Accessing, Mining, and Archiving an On-line Database: The APS Catalog of the POSS I

Roberta M. Humphreys
University of Minnesota

Juan E. Cabanela
Saint Cloud State University

Jeff Kriessler
University of Minnesota

11 January 2001

197th AAS Meeting (San Diego, CA)

#116.06

Abstract

The APS Catalog of the POSS I is an on-line database of over 100 million stars and galaxies (*http://aps.umn.edu/*). A unique subset of this database with over 218,000 galaxies within 30˚ of the North Galactic Pole, the MAPS-NGP, is now available at our web site. This diameter-selected catalog (≥10 arcseconds) is the deepest galaxy catalog constructed over such a large area of the sky (3000 square degrees). The MAPS-NGP includes many additional parameters for the galaxy images not available in the APS Catalog.

Working with members of our computer science department, we have developed a morphological classifier for galaxies that divides our galaxy type into three classes – early, intermediate, and late. We have applied data mining techniques to identify the most useful image parameters for input into a neural network and decision--tree based classifier pipeline.

We are also archiving the APS Catalog for distribution to astronomical data centers including NASA's ADC and SIMBAD at CDS. The extragalactic subset will be integrated into the NASA/IPAC extragalactic database (NED). The MAPS-NGP has already been provided to NED.

The APS is supported by NASA's Applied Information Systems Research Program.

197th AAS Meeting (San Diego)

The APS Project & The APS Catalog of the POSS I

- **The Automated Plate Scanner (APS) Catalog of the POSS I**
  - an on-line database of fundamental data and parameters for about 100 million stars and galaxies
  - derived from digitized scans of glass copies of the blue and red plates of the original, first epoch Palomar Observatory Sky Survey (POSS I).
- **APS Catalog Contents**
  - Contains coordinates, magnitudes, colors, and many other computed image parameters for all matched images on the blue and red plates.
  - Objects down to 21st magnitude (in the blue).
  - **The Image Classifier:** A neural network image classifier used to separate stellar and non-stellar images (Odewahn et al. 1992, 1993, 1995). It has been trained to the faint limit of the photographic plates and has a >90% success rate to within one magnitude of the plate limit.
- **APS Catalog Access**
  - Querying is achieved via a custom-designed database management system called *StarBase*.
  - The completed catalog of objects is available as an on-line database and a Finder Chart service.
  - A complementary image database is also available and includes all of the matched images in the object catalog as well as the unmatched images above the noise threshold on both the blue and red plates.

FOR MORE INFO...

Go to the APS Website at *http://aps.umn.edu/*

197th AAS Meeting (San Diego)

The MAPS-NGP Subset

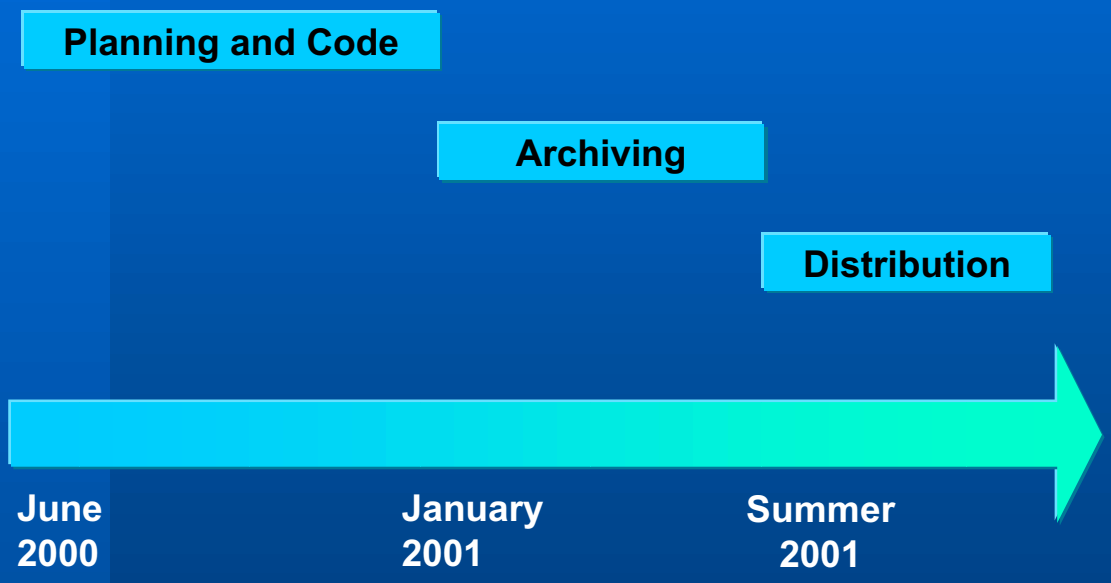
- 217,768 Galaxies covering 3089 square degrees and 97 POSS I fields.
- All known non-stellar images (stars, ghost images, globular clusters, etc.) purged from MAPS-NGP.
- Completeness Limits:
  - O~18.6 and E~16.8.
  - O diameter of 10".
  - $\langle V/V_{lim} \rangle$ indicates presence of local inhomogeneities.
- Currently available online at the APS Website and through NED.

