Database implementation for clinical and computer assisted diagnosis of dermoscopic images

B. S. R. Amorim, T. Mendonca, A. R. S. Marcal
Faculdade de Ciências da Universidade do Porto, Porto, Portugal

J. S. Marques
Instituto Superior Técnico, Lisboa, Portugal

J. Rozeira
Hospital Pedro Hispano, Matosinhos, Portugal

ABSTRACT: Dermoscopy is a non-invasive diagnosis technique for in vivo observation of pigmented skin lesions used in dermatology. There is currently a great interest in the development of computer assisted diagnosis systems, given their great potential to this area of medicine. The standard approach in automatic dermoscopic image analysis can be divided in three stages: image segmentation, feature extraction/selection and lesion classification. In order to validate the algorithms developed for each stage, a great number of reliable images and clinical diagnosis are required. This paper presents a software tool to collect and organize dermoscopic data from hospital databases. It is suitable for clinical daily routine and simultaneously has a data structure to support the development and validation of algorithms created by the researchers to construct the computer assisted diagnosis system. This tool is composed by a database with three related but independent modules: Clinical Module, Processing Module and Statistical Module.

Keywords: dermoscopy, database development, computer assisted diagnosis systems, telemedicine.

1 INTRODUCTION

Dermoscopy is a non-invasive diagnostic technique for in vivo observation of pigmented skin lesions used in dermatology. This diagnostic tool allows the clinicians to have a better visualization of subsurface structures and permits the recognition and evaluation of important morphologic characteristics not distinguishable by the naked eye (Carli et al. 2003). The use of this procedure in daily clinical routine conducts to improved accuracy and robustness of clinical diagnosis, along with an overall improvement of the global health care system. Several benefits can be derived from dermoscopy, namely the earlier screening diagnosis, the restrain of selected cases for exeresis and consequently driving to the decrease of unnecessary surgeries. All of this can be translated into human resources optimization, economical thrift and time effectiveness.

So, attending to the advantages of dermoscopy, there is currently a crescent interest in the development of automatic decision support systems given their great potential for dermoscopy. They should provide meaningful quantitative information to assist the clinical evaluation and perform the desired diagnosis accuractly. Image processing methods, classification algorithms, mathematical criteria and automatic learning algorithms, provide the necessary synergy to achieve this specific challenge (Mendonca et al. 2007).

In the last few years a number of single screening tests have been proposed. These procedures are suitable for health care personal with a minimum clinical training and can reduce number of cases that need to be evaluated by a dermatologist. These screening methods, such as the pattern analysis method (Argenziano and Soyer 2001), the ABCD Rule algorithm (Marghoob and Braun 2004), the 7 point checklist algorithm (Argenziano et al. 2011), the Menzies' method and the Cash algorithm (Henning et al. 2007), are based on human interpretation of dermoscopic images. The common denominator of all these diagnostic methods is particular dermoscopic criteria that represent the backbone for the morphologic diagnosis of pigmented skin lesions. Nevertheless, none of these criteria is widely accepted for fitting the mental dermatologists model of diagnosis. Although there is still considerable work to be done in order to establish a link between the human based criteria and an automatic computer algorithm, advances in these fields may lead to the implementation of a computer based system in the near future.

It is much harder that it might seem at first glance to transpose the medical interpretation to a computer assisted diagnosis system (CADS). There are a number of CADS that perform an evaluation of dermoscopic images but they are still far from having a widespread acceptance from the medical community. The standard approach can be divided into three stages: image segmentation, feature extraction and selection, lesion classification. It is applied to a confined set of data collected in clinical routine. Each of these stages involves a number of challenges by itself. Furthermore, in order to validate the algorithms used in each stage, a considerable number of reference images are required, which is often very hard to obtain.

The propose of this work is to present a software tool developed to collect and organize dermoscopic data from a clinical service in a hospital, suitable for the daily clinical routine use, but also fulfilling the need for data structures and consistency required to support the development and validation of algorithms created for computer assisted diagnosis sustems.

