ELSEVIER

Computer Methods and Programs in Biomedicine 44 (1994) 119–130

computer methods and programs in biomedicine

A knowledge acquisition tool in analytical pathology based on multi-media relational database

Amitava Datta*[a], Karsten Rodenacker[b]

[a]*Computer and Statistical Service Centre, Indian Statistical Institute, 203, B. T. Road, Calcutta 700035, India*
[b]*GSF-Forschungszentrum für Umwelt und Gesundheit, ISS, Neuherberg, FRG*

Accepted 10 May 1994

Abstract

The need for a knowledge based expert system for efficient decision making in the field of pathology has been well accepted. To build the knowledge-base for such an expert system is a painstaking task. This work is an attempt to provide the pathologists a powerful and user-friendly tool that will help them in the process of building the knowledge-base for medical diagnosis by closely looking at the specimen images and their extracted feature values. The tool is based on multi-media relational database and software like VAX Rally (4GL), ILIAD and SAS to provide a wide-range comparative study between feature data and image data; and also their statistical analysis.

Keywords: Expert system; Pathological data; Knowledge-base; Relational database; Image analysis; Query; Feature data; Knowledge acquisition

1. Introduction

In cyto- and histopathology, a major work is to observe the pathological changes in cell and tissue morphology for which cytological specimens and histological sections are investigated. Since these changes are highly complex, their subjective diagnostic or prognostic interpretations are very much laborious and time consuming and may result in some erroneous decision too. To make a qualitative decision more and more quantitative methods are taken into account for improvement of the results and objective decisions. A rule based expert system [1–5] has the advantages that it can formulate the precise problem and can be used quickly and easily, even if the user is not acquainted with the special field concerned. To achieve these goals the characteristic features of histological sections and the microscopical section images are to be closely observed side by side and thereby it is possible to establish the relationships, if any, between these two sources of information. And these relationships lead us to acquire knowledge from the heap of the data on the basis of which we are able to make our future decisions.

The evergrowing pathological data resulting from the qualitative analysis necessitates the development of a complex system for data represen-

* Corresponding author, E-mail: amitava@isical.ernet.in..

0169-2607/94/$07.00 © 1994 Elsevier Science Ireland Ltd. All rights reserved
SSDI 0169-2607(94)01518-K

tation, handling and manipulation. Such a system should have the ability to handle different users having a variety of interests. The data should be represented in such a way that they can be accessed by the users easily and efficiently. Additionally the system should assist in solving decision making problems, and in training the users towards improvement of their expertise. In other words, the objective of the present work is to provide an assisting tool for knowledge acquisition [5,6] from the images and extracted feature data for arriving at a powerful decision support system in the field of cyto- and histopathology (see Fig. 1).

In the field of image analysis of cyto- and histological problems, the present work is a part of a larger project 'Visualization, Training and Development for Image Analytical Cytometry' (VIDIAC). A multi-media database is set up which contains all possible project based clinico-pathological patient data, microscopical images, derived quantitative features as well as the statistical parameters for making decisions. On this database we have developed a comfortable and user-friendly display system for the images and the data, and implemented a multi-user environment for data capturing, updation, query and analysis for knowledge acquisition that finally helps in decision making.

