KNOWLEDGE BASED CLASSIFICATION OF GALAXIES

M. Thonnat
INRIA Sophia Antipolis
2004 Route des Lucioles, F-06565 Valbonne

A. Bijaoui
OCA Observatoire de Nice
Le Mont Gros, BP 139, F-06003 Nice Cedex

1 INTRODUCTION

1.1 Classification as a fundamental tool in observational sciences

The classification of objects belonging to a given category is the essential link between the observational data and the physical hypotheses. The concepts allow to explain these data and to predict the ones connected to other objects. The relevant success of a classification lies in the reduction principle, for which less information must lead to the computation of all the observed ones.

The category is always a fuzzy concept. Generally its limits result from some previous classifications. A given classification proceeds after the definition of a given category of objects in a preceding step. In other words, the classification is always hierarchic , from the general concept of element up to all kinds of category of similar objects.

The classifications must be opened to a new kind of objects, or new observational data, giving a new vision of the categories. The classification building takes into account the observational data and provides a set of classes. Each category object must belong to one or even more of such classes. The belonging can be absolute or fuzzy. In this last case, we obtain a set of belonging probabilities for each object to be classified.

If the classification is well done, the elements belonging to the same class can be considered as similar objects, and they are named identically. The classification is also the way to enrich our language, by introducing new words associated to new classes of objects. We must not forget that if the langage results from the human thinking, the human thinking results also from the langage. In this sense, the object classification is the fundamental tool in observational science.

1.2 Generalities on the classification of galaxies

The classification of galaxies began before the real identification of galaxies as a major category of astronomical objects [58], [36]. The galaxies were a subset of nebulous objects on photographic plates.

Hubble's classification [19] appeared with the use of Mount-Wilson large telescopes. The success of this scheme lies in the small number of bins allowing to classify quite all galaxies, even if peculiar features may mask the structure.

This scheme, and its further improvements [34] [52] allowed to define the language associated to the galaxies and to guide the astronomers for the understanding of this category of celestial objects.

With the amount of new observations, new classes of objects were identified : galaxies with active nucleus, galaxies with a jet, radiogalaxies, lazzy galaxies, etc... Although, the revised Hubble‘s scheme is always considered at the first step to put an observed galaxy in a bin. The other properties are thus considered as a consequence of an abnormal event, more than a defect in the scheme.

The classification of galaxies is defined by the morphology. Four large classes are identified : elliptical, lenticular, spiral and irregular galaxies. These classes are subdivided by taking into account other morphological features : ellipticity, bar or ring. An expertise is necessary to give an accurate type a given galaxy. This work is done for brighter ones [55].

1.3 The need of automated classification

The present classifications are subjective. A same astronomer can attribute different types to the same object, according to the observational conditions. Only a classification based on well-defined rules applied with a computer can give an objective information.

The visual analysis of the image can be long, and the number of classified objects can not increase rapidly. By opposite, the automated analysis is faster, allowing us to process a great number of objects. The visual galaxies to classify means also to examine fainter objects. The visual analysis fails generally in the transfer of the classification rules adapted to bright objects to the fainter ones. A well-built automated system, which takes into account the quantitative information, must give more relevant information.

The morphological classification of galaxies does not take into account all the information known on these objects; the instrumental improvements and the use of a set of large telescope in many wavelength ranges (X, UV, Visible, IR, Radio) allows to obtain new kind of data. The use of multivariate analysis by computer permits to build new statistical classifications based on these data. New classes, new rules could be introduced, leading to another vision of the category of galaxies.

The sophistication of Hubble's classification by the use of new data is only possible using an approach starting from this scheme, and by building new rules allowing to improve the classification. If the statistical analysis can take out these rules, only the Artificial Intelligence methods allow us to activate these rules, in a knowledge based system.

2 ARTIFICIAL INTELLIGENCE TOOLS FOR GALAXY CLASSIFICATION

2.1 The expert systems

In this section we present the artificial intelligence tools which can enable the building of an automated galaxy classification system: the expert systems [17], [16], [40]. A short description of these systems is given showing their originality among the sofware tools and their evolution; then two examples of expert system generators are presented to describe their specificities and their complementarity.

Main components

The first characteristic of expert systems is the separation of the knowledge in an application domain (even limited) and the way to use this knowledge (the control structures).

The expert systems are composed of three main parts : the data corresponding to a special case (or base of facts), the knowledge base and the control structures or the kernel of the expert system.

a)the base of facts

The data contained by the base of facts are the input data of the current problem plus all the deduced data added during the reasoning. These data may be symbolic data, numerical data or more complex structures like for instance a list of an attribute a value and a coefficient of plausibility.

ex :

(color blue 0.8)
(perimeter 122 1)

All the data in the base of facts are usually organised in a list:

ex:

((color blue 0.8)(perimeter 122 1)...)

In recent systems the base of facts is organized in a more structured way, made of objects and subobjects with various slots; their values are filled with the input data or added during the reasoning:

ex :

Name = ngcXXX
ring = unknown
bar= present (0.4), absent (0.6)
shape= spiral (1)
ellipticity = 0.2 (1)

b) the knowledge base

The knowledge bases in the first systems were large sets of production rules. Most of the time the order of the rules in the base was important and could change the reasoning.

The knowledge bases in the recent systems are structured (automatically or not) into small bases to enhance the efficiency. The experts or the knowledge engineers can express their knowledge through various knowledge representation schemes (frames, semantic nets, production rules, procedural knowledge, scripts...) Each of these schemes may be pertinent or well adapted to express different kinds of knowledges.

For instance :

- frames are adapted to describe in a declarative mode the components of an object. (physical or abstract).
- semantic nets are adapted to express the relations between objects.
- production rules are used to express in a dynamic mode heuristic criteria.

For galaxy classification as it will be shown in the next sections we need several knowledge representation schemes to describe the objects (for instance the classes), their relations (hierarchical) and heuristic criteria (for example to diagnose the presence of special features).

c) the control structures

The control structures which work with the knowledge contained in the knowledge base are the kernel of the expert system because they define the class of reasoning which is performed by the expert system.

Among the possible kind of reasoning the existing systems perform, one can notice diagnosis, planning, simulation, classification, conception or interpretation.

For galaxy classification we need of course a classification reasoning to handle the knowledge on galaxy classification but as we also have to process images we need another kind of reasoning (a planning plus control of execution reasoning) to handle the knowledge on the processing of images of galaxies.

Important features of expert systems

As it has been shown in the previous section the expert systems enable to handle symbolic data (as well as numerical ones) and they perform logical reasonings. Another important advantage of expert systems is that they are good development tools for building complex systems. In fact they provide explanations, which is interesting both for a non specialist to understand the logical reasoning but also for the expert developing a large knowledge base. As the knowledge is separated from the control structure, modular and explicit, the modifications in the knowledge base are immediate and easy to perform. Moreover, some extensions of the knowledge base can be added step by step, enabling the building of a huge system as a real galaxy classification system enhancing progressively its power and sophistication but keeping it coherent. Their efficiency being preserved by the automatic translation of the knowledge base into a compact code using software engineering techniques (parsing, compilation,...).

2.2 CLASSIC

In this section we present the main characteristics of CLASSIC [13] an expert systems generator designed for pattern recognition and diagnosis using a classification reasoning. This system developed by ILOG was initially built at INRIA for galaxy classification; since it has been used for several other application domains [31] [50].

The main characteristics of CLASSIC are :
- the organization of the knowledge in description trees (called prototype trees) to explicitly describe the objects to be classified
- the use of deductive rules to represent the dynamic or heuristic knowledge
- and the treatement of both imprecise or uncertain knowledge.

In addition in CLASSIC there are:
- a highly interactive development system including graphic stepping and debugging;
- a simple integration of existing programs written in C or Fortran
- and a structuring and automatic compilation of knowledge guaranting modularity and efficiency.

The Knowledge Base

The knowledge base is made of three main components: the prototypes, the tree of prototypes and the rules.

The prototypes

In Classic a portion of the knowledge is described by prototypes. The prototypes are typical descriptions capable of representing, for instance, physical objects,events or failures.
For example here is the prototype representation of a tiger (an example taken from the animal classification system of P.H. Winston [57]).

example of prototype :
Prototype Tiger
is -a : Carnivor
teeth : pointed
pattern : striped
color : (yellow black)

This kind of concept was traditionnally represented by production rules :

same example using production rule :
If the animal is a carnivore
and animal has stripes
and animal has pointed teeth
Then animal is a tiger

By describing the animal with a prototype, there is an explicit description of the concept. In addition it is declarative, avoiding the procedural dynamic aspect of production rules. Moreover rules have the disadvantage that their fine granularity requires a single concept to be split up into several rules.

The prototype trees

In order to better organize the knowledge, prototypes can be linked to form a specialization tree going from the most general descriptions to the most precise. In our preceding example, the Tiger prototype belongs to the more general Carnivore class. Such an organization fulfills two objectives :

- structuring of the knowledge : due to the inheritance of descriptions from fathers to sons, redundancy is eliminated and modularity is enhanced;

- control of reasoning : the hierarchical organization of the knowledge defines natural reasoning steps and permits efficient control

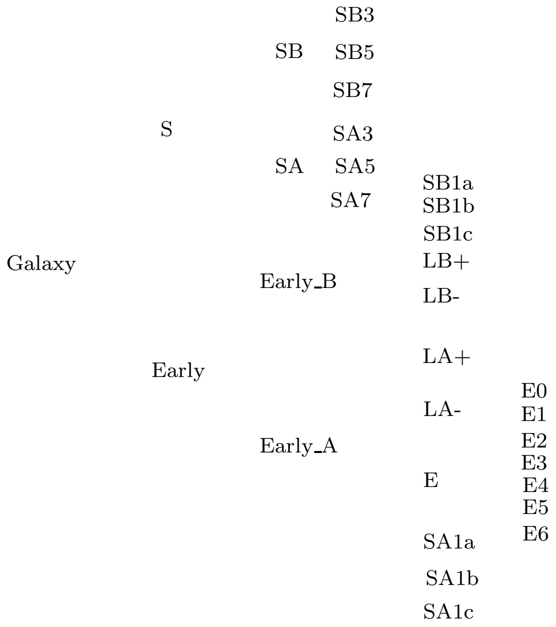


Figure 1: a tree of prototypes

The deductive rules

As well as the explicit representation of concepts, it is necessary to manipulate other kinds of knowledge by rules. In Classic , rules are essential to allow the description of the following knowledge :

- transformation of numerical information to symbolic information :

  if temperature> 38 then fever = high
  if length > width ×10 then shape = elongated

- synthetic knowledge :

  if shape of contour1 member of (average, spiral)
  and shape of contour2 = average
  then arms of galaxy = visible

Classic uses these rules to add information when beginning with initial data.