2 ESTABLISHING A RELIABLE DATA SOURCE

The standard method used in processing dermoscopic images is usually composed of three different parts and it is applied to a restricted set of data collected among specialists. The proposed approach aims to design a dedicated platform (ADDI-Platform) to support the current lack of reliable and organized information sources to feed the developed algorithms. This platform also reunites the different stages of the procedure into one software. A system like this can provide a large amount of meaningful quantitative information to assist clinical evaluation. At a further level, it may also lead to a fully automatic computer assisted diagnosis system for early warning diagnosis of skin lesions. But collecting data and constructing a dedicated platform like this is a complex process, since it is not a simple collection of clinical cases from already existing databases. To achieve this goal, a new phase is added to the standard three phases approach - the construction of a reliable data source using an hospital database. In this particular case, the Hospital Pedro Hispano database was used.

Databases are powerful informatics tools, prepared to storage large collections of data and keep them organized. This way they allow fast and efficient access to the information saved. They also provide a user friendly interface and are easily implemented and adaptable to any professional area. So, implementing a tool based on a database system seemed the logical choice to develop a reliable data source component for the ADDI-Platform. This tool is composed by three database modules: Clinical Module (CM), Processing Module (PM) and Statistical Module (SM).

The construction of each of these components is divided in four stages: system analysis, logical design, physical design and implementation. All stages enclose specific tasks and rules that must be followed strictly. So, in the end of the construction process, data organization, integrity and security will be accomplished. The conception of such data models follows the steps described in this section (CM construction is used to illustrate the process applied to the three modules).

2.1 *System analysis*

First, an analysis of the system which will provide the data has to be done, to set up the main purpose of the work. Here the purpose is the construction of a module to be used as clinical software where all information of patients may be stored and consulted by the authorized clinical staff.

In this first stage it is essential to fully understand the software used in the hospital as well as the daily clinical routine. Furthermore, it is necessary to aggregate all the information among clinical staff and specialists and gather opinions on possible improvements and development of additional features not existing in their current database system. This is a time consuming challenging task and requires a careful approach, since it involves the transfer of delicate information and the collaboration between multidisciplinary areas. In this specific environment the clinical information is mainly based in qualitative and highly conditioned variety of related parameters. All of them have to be converted to quantitative information and discrete data elements.

Following the system analysis stage, the information collected needs to be organized in different groups in such a way they still reflect the real world concepts we are involved with. This drives us to the ER-model construction (Codd 1990). It is important to decide if certain information should be considered an object or an attribute. Then it is necessary to establish the relations between the objects (each object has to be related to at least one other object). All objects have attributes associated to them. The cardinality of each relation has to be analyzed and set also as the participation (total or partial) of the objects in their relations. At last, the primary keys attributes are chosen (underline attributes). This model is represented in Fig. 1.

2.2 *Logical design*

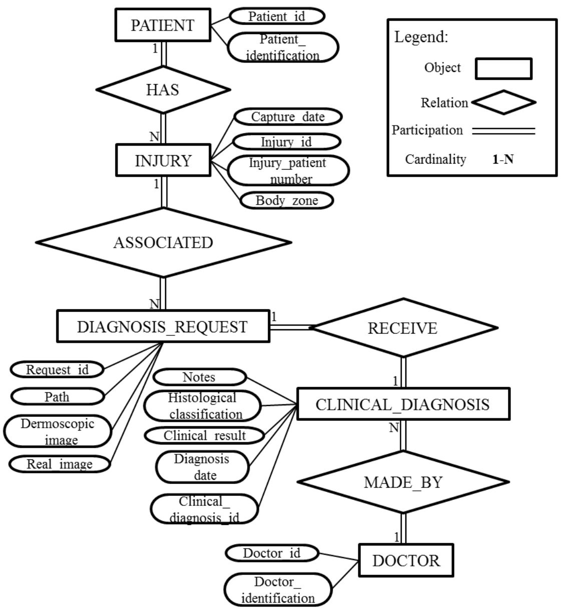


Figure 1: ER-model from Hospital Pedro Hispano Database.

