Publications of the Astronomical Society of the Pacific
107: 1243–1254, 1995 December

The SKICAT System for Processing and Analyzing Digital Imaging Sky Surveys

NICHOLAS WEIR,[1] USAMA M. FAYYAD,[2] S. G. DJORGOVSKI,[1] AND JOSEPH RODEN[2]

California Institute of Technology, Pasadena, California 91125
Electronic mail: weir@fritz.caltech.edu, fayyad@aig.jpl.nasa.gov, george@deimos.caltech.edu, roden@aig.jpl.nasa.gov

Received 1995 April 10; accepted 1995 August 28

ABSTRACT. We describe the design and implementation of a software system for producing, managing, and analyzing catalogs from the digital scans of the Second Palomar Observatory Sky Survey. The system (SKICAT) integrates new and existing packages for performing the full sequence of tasks from raw pixel processing, to object classification, to the matching of multiple, overlapping Schmidt plates and CCD calibration frames. We describe the relevant details of constructing SKICAT plate, CCD, matched, and object catalogs. Plate and CCD catalogs are generated from images, while the latter are derived from existing catalogs. A pair of programs complete the majority of plate and CCD processing in an automated, pipeline fashion, with the user required to execute a minimal number of pre- and post-processing procedures. We apply a modified version of FOCAS for the detection and photometry, and new software for matching catalogs on an object-by-object basis. SKICAT employs modern machine-learning techniques, such as decision trees, to perform automatic star–galaxy–artifact classification with a >90% accuracy down to ~1 mag above the plate detection limit. The system also provides a variety of tools for interactively querying and analyzing the resulting object catalogs.

1. INTRODUCTION

The critical needs of observational astronomers have shifted from the exclusive realm of instrumentation to include that of advanced data analysis. The rate and quality of the data regularly produced by modern instruments frequently overwhelm the tools available to exploit them.

One such data set is the Second Palomar Observatory Sky Survey (POSS-II, Reid et al. 1991). When complete, this photographic northern-sky survey will cover 894 fields spaced 5° apart in three passbands: blue (IIIa-J+GG 395), red (IIIa-F+RG610), and near-infrared (IV-N+RG9). STScI and Caltech have begun a collaborative effort to digitize the complete set of plates (Djorgovski et al. 1992; Lasker et al. 1992; Reid and Djorgovski 1993). Both the photographic survey and the plate scanning are hoped to be >90% complete by the end of 1997. The resulting data set, the Palomar–STScI Digital Sky Survey, will consist of ~3 TB of pixel data: ~1 GB/plate, with 1 arcsec pixels, 2 bytes/pixel, 20340^2 pixels/plate, for all survey fields in all three colors. In conjunction with the plate survey, we are also conducting an intensive program of CCD calibrations using the Palomar 60-in. telescope, using the Gunn–Thuan *gri* bands.

Given the enormous resources devoted to conducting such surveys, it is natural to pay special attention to how, using present-day technology, one can make most effective use of the data once they are available. The traditional means of extracting useful information from imaging surveys is through the construction of object catalogs. Thanks to developments in the field of pattern recognition and machine learning, it is now possible to reliably construct such catalogs objectively and automatically with a higher degree of accuracy than ever before.

Processing of similar data sets have been performed by other groups. In particular, the southern-sky surveys have been digitized by the COSMOS (Heydon-Dumbleton et al. 1989) and APM (Maddox et al. 1990) groups. The APS group in Minnesota has been processing glass copies of the POSS-I survey (Pennington et al. 1993; Odewahn et al. 1992; Odewahn et al. 1993). Here we describe the system used at Caltech for the processing of Digitized POSS-II.

2. OVERALL DESIGN

The Sky Image Cataloging and Analysis System (SKICAT) was designed to facilitate the creation and use of catalogs from large, overlapping imaging surveys, and in particular, the scans of the Palomar–STScI Digital Sky Survey (DPOSS). The purposes of the software utilities comprising SKICAT generally fall into three main categories: catalog construction, catalog management, and catalog analysis. The relationship of these processes is illustrated in Fig. 1. For reducing scans of POSS-II, the first step in SKICAT processing is catalog construction, which results in individual image catalogs. These, in turn, are registered within the SKICAT database management system and matched, object by object, with other catalogs to create a matched catalog of objects appearing in the survey. A matched catalog, or any individual image catalogs, may subsequently be queried in a variety of sophisticated ways to facilitate maintenance or analysis of the data.

While our interest in DPOSS provided the initial motivation for the development of SKICAT, these tools are quite general and applicable to a broad range of data reduction and analysis problems. For example, the catalog construction software could be rather easily adapted to processing large-

[1] Palomar Observatory, Caltech 105-24, Pasadena, CA 91125.
[2] Jet Propulsion Laboratory, Caltech 525-3660, Pasadena, CA 91109.

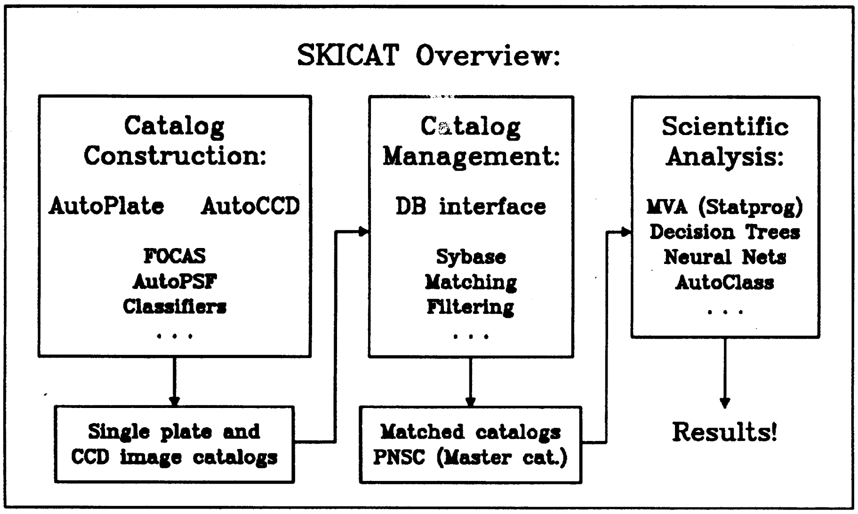

FIG. 1—An overview of the SKICAT system.

scale CCD or infrared imaging surveys. Likewise, the catalog management and analysis tools are useful for integrating and making use of an even wider variety of data sources (e.g., matching radio, infrared, or X-ray sources with their counterparts from optical surveys).

Currently, SKICAT provides utilities for generating catalogs from two types of images, although it was designed to handle virtually any other types of images. One image type consists of a plate scan from the DPOSS survey. The other, a CCD image, is used for photometric calibration and training the star/galaxy classifiers applied to DPOSS catalogs. Step-by-step instructions for processing plates and CCDs from raw pixel into catalog form appear in the *SKICAT Plate and CCD Processing Cookbook* (Weir et al. 1994a) and the *SKICAT User's Manual* (Weir et al. 1994b).

