The CCD/Transit Instrument (CTI) data-analysis system

M. G. M. Cawson, J. T. McGraw, M. J. Keane

Steward Observatory, University of Arizona
Tucson, Arizona 85721

Abstract

The automated software system for archiving, analysing and interrogating data from the CCD/Transit Instrument (CTI) is described.

The CTI collects up to 450 Mbytes of image-data each clear night in the form of a narrow strip of sky observed in two colors. The large data-volumes and the scientific aims of the project make it imperative that the data are analysed within the 24-hour period following the observations. To this end a fully automatic and self-evaluating software system has been developed.

The data are collected from the telescope in real-time and then transported to Tucson for analysis. Verification is performed by visual inspection of random sub-sets of the data and obvious cosmic rays are detected and removed before permanent archival is made to optical disc.

The analysis phase is performed by a pair of linked algorithms, one operating on the absolute pixel-values and the other on the spatial derivative of the data. In this way both isolated and merged images are reliably detected in a single pass. In order to isolate the latter algorithm from the effects of noise spikes a 3x3 Hanning filter is applied to the raw data before the analysis is run. The algorithms reduce the input pixel-data to a database of measured parameters for each image which has been found.

A contrast filter is applied in order to assign a detection-probability to each image and then x-y calibration and intensity calibration are performed using known reference stars in the strip (e.g. AGK-3 stars). These are added to as necessary by secondary standards boot-strapped from the CTI data itself.

The final stages involve merging the new data into the CTI Master-list and History-list and the automatic comparison of each new detection with a set of pre-defined templates in parameter-space to find interesting objects such as supernovae, quasars and variable stars.

Each stage of the processing from verification to interesting image selection is performed under a data-logging system which both controls the pipe-lining of data through the system and records key performance monitor parameters which are built into the software. Furthermore, the data from each stage are stored in databases to facilitate evaluation, and all stages offer the facility to enter keyword-indexed free-format text into the data-logging system. In this way a large measure of certification is built into the system to provide the necessary confidence in the end results.

1. Introduction

The CCD/Transit Instrument (CTI) data analysis system has to be capable of handling up to 450 Mbytes of image data within the 24 hours following each observation being taken. For such a data handling problem novel techniques have had to be devised, the most fruitful of which has been the use of a database approach to all aspects of the project including data representation and storage, data manipulation, and data retrieval and interrogation (see the accompanying paper by Cawson et al.).

Figure 1 shows a block diagram of the CTI daily data analysis system. The analysis system is in the form of a pipe-line through which all the data pass sequentially and this paper is organized in a similar fashion.