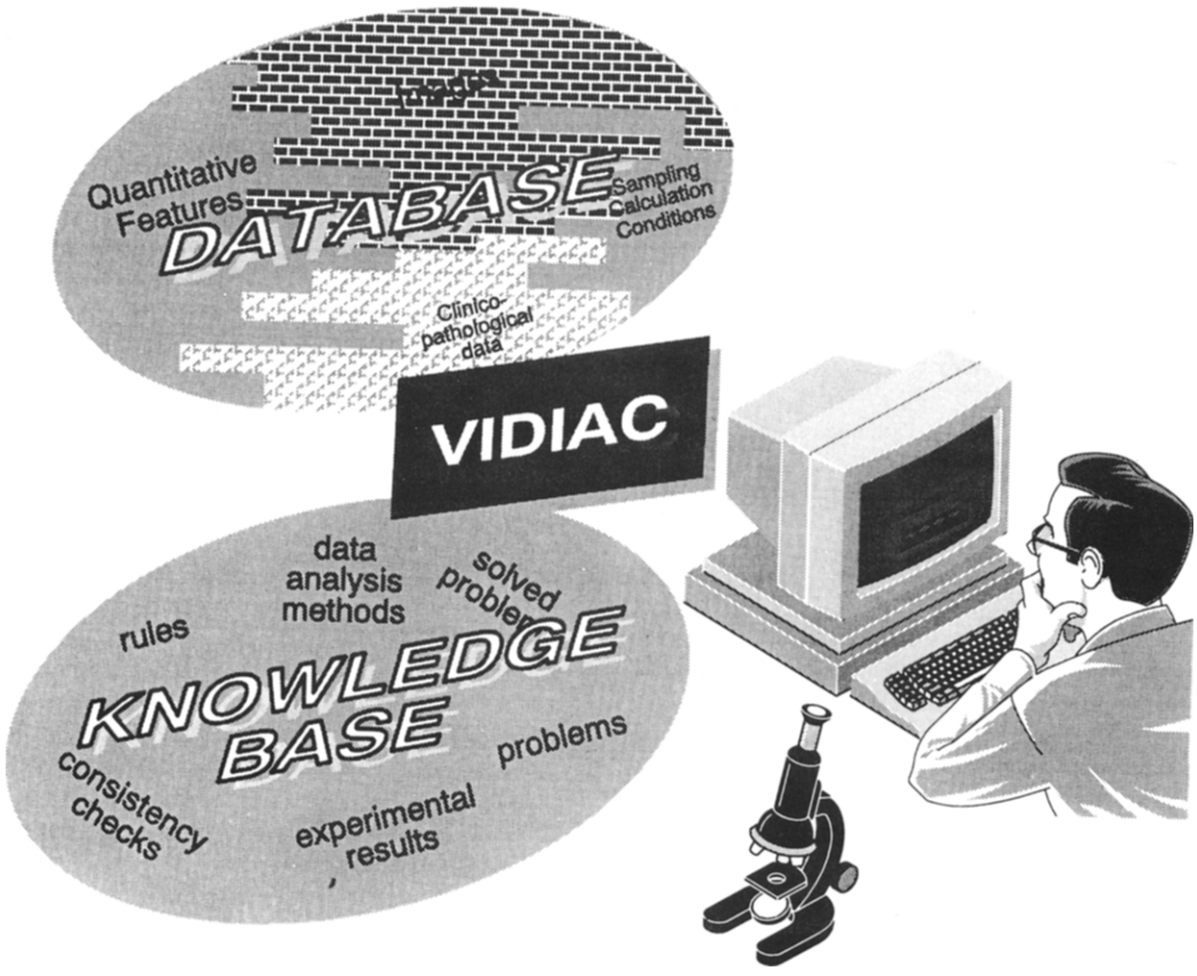

Fig. 1. Objective of VIDIAC: A powerful decision support system for pathologists.

2. The design considerations

The data can originate from different projects. From a specimen obtained from a project, cells (objects of interest) are identified and their feature values [7–9] are evaluated. About 250 cell level features are considered to analyse the smear specimens and histological sections. They are derived from the digitized images using ILIAD programming environment [10]. Not all these features are applicable for all the objects. A particular object may be characterized by a subset of these features and that subset is specified by a set of parameters. The evaluation of feature values are defined and controlled by this set of parameters. These cell level feature values are aggregated by some rules [11] to get the specimen level feature values.

The laboratory at present deals with several projects where in each project there are 50–500 specimens and each specimen has by average 150 cells. To facilitate the process of acquisition and updation of this huge amount of data into a decision support system database with proper consistency and completeness checks, the necessity of developing an integrated relational database was felt. Such a relational database would also help to a great extent in on-line queries on the pathological data as well as in displaying the digitized microscopic-images in a suitably developed interactive and user-friendly manner as required by pathologists to arrive at a knowledge-base.

The framework of the current work is described diagrametically in Fig. 2. As conventional databases are not suitable to store and manage other than formatted text data, multi-media database concept has been adopted. Multi-media database differs from conventional ones (namely hierarchical, network or relational databases) in that it is capable of handling image and vector graphics drawing data besides formatted alphanumeric text data [12]. We use a relational database to store text data in high speed magnetic disk and an image specific subsystem in high capacity optical disk. The bridging between them is done through the relations describing the image file specifications. The advantages are: (1) we can make use of the facilities available in conventional relational

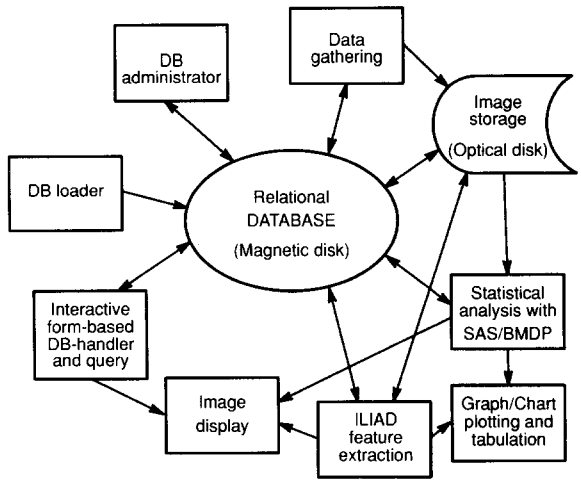

Fig. 2. Scope of the current work.

database tools in formulations of complex queries and easy updation of the data as well as database structures by adding and/or modifing tables and columns; (2) the data can be kept widely open for use to common high level host languages like C, PASCAL, FORTRAN, etc., to fourth generation languages and also to some third party software packages like SAS [13,14]; (3) frequently accessed but less voluminous text data can be stored in high speed magnetic disks while the less frequently accessed but highly voluminous image data can be stored in high capacity optical disk. Moreover, some of the image processing tools which cannot be integrated with database systems as yet, can run independently on the image subsystem.

VAX Rdb/VMS [15] has been chosen as the RDBMS software for development of the system because it is reliable, was easily available and some of the existing programs could be ported directly to the system without much changes. The data acquisition, updation and query programs are developed in fourth generation language RALLY [16] to facilitate future enhancements and modifications. The system has been designed in a modular structure to give it more flexibility and accessibility. It has also been kept in mind during design to have the possibility of adding artificial intelligence components to the current system to arrive at a rule based pathology expert system.

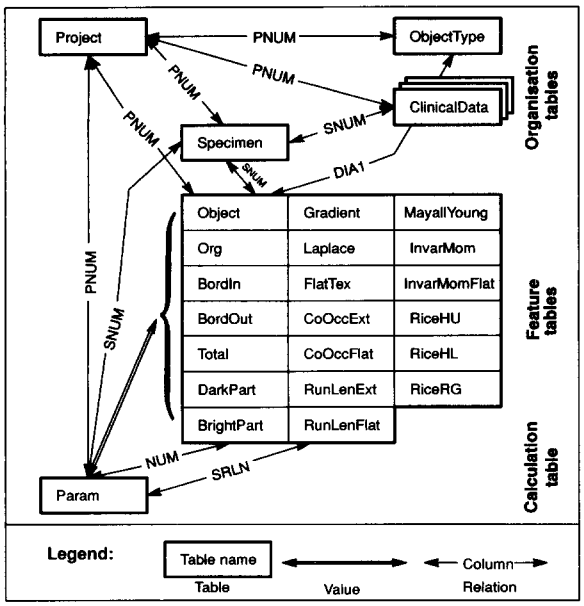

Fig. 3. Database tables with their interrelations.