In this first section, we provide an overview of the steps involved in catalog construction, as well as provide an introduction to the catalog management and analysis tasks supported by SKICAT. In the section which follows, we provide a more detailed discussion of the scientifically relevant details of the plate catalog construction processes. In the final section, we describe how matched and object catalogs are constructed within SKICAT.

2.1 Catalog Construction

2.1.1 Processing plates

The heart of SKICAT is a collection of programs for the quasiautomatic processing of DPOSS plates from raw pixel to classified catalog form. Starting with an 8-mm tape containing a 1-GB digitized plate scan from STScI, SKICAT provides the tools for transferring the pixel data to SKICAT format, measuring the plate sky level and image boundaries, and determining a photographic density-to-intensity relation. The user then initiates a script, AutoPlate, which automates the process of cataloging the plate as a set of overlapping 2048^2 pixel image "footprints." Processing the plate scans in this way serves a dual purpose: it makes the computational task more manageable and it allows us to easily compensate for the variation of the PSF across the plate.

The three most critical elements of plate processing are detection, photometry, and classification. We use the Faint Object Classification and Analysis System (FOCAS, Jarvis and Tyson 1979; Valdes 1982a) for image detection and measurement. By measuring quasiasymptotic magnitudes (FOCAS total), using local sky estimates from annuli surrounding each object, and adapting the measurement thresholds within and across each plate to adjust for differences in sky level, noise, and pixel-to-pixel correlation, we are able to obtain very consistent photometry within the plates and across plate boundaries. Details of our methods for performing photometry and the resulting accuracy appear in Weir et al. (1995a) and Weir (1995).

We use modern machine-learning techniques for object classification. In particular, SKICAT utilizes the GID3* and O-Btree decision-tree induction software (Fayyad 1991; Fayyad and Irani 1992, 1993), together with the Ruler system (Fayyad et al. 1993) for combining multiple decision trees into a robust collection of classification rules. These algorithms work by using measurements of a training set of classified objects and inferring an efficient set of rules for accurately classifying each example. The rules are simply conjunctions of multiple "if...then..." clauses, which condition upon any of eight different object parameters to determine an object's classification. The real advancement in using this type of classifier relative to those used in most large-scale surveys to date is twofold: first, we are able to condition upon a larger and more diverse set of attributes; second, we allow the computer to decide what are the optimal number and form of the rules. Additionally, this technique readily generalizes to other, more difficult forms of classification, such as distinguishing galaxies by their morphology.

We have created separate sets of classification rules for objects from J and F band survey plates, and we are in the process of creating classifiers for N plates, as well. We used CCD calibration data, which generally have superior image quality and higher signal-to-noise ratios at any given magnitude level, to construct the training sets used to train the plate object classifiers. Classifications derived from the CCD data, more reliable than "by eye" estimates from the plates themselves, were matched to plate measurements to form the training sets. The measurements used to perform classification are a set of robust, renormalized object parameters that we found to be distributed in a stable fashion within and across plates. By training the algorithms to classify based on these attributes, we were able to nearly completely remove the effect of PSF variation across a given plate, or even between different plates. Average accuracy of star–galaxy classifications as a function of magnitude may be determined from tests using independent CCD-classified plate data. In both the *J* and *F* bands, we found the accuracy to drop below ~90% at about the same equivalent magnitude level, *B* ~ 21.0 mag. This is ~1 mag above the plate detection limits, and nearly 1 mag better than what was achieved in the past with similar data. This increase in depth effectively doubles the number of galaxies available for scientific analysis, relative to the previous automated Schmidt surveys.

An alternative approach involves use of neural nets (Odewahn et al. 1992; Odewahn et al. 1993). We have also experimented with neural nets, but decided to implement decision trees for our catalog production. The details of our classification methods and results are presented in Weir et al. (1995b).

Plate X, Y to RA,Dec assignment, like object classification, is automatically performed in the final stages of catalog construction. Currently, the astrometric transformation is performed based on the astrometric solutions provided by ST ScI, but improved solutions can be easily implemented. As both astrometric assignment and final object classification rely only upon existing catalog measurements, not raw pixel data, they may be easily repeated at later times using a different set of classification rules or improved astrometric solution coefficients. SKICAT provides database manipulation tools that facilitate the continuous refinement of catalogs as better calibration, or even entirely new algorithms, become available.

2.1.2 Processing CCDs

CCD catalogs are constructed using most of the same tools as are applied to plate data. A script called AutoCCD, analogous to AutoPlate, is used to quasiautomatically process an image from pixel into catalog form. The primary differences between plates and CCDs are in the forms of pre- and post-processing that are applied. The usual CCD calibration procedures (e.g., de-biasing, flat-fielding, photometric calibration, etc.), must be performed before running AutoCCD. In addition, we found FOCAS's built-in classifier to provide very accurate results on the CCDs down to the plate detection limit, which is our magnitude limit of interest. We were, therefore, able to let FOCAS automatically classify each object, with just a quick follow-up check by eye, producing excellent quality data without the need for much human interaction or more sophisticated classification algorithms.

CCD data are used for two purposes in our work with DPOSS. First, they provide "true" object classifications, at very faint levels, for our classifier training sets. Because the CCD images are of higher resolution and signal-to-noise ratio (SNR) than digitized plates, we are able to assign accurate classifications to objects whose morphology is not reliably distinguishable, even by an expert, when looking at the plate image alone. Through the machine learning process, the aim is to train the computer to consistently classify these faint objects, thereby enabling it not just to mimic a human astronomer's performance, but actually *improve* upon it.

The second, most important, purpose for the CCD measurements is to provide photometric calibration for the plate catalogs. We use CCD exposures in the Gunn–Thuan (Thuan and Gunn 1976; Kent 1985) *g, r,* and *i* bands to calibrate the IIIa-*J*, IIIa-*F*, and IV-*N* plate data, respectively. These CCD bandpasses provide a reasonable match to the photographic emulsion plus filter passbands. Details of how we perform our CCD photometry and the level of accuracy we achieve appear in the paper Weir et al. (1995a) and Weir (1995).

2.2 Catalog Management

Once the image catalogs are constructed, they are registered within the SKICAT database. Modifications and updated versions of the catalogs are maintained through database management software and tracked by the SKICAT system. The structure of the SKICAT database was specifically designed to facilitate the creation and classification of image catalogs, comparison of object photometry and classifications, revision of object measurements, and the construction of larger, matched catalogs.

For each plate or CCD image, the catalog construction scripts generate a header and features table, together comprising what we term a SKICAT catalog. The header table consists of columns of parameters used to guide the catalog construction process, the name of the image from which the catalog was derived, the location of the image on offline storage, comments, and other information necessary to identify the data source and reconstruct the catalog from scratch if necessary. The features table contains one row for each detected feature in the image. The columns represent the measured attributes of each feature. Approximately 40 parameters per object are measured and saved in the individual plate and CCD catalogs.

SKICAT system tables maintain a complete description and history of every catalog loaded to date. Catalog revisions, that might result from deriving new and improved plate astrometric solutions or photometric corrections, are also logged. Multiple versions of each image catalog may exist, each reflecting a different processing history. The SKICAT system tables also keep track of which catalogs are currently loaded on-line, or physically loaded on disk.

