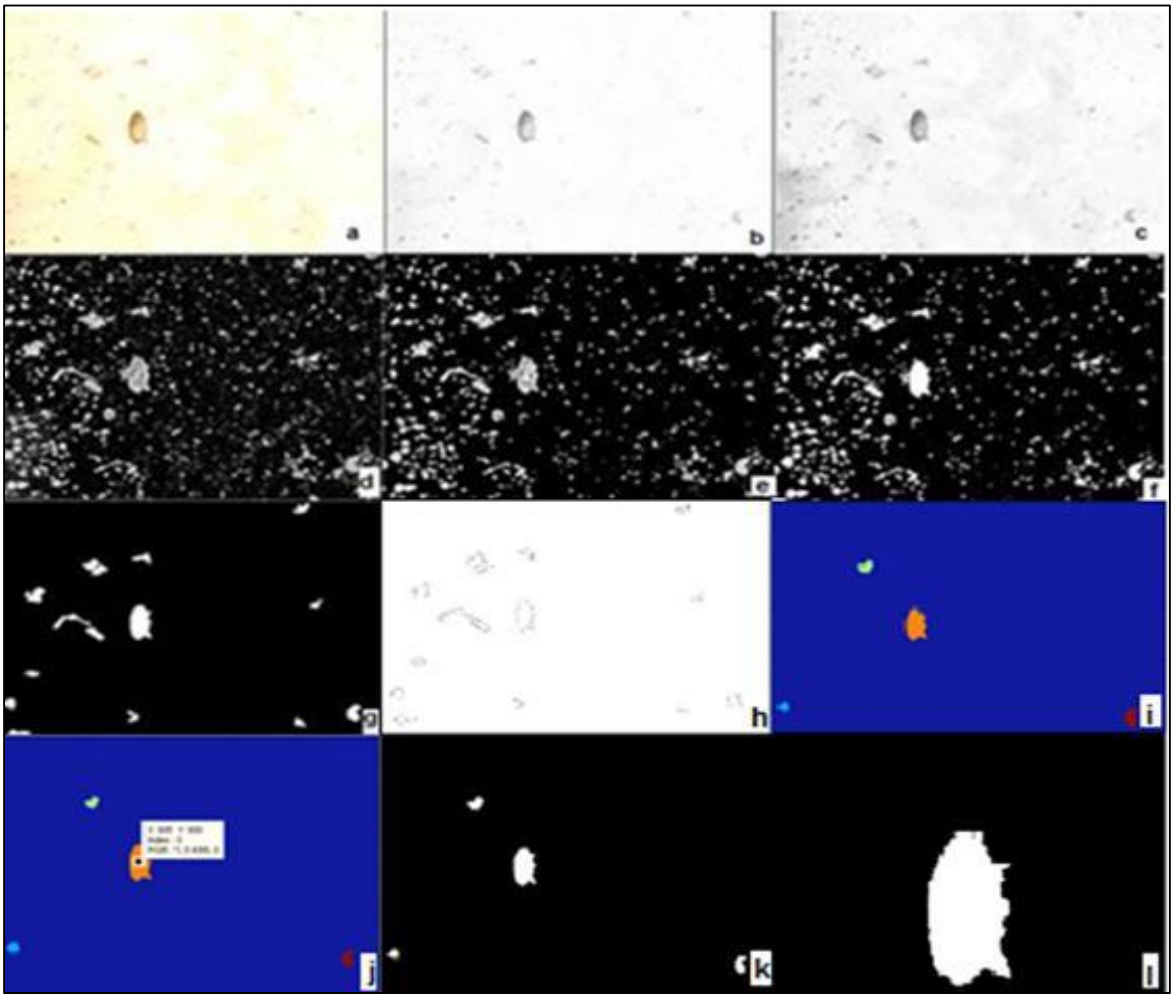

Figure 2. Complete image processing of a *Schistosoma mansoni* egg: a) Image capture, b) Grey-scale image, c) Anisotropic filtering, d-g) Gaussian derivative with image binarization, h) Watershed segmentation, i) Filtering by perimetral compactness and by minimal and maximal area, j) Object labeling and characterization, k) Naïve bayesian classification and l) Zoom of the final image of the egg to be classified.

Based on the isolated structures from the image, the system performs the classification of each object according to its characteristics. To classify the objects, a naive bayesian classifier is utilized regarding properties like area, diameter, eccentricity, compactness, entropy, perimetral

roughness and size of minor/major axis. With this sequence of steps, the first phase of the project was considered completed, and a validation was carried out once again to determine the identification and quantification efficiency of the system against the microbiologist expertise.

