

The Ultimate Sharper Image Catalog

by Jay Aller

The finished catalog will list an estimated 50 million galaxies and 2 billion stars, or several hundred times more information than is contained in the largest existing data sets.

Remember when you first learned to count, and were so proud that you would count from 1 to 100 for anyone who would listen? But soon you found that it took such a loooong time to reach 100 that the fun went out of the counting. Now imagine tallying more than a billion stars, and not just counting them, but also noting their location, brightness, and other vital statistics—it would take an eternity, or at least the entire life spans of many people. Up until now, “approaching the sky-object classification task manually has been forbidding,” to say the least, explained Associate Professor of Astronomy S. George Djorgovski.

But a new computer software system developed jointly by Djorgovski and Nick Weir (PhD '94), and Usama Fayyad and his colleagues in JPL's Artificial Intelligence group, promises to change all that. Called the Sky Image Cataloging and Analysis Tool, or SKICAT (pronounced “sky-cat”) for short, the system is like one of those miracle devices advertised on late-night local television; it's many tools in one: a classifier, a catalog, a database. It stops short of making julienne fries.

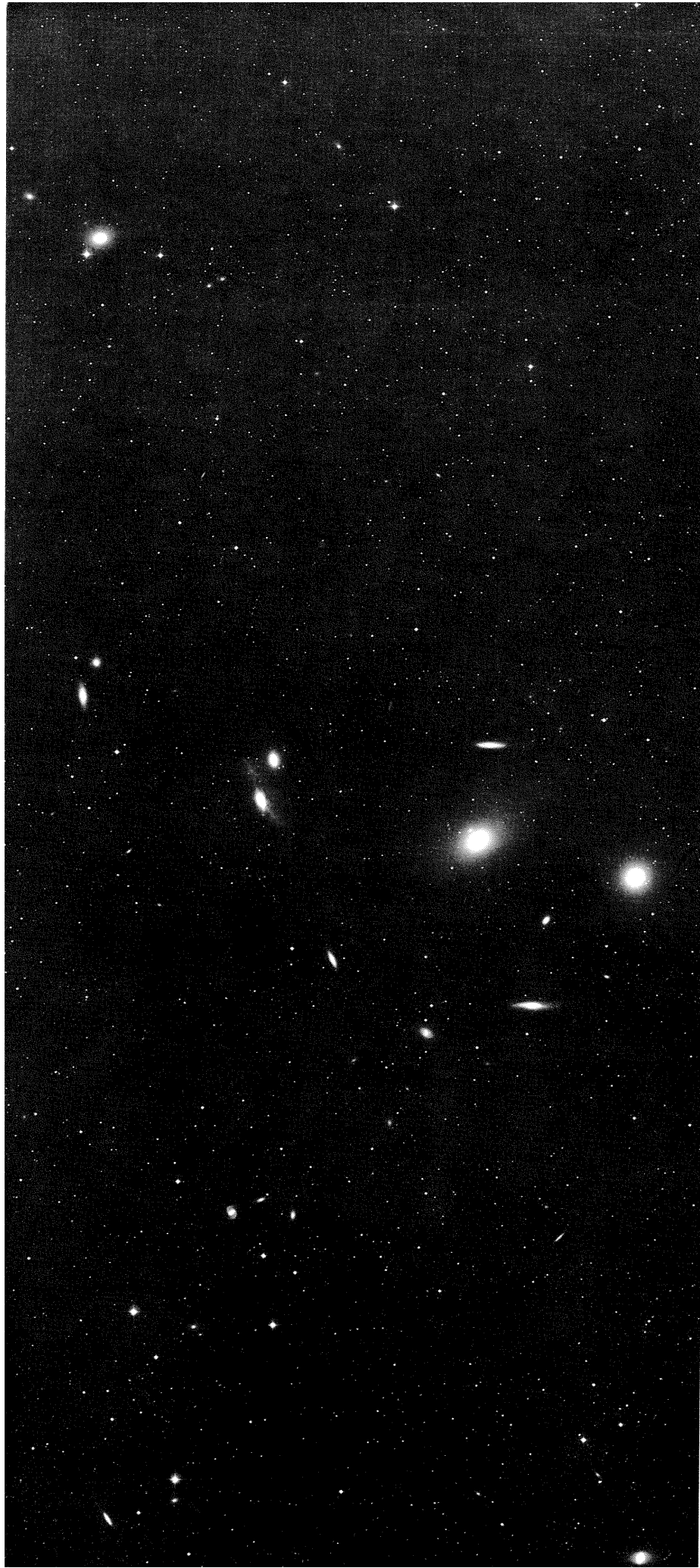
SKICAT's powerful new programs enable it to analyze previously unscalable mountains of data, relieving astronomers from the tedious and visually demanding task of classifying objects by poring over a photograph with a magnifying glass, and freeing them to pursue more challenging problems.