Next the ER-model has to be transformed into an R-model, this is done using normalization. Normalization is a systematic process that taking the ER-model built applies several rules to it (Codd 1990). These rules decompose complex relations into smaller, well-structured ones. They also isolate data, in other words, they allow the operations performed (select, delete, update, insert) to be made in just one table and be propagated to the entire database. For this, a logical arrangement that guarantees the integrity of data and minimizes its redundancy is created. An R-model is completely structured from the logical view point and it is the formal representation of the database. As an example, Fig. 2 shows the normalization of part of the CM.

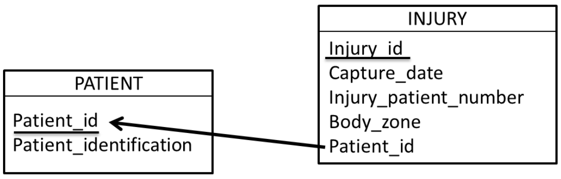


Figure 2: Normalization process result.

2.3 *Physical design and Implementation*

The physical design guarantees that the database performance will be maximized, so methods of efficiently store and access to data have to be defined, such as clustering and partitioning processes (Codd 1990). These tasks have to be performed by someone with full access permissions over the server where the database will be allocated.

The implementation of the formal schema developed previously involves the use of a query language such as SQL to make physical configurations, create tables, define detailed specifications about elements and apply queries to the database (Codd 1990). To develop the user interface any programming language, such as Java or C++, may be applied. Before the database can be release it has to be tested both by medical staff and by researchers, to ensure it fulfills all requirements.

3 ADDI PLATFORM

The software platform in progress joins the three module database system created with the former stages in standard approach. The great advantage of the reliable data source component structure relies on the independency of each module, meaning that, although the three of them are related, each of them works as a stand-alone tool too (Fig. 3). The clinical staff uses the CM as a tool to collect, store and consult patient information's. At the same time researchers access PM to obtain a bank of robust and reliable dermoscopic cases (both images and clinical diagnosis). PM receives a huge and permanent flow of information, since the hospital photofinder software collects daily the real cases and the correspondent dermoscopic and clinical diagnosis. The SM will be accessed later in the image segmentation stage. This module uses QADI web application (Fig. 4) to collect,