Multiple, overlapping catalogs can be matched into a data structure called the matched catalog, which consists of a matched-features table and a table of those catalogs comprising it. The matched-features table contains independent entries for every measurement of every object detected in the constituent catalogs. Because of size and speed considerations, not every attribute may feasibly be saved within the matched catalog, but a sufficiently small subset of parameters is generally more than adequate for most uses of the data. Of course the saved catalogs themselves provide a complete archive of the full list of parameters if they are ever needed.

The matched catalog may be queried using a filtering and output tool to generate a so-called object table, which contains just a single entry per matched object. With this tool, the user may, for example, generate a distributable data product, such as a galaxy list, from the current set of matched-plate catalogs. The tool may also be used to perform consistency checks within catalog-overlap regions, or to perform specialized scientific analysis over large survey regions. For example, a user may request a listing of all stars within a well-defined section of sky covered by multiple J and F plates, specifying exactly which object attributes to report (e.g., magnitude, RA, Dec, etc.) and from which source (specific J plates, average of all F plates, etc.).

Catalogs may be easily altered using a procedure that allows arbitrary operations on table columns. The user simply specifies the C code which describes the computation for the column value as a function of any other column values, external data files, or constants. The utility automatically generates the necessary code for transforming the table and executes it. This utility is used in a number of contexts in the SKICAT system, including the computation of right ascension and declination, as well as for applying the classification rules. In the same way, catalogs may be recalibrated or otherwise adjusted in light of new or improved data.

A catalog may also be modified by using a utility that updates selected columns from corresponding columns in the matched catalog. This procedure would be appropriate if, for example, the entries in a matched catalog were calibrated, and the calibrated measurements needed to be passed back to the original catalogs for archival purposes. An updated catalog could subsequently be reregistered as a new version of the existing catalog. Both the original and new header information would now be saved in the system, maintaining a complete history of catalog revisions. Via this mechanism, SKICAT is designed to maintain a "living," growing database, instead of a data archive fixed for all time.

2.3 Catalog Analysis

The third layer of SKICAT, which is still under development, will consist of a powerful tool box of modern data-analysis algorithms to be applied for survey-data space exploration and the scientific analysis of the catalogs. It will facilitate more sophisticated scientific investigations of these expanding survey data sets, including a multivariate statistical-analysis package, and a wide variety of Bayesian inference tools, objective classifiers, and other advanced-data management and analysis packages and algorithms.

We have also started to explore the potential of *machine-assisted discovery,* where modern, artificial intelligence-based software tools automatically explore large parameter spaces of data and draw a scientist's attention to unusual or rare types of objects, or nonobvious clusters of objects in parameter space. We have begun applying the Autoclass (Cheeseman et al. 1988) unsupervised classification software to DPOSS, with plans to implement this and other Bayesian inference and cluster analysis tools within SKICAT in the future. Some early results are described in de Carvalho et al. (1994).

2.4 Application Environment

The SKICAT system is largely written in C, Unix shell scripts, and FORTRAN, and it is portable across Unix systems. SKICAT is built around and incorporates a number of preexisting software packages: FOCAS routines for image detection and measurement; the GID*/O-Btree/Ruler induction software for object classification; and the Sybase commercial relational database management system (DBMS) for maintaining and accessing the data. While SKICAT was developed using these packages, none are irreplaceable. Each package serves its purpose and, because of the modularity of the system, could be substituted for another which performs the same function. In addition, SKICAT provides quick and easy access to most system utilities through a common X-Windows graphical user interface, while users familiar with Unix can access the same utilities directly from the Unix command line. All of the software except for the commercial Sybase package is freely available.

SKICAT was designed so that all database system operations specific to Sybase would be transparent to the user. The user interfaces and underlying Unix utilities have been designed to allow the user to select and specify subsets of catalogs using a slightly expanded version of the industry standard SQL (Standard Query Language). This extended query language provides additional features of specific interest to users in astronomy. Most database operations controlled through the SKICAT software are implemented using SQL, so that it would be relatively easy to replace the underlying DBMS if the need arose.

3. CONSTRUCTING PLATE CATALOGS

In this section, we provide more detail on the steps involved in constructing a catalog from a DPOSS scan. Aside from the initial preprocessing steps, the process of cataloging a CCD image is very analogous to that for a plate. More details can be found in Weir (1995).

3.1 Preprocessing

Only a few manual steps are required before a plate scan may be pipeline processed using a Unix command-line-based program called AutoPlate, or the X-windows-based graphical user interface to it. A plate scan is provided in the form of pixel data consisting of photographic densities and is 23,040 ×23,040 pixels in size. After defining the plate boundaries, and the sky and saturation densities, the first step is to perform the photographic density to intensity conversion. A SKICAT program automatically retrieves the portion in the

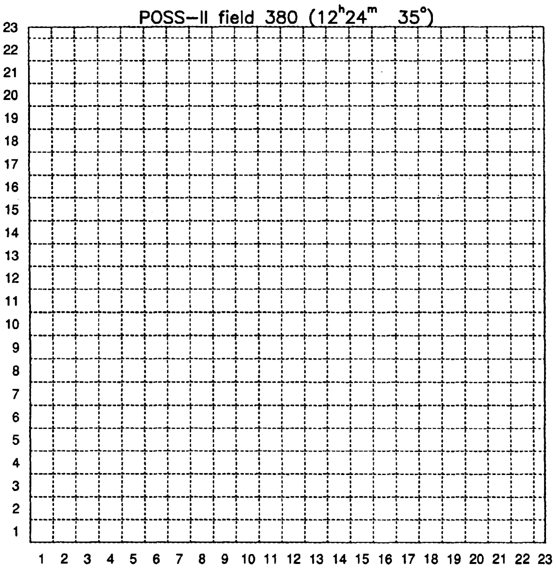

FIG. 2—A plate scan is saved as 23 Vax VMS savesets (rows) of 23 image 'blocks' each. Each image block consists of 1024×1024 pixels, except at the right and top edges, where one dimension is only 512.

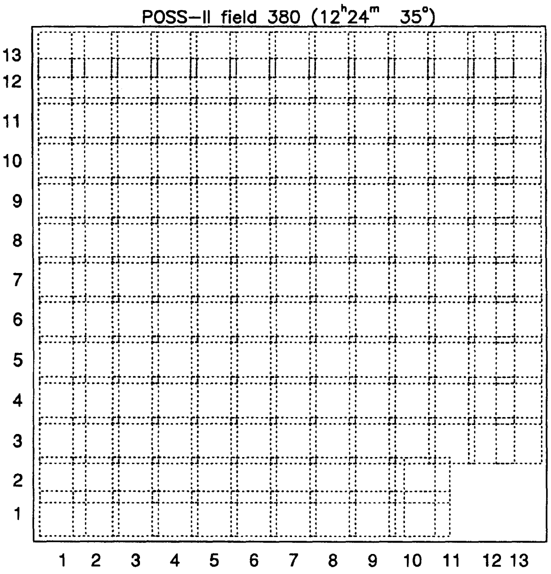

FIG. 3—A plate scan is analyzed as a set of 13×13 overlapping footprint images of 2048^2 pixels each. Not only is this approach computationally convenient, but it provides greater sensitivity to position-dependent plate effects. It also facilitates quality control via the systematic comparison of the overlap regions.

southwest corner of each image that contains the 16 sensitometry spots that appear in each POSS-II plate. This program assists the user in running an IRAF script to measure the 16 spots and compile a list of the densities. It then prompts the user to interactively fit an 'HD' curve to the data points, providing a density to intensity transformation.