3. The system description

3.1. The database

The text data consist of all quantitative information which are relevant for each project, specimen and object combination. They contain descriptions of all clinico-pathological information as well as several calculated features [7–9]. They also contain the information about the methods of analysis and sampling conditions.

The text data are stored in the relational database as a number of tables (relations). Entries concerning a particular featuring method make one feature table whereas the presence of any entry in the feature table depends on the indicator value defined in the corresponding parameter entry in the parameter table. Similarly the clinical data are also stored in the form of tables. The linking between the feature entries, parameter entries and clinical entries are done through the unique com-

Table 1
List of database feature tables (relations)

Feature class	Table name	Feature description
Organisation	OBJECT	Object type and location
Extinction features	ORG	Object and background
	TOTAL	Total image
	BORDIN	Inside border
	BORDOUT	Outside border
	DARKPART	Dark particle from FlatTex
	BRIGHTPART	Bright particle from FlatTex
	RICEHU	Upper half-height from ricefield transformation
	RICEHL	Lower half-height from ricefield transformation
Textural features	GRADIENT	Gradient magnitude features
	LAPLACE	Laplace features
	FLATTEX	Flat texture by median filter
	COOCCEXT	Cooccurrence features from extinction
	COOCCFLAT	Cooccurrence features from FlatTex
	RUNLENEXT	RunLength features from extinction
	RUNLENFLAT	RunLength features from FlatTex
Heuristic textural features	MAYALLYOUNG	Chromatin features after Mayall Young
	INVARMOM	Invariant moments from extinction
	INVARMOM Flat	Invariant moments from FlatTex
Ricefield transformation derivatives	RICERG	Ricefield gradient estimate

bination of project, specimen, object and parameter set number (key columns). To give the readers an overall impression of the database, the tables with their interrelations are shown in Fig. 3 and their short descriptions are given in Table 1.

At present the image data are the digitized images of the microscopic sections. The digitized images are 2-dimensional vector of pixels (picture elements) where each pixel represents a grey level in the range of 0–255 (0 represents black and 255 represents white). The images, often called raster images, are stored in files as contiguous bytes, each byte storing a pixel. They can also contain processed images e.g., filtered or segmented images. It is possible to store color images, TV-images and other types of opto-magnetically stored images in the database.

The linking between text data and image data is simple. It is done by adopting a suitably chosen naming convention for the image file that corresponds to a unique key combination of any of the feature/parameter table in text database. Given any entry in the text database the system is able to construct the corresponding image file name and path and fetch it from the image database for display and vice versa.

3.2. Scopes

As the tool will be used mainly by pathologists who may be inexperienced in computer, the system has been made as user friendly as possible. The entire system is menu-driven and form based. One can just go on selecting the menu items of his/her choice and either reaches to a next lower level menu or a form. A menu is a list of items from which the user can select any one (Fig. 4), while a form is a formatted screen (Fig. 5) documented