The Control Structures

Action of the engine

The inference engine in Classic traverses the prototype tree searching for one or many solutions. At each node in the tree, Classic uses the following basic cycle :

1. collect an initial group of information about the case to be handled;

2. use the production rules to infer as much knowledge as possible from this information;

3. choose the prototype most compatible with the available information;

4. attempt to confirm the choice of prototype with perhaps additional information;

5. continue toward the children of the prototype if it is confirmed, otherwise choose the most compatible prototype.

This process continues until one or all the solutions are found.

Treatment of imprecision and uncertainty

Classic is capable of representing and manipulating imprecise and uncertain knowledge using the results from possibility theory and fuzzy set theory.

•The imprecise fact "fever is high" is characterized by a possibility distribution which corresponds to the characteristic function of a fuzzy set (for example : [38,39])

•The uncertain fact "John seems sick" is characterized by a confidence factor(called a possibility measure) and a doubt factor (which is the certainty of the negation).

Current expert systems generators typically manipulate uncertain information by way of certainty factors but they handle imprecise information.

The use of possibility theory has permitted to define mechanisms to match rule premises with facts or to match prototypes with the object under consideration, neither of which is an "all-or nothing" process.

Rules in Classic can contain fuzzy predicates like ∼=, ∼<, close-to, far-from which permit the manipulation of imprecise numbers. For example, the pattern matching of the premise If temperature close-to 39 with the temperature = 38.8 returns the possibility measure that 38.8 belongs to the fuzzy set defined below is 0.5:

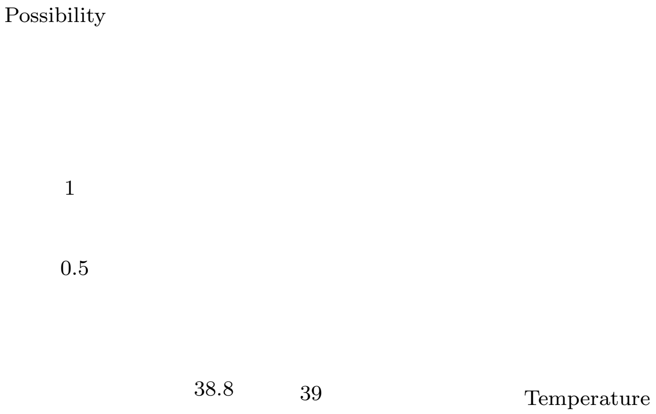


Figure 2: the fuzzy predicate close-to

Control Slots

The control of the search for solutions in the tree is described by specialised slots of the prototype. These slots are used to describe the control of the process of reasoning. Thus, there is no parasitism of the application by introducing control information in the rules or mixing control rules and domain rules. The control slots specify :

•The number of solutions desired. For certain problems or parts of problems, one or all acceptable solutions may be required.

•Thresholds indicating the limits of acceptability of a solution.

•The importance of a prototype relative to the others. This information from the knowledge of the expert, represents the importance in the domain.

•The way to determine components of the current object.

•The order in which the slots of an object must be used and the initial data required once an object is selected.

•The rules to apply once a prototype has been accepted or rejected.

These control slots are automatically filled by Classic with default values. They are constantly editable with the prototype editor.

2.3 OCAPI

Ocapi which is currently developed at INRIA for vision applications [49], [48] is an expert systems generator specialized in monitoring of libraries of procedures.

The main characteristics of OCAPI are :

- the structuration of the knowledge into frames describing the actions performed by the library;
- the use of semantic nets to explicit the decomposition of a processing into several steps;
- the use of specialized production rules (choice between methods, evaluation of results or adjustment of parameters);
- a reasoning of planning (generation of a list of actions to perform) plus control of execution (adjustment of the input parameters and evaluation of the results) see Fig. 3.

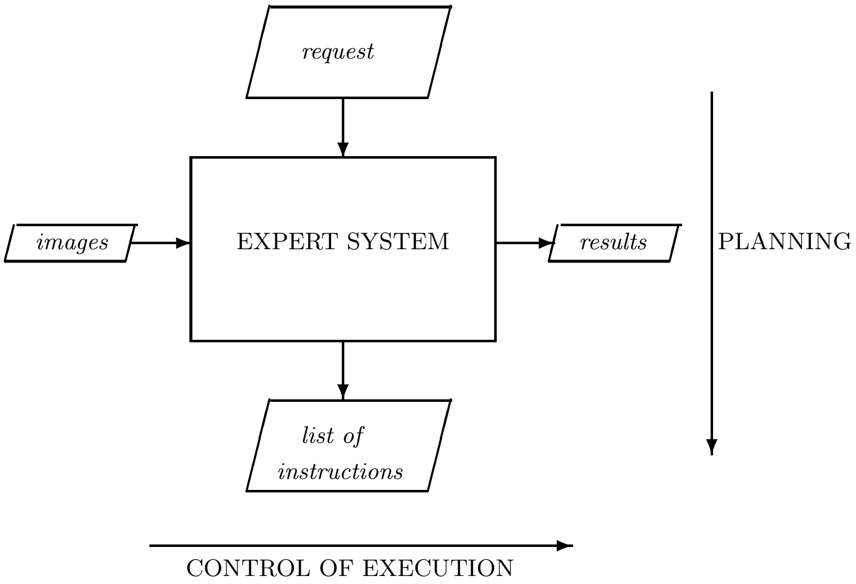


Figure 3: the reasoning

Some other features are :

- the utilization in two modes: automatic or interactive;
- the effective execution of the binary code of the library;
- the possibility of monitoring procedures written in C , fortran or lisp.

The Knowledge Base

The knowledge base is made of goals, operators and requests.

The goals :

Each goal represents an image processing functionality; it is implemented as a frame. The main slots of the object are: the functionality, the input and output parameters (each one is a frame), the knowledge to choose between methods and to test the results (implemented with rules: choice rules and evaluation rules).

Example of a goal:

Goal Filter1

functionality: filtering

input-parameters:

{noisy-image : type image ...}

output-parameters:

{filtered-image : type image ...}

choice-rules:

{ R1 : if noise important
then use recursive-gaussian-filtering
R15 : if noisy-image binary
then use morphological-filtering
R17 : if slopes important
then use median-filtering}

evaluation-rules:

{ R3 : if filtered-image variance > request variance
then noise too-high}

The operators :

Each operator knows how to reach a goal; it is implemented with a frame. The general slots are: the functionality, the input and output parameters, the characteristics which explicit the effects of the operator (in terms of the parameters of the goal), the adjustment which expresses the way to modify the result by changing the values of the input parameters (implemented with rules).

These operators are divided into two subclasses: the elementary or primitive operators and the complex ones.

Primitive operators : They correspond to a program of the library. They inherit the general slots of the operators but they have their particular slots as the syntax and the programming language, to describe the way to run the procedure.

Example of a primitive operator:

Operator med

functionality: filtering

characteristics: {median-filtering}

input-parameters:
{noisy-image : type image ;
ws : type integer }
output-parameters:
{filtered-image : type image ...}
adjustment-rules:
{ R8 : if noise too high
then increase ws
R11 : if $\exists S, \forall object$ object size > S
then ws $< 2 \times S$ }
syntax: med noisy-image filtered-image -s ws

Example of a parameter described by its main slots:

In-parameter ws
type : integer
range : [1,15]
default : 3
function : window-size

Complex operators : They correspond to a particular decomposition of a goal into steps. Each step is a special request of a subgoal. The decomposition is implemented with a semantic net where the nodes are requests and the links are *and* or *then* links.

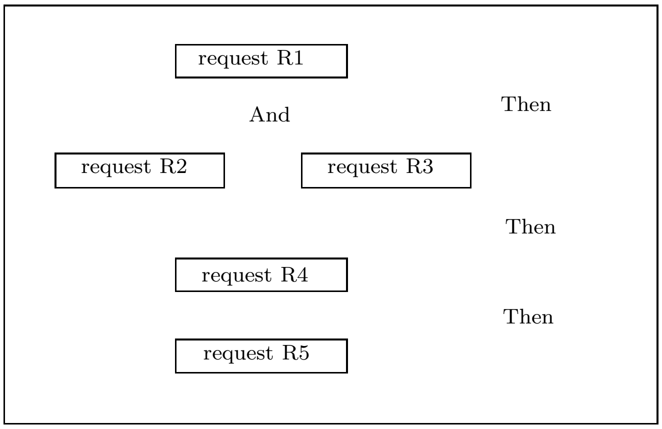


Figure 4: a decomposition of a complex operator

The requests :

The knowledge found in the goals and the operators will be used to satisfy requests, coming from the user or the system. Requests are goals to be reached in a certain restricted context. They are implemented with frames. Their main slots are : the goal and the characteristics of the context, which are the restrictions on the input parameters and the specifications on the results (output parameters of the goal).

Example of a request:

Request R4
median-filtering with *window-size* < 5

The Control Mechanisms

The role of the OCAPI system is to run programs to satisfy requests; i.e. to reach a goal within a certain context. The control mechanisms are built around two tasks: planning and control of execution, each of them using for that the appropriate knowledge found in the knowledge base.