The mathematical formula we use to fit the measured plate densities (D) to relative intensities (I) is

$$\log I = \frac{P(D)}{(D_S - D)\times(D_T - D)}, \tag{1}$$

where $P(D)$ is a polynomial function of the density, and the saturation and toe densities, D_S and D_T, are those corresponding to fully exposed and unexposed portions of the plate, respectively. The polynomial coefficients, together with the toe and saturation values, establish the conversion applied to each pixel value whenever image blocks are subsequently loaded and mosaiced to form larger images. As the average sky density is generally far above the toe level, it is usually desirable to avoid fitting the polynomial to the lowest few intensities, thereby improving the fit in the other portions of the curve. Similarly, the most nearly saturated point or two is also generally ignored. After several iterations adjusting the relevant parameters, we have found it possible to reduce the residual between the fit and all accepted data points to less than 5% in intensity.

There is a long history to efficiently modeling the HD curve. The method employed by STScI (Russel et al. 1990), for example, involves a more complicated formula and averaging many plates together. By their own admission, however, they find the more complicated expression to be overkill for the linear part of the curve of most interest. In addition, we found considerable variation of the curve among different plates, requiring independent fits. As described in Weir et al. (1995a), we also find the instrumental magnitudes resulting from these fits to be extremely consistent from plate to plate, in the sense of only requiring a single zero-point offset to match them. This provides, in our opinion, the most important test of the validity of our linearization scheme.

3.2 AutoPlate Processing

AutoPlate is a C-Shell script which executes a suite of other scripts, C code, and FORTRAN programs to conduct the pipeline processing of plate scans from their raw pixel form to SKICAT catalogs. The steps involved include everything from loading the pixel data from exabyte tape, to image detection and measurement, to catalog construction and quality control. The majority of image processing functions are accomplished using FOCAS routines, while Sybase is used for database management.

3.2.1 Overlapping footprints

Each plate is analyzed as a set of 13×13 overlapping 'footprint' images. After preprocessing, a plate scan exists on exabyte tape as 23 Vax VMS savesets of 23 images each (see Fig. 2). These image blocks are pasted together to form image footprints, which form an overlapping grid covering the entire plate (see Fig. 3). Each footprint is 2048^2 in size, with a minimum overlap between adjacent footprints of 272 pixels, ~4.5 arcmin. The large overlap allows all but the largest objects to be reliably measured in this piecemeal fashion, while providing a quality control check and statistics on footprint-dependent measurement errors. In fact, analysis of these errors indicate that the systematic errors induced by processing the scan in this fashion are at least an order of magnitude below random image-measurement errors.

Footprints are created and processed a row at a time, from bottom (south) to top. As each footprint row is processed, AutoPlate loads the necessary image blocks from tape and deletes unnecessary blocks from disk. Consecutive footprint images, from left (east) to right, are created just prior to their processing. Up to two rows of footprints are always on disk, facilitating the detection of vertical mismatches between footprint tables. Each row of footprint-features tables is saved to the plate-features table only after passing a number of quality-control checks meant to assure uniformity of catalog construction.

3.2.2 Image analysis

Footprint images are analyzed in a few ways prior to object detection, and certain quality-control checks are performed. Next, AutoPlate creates a rebinned version of the image with one pixel per 8×8 in the original. This scale matches that of the 'sky' image produced by the FOCAS detection algorithm. To provide the FOCAS algorithm with a good first guess of the footprint sky, the value is initially estimated by binning the image into blocks of 64^2 pixels each, accumulating the median and quartile sigma[3] for each block, then accumulating the median and quartile sigma for all of the block measurements. Images of the sky median and sky sigma are saved at this reduced (one pixel per 64×64) scale. This robust estimation procedure provides relatively accurate initial sky and sky sigma values, even when relatively large and bright sources exist in the image. Seeded with these values, the FOCAS detection and background estimation procedures have been found to work well.

AutoPlate also estimates the pixel-to-pixel correlation (horizontal and vertical combined) within each footprint. For this measurement, in addition to applying the same binning and median filtering procedure as above, AutoPlate excludes all pixels two and a half sigma above the sky level. This technique was found to provide an extremely robust and accurate measurement for all levels of pixel blurring, even when large saturated objects appear in the image.

3.2.3 Object detection

The basic processes of object detection and measurement are accomplished using only slightly modified versions of the standard FOCAS routines (Jarvis and Tyson 1979; Valdes 1982a). Algorithmic details of these programs may be found in the FOCAS documentation (Valdes 1982b).

Just prior to object detection, a FOCAS catalog is automatically initialized for the current footprint. The appropriate header values are determined in AutoPlate based upon the current footprint row and column numbers, and from information derived from the plate image header. The FOCAS 'detect' command then uses the header parameters for driving its object detection and sky estimation procedure. The result of this command is a catalog of features, or contiguous pixels a certain threshold above the background, and meeting a minimum area and signal-to-noise ratio (SNR) requirement. The FOCAS detect command also produces an estimate of the sky with a one pixel per 8×8 resolution. If this estimate significantly differs from the median sky image computed previously, an error is reported and processing ceases.

For optimal sensitivity, the FOCAS detection algorithm applies a threshold equal to some number of estimated standard deviations (sky sigma) above the locally estimated sky. The assumed sky sigma is the robust value computed for the footprint, as described in the Image Analysis section above. However, because of spatially varying pixel-to-pixel correlation within each plate scan, using the same multiple of sky sigma as the threshold for all footprints would not result in the same detection sensitivity.

To compensate for this effect and approach a common level of sensitivity between and within plates, we sought to derive a factor by which to scale the measured sky sigma so as to make it correspond to approximately one standard deviation in an *unblurred* version of each footprint. To establish this scaling factor as a function of measured blur, we created a simulated footprint image matching the average noise[4] and object number statistics of real footprints, then we convolved it with a series of Gaussians of different width. Given the convolution kernel, we estimate the scale factor to be the square root of the inverse of the sum of squares of the normalized kernel elements. By measuring the pixel-to-pixel R^2 for each image, we are able to empirically derive a mapping from measured (square) correlation to scale factor. We found a sixth-order polynomial to provide a good fit to the relation (see Fig. 4). We also established the relation using a blank simulated sky image and derived virtually identical results, lending confidence in the robustness and accuracy of our correlation estimation procedure.