In addition to classifying billions of objects much faster than humans could, SKICAT also classifies objects better than humans can, in

several ways. For one, it bases classifications on objective criteria, eliminating the biases that creep in when astronomers make judgment calls. It also has a very high correct identification rate of 94 percent. This exceeds the 90 percent necessary for scientific analysis of the data to yield useful results. And, most amazing of all, it is able to detect and categorize objects that appear too faint in the photographs to be recognizable by the human eye.

As SKICAT quickly and accurately classifies the millions of sky objects, it will store them in a new type of astronomical catalog that is revolutionary in both its size and form. The finished catalog will list an estimated 50 million galaxies and 2 billion stars, or several hundred times more information than is contained in the largest existing astronomical data sets. And, unlike other catalogs, which are printed and updated only every few years, the catalog created by SKICAT will always be changing and growing as it is updated with new information. Users will never print the catalog in its entirety, for it would fill roughly 50,000 large volumes, or roughly one floor of Caltech's Millikan Library. Instead they will be able to browse through the billions of entries and sort them by location, magnitude, color, or other properties, all by computer.

The inspiration for the development of SKICAT comes from the Second Palomar Observatory Sky Survey (POSS-II), an effort currently under way to photograph the entire Northern Hemisphere sky. Astronomers are using the Oschin Telescope on Palomar Mountain, a 48-



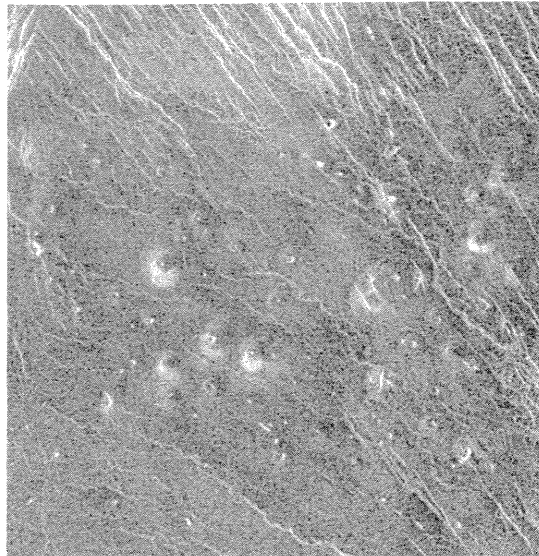
*Multiply those
2,700 photo-
graphic plates by
up to 10 million
objects per plate,
and it's . . . well,
an immense
number.*

**This corner of a
POSS-II plate shows a
section of the Virgo
cluster of galaxies.**

inch instrument also used for the original Palomar Observatory Sky Survey (POSS-I) back in the mid-1950s.

POSS-II will pose a pleasant problem for Caltech astronomers—more data than they can deal with. To enable SKICAT to digest the billions of objects, the survey photographs will be converted into a digital form, which will provide a rich vein for the mining of new information. Using the catalog based on POSS-II, astronomers will be able to map the large-scale structure of the universe and the finer structure of our own Milky Way galaxy, study the evolution of galaxies over billions of years, and pick out large numbers of rare or exciting objects, such as high-redshift quasars. But before scientists can even start any of these interesting projects, the raw data of POSS-II must be transformed into a properly classified catalog. Hence, SKICAT.