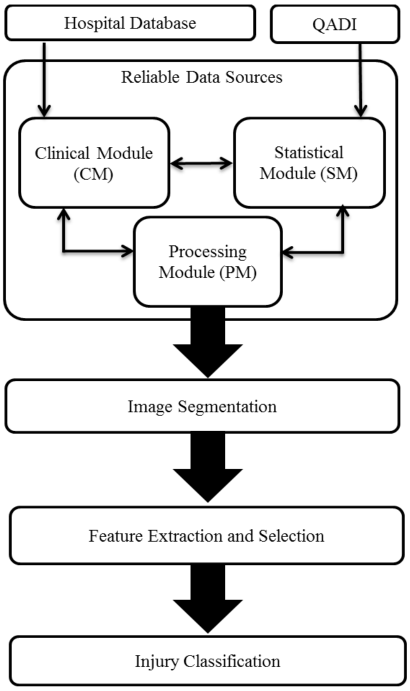


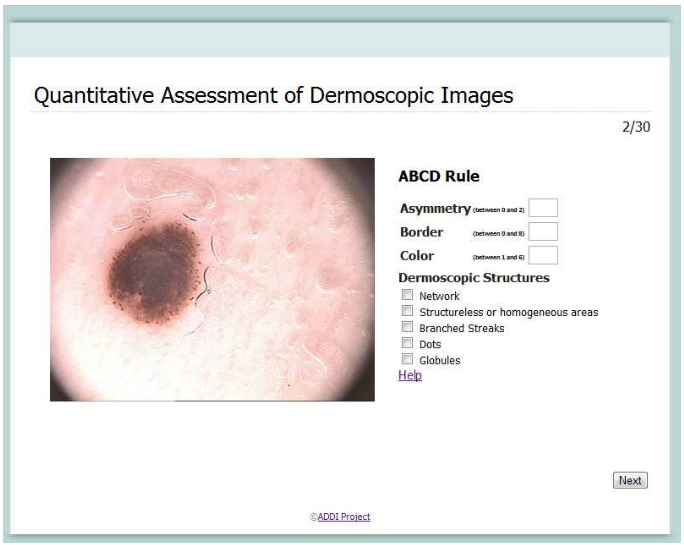
Figure 3: Schematic representation of ADDI Platform

for each case, individual clinical classifications based upon available quantitative tests accepted by the International Dermoscopy Society.

In the image segmentation stage, researchers define a set of restrictions that dermoscopic images must respect in order to successfully integrate the image test group to be applied to the algorithms developed. Then they access PM to get those images. The research team may also access SM in this stage to help them to understand which images gather consensual opinion among specialists. This is a valuable issue, since the development of algorithms and their performance evaluation relies on the existence of well-defined subset of typical images.

In the initial stage of the project, images that do not gather an unanimous opinion among specialists due to their ambiguous appearance, are not elected for training the algorithms, both in the segmentation and classification phases. The reliable data source may also be useful to divide injuries into different sets, such as body zones and other clinical widely accepted standard specifications.

Afterward, feature selection and extraction has to be made. Researchers may apply statistical algorithms to the features information available in PM and SM, in order to choose which features are more likely to produce significant results in lesion classification. Notice that reliable data source structure may be easily modified and extended, since databases own a property called scalability. Furthermore, it may also complete itself, filing some lacks of information using machine learning and data mining algorithms based in all the knowledge acquired with the clinical specialists.



Quantitative Assessment of Dermoscopic Images

2/30

ABCD Rule

Asymmetry (between 0 and 2)

Border (between 0 and 8)

Color (between 1 and 6)

Dermoscopic Structures

□ Network

□ Structureless or homogeneous areas

□ Branched Streaks

□ Dots

□ Globules

Figure 4: ABCD Rule, one of the tests implemented in QADI.

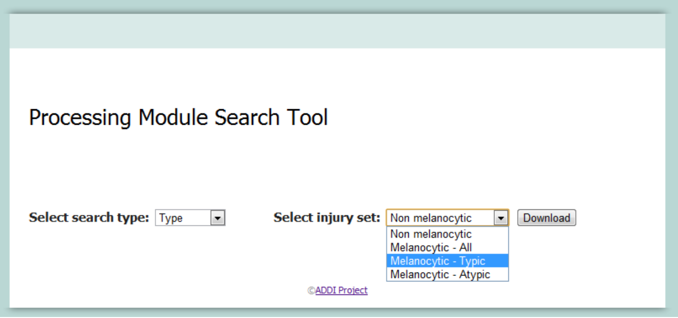
4 RESULTS

The SM web applicattion to collect data is described next:

The QUADI system allows medical experts to analyse dermoscopic images and asks them to answer several questions concerning the dermoscopic information contained in the image (symmetry, color, differential structures, etc). These questions follow the structure of two well known medical algorithms: the ABCD rule and the 7 point checklist. Each question has a finite number of admissible answers. For example the expert is asked to classify the lesion as melanocitic or non-melanocitic (binary decision), and the symmetry as symmetric, partially symmetric or non symmetric (ternary decision). It is well known that experts answers are not always the same for a given image. It is therefore important to characterize the uncertainty associated to each answer and to each expert. It is also known that not all the images are equally difficult. There are easy images in which all the experts will probably agree and there are difficult cases in which they disagree. Again, it would be interesting to characterize this uncertainty.