We then used 2.5 times this scale factor times the estimated sky sigma as our detection threshold. The additional detection parameters required by FOCAS include a minimum object area, "significance limit" for object detection, and predetection blurring kernel. We require every object to comprise six contiguous pixels. We set the significance limit to -100, which is equivalent to turning off this SNR requirement (see the FOCAS manual for details). We used the built-in FOCAS blurring function. The FOCAS detection algorithm works by convolving the image with this kernel, then searching for contiguous pixels with values greater than the locally estimated sky by the specified detection threshold.

Our choice of detection parameters, in particular our scaling correction for pixel-to-pixel correlations, results in relatively consistent sensitivity as a function of plate quality, as

[3] We define a quartile sigma as 0.7415 times the difference between the 75th and 25th percentile values, a robust estimate of the sample standard deviation that is insensitive to outliers. For a Gaussian distribution, this is virtually identical (in the limit of large sample sizes) to the standard deviation defined in the normal way.

[4] The appropriate level of uncorrelated, Gaussian random noise was determined in an iterative fashion. First, we found a Gaussian kernel which, when convolved with the image, produced a degree of blur, as measured by the pixel-to-pixel correlation, closely approximating that of an average footprint. We then found that noise amplitude which, after convolution, resulted in a measured sky sigma closely matching that of an average footprint.

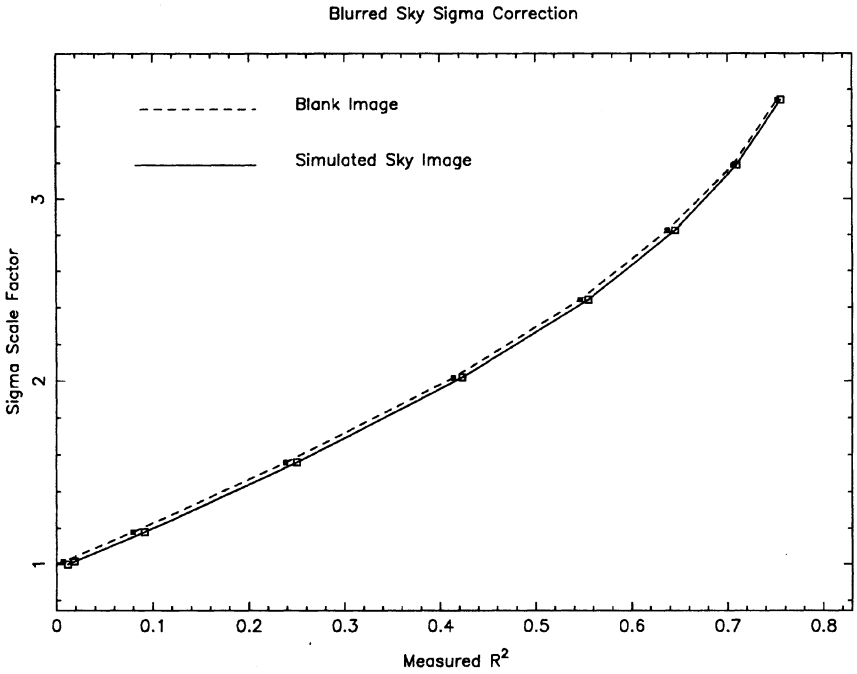

FIG. 4—Given the measured image blur (R^2), we establish the appropriate factor by which to scale the measured sky sigma to approximate that of an unblurred version of the same image.

evidenced by the relative uniformity of object density we detect from footprint to footprint and plate to plate. Our choice of threshold, minimum area, significance limit, and predetection blurring were chosen after extensive tests on both real and simulated images, establishing some feel for the trade-off between completion (percentage of real objects detected) and contamination (percent of detected objects which are not real). On simulated images, this combination of parameters resulted in an average FOCAS detection isophote corresponding to roughly 2.0 *uncorrelated* sky sigma, which is sufficiently far into the noise as to pick up every object readily detectable by eye. It also resulted in what we considered a manageable number of detections per footprint and plate, in excess of the density saved in previous Schmidt plate surveys. Typical galaxy detection limits for the *J* and *F* DPOSS plates are found to be 21.0 to 21.5 mag in *g* and 20.1 to 20.6 mag in *r*, respectively. For point sources, the limit can extend up to half a magnitude fainter.

3.2.4 Object measurement

The local sky brightness for each feature is measured using the FOCAS 'sky' command. It measures the median pixel value in an annular region surrounding each feature, avoiding pixels that are within the detection isophote of another feature. The accuracy and systematic effects of this sky measuring algorithm are addressed in Weir (1995).

After obtaining the sky estimate, additional attributes for each feature are measured using the FOCAS 'evaluate' routine. The total number of measurements number more than 30. The measured magnitudes are instrumental and are computed according to

$$m=30.0-2.5\log L,$$

where L is the luminosity, or sky-subtracted integrated intensity. The offset of 30.0 is arbitrary and was chosen to make the instrumental magnitudes approximate the final calibrated values within a magnitude or two. The aperture magnitudes are computed using a five arcsec radius. The 'total' magnitude and area are computed by 'growing' the detection isophote out a pixel at a time in all directions until the total area is at least twice the original. This magnitude is meant to provide a flux measurement less biased with respect to surface-brightness profile, approximating something like an asymptotic or true total magnitude. The cost for decreased systematic error is greater sensitivity to sky subtraction, integration over more noisy pixels, and hence, increased random error (relative to isophotal or aperture magnitudes).

FOCAS also sets a number of flags for each feature, each of which is saved as an attribute. These flags indicate such things as whether the object touches the edge of the footprint, the object is below the sky level in integrated intensity, the object's size exceeds current FOCAS limits, there are saturated pixels in the object, or the object was not split at any level by the FOCAS deblending routine. Additional useful attributes are obtained by taking nonlinear combinations of some of the attributes. For example, using the intensity-weighted second moments, we can calculate the ellipticity and position angle of each feature. Additional attributes, the so-called 'revised' ones described below, are defined by the position of a feature within the statistical distribution of that footprint's features within some measured parameter space (e.g., within the plane defined by the first radial moment and the total magnitude).

3.2.5 Object deblending

After each feature in a footprint has been evaluated, SKICAT next applies the FOCAS 'splits' command. Effectively, this routine runs the detection algorithm on every existing feature, but using successively higher thresholds. 'Islands' detected at a given threshold are entered into the catalog as distinct features, and all attributes are remeasured for them. The 'parent's' flux is divided between the 'children' according to the ratio of isophotal fluxes obtained using the higher threshold. This process continues recursively until no more islands are detected.

All parents and intermediate children (i.e., a feature's full family tree) are saved within the FOCAS catalog and likewise within SKICAT. Each feature is referenced by an entry and subentry number. A parent and all of its children share the same entry number. Children are distinguished by the hierarchically constructed subentry number: subsequent generations append additional digits to the end. The leaf or leaves in a feature's family tree correspond to indivisible objects and are marked as such by a flag attribute.

We note that improvements can certainly be made to the deblending process. For example, other methods could be used to improve the quality of the photometry of the deblended objects, better take deblending into account when matching overlapping images, handle the extreme crowding conditions to be found in lower Galactic latitude POSS-II plates, etc. Nonetheless, we find the present implementation to be more than sufficient even for detailed analyses of higher latitude plates, and that it at least represents a step above reduction without the use of deblending at all, as in the case of some previous surveys (e.g., APM, Maddox et al. 1990).