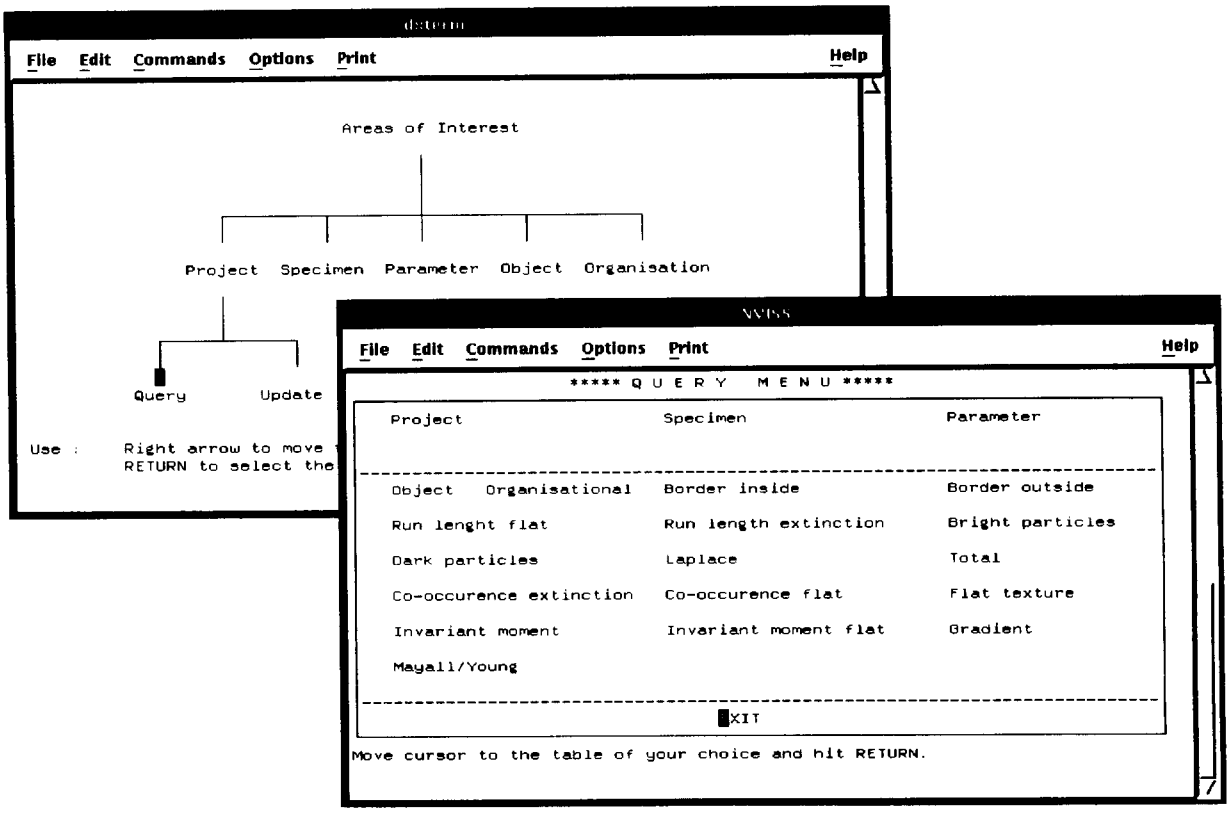

Fig. 4. A typical menu.

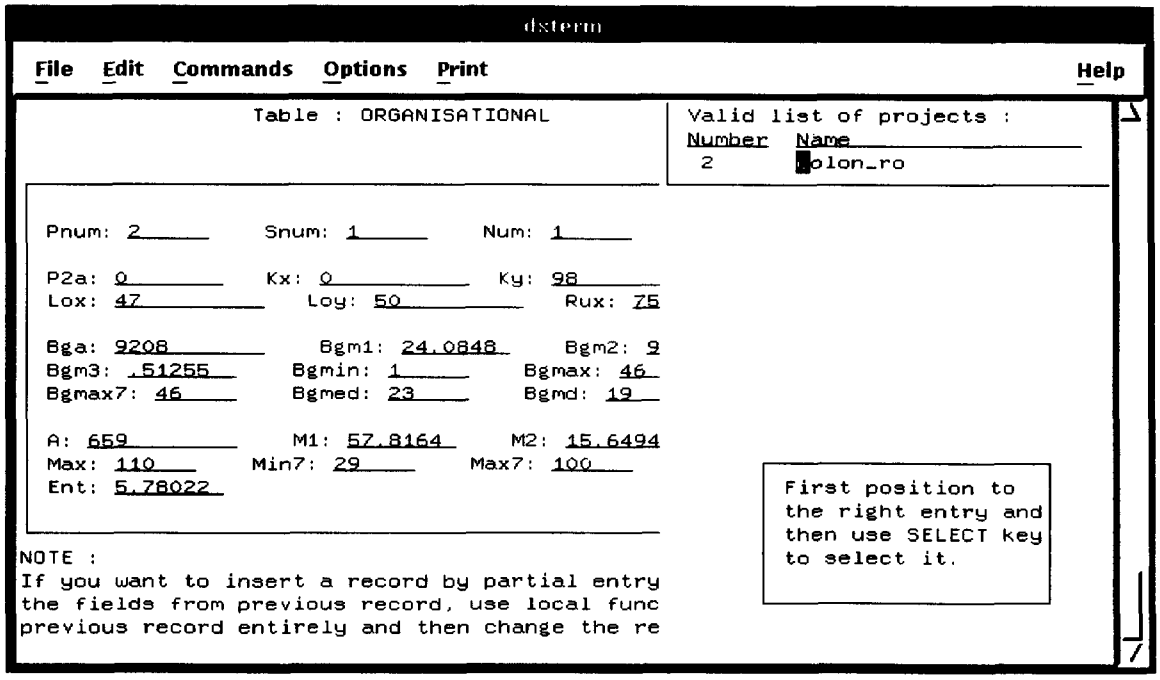

Fig. 5. A typical form with list-of-values and legends.

with suitable texts to be used either for capturing information into the system or retrieving information from the system. The menus are structured in a top-down manner which finally reaches a form as desired by the user.

The scope of the system is divided into three categories depending on the type (or level) of the user. There are three levels of users: (1) end users, (2) data handling users and (3) super users. The end users can only look at the data, the data handlers have read and update accesses to the data and super users besides reading and updating the data can even modify the database structure and the programs. The system is well protected against unwanted damages by keeping secret passwords at different action, table, form and even at field levels. The password also denies unauthorized persons to even look into part of the database which the super user or database manager wants to keep secret. We have divided the scopes for each of these three categories of users and shall discuss them separately.

3.2.1. End users

3.2.1.1. Data and image display. Most of the users are the end users. The end users are those who do not have any editing right to the database and can look at the data on display and read them for further analysis. They can take printout of the results into some predefined formats. They can also execute various types of queries as they desire provided they have the access rights to the relevant tables/fields.

The display of the text data as well as the image are done in windows. Recent advances in window interfaces have made this task easy and efficient [12]. It is possible to display text and the images on the same screen but in different windows. The text information are displayed in a form. The user can either browse through the data sequentially and reach to his/her area of interest, or he/she can enter a query and directly locate the area of interest. The query can be entered by forms, which means the user just has to move to the field(s) on which he/she wants to have some selection criteria and

then key-in the target value and/or the criteria. After the query is executed the form will display the relevant figures on the text window. The user can use complex Boolean expressions with one or more relational operators as the selection criteria. Wild card characters are also allowed for character data types.

