IMAGE ANALYSIS OF SOYBEAN LEAF DISEASES USING METADATA

Thiago Evaristo Wenceslau Moreno[1], Albino Szesz Junior[1], Marcos Monteiro Junior[1], Diulhio Candido de Oliveira[1], Maria Salete Marcon Gomes Vaz[1,3], Lucélia De Souza[2,3], Tatiana Montes Celinski[1], Thalita Scharr Rodrigues[1]

[1] State University of Ponta Grossa. E-mails: thiagowmoreno@gmail.com, albinojr12@gmail.com, marco.monteiro@gmail.com, lolcaos@yahoo.com.br, salete@uepg.br, tmontesc@uepg.br,thalitascharr@gmail.com

[2] State University of West Center. E-mail: lucelia@unicentro.br

[3] Federal University of Parana. Postgraduate Program in Informatics.

Abstract. *The aim of this paper is analyze images of foliar diseases of soybean and extracting metadata to describe aspects of the image and also verify if the leaf is sick or not. To enable this, was developed an integrated environment with routines in C Language. A standard of metadata was created, it was utilized the Canny algorithm to analyze the image and create a prototype interface to see the results. It was possible to detect the edge of images submitted to the algorithm, however there was a necessity to better filtering of noise.*

Keywords. *leaf disease, image processing, metadata.*

1. INTRODUTION

Images are widely used as resources in solving physical problems due to the ease of obtaining content, spread on the Internet. However, often the image processing becomes more difficult due to the loss of quality due to several factors: (i) change of format, e.g. JPEG and BMP extensions, (ii) scanning for various devices such as scanner machines and digital, and (iii) compression, reducing the size of the image.

The computational needs involved in the application must be the way how information is assimilated by the user. The human brain has a great capacity for assimilation, but there are difficulties in differentiating colors. When faced with two very similar colors often do not notice the difference, however you can use computer vision to aid in accurate diagnosis.

Some information can be extracted from the images through metadata (TANNEMBAUM, 2001), but this content should be indexed, using a database. Indexing can be done through a set of metadata, either manually (by user) and/or in automatic mode, where the system identifies the metadata and indexes.

These forms of indexing have their advantages in performance and footprint, where the semi-automatic allows manual indexing and automatically linking of space and memory performance (BOLETTIERI et al, 2007). The semi-automatic indexing will be performed to characterize diseases of soybeans in the context of foliar diseases.

Metadata is used to document and organize in a way structured, data from organizations in order to minimize duplication of efforts and facilitate the maintenance of such data (PRABHAKARAN, 1997).

The process of automatic generation of metadata, without human interaction, can result in wrong information and/or redundant. Therefore, it is necessary to carry out various experiments to minimize these problems. For automatic extraction of metadata, is necessary to define the characteristics of the problem by setting standards and methods.When it comes to images, and the metadata contained in the file header, such as author name and plant pathology and creation date, you must perform a treatment to remove the image metadata/values based on image content. For example: (i) Format: specifies the default format for file storage, such as TIFF, GIF, BMP, JPG, among others, (ii) Compression is the process of reducing the image size by methods such as withdrawal repetitive information or information taken from barely detectable to the human eye. There are some standards for compressing QuickTime, LZM, JBIG, among others, (iii) Type of scanner: scanner, digital camera, among others, (iv) Resolution: 640 x 480, 1366 x 766, among others, (v) Space Color: CMYK, HSV, RGB, (vi) Color Histogram: Represents the distribution of colors in the digital image according each pixel, (vii) Texture: refers to the repetition of basic elements (*texels*), and (viii) Brightness: proportional to the integral of the product curve and the efficiency function of luminosity (FOLEY et al, 1990).

With specialized algorithms, these characteristics can be manipulated for human understanding, resulting in benefits for your application. An example is the soybean rust, a disease where your first symptoms are tiny, with points around 1 mm in diameter, darker than the healthy tissue of the leaf (SINCLAIR; HARTMAN, 1999).

With trained algorithms, it is possible to obtain a proportion of contaminated tissue, and may help in decision making to determine preventive control actions or emergency, resulting in cure or reduce the effects caused.