The scientists' present goal is not only to recreate the 1950s sky survey, but also to make a better survey, better both in sensitivity and accessibility. The new survey is able to detect objects approximately 1 to 1.5 magnitudes fainter than the original, due mainly to the new fine-grain emulsion film and the better image quality of the improved telescope optics. These advances also make classification of faint objects as either stars or galaxies possible to at least 1 to 2 magnitudes fainter. The dimmest detectable objects are near 22nd magnitude, or a few million times fainter than can be seen under optimal conditions by the naked eye. The magnitude gain would be larger, except that the sky above Palomar is now much brighter than it was 40



This digital image of Venus from the Magellan spacecraft shows numerous small volcanoes, some of the perhaps million volcanoes on the planet's surface. How many can you spot? See page 35.

years ago, due to the encroaching lights of San Diego County. And since SKICAT can classify sky objects that are too faint for humans to recognize, the resulting catalog will contain a wealth of new information.

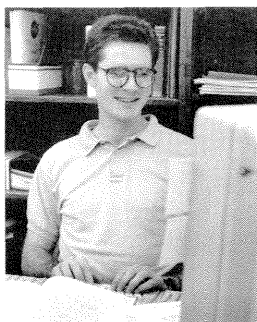
Astronomers also want to make data from POSS-II more accessible for themselves and their colleagues. The 1950s sky survey remains a remarkable accomplishment, beautifully preserved in atlases, but much of the data is in practice inaccessible, simply because there is so much of it. Poring over the photographic plates to pick out certain types of objects would require absurd lengths of time. POSS-II will present the same problem. In fact, due to the greater sensitivity of the present survey, there will be even more objects to sort through. POSS-II is expected to photograph four times more objects than were seen in the 1950s. So Djorgovski and Weir turned to Richard Doyle and his colleagues in the Artificial Intelligence Group at JPL for assistance in creating a catalog from the POSS-II data. Working with Usama Fayyad, Weir developed features in SKICAT that enable it not only to process immense data sets, but that also make the resulting catalog easy to use.

The 48-inch Oschin Telescope can photograph the entire Northern Hemisphere sky with about 900 exposures. Each exposure is made on a square photographic plate, 14 inches on a side, and each plate contains up to 10 million objects. Because the scientists want to see as many faint objects as possible, they expose each plate for up to an hour or two. On a good night, astronomers can expose three plates. To gauge the stars'

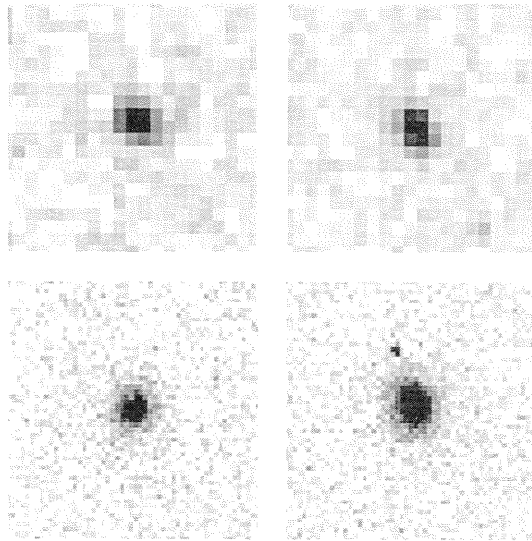
colors, which are all-important in determining their temperature, each square of sky is photographed not once but three times, in the colors of blue-green, red, and near infrared—wavelengths of 480, 650, and 850 nanometers—so the entire survey will produce some 2,700 photographic plates.