After getting the text information displayed one can very easily look at the corresponding image just by pressing a single functional key. In addition to the existing text window another window containing the relevant images (as well as possibly the transformed images) appear on the same screen and thus one can comfortably study the interrelationships between the feature measurements and the visual perceptions to improve the judgements. The users can observe the quantitative measurements from the results of their query and also their interpretations in the actual image. They can modify some feature values and study the effect of the change in the corresponding images. Moreover, they can even directly go back to the original section image at the microscope in an automatic way with the help of X- and Y-coordinate values available in the database. This is essential when the digitized images stored in database are not of good quality due to some reasons, or some particular information are not clear in the digitized images. It is possible to have more than one (at present four) window for displaying sets of images per window. In addition, relevant information about the optical disk volume where a particular image file is stored are also kept in the database to guide the user in mounting the proper volume depending on the query he/she makes at time to time. This is required because the number of optical disk volumes containing the image files is quite large and only a few of them can be kept mounted at a time.

3.2.1.2. Data analysis. The database can be accessed by the end users through their own programs as well as through third party software (e.g. ILIAD, SAS). Users can run their own programs written in conventional host languages (like C, PASCAL, FORTRAN, COBOL, etc.) to read data from the database using the available interface between the host languages and Rdb/VMS. One can also use SAS-Rdb/VMS interface [14] for statistical tabulation, graph/chart plotting, histogram generation, etc. for statistical analysis. A library consisting of interfacing routines has been developed to provide the users ready accesses to the database from other programming environments.

3.2.2. Data handling users

3.2.2.1. Loading the database. The laboratory is already working with a variety of pathological data as discussed in section 2, which are available in the form of conventional files. To convert them into the relational database the system has been equipped with database loading facilities which also perform all possible consistency checks and thus ensure the reliability of the database. This loading subsystem can also be useful to port data from other hardware and operating system and vice versa. This module will not be required for new projects. New users can use the on-line facility from the very beginning.

3.2.2.2. On-line data acquisition User can directly feed their data into the system by filling up the forms displayed on the terminal. He/she need not know anything about the database design or structure, or even the internal dataset/program or field names. The form displayed gives suitable texts which are commonly used by pathologists and are self explanatory. In case of any difficulties more information are available as run-time help which are obtained by just pressing the 'help' key. Additionally, the 'legend' display (Fig. 5) provides the users with a constant guidance. One cannot enter wrong entry as cross checks like field type checks, range checks, uniqueness checks are in-built with the system. Checks against a valid 'list-of-values' are also performed wherever necessary. From run-time 'list-of values' (Fig. 5) the user can easily find the correct entry in case he/she does not remember it.

3.2.2.3. Data updation. In a similar manner as in data acquisition, the data can be updated after choosing the proper form for the table required to be updated. The consistency checks and run-time helps that are available in data entry operation are also made available here. In case of deletions the child entries are automatically deleted when the parent entry is deleted from the database by

proper use of 'triggers'. Similar concept has been applied in case of updation so that when an entry, on which some other entries are dependable, is updated the corresponding updations/deletions for the dependable entries are also done. This allows the user to remain transparent about the internal relationships among the different relations (tables) of the database and the consistency of the database is maintained throughout automatically.

3.2.3. Super user

3.2.3.1. Data security. A super user can control users' access rights to the database. He/she can assign and deassign passwords to users of different levels according to his/her choice. It is possible to restrict one user from using a menu or a form or even a particular field in the form. The super user can also put some restrictions on the type of usage like query, update, insert, etc.

3.2.3.2. Editing the application. The super user plays the role of the database administrator. He/she can edit any part or whole of the application. He/she is capable of changing any menu or form that are currently being used in the application and can even edit the structure of the database. The super user is also responsible for monitoring the application.

4. Examples and results

To illustrate the usefulness of the system, a typical project colon carcinoma is chosen. At a clinic a set of tissue sections from the carcinoma are prepared from several patients. Each specimen is accompanied by anamnestic data as well as clinico-pathological data. The aim of the project is to find out the relationship between specific properties of the recognized malignant tumors and the quantitative features extracted from the image data obtained from the tissue sections. After deciding sampling and calculation parameters, a certain number of cells are identified from each specimen and the relevant feature data are calculated and stored into the database for the corresponding project, specimen and object (cell) combinations.

The data analysis starts with the first step by evaluating the statistical properties of the data gathered. Subjectively certain features are observed by their frequency distribution as well as corresponding cell images. For example, in cases of smears, instead of tissue sections, the integrated optical density is taken as an estimate of the DNA-distribution of a specimen. As a next intermediate step it is possible to define subclasses allowing a