The Planning :

Given goals expressed with requests, the role of the planning is to find the operators (the plan) to satisfy the requests.

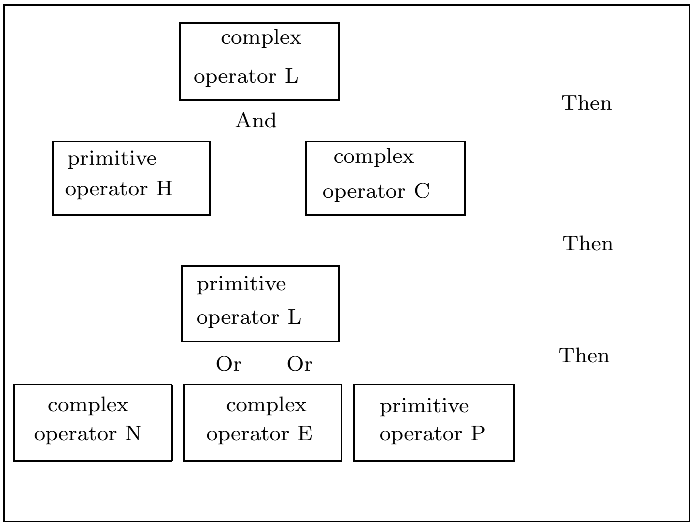


Figure 5: a plan

The planning proceeds in two steps: a simple selection phase and a more complex phase using inferences.

The Control of Execution :

The *Control of execution* tries to execute the plan in such a way that the specifications on the results expressed in the request are satisfied. After the choice of the next operator to execute and the initialization of each of its input parameter, the control will next proceed to the effective execution of the operator, according to its type: *for a primitive operator:* control calls the binary code of the library procedure; *for a complex operator:* control decomposes it into less complex requests.
Then, *the evaluation of the results* is achieved automatically according to the specifications of the request or by graphic presentation of the results to the user who evaluates them interactively. In case of unsatisfying results, another execution will be performed after adjustment of the input parameters of the operator or after choice of another operator.

3 THE KNOWLEDGE TO EXPRESS

A really "intelligent" system in galaxy classification must handle several kinds of knowledge as the astronom does. First a classification to be available must take into account the nature of the context of the observations and of the images formation. Then an automatic system has to process the data contained in the images using an expertise in data reduction (preprocessing and object description). Then the complete description of the processed galaxy is matched with the classes using the knowledge on the description of the 3-D models of galaxies.

For each of these kinds of knowledge (context, preprocessing, object description, class description) we detail the various elements of their contents, their use and eventually how they can be implemented into knowledge representational schemes.

3.1 The context

The contents

The elements of this knowledge are :
- the observed region (is it a region with special objects: cluster, resolved stars...?)
- the observation instrument (is it a large field instrument: Schmidt camera or not, what are the resolution, the quality of these instruments in general, of this particular instrument?)
- the physical support (what kind of support is it: a photographical plate, CCD..., what are the specificities of CCD...?)
- the wave length (is it a special wavelength: H_α, V, IR..., or a set of them, how to process them, to interpret them, to mix them...?)
- the exposure time (the value in minutes, the symbolic value :short, normal or long, or the knowledge to deduce from the numerical value the symbolic one)
- the observational conditions (seing, sky background and absorption expressed with a scale of qualities: poor, average, good, exceptional)
- the digitizer (the type of digitizer: microdensitometer, CCD, Image Photon Counting System..., the characteristics: noise, resolution)
- the calibration (presence of data, type of calibration: receiver, astronomical..)

The use

Essentially the context is used as a filter.

1. A filter to check if a value belongs to an interval or a set of predefined values.

example :

Wave length $\in$ { B, V, R }

This feature allows to explicit the actual stage of development and skill of the knowledge base;

2. A filter to choose between several processings;

example :

if calibration-data present
and data-type receiver-calibration
and receiver-type photographical plate
then execute calibration-chain photo

3. A filter to choose between several thresholds in an inference rule.

example :

```
    if exposure-time short
and ellipse-error > 0.4
then contour-shape spiral
```

```
    if exposure-time normal
and ellipse-error > 0.2
then contour-shape spiral
```

In fact such a knowledge (if it is supposed that this example is pertinent) has to be split into :

- context rules to define the threshold

```
if exposure-time short
then err-ellipse-thr = 0.4
and
if exposure-time normal
then err-ellipse-thr = 0.2
```

- description rules to use the threshold

```
if ellipse-error > err-ellipse-thr
then contour-shape spiral
```

4. An evaluation of the quality of the input-data and thus a way to measure the quality of the further inferences.
example :

```
    if observational-conditions ∈ { 2,3 }
and telescope = Schmidt Mount Palomar
then validity-data correct
```

5. Finally the context data are useful to perform multisensor data analysis. This filter does not work on the values of the data but on the presence of certain type of information.
example :

```
    if ∃ data-wavelength in B
and ∃ data-wavelength in R
then stars-age computable
```

3.2 The image processing

The Contents

It is a bit arbitrar to separate among the data processings those which belong to preprocessing to those which belong to the object description. Moreover all this kind of knowledge has the same nature. According to a goal to be reached (description of images of galaxies) several steps have to be defined and the available and pertinent procedures have to be selected, run and controlled. We recognize for this part of the knowledge the OCAPI notions of planning and control of execution (see section 2.3) or more generally a knowledge in automatic monitoring of procedures.

Descriptive and relational knowledge

The first knowledge to describe here is the decomposition of the image processing into the always needed steps:

- the sky background reduction (background cartography)
- the noise reduction
- the galaxy detection
- the galaxy location (center of the bulge if visible)
- the galaxy isolation (detectable limits of the galaxy)
- the star removal (and the removal of all objects but the galaxy)
- the galaxy description
- the galaxy classification

These steps represent goals to reach; so one operator galaxy-processing could be implemented using the OCAPI terminology like this:

Ex:

```
Operator Gal-proc1
        functionality:              galaxy-processing
        characteristics:      {unique-input-image}
        input-parameters:
                               {input-image : type image ;
                               context : type frame }
        output-parameters:
                               {class : type symbolic}
        decomposition:{
                                          sky-background-reduction
                               then
                                          noise-reduction
                               then
                                          galaxy-detection
                               then
                                          galaxy-location
                               then
                                          galaxy-isolation
                               then
                                          star-removal
                               then
                                          galaxy-description
                               then
                                          galaxy-classification
                                }
```

Descriptive and heuristic knowledge

In fact, for the same functionality (or goal) i.e.: galaxy classification several operators can be implemented according to the nature of the observed images (one, two or several images) and of the calibration data.

Among the steps of the decomposition of the operator some of them can be quite complex; for instance star removal can be performed using various methods:
by the shape (correlation with the point spread function),
by the intensity (detection of the bright objects),
by the color (if the wavelength is different of those of the HII regions),
by the position (if some stars are identified in the field).

Such a knowledge can be implemented using production rules; these rules are the choice rules attached to the goal star-removal:

Ex:

Goal Star-removal
  functionality: star-removal
  input-parameters:
    {spoiled-image : type image ...}
  output-parameters:
    {cleaned-image : type image ...}
  choice-rules:
    { R1 : if ∃ identified objects
      then use remove-by-position-and-shape
    R2 : if wavelength $= H_\alpha$
      then use remove-by-intensity-and-shape
    Rn : if ....
      then use ...}

For each of these steps we need to have a good description of all the available operators and of the way to choose between them; since some of them can be subdivided into distinct parts we can still have complex operators for different levels of abstraction.

For instance an operator which performs the galaxy description can be implemented as a complex operator with a decomposition in a first step of photometrical description and a second one of morphological description.

To automate the image processing we also need to introduce knowledge on the evaluation of the results for the various levels of the requests to subgoals; this can be done using heuristic criteria implemented with rules attached to the goals. For instance if we suppose that the goal galaxy-detection provides as output a segmented image and a measure of the size of the detected object, an heuristic rule can evaluate the quality of the detection (which depends of thresholds difficult to initialize) by testing the size of the detected object.

Exemple of goal with evaluation rule :

Goal Galaxydetection
  functionality: galaxy-detection
  input-parameters:
    {complex-image : type image ...}
  output-parameters:
    {detected-object : type frame}
  evaluation-rules:
    { R3 : if object size ≤ star size

then detected-object ambiguous}

Once the request has failed we need to use some knowledge to readapt the processing according to this evaluation. This can be done using heuristic criteria attached to the operators available to perform the desired goal. For instance if the galaxy detection is performed by the operator Galdet1 adjustment rules can expressed how to modify the input parameter in such a situation:

Exemple of an operator with adjustment rule :

Operator Galdet1

functionality: galaxy-detection

input-parameters:

{complex-image : type image ...}

output-parameters:

{detected-object : type frame}

adjustment-rules:

{ R1 : if detected-object ambiguous
then less strict thresholding}

decomposition:

{ ....

then

thresholding

then

.............}

In one of the operators performing the thresholding we know through other adjustment rules how a less strict thresholding can be obtained:

Exemple of an operator with adjustment rule :

Operator Muls

functionality: thresholding

input-parameters:

{greylevel-image : type image
thr : type float ...}

output-parameters:

{binary-image: type image}

adjustment-rules:

{ R19 : if less strict thresholding
then increase thr}

The Steps

Now we detail the various steps:

Sky background reduction

The determination of the primitives for the morphological analysis is very sensitive to the sky background reduction. Many techniques were purposed to remove the background.

One of the first of them, the numerical mapping [22], is based on an iterative adjustment of the data with 2D homogeneous polynomial. At each step, we reject the pixels for which the deviation is greater than 3 times the standard deviation. After some steps (3 to 4) we obtain a clean map.

Some improvements were done using a division of the image into blocks. In each block a first background value is determined using the intensity histogram (modal, median or other estimators). That leads to a fast convergence. The spline functions (B_splines, pseudo or true splines) are often used to compute the map, after the determination of the background in each block [2] [24].