The objective of this research was to create an integrated environment with routines in C language for image analysis of foliar diseases of soybean and extracting metadata that describe aspects of the image and also to verify that the sheet is ill or not. With this information, it is expected to provide a decision-making in combat and/or control of soybean rust.

This paper is organized how follow. Besides this introductory section, the second section describes about metadata and standard in the agricultural environment, the third section shows the standard of metadata that is proposed, the fourth section describes the algorithm of image processing, and the fifth section relates the integrated environment, following of the conclusions.

2. METADATA AND STANDARD IN THE AGRICULTURAL ENVIRONMENT

Some of the main reasons that justify the use of metadata for image description, usually, involve the difficulty to do the exact match patterns; the difficulty of research-based content, since there is need for analysis in large data set; and, when a research-based content is done, it can not be done often due to performance reasons.

In the literature, metadata are generally defined as data about data, or data describing the attributes of resources, facilitating its location, discovery, documentation, selection and retrieval (DUVAL et al, 2002). Above all, the metadata form as an instrument that helps for the processing of the raw data in knowledge.

In an attempt to organize the metadata to facilitate these activities, there were several schemes, called Metadata Standards, as the context of applications, defining the information to describe the data, establishing the mandatory and optional, the semantics and syntax of the same (NISO, 2004).

The metadata available, describe, locate and assist in the understanding of the data, transforming them in knowledge. By having knowledge of what data are available, understand the context and where they are located, accurate information is obtained and better decisions can be made.

Metadata can also be classified as (TANNEMBAUM, 2001): *administrative*, being used in the management of information resources; *descriptive*, describing characteristics of a document, facilitating identification, research and information management; *preservation*, being those who safeguard the information; *technical* and related to system operation and behavior of metadata; *structural*, describing the way how the objects are interconnected and, finally, *use* metadata, which correlate with the type of use of technological resources.

The set of information corresponds to a metadata schema defined for any given context. By identifying problems in the storage and retrieval of information by lack of standardization, various schemes are created to serve different purposes, called metadata standards (NISO, 2004).

The use of standards has been viewed as a limitation of the developer community. Currently, with the growth of stored data, standards are seen as partners for documentation of resources. The use of standards results in benefits to the community, to facilitate the activity of analysis because, generally, are widely documented, and facilitate communication between users, providing uniformity and integration between solutions.

The use of metadata standards facilitates the decision of which metadata should be collected and maintained. Also, communication and interoperability are facilitated when different communities use the same standard and do not propose different types of metadata and/or adopt different vocabularies.

After research carried out in search of metadata standards used for multimedia platforms, we found the following standards: CSDGM - Content

Standard for Digital Geospatial Metadata (WESTBROOKS, 2003) and Dublin Core (ONYANCHA, 2001).

The CSDGM was developed by the agency of the U.S. government, Federal Geographic Data Committee - FGDC, was approved in 1994, coordinating the development, use, sharing and dissemination of geographic data.

The FGDC has established a single set of metadata, with 245 metadata elements, for the documentation of digital geospatial data, establishing the names of metadata and groupings, plus information about the amounts to be provided for each metadata element.

On the other hand, the Dublin Core is a standard of metadata utilized in the description of the electronic sources: documents, audios, images and clips video, showing characteristics that make one of the standards more utilized in the area.

The Dublin Core has two levels: simple and qualified. The simple level consists of fifteen elements of metadata, being: *Title*, *Creator*, *Subject*, *Description*, *Publisher*, *Contributor*, *Date*, *Type*, *Format*, *Identifier*, *Source*, *Language*, *Relation*, *Coverage* and *Rights*. The qualified level has three additional elements: *Audience*, *Provenance* and *Rights Holder* and a group of element refinements that are the qualifiers to refine the semantics of the elements facilitating the resource discovery.

In this paper, the standard was created based on the Dublin Core, which works with images, suitable for dealing with images of cultures. The Dublin Core has a schema for describing metadata that facilitates the representation of electronic resources, making them more visible to search engines and retrieval systems. The next section describes about the standard of metadata that was created in this work for description of the agricultural images.