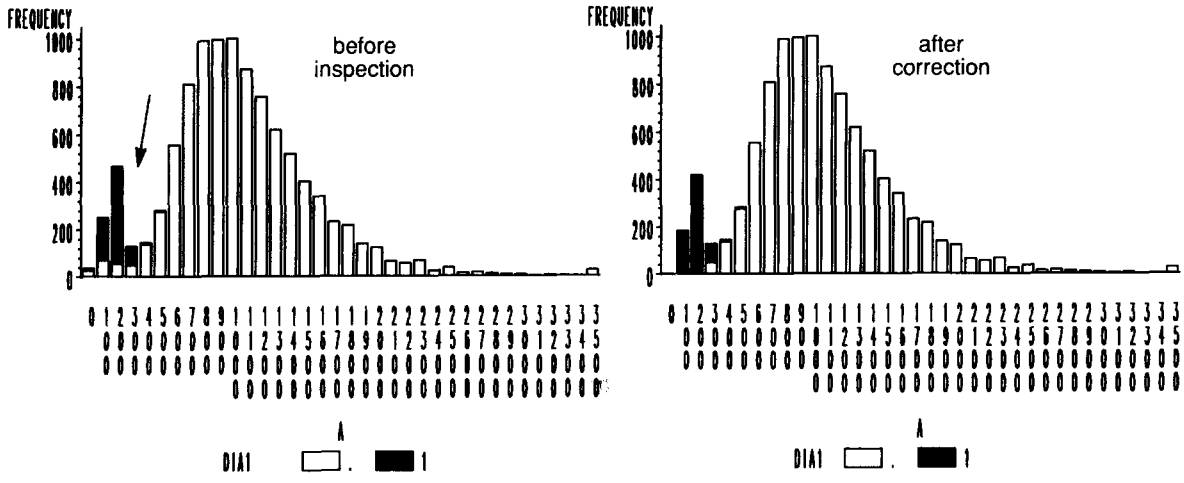

Fig. 6. Frequency distribution of cell (object) sizes A before inspection and after correction.

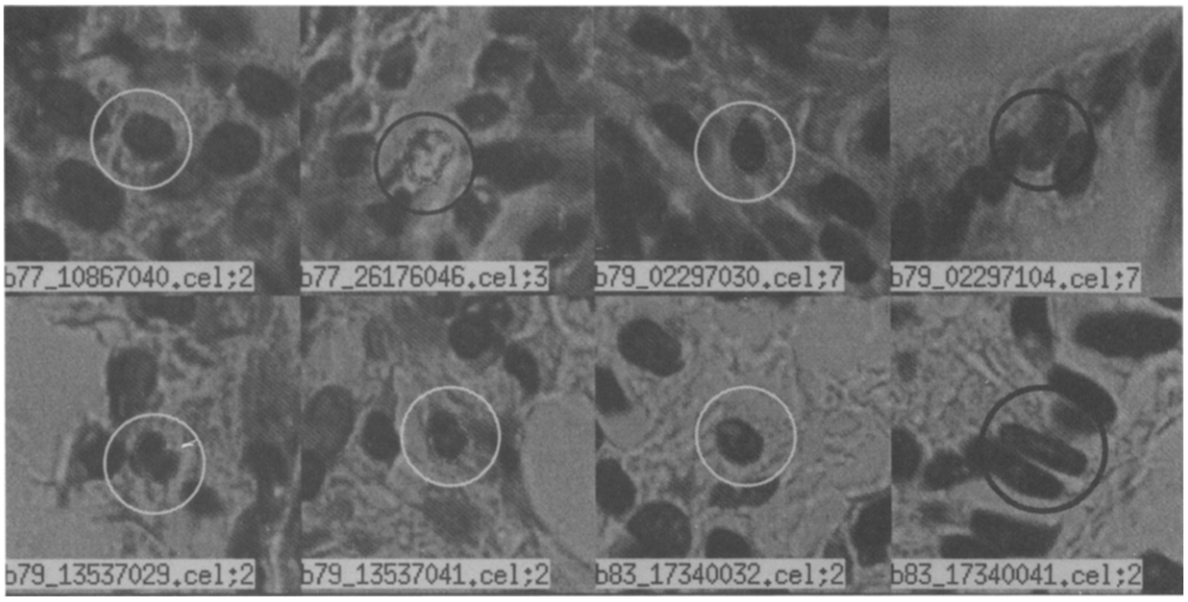

Fig. 7. Images as selected from frequency distribution in Fig. 6. Objects circled in white are lymphocytes, the other cells are misclassified epithelial cells.

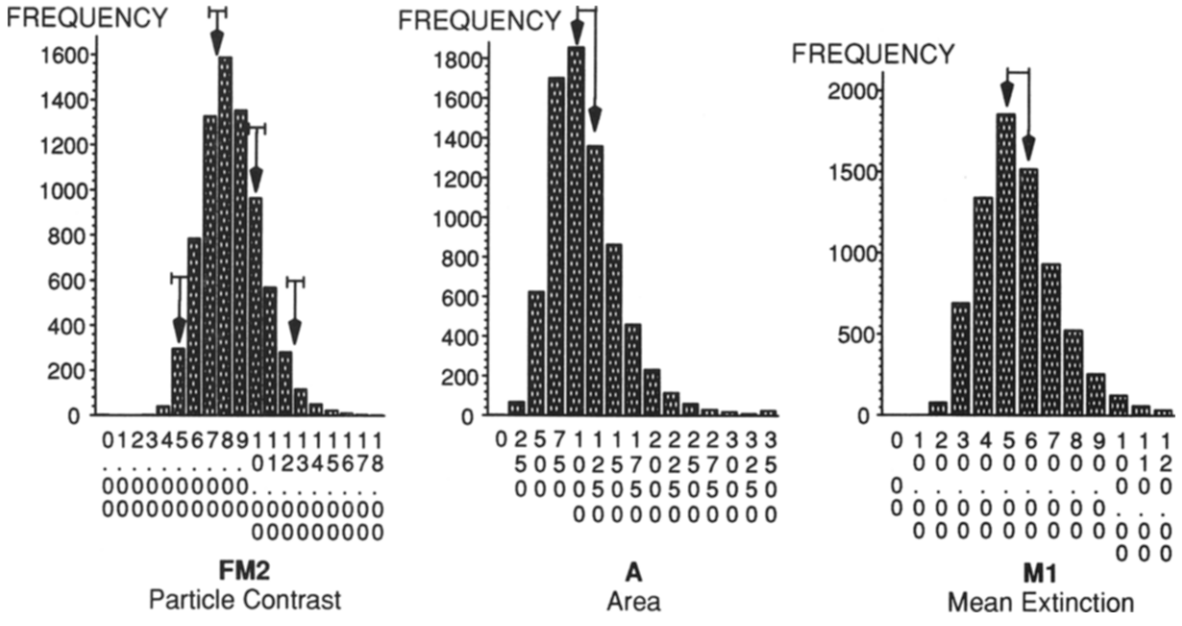

Fig. 8. Frequency distribution of features particle contrast FM2, area A and mean extinction M1. Selected intervals are indicated by arrows.