In the mapping, the galaxy region is removed, in order to suppress the influence of the galaxy onto the background mapping.

Other methods are also used, based on the star removal and cicatrisation [45]. Some systematic errors could result from a bad determination of the background in the wings of the objects.

Noise reduction

The morphological analysis and the parameters extraction depend on the noise reduction done on the image. Generally, a classical smoothing with a gaussian profile allows to obtain the best compromise. The gaussian profile is chosen near the stellar one. This smoothing keeps the flux but it modifies some structural parameters : the central elongation, the gradients, etc... The median filter, at the opposite, does not keep the flux, but keeps the mean slopes. This filter may suppress important structural features, like hot spots or the galaxy nucleus. Other enhancements, like non linear deconvolution [42], could be used, in order to improve the detection of arms, bar, ring or other kind of patterns.

Galaxy detection

The bright galaxies are identified for many years, it is easy to localize them using a computerised catalogue. We can have some problems to locate automatically them on a large digital image. Taking into account the background mapping and the classical detection technique we often detect them as a set of objects, and not as a whole.

For fainter galaxies, the identification strategy is very different [27] [24] [39]. After a classical preprocessing, a detection threshold is determined. The objects are extracted using the image segmentation, allowing to build all the connected fields. Some primitives are computed at this step : area, integrated flux, barycentral position, various moments or extrema.

The separation of the galaxies from the stars is performed by bayesian classifiers. The area, the integrated magnitude, a concentration parameter derived from the weighted moments are the variables used generally in the analysis.

It is evident that the detection based only on the appearance eliminates compact objects. Only a careful photometric analysis, spectographic or radio observations lead to detect galaxies with a quasi stellar shape.

Galaxy location

The image segmentation [33] leads to roughly locate the galaxy. When the galaxy is very large, the center obtained by that technique (weighted or not) may be far from the real center.

The determination of the intensity maximum gives another position which can be better than the barycenter, but can also be very far than the real center (superimposed object, hot spot, etc.)

The center can be determined using a set of rules, taking into account the distance to an isophote, or the relative intensities of the internal maxima.

A multi threshold image segmentation can give also some information on the structure of the galaxy, giving an explanation on the divergence between the barycenter and the maximum multiple object, fuzzy center, etc...

Limits of the galaxy

The morphological analysis depends strongly of the limits given to the galaxy.

The threshold at a given level never gives good limits : if the threshold corresponds to a magnitude of 23"^{2}, it is too high and the photometrical parameters are underestimated, for a magnitude of 25"^{2} the noise is too high to obtain a good contour, and the morphological parameters are badly estimated.

A good method comes from the radial profile [25]. The accuracy increases with the distance , and we can determine the limit at a high magnitude, like 26.5"^{2}.

In many cases, particularly in clusters of galaxies, two galaxies can be so close that one galaxy contaminates the other one. Very few works have been published on the way to process the corresponding images. It is evident that we can not remove a galaxy without a good knowledge on the profile.

Star removal

After the detection and the determination of the limits of the galaxy, it is necessary to remove the stars and the possible superimposed objects.

Identification

It is not evident to identify what objects belong to the galaxy, and what can be considered as field stars. A stellar like object can be a faint globular cluster, an OB association or a superassociation, an HII region, an *hot spot*, a supernova, etc... In each case, the presence of this object can bring an important information on the galaxy classification, we need so to try to eliminate only the features external to the galaxy, essentially galactic stars.

The position of the objects is the first criterium. A bright object far from the galaxy center is generally a superimposed star. The photometrical analysis may roughly separate the normal galactic stars from some features belonging to the studied galaxy (blue associations, globular clusters, etc...) The spectographic observation, giving the radial velocity, is the perfect tool to indicate what objects belong or not to the galaxy. Some observations with narrow-band filters, or scanning Perot-Fabry interferometers may bring also this information.

The study of the light variabilities, the polarisation, or some radio spectroscopy or radio mapping allow us sometimes to obtain the good criterium.

Number of objects to remove

The number of objects depends on many parameters : galaxy size, limiting magnitude, celestial position, etc.... The position, with regard to the galactic plane and the galactic longitude, is an essential parameter. If the galaxy is high on the galactic plane the absorption and the stellar density are small, on the contrary it is very difficult to detect and to identify a galaxy in the galactic plane, especially in a direction close to the Galaxy center.

The number of objects to remove is typically between ten to thirty.

Reduction technique

Some various techniques were proposed to remove point-like or extended objects.

For stellar shape objects, the theorical best technique is based on a local adjustement of a stellar profile. Then the resulting profile is substracted. The residues are around 3 magnitudes higher than the surface magnitude before substraction [7].

Another classical method uses the cicatrisation [44]. The object to remove is circled, and the internal field is step by step filled, using a linear interpolation with the values on the previous contour. Some small defects are observed, especially a smoother internal texture. That tool can be applied to point-like or extended objects.

The faint stars, or some artefacts can be removed using a classical median filter [43]. This operation does not keep the flux, but the morphology is not really modified. Another advantage of this filter is to modify the values only on a small scale.

In the future data analysis, it can be necessary to examine the influence of the removed regions on the computed parameters.

3.3 The galaxy description

The galaxies as physical objects

The identification of the real nature of the galaxies [20] begins at the years 1920s with the first spectrographic observations.

Generally a galaxy is not an isolated object in the Universe, it is a concentration of matter included in a group or a cluster. It is clear that the clusters themselves are correlated, and it exists superclusters and other huge structures [53] [30] [29]. But why do so many galactic patterns exist?

The first fundamental parameter is the total mass. The classical elliptical galaxies are more massive than the spiral ones [4]. The irregular galaxies are generally the less massive objects, with the dwarf elliptical ones. The mass is not the alone parameter which determines the morphology. A galaxy is a mixing of stars, gas, dust and probably non baryonic dark matter [9]. The morphology is strongly correlated with the matter repartition.

A normal or a dwarf elliptical galaxy has only a few percentage of matter as gas or dust. Quite all the luminosity is due to old stars. In opposite the spiral galaxies have a large fraction of matter as gas and dust. It exists evidence of newly formed stars. The irregular galaxies show also recent star formation.

The following scheme tries to explain the structure of the galaxies [11] [41]. A large mass of matter collapses (gravothermal instability), in the core a first population of stars is formed, embedded in a large halo of gas. The stars evolve, the biggest rapidly, the smallest very slowly. It exists an equilibrium between the gas ejected from the stars and the formation of the new ones. The galaxy environment plays a very important part in the gas content, and the evolution of the mass [41]. Many phenomena can be identified: gas striping by a close encounter, galaxy merging, tidal interactions,cannibalism of small galaxies by giant ones, etc... If the gas rate is too small, no star formation is done, the population becomes aged, is mixed and we have an elliptical object; if the gas rate is high a stellar formation is possible. Generally the stars are formed by the interaction of a density wave with the gas. This perturbation has a spiral shape [26]. Other perturbations are possible leading to irregularly distributed newly formed stars.

The studies of the morphological types in the cluster of galaxies have proved the importance of the environment [8]: the elliptical galaxies are in the dense regions while the spiral galaxies are in sparse ones. Quite all the curious morphological objects have been explained by encounter, between two or even more objects.

Another phenomenon plays an important part : the nucleus activity [37]. A classification was done only on its appearance. Photometric and spectroscopic observations have shown the real importance of the nucleus even for normal galaxies. Its properties are not completely understood in the classical scheme. The most active nuclei are always associated to giant elliptical galaxies (when the galaxy can be seen).

The light distribution results from the stellar dynamic [11]. N_body computations tried to explain the different distributions [28]. For elliptical galaxies, a $r^{\frac{1}{4}}$ law [51] is observed. It is interpreted by a truncated isothermal distribution. The S0 galaxies show a mean exponential decreasing [21]. For many galaxies, we can have a superposition of two laws, corresponding to two different populations (disk and bulge) [38].

The morphological classification must take into account all this knowledge. The choice of primitives is guided by the interpretation : if the object is smoothed, it is probably elliptical; if it has granulation, that is due to the stellar formation and it can be spiral or irregular. If the profile decreases rapidly, it is a spiral; if it follows the $r^{\frac{1}{4}}$ law it is elliptical, etc...

The photometrical measures

The profiles

The radial profile is the main photometrical characteristic. It is easy to estimate a profile, but its quality depends i) of the manner that the stars have been removed, ii) of the way to take into account the discrete geometry of the digital image. The spline interpolation may reduce the sampling effects. A fuzzy integration [3] permits also to compute realistic profiles, without an image extension.

The profile can be computed in elliptical rings, in order to have in each ring a quite uniform value. The axis ratio and the orientation are computed on a given typical isophote.

The profile projected on the major and the minor axis can also be easily estimated [56].

Parameters derived from the profiles

The radius r_l from a given magnitude m_l is easy to obtain, when the field is not too crowded.

With the radius determination, we compute the integrated magnitude m_{il} for a given threshold. A surface brightness m_{sl} is obtained by :

$$m_{sl} = m_{il} + 5log_{10}(r_l) + 2.5log\pi \tag{1}$$

The surface brightness is an important parameter in the galaxy classification.

The total magnitude m_t is estimated by the study of the profile convergence.

It is interesting in order to compute the color indices to obtain the magnitudes at given radius m_r.

The profile is also a characteristic of the morphological type. For elliptical galaxies, the profile follows de Vaucouleurs'law :

$$m = ar^{\frac{1}{4}} + b. \tag{2}$$

For S0 galaxies, the mean profile follows an exponential law :

$$m = ar + b. \tag{3}$$

The parameters a and b are related to the equivalent radius r_e and the equivalent magnitude m_e. They can be computed with the least mean square estimator.

Other laws are often used like Hubble's or King's profiles [18] [23].