Multiply those 2,700 photographic plates by up to 10 million objects per plate, and it's . . . well, an immense number. Before SKICAT can classify these myriad points of light, the information must be converted into digital form. So Palomar sends the 14-inch square plates to the Space Telescope Science Institute (STScI) in Baltimore, where each plate is converted into an electronic image containing more than 500 million picture elements, or pixels, in a 23,040 by 23,040 grid. To give an idea how sharp this resolution is, a television or computer screen contains only about 250,000 pixels, on a 512 by 512 grid. With each plate digitized into almost 100 billion bits of information, the total amount of data quickly becomes, (dare I say it?), astronomical. To give some idea of how much total data will be collected, the estimated three terabytes (24,000,000,000,000 bits) of information is several hundred times more than that gathered by the Infrared Astronomical Satellite (IRAS), one of the largest collections of data ever.

After STScI records the digitized information on tape, it sends the tapes back to Caltech, where the data are fed into SKICAT, which will automatically process the roughly 24 trillion electronic bits of image data to produce a comprehensive catalog in the form of a computer database con-



Nick Weir counts galaxies to see how they evolve.



The top two images are digitized objects that look virtually identical, from one of the POSS-II plates. Below them are the same two objects, but from higher-resolution CCD frames. SKICAT correctly identified the one on the upper left as a star and the one on the upper right as a galaxy.

Some classification tasks performed in the past over a period of years could now be finished in a few hours with SKICAT.

taining an estimated two billion entries. This analysis is hundreds of times faster than present methods. "Some classification tasks performed in the past over a period of years could now be finished in a few hours with SKICAT," Nick Weir said.

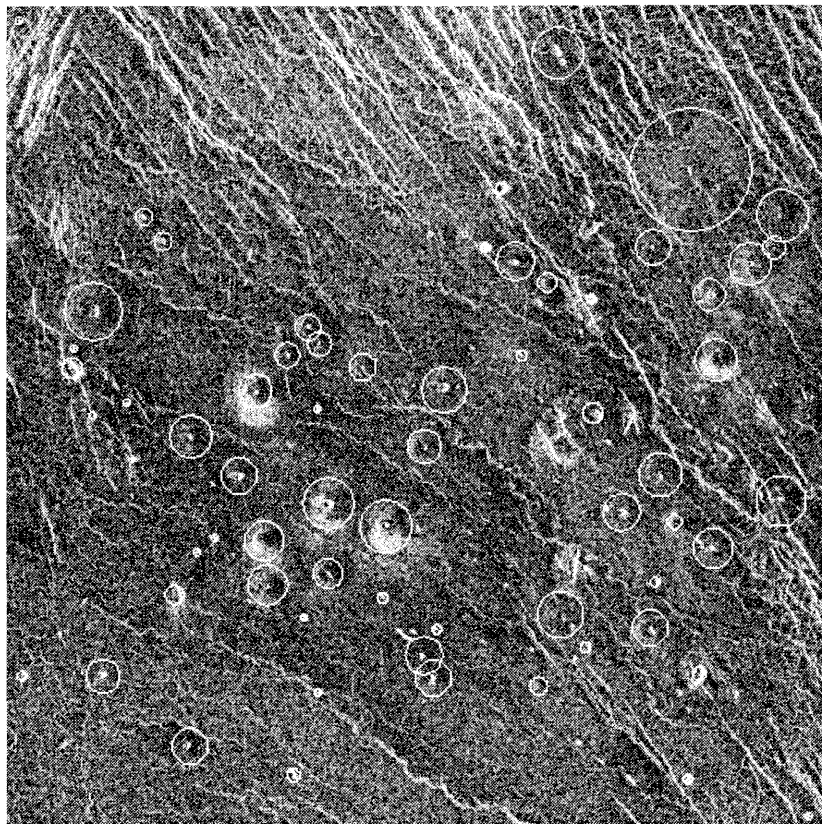
Astronomers estimate that the survey, begun in late 1986, will be 90 to 95 percent complete by late 1996. (All-sky surveys take a long time; the original one, with two-thirds the number of plates of this time around, took eight years—from 1949 to 1957.) Digitizing the data didn't start until 1992, but is proceeding more quickly than the survey itself, so the two tasks should be finished within a year or two of each other. Estimated completion time of the digitizing is roughly 1995 to 1997. The resulting Palomar Northern Sky Catalog (PNSC), a continuously updated database accessible by computer network links, will be an entirely new type of astronomical catalog. A partial release of the PNSC may start in 1995, with a nearly complete release planned for 1998. The resulting data set will not be surpassed in scope for the next decade.