3. STANDARD OF METADATA FOR DESCRIPTION OF DISEASE IN PLANTS

The standard of metadata for description of the agricultural images was defined considering three categories: Content, Intellectual Properties and Temporal (Table 1), including additional metadata for agricultural images. With this standard of metadata, will be possible doing image analysis of foliar diseases of soybean and extracting metadata to describe aspects of the image and also verify if the leaf is sick or not.

The category of *Content* metadata is described as follows. *Coverage* metadata corresponds to the range, location, extension of the data (geographic coordinates or place name). With the *Description* metadata, can be used vocabulary, which is used to describe the corresponding item in question. Generally, can be an index, figures and/or textual descriptions, in other words, all that is needed to describe the data. The *Type* metadata describes the nature or genre of the data.

Table 1: Standard of metadata for description of disease in plants.

Content	Intellectual Properties	Temporal
Coverage	Contributor	Date
Description	Creator	Format
Type	Publisher	Identifier
Relation	Rights	
Subject		
Source		
Title		
Target Audience		
Name of Disease		
Level of Severidade		
Leaves face		
Color		
Percentage		
Language		
Equipment		

The *Relation* metadata tells whether the data is part of, version, and format, is based on, or demands something. Already the Subject metadata defines the keywords given. It is recommended vocabulary.

The *Source* metadata tells where the source of the data. The Title metadata, match in which name the data is known. The *Target Audience* metadata determines which person (or group of people) is related. It is recommend use of vocabulary.

The *Name of Disease* metadata determines the name of the leaf disease. The *Level of Severity* metadata specifies the percentage of attacked leaf, defining a level (grade) of severity. This value is removed via software.

In the *Leaves Face* metadata, the attacked leaf (top or bottom) is described. It is because some diseases only attack one side of the sheet. The percentages of colors in the image are specified in the *Color* metadata, while the *Percent* metadata specifies the percentage of diseased leaf.

The *Language* metadata defines the language of the content, or the creators of the data. The Equipment metadata contains the image capture equipment which can be provided for future search and retrieval.

The category of *Intellectual Property* metadata is described as follows. The *Contributor* metadata defines the organization or person that contributes to the creation of the data. The *Creator* metadata defines the responsible by creation of the data. It might be company, person, software and/or equipment.

The *Publisher* metadata defines responsible for making the data accessible and can be company, person, software and/or equipment. The last type of this category is *Rights* metadata, which defines the copyright on the data, receiving the full text of the same copyright.

The category of *Temporal* metadata is defined as follows. The *Date* metadata refers to the date of creation or availability of data. The *Format* metadata defines the physical format or digital data and should be used with vocabularies. The *Identifier* metadata defines the unique identifier. Find a formal feature identifier for desired data.

The next section reports on the algorithm of image processing that was utilized in this work to edge detection.

4. ALGORITHM OF IMAGE PROCESSING

To satisfy an edge detection algorithm, some requirements are needed as (Canny, 1986): (i) error rate, where the edge detector to detect and find just the edges; no edge is missing, (ii) location where the distance between the edge pixels found by the edge detector and the current border should be as small as possible, and (iii) response, where the edge detector should not identify multiple edge pixels if there is only a single pixel. After filtering, we created another image showing the result of processing as to the delimitation of the border.

Among the algorithms of image processing found in the literature, o Canny algorithm (Canny, 1986) was choose because their characteristics that including simplicity, be widely utilized and efficient.

To doing edge detection, the Canny algorithm consists of the four steps: 1. Gaussian filter smoothing image to restrain noise; 2. finite difference of the first order partial derivative for calculating both the gradient magnitude and the direction of the image; 3. non-maximum suppression for the gradient magnitude and 4. to detect and connected edges, it is used the dual-threshold algorithm.

In this paper, besides of the Canny algorithm have many modified versions, was utilized the original version. In order to use the Canny algorithm, were selected images of soybean leaves to the required tests. After loading the image, we make a copy of the same in grayscale and this is subjected to the edge detection algorithm.

An integrated environment was developed in this work for image analysis which is described bellow.