Section 2 gives a description of the control system which holds the various stages of processing into a pipe-line capable of being operated by non-experts. Section 3 describes the various stages in the analysis pipe-line, in particular the actual extraction of information from the pixel data in Section 3.6, the definition of the CTI knowledge base which is produced by the system in Section 3.9 and the extraction of astronomical information from the knowledge base in Section 3.11. Section 4 covers the archival of the raw pixel data and the extraction of sub-sets of data from the archive. Section 5 describes the coaddition of pixel data from many observations and the subsequent analysis of the coadded data. Conclusions are drawn in Section 6, references quoted in Section 7 and
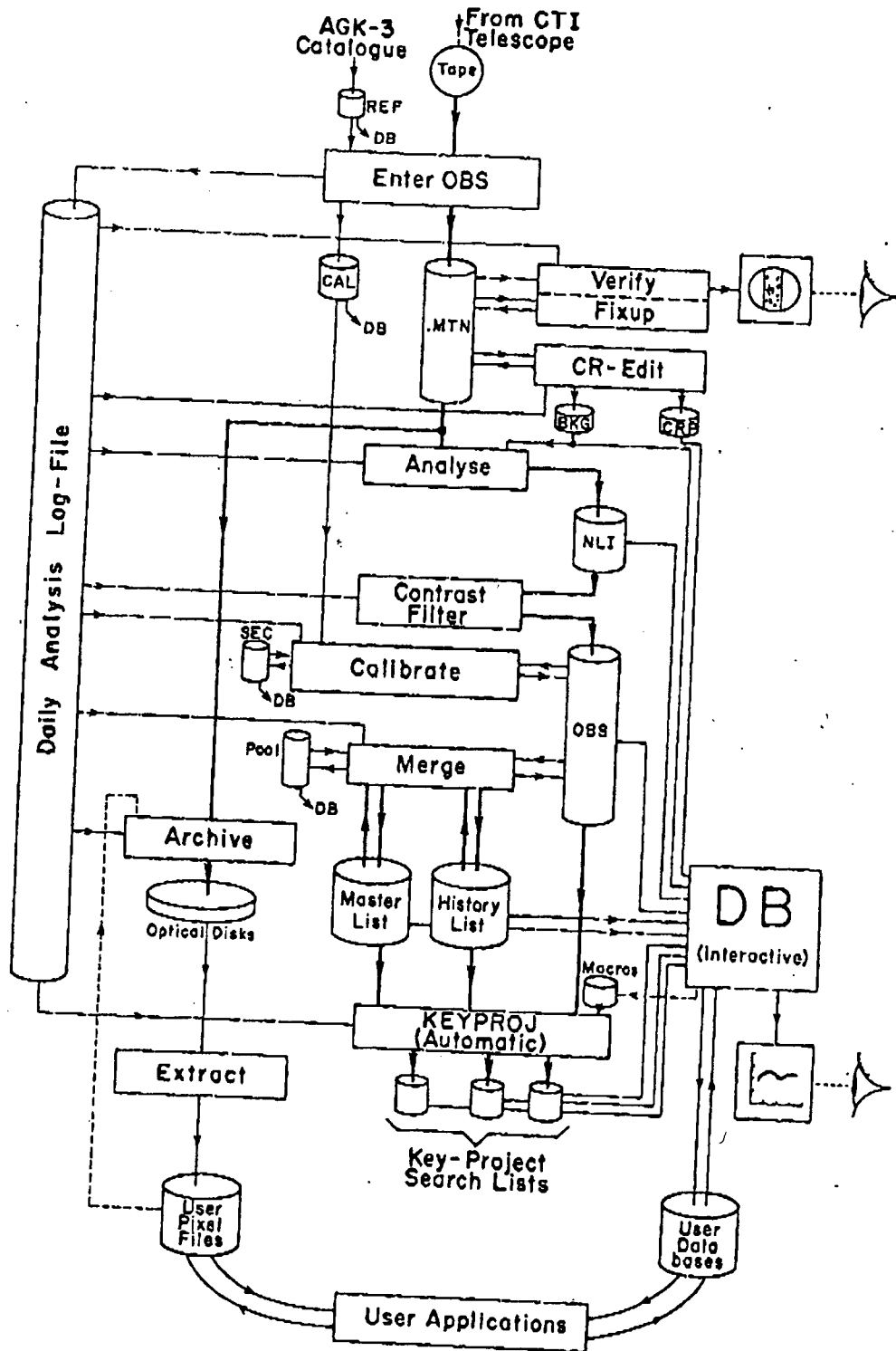
Figure 1.   Block diagram of the CTI data-analysis system.

acknowledgements of support given in Section 8.

2. Analysis control and logging system

All observational data from the CTI telescope pass through the analysis pipe-line sequentially under the control of a Daily Analysis Logging system. This control system is actually a database (known as the 'DAL-file') with one entry for each observation made with the CTI telescope. As each observation passes through the data reduction system the entry in the database is updated and key software performance diagnostics are also entered into the DAL-file to monitor the stability of the reduction system to data taken under widely differing observing conditions (e.g. phase of the moon, seeing, galactic latitude, time of night etc.). The DAL-file also prevents accidental operator error as one stage of the reduction cannot be initialised unless the DAL-file shows that the necessary previous stages have been completed successfully. Similarly, a stage cannot be repeated accidentally.