3.2.6 Classification related measurements

An additional set of attributes are measured solely for the purpose of facilitating feature classification. Four revised attributes are determined by automatically estimating and subtracting the 'stellar locus' from the parameters $M_{\rm core}$, the magnitude of the brightest 3×3 pixel region, of total intensity $L_{\rm core}$; the log of the isophotal area, $\log A$; the intensity-weighted first-moment radius, r_1; and S, where

$$S=\frac{A}{\log[L_{\rm core}/(9\times I)]}$$

and I is the average intensity of the detection isophote. The stellar locus is the attribute value as a function of magnitude around which point sources are fairly narrowly distributed, at least at brighter magnitudes. As described in Weir et al. (1995b), we have found that the resulting revised attributes are relatively insensitive to footprint-to-footprint, and even plate-to-plate, variations.

In order to derive even more powerful classification attributes, we form an empirical estimate of the PSF for each footprint. Along with magnitude and ellipticity, the four revised attributes are fed as input to a decision-tree classifier, which culls out a list of 'sure-thing' stars. FOCAS routine then adds images of these stars to form a two-dimensional PSF template.

Using the PSF template, the FOCAS 'resolution' routine determines the best-fitting 'scale' (α) and 'fraction' (β) values, which parametrize the fit of a blurred (or sharpened) version of the PSF to each feature (Valdes 1982a). The template used to model each feature is of the form:

$$t(r_i)=\beta s(r_i/\alpha)+(1-\beta)s(r_i),$$

where r_i is the position of pixel i, α is the broadening (sharpening) parameter, and β is the fraction of broadened PSF. This template-based approach is the core of FOCAS's Bayesian classification method. Objects are classified as stars, galaxies, artifacts, etc., according to their maximum likelihood (best-fitting) location within two-dimensional scale and fraction space. Extensive tests performed by Valdes (1982a) indicate that one can achieve significantly higher accuracy in star/galaxy separation with this template-fitting approach versus simpler approaches employed previously. Weir and Picard (1991) explicitly tested the use of these two techniques on digitized Schmidt plate data and confirmed this result.

In this present version of SKICAT, we combine these resolution parameters along with total magnitude, ellipticity, and the four revised attributes described above to form an even higher dimensional space in which to perform feature classification. Actual classification is run as a post-processing procedure, using the measured attributes within the plate catalog. One can thereby alter the existing, or create an entirely new, classifier and apply it to a catalog at any future date. The classifier currently applied to plate features within SKICAT was generated using the GID3*/O-Btree and Ruler decision-tree induction programs. A full description of how it was created and the results we have achieved on actual plate data appears in Weir et al. (1995b). The net effect is that by employing this new technology, we are able to go about a magnitude deeper in achieving accurate object classifications, resulting in approximately three times larger classified-object catalogs than in previous surveys using comparable data.

3.2.7 Quality control tests

Each individual footprint FOCAS catalog, and its corresponding revised attribute list, is joined into a Sybase table for subsequent processing. As a quality-control check, the current footprint-features table is matched with the tables of the footprints to its left and bottom, if they exist. If any major discrepancies are detected in the mean or standard deviation of measurements in the overlap, processing is halted and an error reported. Otherwise, AutoPlate appends these results to a summary file characterizing the footprint row.

After a row is complete, Autoplate searches the footprint summary file for outliers and trends, halting the program if it encounters any problems. If none are found, the previous row of footprints is added to the Sybase plate catalog and any auxiliary files are saved. First, the row's footprint summary file is appended to the corresponding file for the plate. Next, each footprint's compressed original, sky, median sky, and sky sigma images are pasted into corresponding composite images for the entire plate. Footprint specific parameters are appended to a footprints file. All features with central coordinates in a nonredundant portion of the plate image are added to the plate features table, while features whose outer isophotes extend beyond any single footprint's boundaries are saved to a border objects list. Generally these are features which appear at the edge of the plate. In addition, AutoPlate appends to a list of footprint overlap statistics, and summary thereof. Data for the previous row are deleted after each of these operations is complete.

After all rows have been processed, the system checks the footprint summary file for outliers and trends among footprint statistics in the vertical direction. Provided none are found, catalog generation is complete, a plate catalog header is created (if it was not already) and all remaining footprints and image blocks are deleted.

3.2.8 Data products

The final products of an AutoPlate run are a SKICAT catalog, consisting of a Sybase format features table and header table, and several auxiliary files. The plate catalog resides on the Sybase disk partition while the auxiliary files are saved within a Unix directory hierarchy created specifically for that plate. The auxiliary files include the following images: a re-binned version of the plate scan containing the average of every 8×8 pixels in the original; the 'sky' image produced by the FOCAS detection algorithm at the same scale; images of the median and quartile sigma of the plate scan at a one pixel per 64×64 scale. Besides providing an overall reality check of the AutoPlate process, these images may be useful for scientific programs, such as searches for low surface-brightness galaxies.

In addition, SKICAT saves each of the FOCAS 'areas' files produced for each footprint. These files contain a run-length encoding of all the pixels comprising every feature in each image. This information may prove useful in the future for locating the precise extent of a feature when all of the imagery data, in addition to catalogs, are readily available online for querying and analysis.

The other auxiliary files produced by AutoPlate are those produced and used for quality-control purposes. They include a footprint-statistics file, containing lists of statistics measured for each footprint (e.g., number of features detected, average sky level, etc.) which are used to detect trends and outliers among the footprints along any given plate row or column. The other quality-control file contains lists of all of the overlap statistics measured between adjoining footprints.

3.3 Post-processing

After a plate catalog has been created by AutoPlate, there are still a few operations which must be performed as a part of the plate's standard pipeline processing. These include the assignment of Right Ascension and Declination (RA,Dec) to each object, as well as classification. As neither of these operations require access to the pixel data themselves, one is able to re-run either of these multiple times in the future using new and better coefficients or algorithms.

3.3.1 Astrometric transformation

The J2000 RA and Dec of the central pixel (specified in plate standard coordinates by the XC and YC attributes) of each feature is calculated using coefficients in the plate catalog header. These coefficients are initially provided by STScI and are supposed to be good to ~0.5 arcsec rms accuracy over scales less than about a square degree. When in the future better plate solution coefficients are available, it is simply a matter of entering them in the catalog header, then re-executing a catalog modifying procedure to assign a new RA and Dec to each feature.

3.3.2 Classification

The plate features classifier provided with SKICAT was generated using the GID3*/O-Btree and Ruler programs, and is implemented as a procedure executed by a more general utility for modifying columns within a database table. By applying a set of rules that condition upon a subset of the parameters in a plate features table, the procedure provides a classification to each object. An entry within the plate's header table specifies the classifier rules file to use.