Archiving an On-line Database: Strategy and Distribution

- **Motivation**
  - Planning for possible construction in the lab containing the APS computers.
  - **Planning for the survival of the APS dataset beyond the survival our laboratory and computers.**
- **Distribution**
  - The APS Catalog of the POSS I will be distributed on approximately 5 to 6 DVDs with data divided by POSS field.
  - Data will be stored in flat binary files, with subroutines provided for reading the data files in C, Fortran 77, and Perl.
  - Documentation for the archival version of the APS Catalog will be provided on the DVDs.
  - Redistribute to multiple online astronomical resources: Each with their own focus:
    - *NED:* Will include all “galaxies” from the APS Catalog
    - *The Astrophysical Data Center:* Will archive the entire catalog
    - *CDS:* SIMBAD will archive the entire catalog

Archiving Schedule

- **Current APS Catalog Archiving Schedule**



Current Status: We have nearly completed writing the code and are now preparing for creation of the beta version of the archive.

Image Parameters Archived

Variable	Plate	Description
starnum	O,E	image raster number
ra		Right Ascension (1950) (stored in seconds)
dec		Declination (1950) (stored in arcseconds)
Xsct	O,E	X centriod position on plate (in ERE)
Ysct	O,E	Y centriod position on plate (in ERE)
dia	O,E	Major-Axis diameter (in arcseconds)
magi	O	integrated O magnitude
magd	O	D-M O magnitude (appropriate for stellar images)
colori	O,E	O-E color from integrated magnitudes
colord	O,E	O-E color from D-M magnitudes
mean_sb	O,E	mean surface brightness
theta	O,E	position angle (from moments analysis)
ell	O,E	image ellipticity (from moments analysis)
galnod	O,E	galaxy node value from classifier
Psat	O,E	Percent saturation of image
Tavg	O,E	Average transmittance
Tsky	O,E	Sky transmittance(from MBACK scan)
Reff	O,E	Effective (half-light) radius
C42	O,E	C(r100/r50) concentration index
C32	O,E	C(r75/r50) concentration index
Mir1	O,E	First Moment of Image
Mir2	O,E	Second Moment of Image
flag		O and E imgpars flags (10*Obad + Ebad) 0: Image OK 1-2: Unexpected Machine Noise 3: Likely Scratch (ellipticity > 0.95) 4: Clipped Image 5: no MBACK coverage for this image 6: Negative Sky Intensity computed

Automated Morphological Classification of Galaxies: Why?

- **Motivation**
  - For many astrophysical problems the actual morphological type of the galaxy is very important, especially for studies of galaxy formation and evolution and large-scale structure in the universe.
- **The Problem**
  - The classification of galaxies is typically performed by visual inspection of the images requiring a great deal of practice and time on the part of the classifier. With today's large all-sky surveys, generating millions of galaxy images, **human classification is no longer a viable option.**
  - Furthermore, **human classifications tend to be subjective**; studies show that morphological catalogs of galaxies produced by even the best human classifiers disagree between 10% and 20% of the time.
  - Therefore, in order to produce large, objective catalogs of morphological types, **computer generated classifications are required.**

- *We have recently had some success applying data mining and pattern recognition codes to identifying the most useful parameters for automating the classification of the galaxy images by their morphological types.*

Automated Morphological Classification of Galaxies: Our Approach

- **Test Sample**
  - visually classified some 1500 galaxy images from the APS database in the region of the north galactic pole. Galaxies which were hard to classify or with uncertain types (<1%) were removed from this sample.
  - we have calculated over 500 image parameters (two colors) for each galaxy.