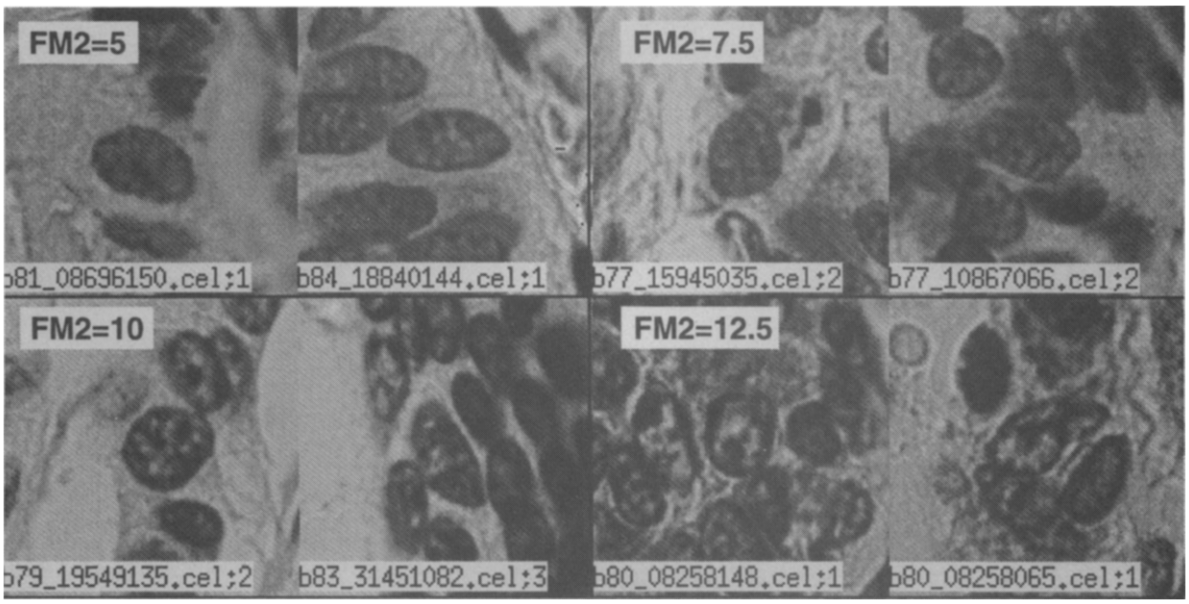

Fig. 9. Some of the cell images from the intervals selected in Fig. 8. Objects are always in the centre of each image.

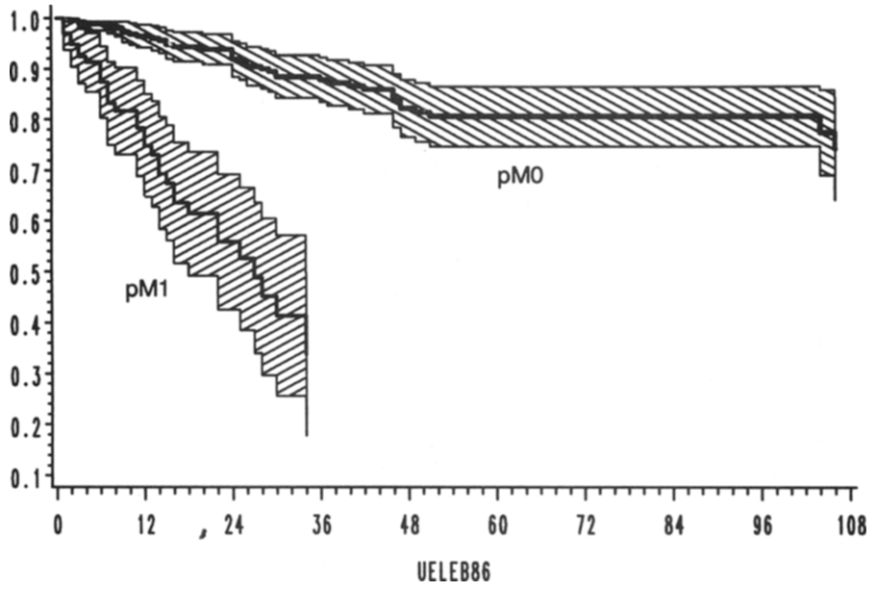

Fig. 10. Kaplan-Meyer survival probability curve for invasive tumors with (pM1) and without (pM0) metastasies (within 95% confidence limits). Survival time UELEB86 is given in month.

refined analysis on the basis of those preliminary observations.

Without going into more complex methods of analysis we illustrate some simple examples of access to the data in the next sequence of figures.

(a) Data reliability: In the frequency distribution of cell size (feature A(= area)) per epithelial cell (.) and lymphocyte (1), let us suspect that several cells were misclassified (marked in Fig. 6 by an arrow). These cells can easily be selected by the the 'Query' facility of the system and can be displayed (Fig. 7) as described in Section 3.2.1.1. The small dark cells are lymphocyte, not epithelial cells. The correct classification of cell type can be set into the database using the 'Update' facility (Section 3.2.2.3).