3.3.3 Bright object editing

In the initial version of SKICAT, the user was required to hand create a list of the 'bad regions' within the plate, such as areas corrupted by bright stars. One detected the bad regions by analyzing the 8×8 binned average of the full scan image produced by AutoPlate. By displaying this image, the user could easily pick out and mark the 100 or so brightest objects in the scan which had been poorly processed by AutoPlate. It is particularly important to mark the regions surrounding bright stars, as their halos and spikes are split into sometimes hundreds of small artifacts which may be mistaken for real objects in the catalog (e.g., see Fig. 5). At this time, the bad-regions list is not used to filter or flag entries in the SKICAT plate catalog itself, but rather for subsequent filtering of ASCII data files generated by queries of the plate or matched catalog. The entire process of bright-object detection has recently been automated, however, and will be a documented component in the next version of SKICAT.

3.3.4 Catalog registration

Once all of the aforementioned processes are complete, the plate catalog is ready for registration into the SKICAT catalog management system. This loads the catalog header information into the SKICAT System Tables, allowing it to be matched with other catalogs or saved to/loaded from tape. At this time, the plate catalog, along with the auxiliary files, are generally saved on an archival tape, and plate processing is complete.

4. CONSTRUCTING MATCHED AND OBJECT CATALOGS

4.1 Matched Catalogs

SKICAT provides the ability to match features from multiple plate and CCD catalogs based on the similarity of their measured positions in celestial (RA,Dec) coordinates. This

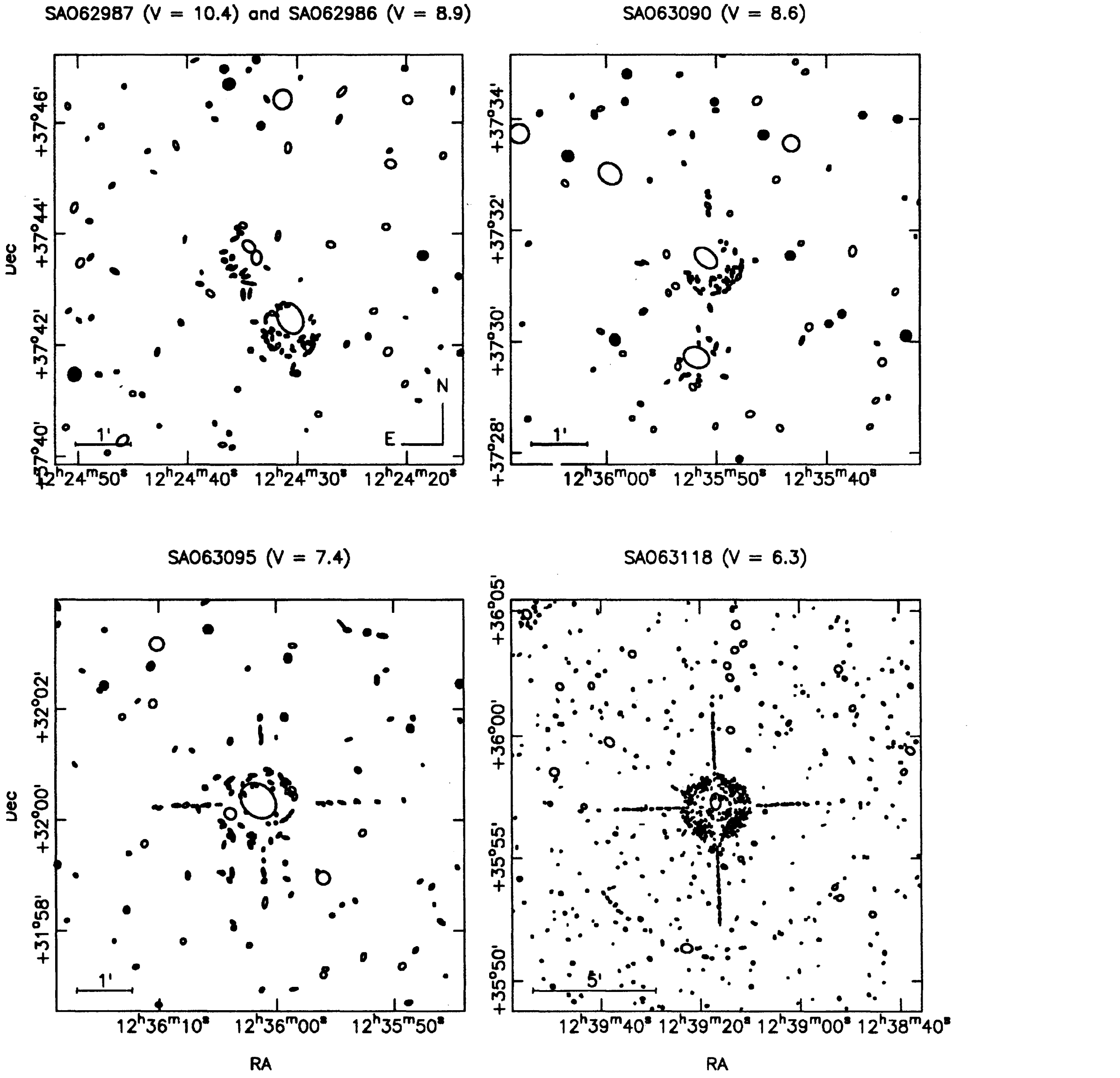

FIG. 5—The regions surrounding bright stars must be avoided when analyzing the plate catalogs generated by SKICAT, as it typically splits these objects into dozens, or even hundreds, of spurious artifacts.

procedure is essential for analyzing objects measured in multiple bandpasses, such as finding optical IDs of nonoptical sources; constructing object list spanning multiple overlapping images; and for performing consistency checks of object measurements and classifications.

The process of adding a catalog to the matched catalog involves matching each feature in the catalog to the nearest object meeting certain criteria within the matched catalog, after solving for a small systematic *X,Y* offset between the two. To perform this matching, the filtered-source catalog is broken down into a user-specified number of solid angle 'segments'. A best fit transformation in *X* and *Y* is solved for using a robust fitting algorithm and applied to each segment when it is matched. To optimize this process, the catalog should be split into as many segments as necessary to allow for systematic deviations in its astrometric accuracy.

For each segment, the matcher attempts to minimize the overall match error (defined as the average matched feature difference) separately in *X* and *Y* by repeating the matching process until the errors meet specified criteria. For each feature in a segment, the matcher attempts to find the closest feature within some search radius within the matched catalog, offsetting by the previous iteration's match error in *X* and *Y*. These errors are accumulated over each iteration to form a mean offset. The initial search analysis is given by the user; subsequently it is determined as some multiple of the measured standard deviation in the previous iteration's offsets. These average offsets and the standard deviations are computed only for a quartile-sigma clipped fraction of the matches from the previous iteration, in order to exclude out-

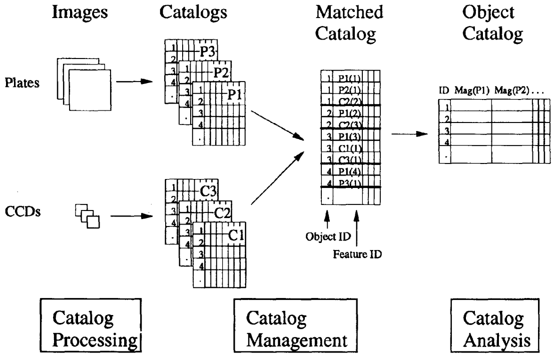

FIG. 6—An overview of the SKICAT object matching process.

liers from the estimate. This matching and estimation procedure repeats until the iteration's match error in both X and Y is less than some multiple of the estimated error in the mean offset. The matcher then performs a final pruning of the matched-object list, passing only those matches with a residual Chi-squared error less than some threshold.