5. INTEGRATED ENVIRONMENT

To implement the algorithm of image processing Canny (Canny, 1986), we used the NetBeans 6.9.1 IDE, with a nice interface and easy to use for programming the software routines.

The OpenCV Library was used, having License BSD (Berkeley Software Distribution), and free for academic use and commercial language written in C and C++ interface. It has several algorithms for image manipulation, allowing image filters, camera calibration and object recognition.

In this work, was developed an integrated environment (Figure 1) with routines in C language for image analysis of foliar diseases of soybean and extracting metadata that describe aspects of the image and also to verify that the sheet is ill or not.

It was observed with the results that finding the edges, there are still noise in the image, turning the analysis of leaf area complicated (MONTEIRO et al, 2011). The image processing produces a satisfactory effect, where the computer vision produces a more accurate diagnosis.

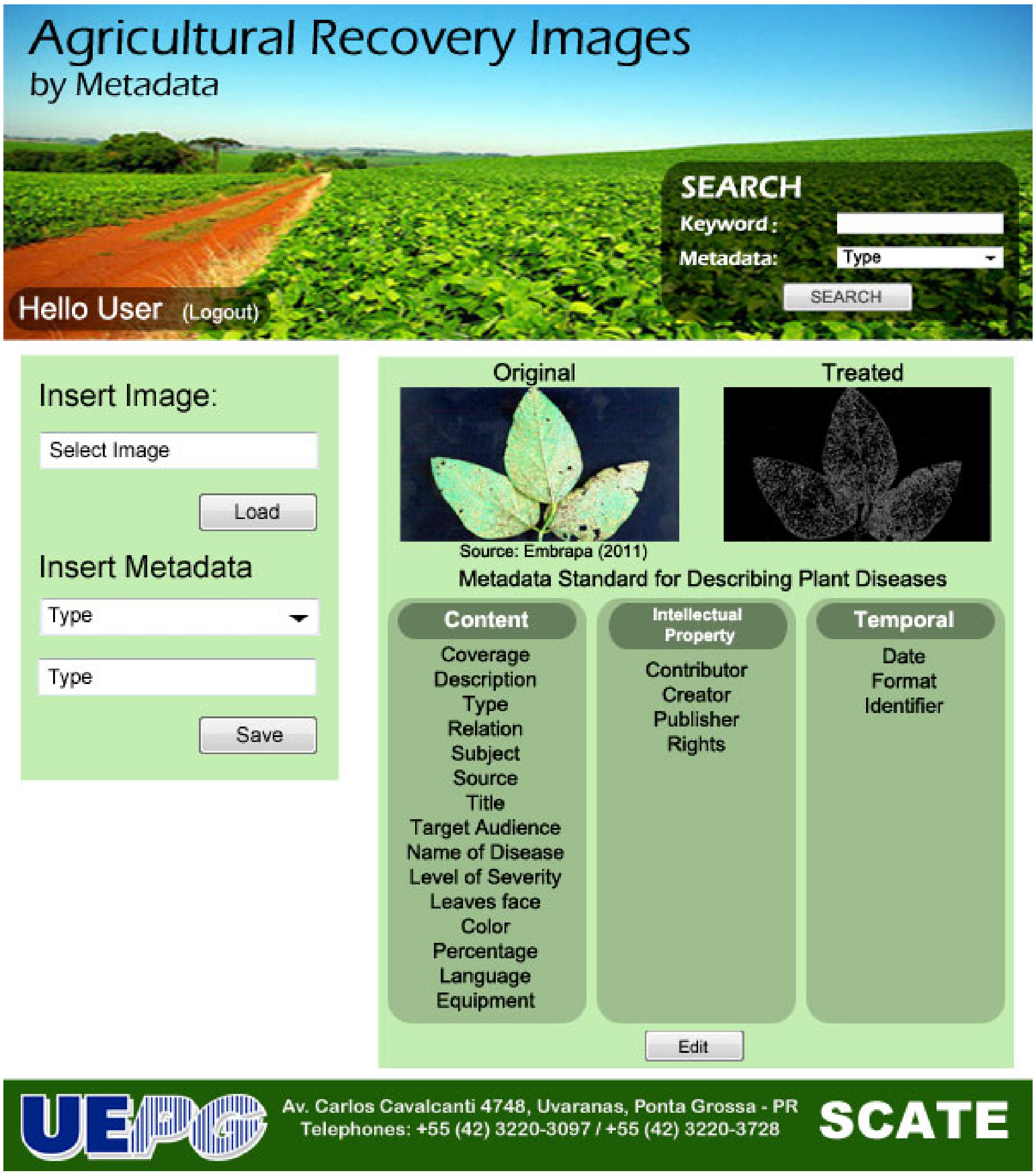

Figure 1 – Prototype of interface to recuperation of agricultural images.

However, the algorithm utilized does not yet meet all kinds of image, where some images with low resolution or low quality after using the Canny algorithm (Canny, 1986) had unwanted imperfections, not producing an accurate diagnosis. With this information, it is expected to provide a decision-making in combat and/or control of soybean rust. After treatment performed on the images, we obtained the results of Figures 2 and 3.

Figure 2 – Original Image

Figure 3 – Image After Treatment

The image can be treated overwrite the original image to define the working area of the sheet in order to be subjected to a segmentation algorithm to extract the percentage by color comparison between healthy and ill considered, generating the metadata type in question. The prototype of the interface (Figure 3) integrates the fields input, output and search on a single screen for easier viewing of information.

6. CONCLUSIONS

One of the most important tasks of the systems in multimedia database is the extraction of metadata values of the images dynamically. The creation of metadata standards for description of the images allows the formal description of the characteristics of soybean diseases.