System validation

A first validation was made for the system working without the Gaussian derivative procedure. Samples from three different wastewater quality levels were used and considered as class I for water containing less than 10 mg/L of total solids, class II for water containing 10 to 150 mg/L of total solids, and class III for total solids content above 150 mg/L. According to the amount of solids, the validation tests were directly comparable with the wastewater source. A comparative quantification was made between the traditional technique of direct observation, and the performance of identification of the system. Direct microscope quantification (Carl Zeiss, model Axiolab A1) was made by four different microbiology expert observers, and simultaneously pictures of the same observed samples were taken and analyzed by the system. The parameters used to evaluate the proficiency of the system were:

Sensibility: It corresponds to the percentage of true positives as follows:

$$Se = T_p / (T_p + F_n)$$

Where T_p is the number of true positives and F_n is the number of false negatives, regarding earlier as the number of helminth eggs correctly identified and the latter as the number of existent helminth eggs that were not identified correctly in the sample.

Specificity: It is the percentage of the true negatives as follows:

$$Sp = T_n / (T_n + F_p)$$

Where, T_n is the number of true negatives and F_p is the number of false positives, regarding the earlier as the number of other objects different than helminth eggs that were identified correctly, and the latter as the number of other objects different than helminth eggs that were wrong identified as eggs.

The results obtained for the system without the Gaussian derivative, were for those samples of water class I, where the eggs were almost completely free of other confusing objects, and for samples of water class II, which had an amount of solids equivalent to an effluent of the clarification step of a wastewater treatment plant. For both cases, the average specificity was considerably high (0.99 and 0.98, respectively) indicating that the system was capable of distinguish with significant accuracy between a helminth egg and other different objects. The sensibility results were lower than specificity (0.83 and 0.80, respectively), indicating that the ability of the system to identify one species among the others was at least 80% effective.
Although the specificity and sensibility values were not so different for water class I and II, the results yielded that the system might be limited by the amount of solids present in the sample. This was better evidenced for the case of water class III (raw wastewater), where the amount of solids was so high that the eggs recognition became much more difficult, and the identification effectiveness was significantly lower (less than 15%) than for those samples class I or II.

After the introduction of the Gaussian derivative to the system, a sensibility of 0.90 and a specificity of 0.99 was obtained for the water class I, proving the effectiveness of the last changes that have been described in the improvement section. So, according to the system

current conditions, if it was necessary to analyze high solid content wastewaters (class III), a previous dilution of the samples should be performed to obtain better identification results.
In summary, the project team has created a system capable of detecting helminth eggs with 99% of effectiveness distinguishing between an egg and other objects, and of 90% distinguishing between a specific egg species from another egg species, which is a significant achievement at this point. The table 1 shows the advantages of the software procedure over the conventional identification method.
The main advantage by using this system is how fast a sample can be identified and quantified, and that an expert parasitologist is not required to perform this task. The time consuming is significantly reduced with no more than 10 minutes per sample using the system, in comparison with the hours that only one sample might take to a technician exclusively to carry out the identifying and counting step. Thus, the limitations of the human expertise, either in terms of ability as of availability, are covered leading to a decrease on the identification costs, and bringing to a much broader group of instances and communities the possibility to have prompt and reliable helminth eggs analysis.
This system will be of the utmost convenience for the water quality surveillance facilities and organizations of all countries around the world, but especially to low income ones in which the incidence of this kind parasites is higher, playing an important role in the life's quality improvement of such regions that nowadays represent more than 50% of the world's population.

Challenges

New changes and strategies could be taken to increase even more the capabilities of this system and, though we are still in the beginning, we are thinking beyond to upgrade such capabilities to be able to process even sludge, biosolids and excreta samples.
We expect that this development shall be very useful in many regions, especially where water quality assurance is yet a challenge.

Other sources of project support

Schistosoma genera ova were kindly donated by a Brazilian laboratory; a small spend related to this is included in the proper section at the final financial report, as well as the over budget expenditures that were necessary to continue the project, and that were all covered by the University.

Table 1. Advantages of the procedure using the developed software compared to the conventional identification method.

Issue	**Conventional method**	**Procedure with the developed software**
Skills required for identification	2 to 3 weeks of personnel training at minimum	Not required
Time consuming for identification step	Clean samples 1 hour Dirty samples 3 hours or more	One sample each 10 minutes maximum
Initial cost of the system	USD $ 20,000	USD $30,000
Cost for the identification step	Clear samples USD $10 Dirty samples USD $35	USD $2 at maximum (any kind of sample)
Sensibility (Se)	Human skill dependent	>90%
Specificity (Sp)	Not determined	99%