A mathematical model to translate and represent the information collected by QADI is being developed. It will be tested with a sample of cases collected among specialists. The results will be analyzed in order to classify the lesions as having consensual or not consensual punctuation and consequently understand if they are good lesions to be submitted to processing image algorithms or not.

The Processing Module Search Tool (PMST) application is a tool that allows all the research team members to have access to the clinical data collected from HPH database. It is a form where the user chooses the criteria he wants to evaluate and download automatically an image set respecting those criteria. The download file is a text file including a list of image names present in PM database. This downloaded file is easily introduced into Matlab, which is the software used to develop the image processing algorithms. The txt extension can also be easily imported to algorithms developed in other lower level languages such as C or Java.(Fig. 5).



Processing Module Search Tool

Select search type: Type

Select injury set: Non melanocytic

Figure 5: Processing Module Search Tool interface

5 CONCLUSIONS AND FUTURE WORK

Computer assisted diagnosis systems have great potential for dermoscopy. These systems provide meaningful quantitative information to assist clinical diagnosis and, at a further level, to perform an automatic early screening of skin lesions. A number of computer assisted diagnosis systems are able to classify skin lesions with high sensitivity. The access to a reliable data source is a capital gain in terms of time and accessibility to clinical information.

The development of a database divided in modules like the one here presented allow the access to different sets of information from different groups of people ensuring the security and confidentiality of the data. It also allows other developed applications to be connected to the reliable data source with the purpose of increasing their computational power.

A platform such as ADDI-Platform provides several benefits both to teledermoscopy and telemedicine, since computer assisted diagnosis systems are essential tools in these areas. However, further developments are still required in order to have a robust and reliable complete platform.

In the near future more functionalities will be developed. A sinchronization tool will be implemented, which will allow PM and SM to be updated real-time everytime a new case is introduced in CM at the hospital. PMST and PM will be extended to include information whether the lesions are acral, facial or other, since acral and facial lesions have an altogether different process of diagnosis. The QADI mathematical model will also be optimized and the web application will be povoated with new images, over time, in order to provide even more quantitative information to the research team. An user friendly interface to access CM will also be soon ready for testing by clinical team.

6 ACKNOWLEDGMENTS

The authors would like to thank Fundação para a Ciência e Tecnologia (FCT) who partially funded the research present on this paper through the project ADDI (reference PTDC/SAU-BEB/103471/2008), as well as the clinical team from the Hospital Pedro Hispano: Dr. Jorge Rozeira, Dra. Marta Pereira, Dra. Marta Teixeira, Dra. Ofélia Morais and Enf. Lina Santos for all the support provided on clinical issues.

REFERENCES

Argenziano, G., C. Catricala, A. Ardigo, P. Buccini, P. De Simone, L. Eibenschutz, A. Ferrari, G. Mariani, V. Silip, I. Sperduti, and I. Zalaudek (2011). Seven-point checklist of dermoscopy revisited.

Argenziano, G. and P. Soyer (2001). Dermoscopy of pigmented skin lesions - a valuable tool for early.

Carli, P., E. Quercioli, S. Sestini, M. Stante, L. Ricci, G. Brunasso, and V. De Giorgi (2003). Pattern analysis, not simplified algorithms, is the most reliable method for teaching.

Codd, E. F. (1990). *The relational model for database management: version 2*. Boston: Addison-Wesley Longman Publishing Co., Inc.

Henning, J., S. Dusza, S. Wang, A. Marghoob, H. Rabinovitz, D. Polsky, and A. Kopf (2007). The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy.

Marghoob, A. and R. Braun (2004). *Atlas of dermoscopy*. Taylor and Francis Taylor and Francis Group.

Mendonca, T., A. R. S. Marcal, A. Vieira, J. Nascimento, M. Silveira, J. S. Marques, and J. Rozeira (2007). Automatic analysis of dermoscopy images - a review.