The central part of the profile is difficult to correctly estimate. The deviation to de Vaucouleurs' law may be interpreted as the existence of a massive central attractor [23].

Bulge/Disk Ratio

For spiral and lenticular galaxies, a classical scheme lies in the superposition of two populations. The older one, the bulge population, follows de Vaucouleurs'law, while the younger one corresponds to an exponential decreasing.

The profile can be decomposed in two parts [38]. The ratio between the two populations gives another classification criterium.

Generally the bulge is near round, while the disk is not seen frontly. The ratio between the populations depends on the angular position. Two profiles are computed in separate regions, taking into account the axis of the disk population.

Photometrical indices

The galaxy evolution can quite perfectly explain the value of the color indices [32]. Roughly, the elliptical galaxies are red because the old red stars give the main part of the flux, while the newly formed bright stars give a blue indice for the spiral galaxies.

It is important to estimate the indice in the same area, and not for the same limiting magnitude.

Flux/Area measures

We have seen that we derive a surface magnitude from the radial profile. We can obtain a similar parameter from the integration on the field limited to a given isophote. The difference is small for regular elliptical galaxies, while the result can be different from spiral or irregular ones.

The morphological parameters

Different morphological features can be extracted from the images :

First of all, some global parameters have to be computed to describe the morphology of the galaxy (the orientation of the main axis, the elongation, the surface, the perimeter, the first moments [15] [14] [12] [1].

This description can be completed by other special features detection (like the texture, the bar, the arms....).

Texture

The description of the texture is very important for a good classification. It can be a simple coarse distinction between a highly smooth or not texture (computed just using variance) but even for a simple measure we need to take care of the quality of the noise reduction. A finer texture description would be useful (but it has not yet been done) to describe the nature of the inhomogeneities in the disk; this texture is of course related to the morphological type but also to the resolution.

Bar

Several ways can be used to detect the bar which is a bright elongated feature accross the bulge [46]. Among them we need to measure the elongation for several isophotes representing the various regions in the galaxy. These elongations have to be compared with the global elongation of the galaxy. Another mean to detect the presence of the bar is to measure for several isophotes their orientation and to compare them with the global orientation of the galaxy.

Arms

Most of the time the arms have very complex shapes, so the isophotes describing the various regions in the galaxy are far from elliptical shapes. A measure for each isophote of the distance from the best ellipse together with the compacity leads to an evaluation of the presence of a spiral shape [46]. But a lot of work still remain to do to describe the structure of the arms. For that ridge detectors [10] adapted to galaxies images are currently studied.

Ring

For bright galaxies the description of the morphology needs the detection of the eventual internal or external rings. These features are not so easy to isolate; they can be circular, elliptical in function of the orientation or with a 8 shape; they can be pertubed by the proximity of the arms, even disrupted. Presently they are essentially visually detected and described [5]. For an automated detection peak detectors algorithms have to be implemented.

Symetries

Another important characteristic , easy to measure, is the relative symetry of the shape or of some subparts (main axis profile asymetries). This is a good way to detect irregular galaxies (magellanic galaxies).

3.4 The classification system

The goal of this section is not to describe the various galaxy classification systems, but to discuss of the knowledge on the classification system that has to be expressed in the knowledge base.

Descriptive knowledge

The first kind of knowledge to express is the description of the various classes: what is a spiral galaxy, what are the various subparts and how to qualify them? This knowledge is a declarative one, and can be represented using structured objects (frames)

example :

```
        CLASS: spiral
        BULGE:{SHAPE:spherical
              TEXTURE: homogeneous
              PHOTOMETRY:r1/4
              LOCALISATION: central}
        DISK:{SHAPE: flat
              TEXTURE: unhomogeneous
              PHOTOMETRY:linear
              LOCALISATION: ....}
        .....
```

Relational knowledge

Of course this description can be more or less complete and precise; thus it is useful to organise these frames showing the hierarchical links existing between the various classes. Schematically the class Galaxy can be subdivided into three subclasses: elliptical, lenticular and spiral; each of these classes can be subdivided into different subclasses and so on. This knowledge can be easily represented using semantic nets with the classes as nodes and the hierarchical dependencies as links.

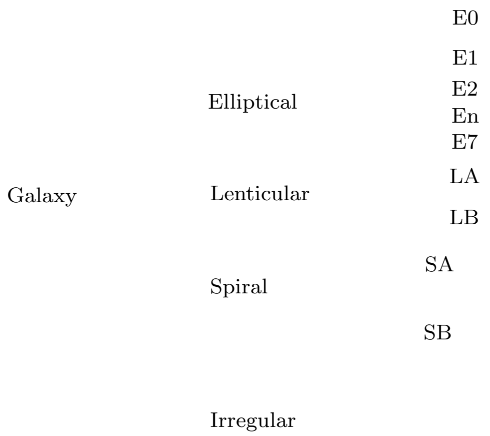


Figure 6: A semantic net describing the various classes

Deductive knowledge

This descriptive knowledge is not sufficient to be able to actually recognize the object detected in the image (or images) as a galaxy belonging to one of these classes. Another kind of knowledge more operational is needed to perform the connection between the numerical or symbolic parameters describing the object of interest and the features used in the description of the classes. This can be done with production rules expressing heuristic criteria as:

if ellipse-error > 0.5
then contour-shape = distorted

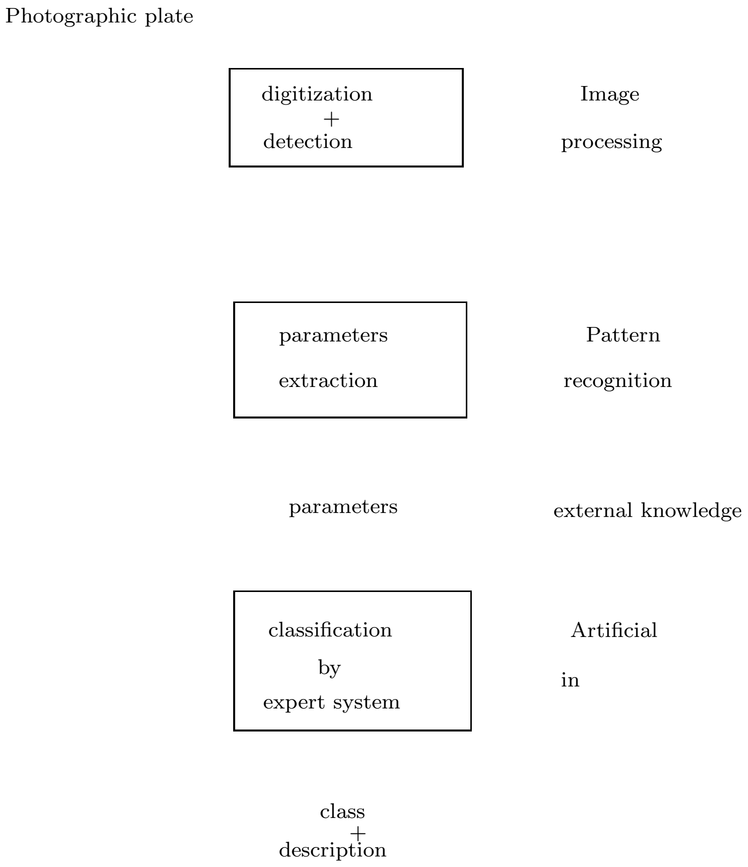
if variance > 6
and contour-shape distorted
then texture unhomogeneous

4 FEASABILITY OF SUCH SYSTEMS

4.1 Introduction

In this section we describe SYGAL an attempt to the building of such a system. This work is described in [46] and [47].

Figure 7 shows the global synoptic of this system



Among the properties of such images, we have to notice that:
These images have a very low signal to noise ratio, which becomes critical for faint objects.
There is no clear limit discriminating the galaxy from the background.
Other objects may be close, or even partially overlapping the galaxy.

4.2 Preprocessing

The preprocessing phase was a fully automatic phase which was performed purely algorithmically; i.e. in a deterministic way and not using an expert sytem supervisor. This phase performed the following functions:

- location of the exact position of the galaxy;
- determination of the limits of the galaxy;
- building and supression of a cartography of the background;
- removal of the objects in the neighbourhood;
- elimination of the noise.

4.3 Parameters extraction

In this section we present the various parameters extracted from the galaxies to describe their shape.

Principal axis

The first parameter computed is the orientation of the principal axis of the galaxy. The value of its direction is given by the angle θ which minimizes the moment of inertia J:

$$J = \sum_{x,y} d(x,y)(y\cos\theta - x\sin\theta)^2$$

Let θ_1 and θ_2 be the two solutions of

$$\tan(2\theta) = 2\frac{\sum_x \sum_y d(x,y)xy}{\sum_x \sum_y d(x,y)(x^2-y^2)} with J(\theta_1) < J(\theta_2)$$

- θ_1 is the orientation of the principal axis;
- θ_2 is the orientation of the small axis (axis orthogonal to the principal axis).

A set of curves representing the intensity of the galaxy from the center in different directions are built. These directions are the two principal half-axis and the two small half-axis.

Ellipticity

The shape of the galaxy is very dependent on the view-point under which the object is observed. Therefore, an important measure is the apparent ellipticity of the galaxy: $e = 1 - \frac{b}{a}$ with a and b respectively the width of the principal axis and the width of the orthogonal axis.

This value is provided by the previous curves indicating the width of the galaxy in the principal axis and in the small axis directions. As these curves are noisy and have a smooth slope due to the diffusion in the emulsion, we use an integrated width value. Let F(x) be the average distribution of the intensity along the principal axis, and f(x) the average distribution of the intensity along the small axis; let w_1 and w_2 respectively the width of the galaxy in the two directions:

The widths are deduced from:

$$\int_0^{w_1} F(x)dx = \frac{3}{4}\int_0^{\infty} F(x)dx$$

and

$$\int_0^{w_2} f(x)dx = \frac{3}{4}\int_0^{\infty} f(x)dx$$

Finally the apparent ellipticity is given by:

$$e = 1 - \frac{w_2}{w_1}$$

Size

In order to estimate the reliability of the measured parameters we compute the size of the galaxy S with the hypothesis of an elliptical shape:

$$S = \pi w_1 w_2$$

Average profile

Using the measure of the apparent ellipticity e we build the curve of the average distribution of the intensity from the center g(x).