(b) Data visualization: Most of the quantitative features describing textural properties cannot be directly visualized from the images without prior knowledge or training. The next example shows cell images selected from four intervals of a texture feature 'particle contrast' (FM2). In Fig. 8 the frequency distributions of features FM2, A and M1 are displayed indicating the selection interval boundaries by arrows. Fig. 9 shows some of the corresponding cell images. These are done by the 'Query' facility and then by invoking image display and SAS interface.

(c) Data analysis: From clinico-pathological data in retrospective projects the survival probability curve can be obtained for each item or combination of items. Fig. 10 shows the Kaplan-Meyer survival probability curve for invasive tumors with (pM1) and without (pM0) metastasies enclosed by 95% confidence boundaries, calculated with SAS/STAT procedure LIFETEST. This graph shows the prognostic importance of the existence of metastasies. Immediately the question arises to relate occurrences of metastasies with features from cells gathered from the primary tumor. Subsequent statistical analysis can show how such a relation can be stated.

The above examples show partly the power of the integrated system providing interactive analysis and comparative study of the data and thereby helping the users in extracting knowledge from the feature and image data.

5. Future plans and conclusion

Before concluding we shall mention some of the advantages of the system. By using database data redundancies and inconsistencies are reduced [17]. The data can be shared by a number of users simultaneously with proper securities and from various other software. View facilities on tables provide independency of data, more security and more ease of retrieval. Using SQL and QBE data retrievals and manipulations became very easy. With multi-media database, advanced window manager and menu/form manager the system has been made efficient, comfortable and user-friendly.

Classical visual cytology and histology are increasingly supported by analytical methods because they provide quantitative information for the improvement of diagnosis. No doubt, it is useful to combine the knowledge gained by visual perceptions, cytometric measurements and clinico-pathological data. This helps to improve and economize data acquisition and optimizes the quality of knowledge in respective fields by interaction between the human expert and the computerized system and thus increases the decision power. The goal of the project VIDIAC is to achieve a powerful Decision Support System which will serve for the training of users as well as for further development of cyto- and histometric techniques and strategies. The present work is the initial part of VIDIAC. It develops an intelligent work station for the display of images, feature descriptions and measured data by establishing a tight link between them. The later and future part aims at finding ways to obtain the interrelationships between the two sources of information namely, visual perceptions and feature measurements. This will lead to the knowledge-base, prerequisite for an expert system, to decide at which point of decision making which kind of cyto- and histometric measurements are necessary and to what extent.

Acknowledgement

The authors are grateful to Late Dr. G. Burger for motivating them towards the work.

References

[1] W.J. Clancey and E.H. Shortliffe, Readings in Medical Artificial Intelligence (Addison Wesley, 1984).

[2] M. Ramoni, M. Stefanelli, L. Magnani and G. Barosi, An epistemological framework for medical knowledge based systems, IEEE Trans. Syst. Man Cybernet., to appear.

[3] G. Lanzola and M. Stefanelli, A specialized framework for medical diagnostic knowledge based system, Comput. Biomed. Res. 25 (1992) 351.

[4] J.P.A. Baak and P.H.J. Kurver, Development and use of a rule based pathology expert consultation system, Anal. Quant. Cytol. Histol. 10 (1988) 214–218.

[5] Astrid M. van Ginneken, W. Jansen, A.W.M. Smeulders, J. Vander Lei and J.P.A. Baak, A method for the acquisition of formalized knowledge in pathology, Methods Inf. Med. 29 (1990) 182–192.

[6] M.A. Musen, Automated Generation of Model-Based Knowledge-Acquisition Tools (Morgan Kaufmann, London, 1989).

[7] B.B. Chaudhuri, K. Rodenacker and G. Burger, Characterization and featuring of histological section images, Pattern Recog. Lett. 7 (1988) 245–252.

[8] K. Rodenacker, M. Aubele, U. Jütting, P. Gais, G. Burger, W. Gössner and M. Oberholzer, Image cytometry in histological sections, Acta Stereol, 11 (Suppl. I) (1992) 249–254.

[9] K. Rodenacker, M. Aubele, G. Burger, P. Gais, U. Jütting, W. Gössner and M. Oberholzer, Cytometry in histological sections of colon carcinoma, Pathol. Res. Pract. 188 (1992) 556–560.

[10] O. Eriksson, E. Bengtsson, T. Jarkrans, B. Nordin and B. Stenkvist, Tuning of an interactive software system for image analysis to quantitative microscopy, Proc. Int. Symp. Medical Imaging and Image Interpretation (ISMII 82), Berlin (1982) 549–553.

[11] G. Burger, M. Aubele, U. Jütting and G. Auer, Interactive cytometry, chance or evil of bias?, Pathol. Res. Pract. 188 (1992) 391–395.

[12] F. Beltrame, G. Marcenaro and M. Sassoli, Biocell: an integrated and interactive tool for the management of linguistic and eidetic biological data, in: Lecture Notes in Medical Informatics 45, K.P. Adlassnig, G. Grabner, S. Bengtsson, R. Hansen, eds., pp. 673–677 (Springer-Verlag, Berlin, Heidelberg, 1991).

[13] SAS/STAT User's Guide Version 6 (SAS Institute, Cary, USA, 1990).

[14] SAS/ACCESS Interface to Rdb/VMS Version 6 (SAS Institute, Cary, USA, 1990).

[15] VAX Rdb/VMS Reference Manual Version 4.0 (1990).

[16] VAX RALLY Reference Manual and Master Index Version 2.2 (1990).

[17] C.J. Date, An Introduction to Database Systems (Addison Wesley, 1981).