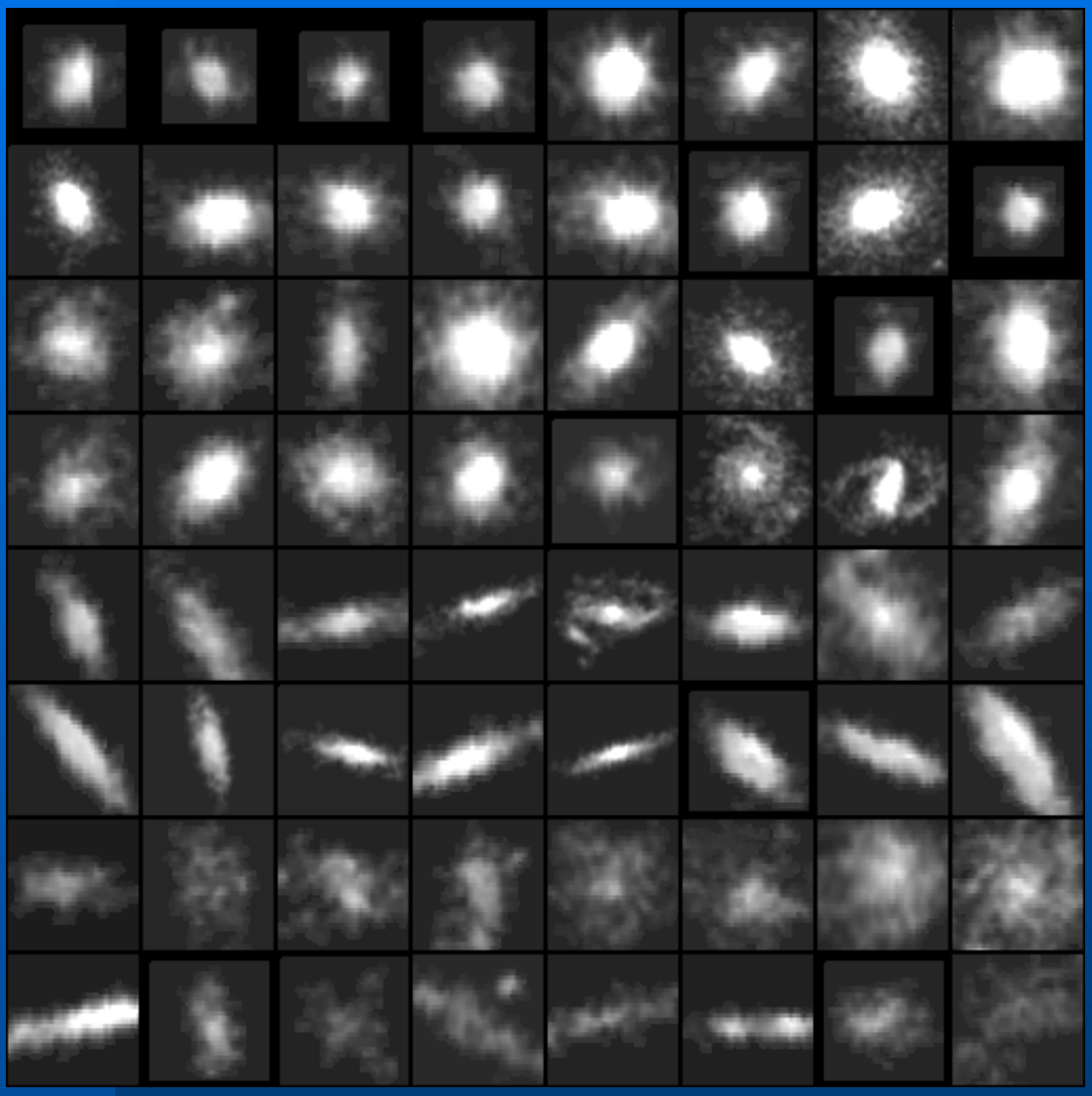
- **Parameter Selection**
  - We cannot simply present all of these parameters to a neural network and let the training algorithm determine which are the most important. We would merely end up with a network that has memorized the training sample perfectly, but performs poorly on samples not seen during training.
  - ***For practical reasons we must limit the number of parameters presented to the neural network.***
  - ***Finding clusters in large dimensional spaces falls within the sphere of data mining.***

Automated Morphological Classification of Galaxies: Data Mining

- **First Results:** Working with the data mining group in the Computer Science Dept. at the University of Minnesota
  - We have experimented with two different software packages on a simplified three class system:
    - ***Early: ellipticals and S0's***
    - ***Intermediate: spirals (Sa, Sb, Sc)***
    - ***Late: (Sd, Im)***
  - ***MineSet*** (commercially available code from SGI) allows the quick evaluation and ranking of the parameters as well as creating a decision tree classifier. Using the 10 best parameters we have created a classifier with a 85% success rate for all three classes.
  - ***c4.5*** a publicly available code that is similar to *MineSet* and also creates a decision tree classifier. we have run several different tests with this package. Separating the early types from all "others" we also get an 86% success and with an intermediate/late separation better than 80% is achieved.

- **Preliminary results**: We are continuing to work on improving the success of the classifier, but given that the best human classifiers disagree 10 – 20% of the time based on classifications primarily from photographic plates, **a success rate of 85% from the POSS I may be realistic.**
  - Some of the most "promising" parameters from the different tests are listed in Table 1.

Automated Morphological Classification of Galaxies: Sample Images from A2151



APS images of galaxies in A2151 order by morphological classification. A decision tree was used to classify the images as ellipticals/S0s (top 3 rows), spirals (rows 4,5, and 6) and irregulars (bottom two rows). A separate classifier was used for high ellipticity images.

197th AAS Meeting (San Diego)

Automated Morphological Classification of Galaxies: Successful Parameters

TABLE 1: The Most "Promising" Parameters from the Different Automated Morphological Classification Tests.

O–E	**O–E color**
$\Delta\mu 32_O$	**difference in surface brightness between R_{75} and R_{50} for the blue image**
Σa_E	**sum of Fourier amplitudes on red plate**
$C32_O$	**concentration index between the 75% and 50% flux level of the blue image**
$C32_E$	**concentration index between the 75% and 50% flux level of the red image**
ε_E	**ellipticity on the red plate**
μ_E	**surface brightness at R_{25} on the red plate**
$O–E_i$	**color from integration of surface brightness profile**
G_O	**slope of the elliptically averaged intensity on blue plate**