The curve g(x) is the radial distribution of the intensity in the new referential

$$(\mathrm{X},\, Y = \frac{w_1}{w_2 y})$$

with X the coordonate on the principal axis and Y the coordinate on the small axis

.

Projected profile

Theorical studies [54] [56] have shown that the radial variation of the luminosity from the center of the galaxies is the sum of two functions;

The first function is the spheroidal component characterizing the elliptical galaxies.

- the intensity I is given in terms of the radius r by:

$$I_1(r) = a_1 e^{-b_1} r^{\frac{1}{4}}$$

- the density $d = \log(\frac{I}{I_0})$ is given in terms of the radius by:

$$d_1(r) = \alpha_1 - \beta_1 r^{\frac{1}{4}}$$

The second function is the flat component characterizing the disk of the lenticular and spiral galaxies.

- the intensity I is given in terms of the radius r by:

$$I_2(r) = a_2 e^{-b_2 r}$$

- the density $d = \log(\frac{I}{I_0})$ is given in terms of the radius by:

$$d_2(r) = \alpha_2 - \beta_2 r$$

In order to avoid the influence of the viewpoint under which the galaxy is observed we build a new curve: the projected profile d_p. This curve is obtained by orthogonal projection of the densities d of the galaxy along the principal axis.

$$d_p(u) = \sum_{x,y,u=x\cos\theta+y\sin\theta} d(x,y)$$

With θ the direction of the principal axis, x and y the coordinates of the pixels belonging to the galaxy. From this curve we extract two parameters:

- the parameter (profile) which is the ratio of the mean square errors made by approximating the projected profile respectively with d_1 and d_2 for r greater than the radius of the bulge. The estimation of the radius of the bulge is obtained from the average profile previously built.

$$profile = \frac{\sum_{r=r_{bulge}}^{r_{limit}} (d_p(r)-d_2(r))^2}{\sum_{r=r_{bulge}}^{r_{limit}} (d_p(r)-d_1(r))^2}$$

- the parameter (linear_err) measuring directly the mean square error made by approximating the complete projected profile for r=0 to $r = r_{max}$ with a linear function.

$$linear\_err = \sum_{r=0}^{r_{limit}} (d_p(r) - d_1(r))^2$$

Contours

Contour building

In order to describe the variation of the structure in the different regions of the galaxy, we need to extract several isophotes. The isophotes must be completely representative of each region in the galaxy and they must be built without interactive process. First we compute five thresholds from the distribution of the density along the principal axis F(x), then these thresholds are used to obtain five binary images on which we apply an edge detector algorithm.

The five thresholds $t_i, i\epsilon[1,...,5]$ are given by:

$$\int_0^{t_i} F(x)dx = \frac{1}{n_i}\int_0^{\infty} F(x)dx$$

with n_i respectively equal to 0.20, 0.50, 0.75, 0.85 and 0.90

For each thresholded image, we use a Sobel edge detector, then we exhibit the maximal chain issue from the contour chaining algorithm of the INRIMAGE [6] library.

Parameters extracted from these contours

From each of these contours we extract five parameters characterizing its shape: ellipticity, angle of the principal axis, relative position of the center, compacity and distance of the closest ellipse.

the angle of the principal axis :
In the same way as for the image, we compute the angle θ_i which minimizes the inertial moment J_i and the orthogonal angle ψ_i.

the ellipticity :
For each contour we compute the ellipticity (the description has already been made).

the relative position of the center :
For each contour we compute the euclidian distance (center_err) between the center of the bulge and the center of gravity of the contour g_i.

the compacity :
The compacity C measures the roundness of a shape and is minimal for a circle ($c \leq 4\pi$).

$$c = \frac{Perimeter^2}{area}$$

In fact we take into account the value of the estimated ellipticity $e = 1 - \frac{b}{a}$ and we measure the quantity:

$$C = \frac{b}{a}\frac{c}{e\pi}$$

the distance of the closest ellipse :

The last parameter (ellipse_err) is the distance between each contour and the ellipse which as its main axis in the θ_i direction, an ellipticity e and is centered in the center of gravity g_i.

Normalization is performed as follows:
Let S_E the area of the ellipse, S_C the area of the contour and S_{dif} the sum of the areas between the two closed curves

$$ellipse\_err = \frac{S_{dif}}{S_E + S_C}$$

Sd is approximated by dividing the area into small triangles, the vertices of which are respectively, the points of the contour C_i, the projections of these points on the ellipse p_i and the intersections of the contour and the ellipse P_j. Let x_{p_i} and y_{p_i} the coordinates of the projections p_i and x_i and y_i the coordinates of the points C_i belonging to the contour.

$$x_{p_i} = a\cos(\arctan(\tfrac{ay_i}{bx_i}))$$

and

$$y_{p_i} = b\sin(\arctan(\tfrac{ay_i}{bx_i}))$$

4.4 Example of measured parameters

We display the parameters extracted from NGC4569.

orientation :	79.33
ellipticity :	0.50
linear_err :	0.039
profile :	2.77
area :	38013.3

Contours :	*center_err*	*ellipse_err*	*compacity*	*angle*	*ellipticity*
*c*1 :	1	0.06	1.6	−6.3	0.09
*c*2 :	2	0.21	5.4	70.8	0.36
*c*3 :	11	0.16	7.1	83.6	0.58
*c*4 :	9	0.21	11.8	76.8	0.47
*c*5 :	7	0.14	8.6	76.1	0.50

4.5 Automatic classification

This phase was performed using an expert system

The engine

The engine used for the expert system was the first version of CLASSIC (see section 2.2) which has been built for this purpose. This version did not allow the call of external procedures (written in C or fortran) so it was not possible to measure new parameters in the image during the reasoning. Some other interesting features as a large number of weights (to balance the rules, the descriptors in the prototypes...) and the multi-window interface was not yet provided in this version of CLASSIC.

The knowledge base

The symbolic parameters

The description of a class is made by the astronomer in terms of symbolic parameters which represent the different structural patterns of the galaxies. We have chosen to introduce such parameters in the knowledge base in addition to the quantified parameters previously described.

- *The symbolic parameter H :*
  This symbolic parameter corresponds to the hierarchical label coding; it takes the values of the possible classes:

  H: E, L, S, E0, E1, ..., E6, LA, LB, LA-, LB- ,LA+ ,LB+ , SA, SB, SA0, ...SA7, SB0, ..SB7

  As the irregular galaxies are not yet classified the associated values are not available.

- *The symbolic parameter T :*
  This parameter corresponds to the continuous label coding; though it takes numerical values, it is not a measured parameter, but each value represents a label:

  T: -5, -3, -2, -1, 0, 1, ...,7

  The values greater than 7 are not considered as they correspond to irregular galaxies.

- *The symbolic parameter bar :*
  This parameter specifies the knowledge about the possible presence of a transversal bar; so, the symbolic values are:

  bar: present, absent, unknown

- *The symbolic parameter shape :*
  This parameter qualifies the general structure of the galaxy; it may take the values:

  shape: elliptical, average, spiral

- *The symbolic parameter isophotes :*
  This parameter points out the degree of perturbation of the contours; the possible values are:

  isophotes: smooth, normal, distorted

- *The symbolic parameter arms :*
  This parameter qualifies the arms of the galaxy; the values are:

  arms: absent, incipient, evident, branched

- *The symbolic parameter bulge :*
  This parameter mentions if the central bulge can be detected on the projected profile; so, the values are:

  bulge: visible, unvisible

- *The symbolic parameter flatness :*
  This parameter qualifies the apparent ellipticity of the galaxy.

  flatness: null, negligible, very_faint, faint, light, average

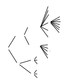
- *The symbolic parameter centring :*
  The parameter centring measures the difference between the centers of the galaxy and the center of the contours.

  centring: good, average, mediocre, indifferent

- *The symbolic parameter profile_concavity :*
  This parameter precises if the concavity of the curve of the projected profile is great or not.

  profile_concavity: great, average, null

- *The symbolic parameter validity :*
  This important parameter specifies the quality of the observation and thus the quality of the measured parameters; it is function of the size of the image of the galaxy. The values are:



validity: good, bad

We have seen in the previous section that for each processed galaxy five contours are built; these contours are described with both symbolic and quantified parameters. The quantified parameters correspond to the parameters which have been extracted from each contour: center_err, ellipse_err, compacity, angle, ellipticity. Two symbolic parameters describe also the contours in the different regions of the galaxy.

- *The symbolic parameter contour c_i: shape :*
  This parameter qualifies the shape of each contour, in the same way as the symbolic parameter shape; it takes the same values as the global parameter:

  contour c_i: shape: elliptical, average, spiral

- *The symbolic parameter contour c_i: isophotes :*
  This parameter indicates the degree of perturbation of each contour; like the global parameter associated to the whole galaxy it can take the values:

  contour c_i: isophotes: smooth, normal, distorted

The symbolic parameter contour c_i: flatness :
This parameter indicates if the ellipticity of each contour is not too great

contour c_i: flatness: valid

<u>*The prototypes*</u>

There are 37 prototypes used to define the various classes; they are organized in a hierarchy which strictly reflects the hierarchy of the classes.