The matcher then assigns each feature an identification number according to the match results. Features with no corresponding object in the matched catalog are assigned the default next ID, which is then incremented. For each feature from the segment, a row including a user-specified subset of attribute columns is appended to the matched catalog's features table. The match and converge process is repeated for each segment of the catalog. After each segment has been matched, information about the input catalog is added to system files detailing the contents of the matched catalog.

4.2 Object Catalogs

While the matched catalog is the most comprehensive form of database produced by SKICAT, it is generally too unwieldy for direct use in large scale survey analysis. By allowing a virtually unlimited number of independent feature entries per object, very little data reduction actually takes place in the matching process. Although in practice, one generally limits the number of attributes saved in the matched catalog, this still leaves unsolved the problem of combining the multiple measurements that are usually present for any given attribute and feature.

To provide the user with power and flexibility in accessing the matched catalog for scientific analysis and calibration, we developed a sophisticated database-querying mechanism. This program summarizes data from the matched catalog to form an object catalog, which by our definition contains just one entry per object. The flow of data from image to final object catalog form is illustrated in Fig. 6. The query program has two primary inputs: a filter and an output specification file. The filter basically defines the conditions that an object, or its constituent features, must meet in order to be passed on for output. A full description of the filter language appears in the *SKICAT Users Manual* and specific useful examples appear in the Query section of the *SKICAT Plate and CCD Processing Cookbook.* These filter conditions might include a requirement on the number of features measured per object, that an object be measured in a particular catalog, that an object not be measured in a particular passband, that an object's magnitude falls within a certain range, etc. The most important filter specification is of an allowable RA and Dec range, as the matched catalog is sorted on those fields.

Using the query program, the user can combine the data in the matched catalog in most ways needed for subsequent scientific use. To facilitate the construction of the filter and output specification files, we created an X-windows interface to the program. The user has the option of producing another Sybase table or an ASCII text file. The former is of use if the user might wish to perform subsequent queries of the resulting table using any of the available Sybase database management tools.

After the successive application of each of these tools, from creating individual plate and CCD catalogs, to matched catalog construction, to the generation of user-specified object catalogs, the user will have reduced the raw pixel data into a form suitable for systematic study.

This work was supported at Caltech in part by NASA AISRP Contract No. NAS5-31348, the Caltech President's fund, and NSF PYI award AST-9157412, and at JPL under a contract with NASA. The POSS-II is partially funded by grants to Caltech from the Eastman Kodak Co., the National Geographic Society, the Samuel Oschin Foundation, NSF Grants No. AST 84-08225 and No. AST 87-19465, and NASA Grants No. NGL 05002140 and No. NAGW 1710. We acknowledge the efforts of the POSS-II team at Palomar, the scanning team at STScI, and the rest of the SKICAT team at JPL.

REFERENCES

Cheeseman, P. et al. 1988, in Proc. Fifth Machine Learning Workshop (San Mateo, Morgan Kaufmann), p. 54

deCarvalho, R., Djorgovski, S., Weir, N., N. Fayyad, U., Cherkauer, K., Roden, J., and Gray, A. 1994, in Astronomical Data Analysis Software and Systems IV, ed. R. Hanisch, R. Brissenden, and J. Barnes, A.S.P. Conf. Ser. (in press)

Djorgovski, S., Lasker, B., Weir, N., Postman, M., Reid, I., and Laidler, V. 1992, BAAS, 24, 750

Ellis, R. 1987, in Observational Cosmology, IAU Symp. 124, ed. A. Hewitt, G. Burbidge, and L. Z. Fang (Dordrecht: Reidel), p. 367

Fayyad, U. 1991. Ph.D. thesis, EECS Dept., The University of Michigan

Fayyad, U. and Irani, K. 1992, in Proceedings of the Tenth National Conference on Artificial Intelligence AAAI-92, San Jose, CA

Fayyad, U. and Irani, K. 1993, in Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93) (Chambery, France, Morgan Kauffman) (in press)

Fayyad, U., Weir, N., and Djorgovski, S. 1993, in Proceedings of AAAI-93 Workshop on Knowledge Discovery in Databases, Washington, D.C., ed. G. Piatetsky-Shapiro (AAAI/MIT Press), p. 1

Heydon-Dumbleton, N. H., Collins, C. A., and MacGillivray, H. T. 1989, MNRAS, 238, 379

Jarvis, J. and Tyson, J. 1979, SPIE Proc. on Instrumentation in Astronomy, 172, 422

Kent, S. M. 1985, PASP, 97, 165

Lasker, B., Djorgovski, S., Postman, M., Laidler, V., Weir, N., Reid, I., and Sturch, C. 1992, BAAS, 24, 741

Maddox, S., Sutherland, W., Efstathiou, G., and Loveday, J. 1990, MNRAS, 243, 692

Odewahn, S., Humphreys, R., Aldering, G., and Thurmes, P. 1993, PASP, 105, 1354

Odewahn, S., Stockwell, E., Pennington, R., Humphreys, R., and Zumach, W. 1992, AJ, 103, 318

Pennington, R., Humphreys, R., Odewahn, S., Zumach, W., and Thurmes, P. 1983, PASP, 105, 521

Reid, I., et al., 1991, PASP, 331, 465

Reid, N., and Djorgovski, S. 1993, in Sky Surveys: Protostars to Protogalaxies, ed. B. T. Soifer, A.S.P. Conf. Ser. #43, 125

Russell, J. L., Lasker, B. M., McLean, B. J., Sturch, C. R., and Jenker, H. 1990, AJ, 99, 2059

Thuan, T. X. and Gunn, J. 1976, PASP, 88, 543

Valdes, F. 1982a, SPIE Proc. on Instrumentation in Astronomy IV, 331, 465

Valdes, F. 1982b, FOCAS User's Manual (Tucson, NOAO)

Weir, N. 1995. Ph.D. thesis, California Institute of Technology

Weir, N., Djorgovski, S., and Fayyad, U. 1995a, AJ, 110, 1

Weir, N., Djorgovski, S., Fayyad, U., and Roden, J. 1994a, SKICAT Plate and CCD Processing Cookbook (Pasadena, JPL/Caltech)

Weir, N., Fayyad, U., and Djorgovski, S. 1995b, AJ, 109, 2401

Weir, N., Fayyad, U., Roden, J., and Djorgovski, S. 1994b, SKICAT User's Manual (Pasadena, JPL/Caltech)

Weir, N., and Picard, A. 1991: in Digitised Optical Sky Surveys, ed., H. T. MacGillivray and E. B. Thomson (Dordrecht, Kluwer), p. 255