The computer vision produces a more accurate diagnosis, when compared to human perception, providing the opportunity to analyze different parameters of the images and collaborating in the identification of diseases.

As future work, based on tests performed to detect the edges, it is concluded that the algorithm Canny that was utilized needs improvement in the filtering of the pixels to minimize the noise found in agricultural images used.

REFERENCES

BOLETTIERI, P.; FALCHI, F.; GENNARO, C.; RABITTI, F. **Automatic metadata extraction and indexing for reusing e-learning multimedia objects.** MS´07 Workshop on multimedia information retrieval on the many faces of multimedia semantics, 2007.

CANNY, J. **A computational approach to edge detection**. IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679-698, 1986.

DUVAL, E.; HODGINS, W.; SUTTON, S.; WEIBEL S. **Metadata Principles and Practicalities**. D-Lib Magazine 8(4). 2002. URL: <http://www.dlib.org/dlib/april02/weibel/04weibel.html>. Access in november/11.

FOLEY, J. D.; VAN DAM A.; FEINER, S. K.; HUGLES, J. F. **Computer Graphics: Principles and Practice**. Addison-Wesley Publishing Company, 1990.

MORENO, T. E. W.; MONTEIRO J. R. M.; OLIVEIRA, D. C.; SZESZ J. A.; YAMADA, M. A.; VAZ, M. S. M. G.; CELINSKI, T. M.; SOUZA, L.; RODRIGUES, T. S. **Ambiente Integrado Para Análise de Imagens de Doenças Foliares da Soja Utilizando Metadados.** In: VIII Congresso Brasileiro de Agroinformática, 2011, Bento Gonçalves, RS. Anais do VIII Congresso Brasileiro de Agroinformática, 2011.

NISO, National Information Standards Organization. **Understanding Metadata**. NISO Press. 2004.

ONYANCHA, I.; KEIZER, J.; KATZ, S. **A Dublin Core Application Profile in the Agricultural Domain.** Proc. Int'l. Conf. on Dublin Core and Metadata Applications, 2001.

PRABHAKARAN, B. **Multimedia Database Management Systems.** Kluwer Academic Publishers, 1997.

SINCLAIR, J.B.; HARTMAN, G.L. Soybean rust. In: Hartman, G.L., Sinclair, J.B. & Rupe, J.C. (Eds.). **Compendium of soybean diseases**. 4. ed. St. Paul, Minnesota: American Phytopathological Society. 1999. p.25-26.

TANNENBAUM, A**. Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand.** Addison Wesley, 2001.