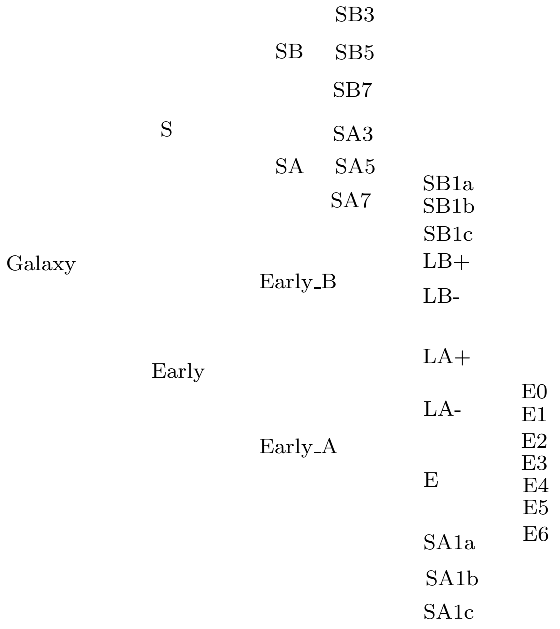


Figure 8: the tree of the prototypes

The descriptors of a prototype characterize the objects belonging to the corresponding class. Symbolic, numeric and complex structures are represented by different data structures. Some descriptors may have complex structures as shown in the following example:

Class contour
compacity : $[0, 10000]$;
ellipticity : $[-0.9, 1]$;
angle : $[-90, 90]$;
ellipse_err : $[0, 10000]$;
center_err : $[0, 10000]$;
shape : *elliptical, average, spiral*;
isophotes : *unknown*;
flatness : *unknown*.

We show here some examples of prototypes describing the classes Galaxy, SB7 and E0.

Class Galaxy
T : $[-5, 10]$;
profile : $[0, 10000]$;
ellipticity : $[-0.9, 1]$;
orientation : $[-90, 90]$;
linear_err : $[0, 100000]$;
area : $[0, 900000]$;
$c1$: *Class contour*;
$c2$: *Class contour*;
$c3$: *Class contour*;
$c4$: *Class contour*;
$c5$: *Class contour*.

Class SB7 *upperclassSB*
H : $SB7$;
T : $[7, 7]$;
shape : *spiral*;
validity : *good*;
bar : *present*;
centring : *indifferent*;
arms : *late*.

Class E0 *upperclassE*
H : $E0$;
T : $[-5, -5]$;
shape : *elliptical*;
bulge : *visible*;
validity : *good*;
profile_concavity : *great*;
flatness : *null*
bar : *absent*;
isophotes : *smooth*;
centring : *good*;
arms : *absent*.

<u>*The rules*</u>

There are 106 production rules. They represent the operating knowledge of the expert. The rules are used to affect a symbolic value to a descriptor of the object.

Example of rule:
Rule 8:

> **If** ellipticity of contour3 $\gg$ ellipticity 0.1
> ellipticity of contour3 > 0.4
> **then** bar be present

"As the ellipticity of the third contour is greater than the global ellipticity of the galaxy and is important, the bar is present."

There are several kinds of rules:

- the rules used to interpret the photometrical parameters (as linear_err or profile) and to affect a symbolic value to the parameters bulge or profile_concavity;
- the rules used to interpret the shape of each contour according to the values of the parameters ellipticity and compacity;
- the rules used to interpret the global shape of the galaxy according to the shape of each contour and the parameter center_err of each contour;
- the rules used to detect the presence of a bar according to the parameters orientation and ellipticity of the galaxy and angle and ellipticity of the contours;
- the rules used to precise the quality of the input parameters and thus of the classification according to the parameter area.

<u>The base of facts</u>

The facts correspond to the data of the current case and are temporarily added in the base ; for our application, the facts represent the information associated to one galaxy. The base of facts consists only of the object to be classified. This object has the same structure as the root prototype. At the beginning, all the values of its descriptors are unknown, but those corresponding to the measured parameters. At the end of the procedure, the symbolic descriptors of the object have the same values as the prototype representing the class of the object.

4.6 Exemple of session

Figure 9 shows the path in the hierarchy of the classes during the classification of NGC4473.
The initial description of NGC4473 is: (measured parameters)

orientation :	−3.18
ellipticity :	0.4
linear_err :	0.061
profile :	0.35
area :	10935

contours :	*center_err*	*ellipse_err*	*compacity*	*angle*	*ellipticity*
*c*1 :	0	0.07	1.6	−1.8	0.41
*c*2 :	1	0.04	1.9	−1.5	0.35
*c*3 :	0	0.05	3.2	−2.3	0.41
*c*4 :	0	0.05	3.8	−1.4	0.43
*c*5 :	5	0.09	6	−3	0.42

24 production rules (among the 106 rules) have been activated in order to complete the description of the measured galaxy; the final symbolic description of the galaxy is:

$$\begin{array}{ll}
GalaxyNGC4473: & \\
class: & E4(compatibility:total, incompatibility:null) \\
validity: & good(certain) \\
arms: & absent(certain) \\
shape: & elliptical(certain) \\
bulge: & visible(certain) \\
flatness: & faint(certain) \\
profile\_concavity: & important(certain) \\
centring: & good(certain) \\
bar: & absent(certain) \\
isophotes: & smooth(certain)
\end{array}$$

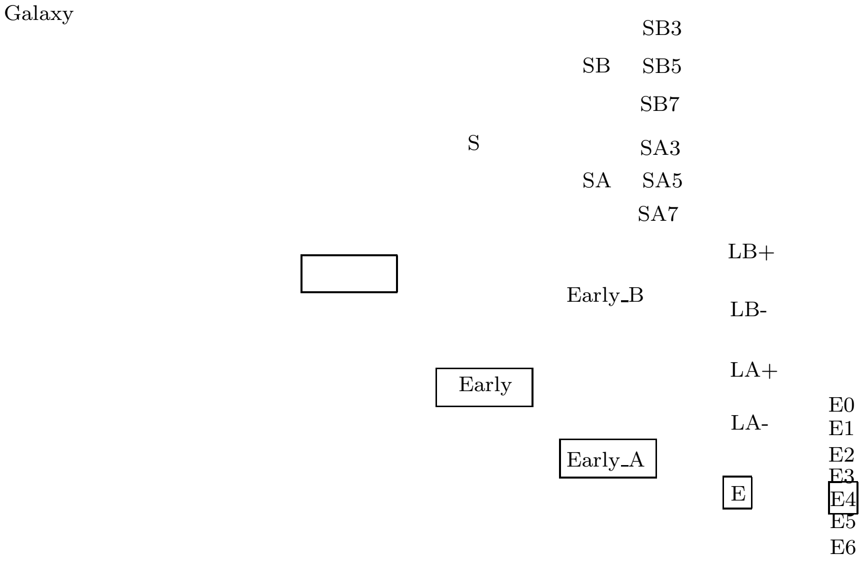


Figure 9: the path in the tree during the reasoning

4.7 Results

Reliability tests

A knowledge base of 37 prototypes and 106 rules has been built. The system has been tested on a set of 21 galaxies in the Virgo cluster digitized from the same Schmidt plate. The results of the automatic classification of these galaxies, is presented in the following table:

Name	Real class	Result	Comment
NGC4459	LAR+	LA+	correct
NGC4569	SXT2	SX3	correct
NGC4474	L...P*	LA+/LB+	correct
NGC4473	E.5..	E4	correct
NGC4477	LBS.*$	L	correct
NGC4571	SAR7	SA7	correct
NGC4468	LA...*	bad validity	small area
NGC4531	S..1	LA+	slight difference
NGC4419	SBS0	SB1/SA1	almost correct
IC3392	SA.3	SB1	almost correct (dust lanes)
M91	SBT3	SB1/SB3	correct
NGC4461	LBS+*	LA/LB	almost correct
NGC4438	SAS0	LB/SB1	correct (dust lanes)
NGC4421	SBS0	LB/SB1	correct
IC800	SBT4P$	SB5	correct
NGC4523	SBS8	SB	irregular
NGC4595	SXT3$	SA1/SA3	correct
NGC4639	SXT4	SB3	correct
NGC4540	SXT6	SB7	correct
M86	.E.3..	E3	correct
M84	.E.1..	E0	correct

Although the results are quiet good, it remains a problem of false detection of a bar in presence of transversal dust lanes.

Tests have shown that the method needs a minimal size for the image of the galaxy (about 50×50). So, the rule:

if area < 2500
then validity bad

is activated to prevent a misclassification; in order to decrease the computing time, the parameter validity is set to good for each successor node of the root node Galaxy; so, if the previous rule is activated, no successor may be validated and the classification returns class unknown whatever are the other values of the measured parameters.

The results obtained on a set of galaxies observed with another telescope (Schmidt telescope of Observatoire de Haute-Provence) and with different exposure times show that the knowledge base must contain information about these symbolic parameters (telescope and exposure_time). In function of the values of these parameters different branches of the tree of the prototypes must be scanned.

In its current state the system classify a galaxy in approximately 122 seconds: 120 seconds for the extraction of the parameters (which is of course highly dependent on the size of the image) and 2 seconds for the inference phase.

4.8 Limitations

SYGAL is a complete automatic method to classify complex objects (galaxies), using an expert system approach. This system which has a knowledge base made of 37 prototypes and 106 rules must be completed to enhance the precision of the classification but the classes obtained by this method are quiet good.

Yet SYGAL, at its stage of development, has still strong limitations to be an instance of a real galaxy classification system; Several kinds of limitations can be distinguished:

Artificial intelligence tool

The artificial intelligence tool which was used was a first version of CLASSIC , for which there was not the possibility to go back to the low level during the classification reasoning to run specialized procedures.

In SYGAL the image processing was not supervised by an intelligent system (like an OCAPI expert system) specialized in monitoring.

Algorithms

There are no peak or ridge detection algorithms:
there are no real arms detection but the analysis of the contour shape; so, there is no arm/interarm measure;
there are no ring pattern detection or extraction
in the same way there is not a real bulge/disk ratio measure but a measure of the degree of visibility of the bulge on the projected profile and the analysis of the contour shape.