SKICAT's machine intelligence comes into play after the digitized information returns from Baltimore and is fed into the system. SKICAT scans the tapes and uses its built-in artificial intelligence to decide how to classify all the millions of objects. First it must learn how to classify by practicing on a "training set" of objects. This is a set of images taken using a charge-coupled device (CCD) instead of a photographic plate. CCDs are much more sensitive than photographic film to faint light, and give

higher-resolution images. The high-resolution objects in the training set can be classified fairly easily, one by one, by a program (which is then checked by an astronomer) into one of four categories: star (s), star with fuzz (sf), galaxy (g), or artifact (long), a sort of catch-all class for objects that don't fit neatly into any of the other three categories. These training data, along with the classes, are fed to SKICAT, which in its computer manner notes the category and examines the properties of each object, for example its sharpness, color, shape, brightness, etc. It then automatically makes up a "decision tree."

At each fork in a branch of the decision tree is a question about some distinguishing feature of an object, such as its diameter, or its core magnitude. For example, SKICAT might "say" to a group of objects, if your diameter is bigger than two arc seconds, branch left, otherwise, branch right. And then one subgroup might reach a fork in the next higher branch, where the system would say, if your core is brighter than 19th magnitude, branch left again, if not, branch right this time. The other subgroup might get the same treatment, or might be sorted according to how elliptical or circular they are. At the end of each branch of the decision tree are clusters of objects correctly grouped into the same category: galaxies, or stars with fuzz, etc.

One of the most powerful features of SKICAT is its ability to automatically create a decision tree, which it does in a matter of seconds, that correctly classifies all of the objects in a data set. The core of the new system includes two machine learning algorithms, called GID3* and O-Btree,



SKICAT was able to pick out 68 small volcanoes on this slice of the Venusian surface.

which automatically create decision trees based on the training data. SKICAT tests the tree on another set of objects, which have also been classified by hand, but which the computer system hasn't seen before. It categorizes objects correctly about 94 percent of the time. By contrast, the best performance of a commercially available learning algorithm is about 75 percent. The main goal is to automate the process of transferring data from photographic plates to a catalog. But an added benefit is that by training the learning algorithms to predict classes for faint objects, the algorithms can learn to classify objects that are too faint for humans to identify.

SKICAT is not limited to analyzing data about sky objects. It has also been used by scientists at JPL to pick out small volcanoes on the surface of Venus in images sent back by the Magellan spacecraft. The problem, as with the Palomar Sky Survey, was an overwhelming amount of data. Magellan's map of Venus is contained in some 30,000 images stored on more than 100 CD-ROM disks, and planetary scientists estimate that as many as a million small volcanoes less than 15 kilometers in diameter may be scattered over the planet's surface. The Magellan team took an approach much like the one used in the sky-object classification problem, utilizing SKICAT's artificial intelligence, but teaching it with a different training set. Unlike astronomers, who can measure magnitude, area, and other properties of a star or galaxy, planetary geologists do not have a good set of features to measure for the volcanoes. So, much of their work deals with automatically extracting features from pixels in order to rapidly identify all the tiny volcanoes.

The Palomar Observatory Sky Survey and the Magellan satellite both present a problem confronted by many scientists as research becomes computerized: a computer's ability to store data has far outpaced our ability to analyze it. SKICAT may prove useful not only for separating stars from galaxies and picking out Venusian volcanoes, but may be applicable to a wide variety of chores. "We view SKICAT as a step toward the development of the next generation of tools for the astronomy of the turn of the century and beyond," Djorgovski said.

Jay Aller has been the science writer in Caltech's Office of Media Relations since 1992. He holds a BS in astronomy from Whitman College and completed the graduate science-writing program at UC Santa Cruz.