Knowledge base

In fact, the knowledge base does not yet contain information about the irregular, the peculiar galaxies and about the ring pattern structures. This lack of information is mainly due to the absence of such galaxies in the sampling set, so some study are needed in this domain. Because of the resolution of the images used for the development of SYGAL, there was no knowledge on the arm/interarm ratio and on the structure of the arms.

5 VALIDATION

A real automated system for the classification of galaxies need to be validated with strictness. We can distinguish four kinds of validations:

Algorithms validation

This validation concerns the expertise in vision and image processing. The generality, the applicability, the robustness, the precision of the results and the meaning of the results of each procedure have to be checked.

Extracted parameter interpretation validation

The output of the image processing algorithms are used in the production rules or even in the prototypes; the validation of the knowledge which performs the passage of these measured parameters to the astronomical domain imply an expertise in both vision (for a good interpretation of the output of the algorithms) and in astronomy (for the understanding of the astronomical features).

Classification knowledge validation

The knowledge in classification expressed in the tree of the prototypes and in the prototypes is an astronomical knowledge; thus this part needs to be validated by astronomers specialized in galaxy classification. This validation has to check the completeness, the correctness and the coherency of the expertise.

Images validation

The validation of an expert system (even if each part of the knowledge is checked by the experts of the corresponding skillness) is done according to one or several sets of data (images). So, a good validation needs large sets of homogeneous data with their context (observational conditions, image formation conditions,....), their class and their calibration data. This particular skillness is an observational expertise.

6 CONCLUSION

We have seen that the classification of galaxies is a very complex task, but that it is not really impossible to automate it. This can be useful both to process objectively large sets of data and to centralize and structure different kinds of expertise (in image processing, observation, classification); this last point becoming crucial with the multiplication of the sensors used and the difficulty to perform multisensor analysis.

The development of the instrumentation (an increasing number of galaxies are studied with CCD) will give us a lot of new informations. The knowledge on the morphology of these objects will seriously increase. That will surely lead us to refine the classification scheme.

This method can be extended to the study of others objects instead of galaxies; for instance for the study of clusters.

References

[1] Accomazzi A. Bordogna G. Mussio P. Rampini A. Rule based description and plausible classification of object in digitized images *3rd International workshop on data analysis in astronomy* Ed. di Gesu Plenum Press Erice, Italy June 1988

[2] Bijaoui A. 1980 *Astron. Astroph.* **84** 81-84.

[3] Bijaoui A. 1988 *12nd IMACS Conference* II p.316 Ed. R.Vichnevetsky, P.Borne, J.Vignes.

[4] Binggeli B., Sandage A., Tamman G.A. 1988 *Annual Rev. of Astron. Astroph.* 26 pp 509-650.

[5] Buta R. "Galaxies morphology and classification" *in Le Monde des Galaxies* april 1988 Springer verlag

[6] Cipiere P. "Manuel de reference du logiciel INRIMAGE" *INRIA* 1984

[7] Debray B. 1988 *Thesis University Aix-Marseille.*

[8] Dressler A. 1980 *Ap.J.* **236** p.351.

[9] Faber S. 1987 *in Dark Matter in the Universe* ed. J.Kormendy, G.K.Knapp pp.1-16.

[10] Fairfield J. 1989 A nearly invisible grid *Computer Vision and Pattern Recognition* CVPR 89 (submitted)

[11] Freeman K.C. 1975 *in Galaxies and The Universe* ed. A.Sandage, M.Sandage and J.Kristian pp.409-508 The University of Chicago Press Chicago.

[12] di Gesu V. and Maccarone C. *ICPR* 1986

[13] Granger C. and de Mongareuil A. "The reference manual" *Ilog* 1987

[14] Grosbol P. "Extraction of morphological parameters from images of galaxies" in *Astronomy for large databases* Ed. F. Murtagh and A. Heck 1987

[15] J.F. Jarvis and J.A. Tyson "Focas:Faint Object Classification and Analysis System" *Astron. J.* 1981 Vol. 86 (p. 476)

[16] Hanson A. and Riseman E. "VISIONS: A Computer System for Interpreting Scenes" *Computer Vision Systems* A. Hanson and E. Riseman, eds Academic Press 1978 (p. 303)

[17] Hayes-Roth F. Waterman D. A. and Lenat D. B. 1983 in *Building expert systems* Reading, Mass: Addison-Wesley

[18] Holmberg E. 1975 *in Galaxies and The Universe* ed. A.Sandage, M.Sandage and J.Kristian pp.123-158 The University of Chicago Press Chicago.

[19] Hubble E. P. 1926 *Ap. J.* **64**, 321

[20] Hubble E. 1936 *The Realm of the Nebulae* Yale University Press New Haven

[21] Johnson H.M. 1961 *Ap.J.* **133** p.314.

[22] Jones W.B., Obbits D.L., Gallet R.M., De Vaucouleurs G. 1967 *Publ. Dpt Astr.* Univesity of Texas, Austin Ser. II 1 n. 8.

[23] Kormendy J. 1980 *in Two Dimensional Photometry* ed. P.Crane and Kjär p.191-238 E.S.O. Munich.

[24] Kurtz M.J., Huchra J.P., Beers T.C., Geller M.J., Gioia, Maccaro T., Schild R.E., Stauffer J.R. 1985 *Ap.J.* **90** pp. 1665-1675.

[25] Le Fèvre O., Bijaoui A., Mathez G., Picat J.P., Lelièvre G. 1986 *Astron. Astrophys.* **154** p.92-99.

[26] Lin C.C., Yuan C., Shu F. 1969 *Ap.J.* **155** P.721.

[27] MacGillivray H.T., Stobie R.S. 1984 *Vistas in Astronomy* **27** p.433.

[28] Martinet L. 1985 *in New Aspects of Galaxy Photometry* ed. J.L.Nieto p.121 Springer Verlag Berlin.

[29] Oort J. 1984 *in The Big Bang and Georges Lemaitre* ed. A.Berger p.299 Reidel Dordrecht.

[30] Peebles P.J.E. 1980 *The Large Scale Structure of the Universe* Princeton University Press, Princeton NewJersey.

[31] Pouget S., Thonnat M., Clement V. and de Fombelle A. J. "Automatic diagnosis of a technical device by classification, the expert system DANTE: an application to antenna failures" em 6th international conference on reliability and maintenability 1988 Strasbourg, France

[32] Rocca-Volmorange B. 1985 *in New Aspects of Galaxy Photometry* ed. J.L.Nieto p.199 Springer Verlag.

[33] Rosenfeld A. 1969 *Picture Processing by Computer* p.127 MacGrawHill New York.

[34] Sandage A. 1961 *Hubble Atlas of Galaxies* Carnegie Institution of Washington Pub. 618 Washington.

[35] Sandage A. 1969 *Ap. J.* **157**, 515

[36] Sandage A. 1975, in *Galaxies and the Universe, Stars and Stellars Systems* Vol. IX, ed. A. Sandage and Kristian J. (University of Chicago Press)

[37] Sersic J.L. 1982 *Extragalactic Astronomy Reidel Dordrecht.*

[38] Simien F., De Vaucouleurs G. 1986 *Ap.J.* **302** pp.564-578.

[39] Slezak E., Bijaoui A., Mars G. 1988 *Astron. Astroph.* **201** p.9-20.

[40] Shortliffe E. H. 1976 in *Computer-based medical consultation: MYCIN* New York: American-Elsevier

[41] Silk J. 1986 *Sky and Telescope* p.582-587.

[42] Titterington D.M. 1985 *Astron. Astrophys.* **144** pp.381-387.

[43] Thonnat M., Llebaria A. 1981 *in Astronomical Photography* ed.J.L.Heudier et M.Sim pp.251-260 CNRS-INAG Paris.

[44] Thonnat M. 1982 *in Traitement Numérique des Images Astronomiques* ed.A.Bijaoui pp.255-279.

[45] Thonnat M., Llebaria A. 1983 *Applications of Digital Image Processing* Proc. SPIE v.397 pp 454-461.

[46] Thonnat M. "Automatic morphological description of galaxies and classification by expert system" *Rapport de recherche INRIA* march 1985 No 387

[47] Thonnat M. "Toward an automatic classification of galaxies" *in Le Monde des Galaxies* april 1988 Springer verlag

[48] Thonnat M. and Clement V. "OCAPI: an artificial tool for the automatic selection and control of image processing procedures" *3rd International workshop on data analysis in astronomy* Ed. di Gesu Plenum Press Erice, Italy June 1988

[49] Thonnat M. and Clement V. "OCAPI: a monitoring tool for the automatic control of image processing procedures" *12th World congress on scientific computation IMACS* II p.318 Ed. R.Vichnevetsky, P.Borne, J.Vignes.

[50] Thonnat M. and Gandelin M.H. "An expert system for the automatic classification and description of zooplanktons from monocular images" *9th International Conference on Pattern Recognition ICPR9, Roma, Italy, November 1988*

[51] de Vaucouleurs G. 1948 *Ann.Astroph.* **11** p.247.

[52] de Vaucouleurs G. "Classification and Morphology of External Galaxies' *Handbuch der Physik* 1959 Vol. 53 (p. 275)

[53] de Vaucouleurs G. 1970 *Science* n.3922 v.167 pp.1203-1213.

[54] de Vaucouleurs G. in *Le monde des galaxies* Obs. de Besançon et lab. d'astronomie de la Faculté des Sciences 1976

[55] de Vaucouleurs G. de Vaucouleurs A. and Corwin H. G. Jr. 1976 *Second Reference Catalogue of Bright Galaxies* Texas University Press, Austin (RC2)

[56] Watanabe M. K. Kodaira and Okamura S. "Digital surface photometry of galaxies toward a quantitative classifica tion" *Astrophys. J. Suppl. Series* 50 pp.1-22 1982

[57] Winston P. H. in *Artificial Intelligence*, Addison-Wesley, Readings, Massachussets (1977)

[58] Wolf M. 1908 *Pub. Ap. Inst. König, Heidelberg*, Vol. **3** No 5