Because the DAL-file is a database it may be interrogated in the same manner as any other database, however, in this case the types of questions that are answered by the data in the DAL-file are of the form: "How far has a particular observation passed through the system?", "What is the next observation to be processed by a particular stage?"; "What was the mean image (or cosmic ray) density detected in a specified observation?"; "How stable are the geometric calibration coefficients over different time scales?" etc. As data are collected over longer periods of time so the DAL-file will also be useful for answering questions on the percentages of night time spent observing in different colours or not observing due to cloud etc.

3. Description of the CTI daily data analysis system

3.1 Entering observations into the system

Each clear night between 200 and 450 Mbytes of pixel data are transported from the telescope on Kitt Peak Mountain to Steward Observatory on magnetic tapes. The characteristics of these data are given in the accompanying paper by McGraw et al.

The first stage of the data analysis system involves entering the observations into the system. The data tapes from the telescope are mounted on the CTI Data General MV/10000 computer and a brief inspection of the data at the beginning of each tape is made using the IIS image display hardware. The UT at the start of the observation is read from the tape header and the positions of a few AGK-3 stars are marked with the cursor to achieve approximate geometric calibration.

The system then takes over automatic control and the entire observation is copied onto magnetic disk for further processing. As data around each reference star are read from tape the positions of the stars are measured automatically by an algorithm which centres on the steepest gradients in the pixel data. This is done to avoid using pixels at the image centre itself which are likely to be highly saturated. Image centres are measured to an accuracy of less than quarter of an arc-second and the rms residuals from using all AGK-3 stars observed (typically over one hundred) is less than one arc-second (more accurate calibration is performed using thousands of secondary standards at a subsequent stage -- see below).

Several automatic tests are performed on the data as they are copied from tape to disk and assuming the data passes these tests a new entry is made into the DAL-file for the observation. This entry includes the date, colour and duration of the observation and its approximate geometric calibration.

3.2 Data verification

The second stage of processing is the formal verification of the data by visual inspection of sub-sets of the data at a number of randomly chosen locations. Visual inspection has proved to be the most efficient way to verify:

(1) that the flat-fielding has been correctly applied resulting in image data free from structure across the field (background variations, linear gradients and simple curvature are permitted and can be caused by diffuse light from the moon or the sun near twilight);

(2) that the structure of images in the data is good (sharp point spread function with no elongation caused by incorrect dewar rotation etc.);

(3) that the mean sky background values are reasonable (at different angles from the moon it is necessary to change filters to keep the photon noise comparable with the

readout noise of the CCD);

(4) that residual charge in the CCD was correctly flushed at the start of the observation; etc.

If any flaws are found in the data, notes can be entered into an on-line 'jotter' file by the operator -- the system automatically logs the file name refered to, and the date, program and operator name along with the text in order to improve diagnosis of the problem. Depending on the nature of any problems the progress of the observation through the system may be terminated, the data may need to be 'fixed up', or processing can continue. The DAL-file is updated accordingly.

3.3 Correction of any simple problems, if necessary

If an observation has failed to pass the verification stage then it may be possible to correct the data, for instance by re-applying a flat-field from a different night or even by calculating a new flat-field from the data itself. This is possible with the Time-Delay and Integrate (TDI) mode of operation of the CCD because the flat-field function is one-dimensional and an observation provides as many measurements of the function as there are rows of data -- a histogram of pixel values for each column can be used to find the modal value and when normalised these values are the required flat-field function.

After an observation has been 'fixed up' then it has to be verified again before passing on the the next stage. This loop could potentially be repeated many times but in practise this would only be done as a learning exercise to try to track down the cause of any data problems.

3.4 Background following

Data that has passed the verification stage is then scanned and a smooth background is fitted to the entire observation. This is achieved by dividing the strip into five semi-overlapping sub-strips of 106 columns wide and forming the running histogram of 106 rows in each sub-strip. The median pixel values and the 80-percentile points are found in each histogram, and the medians are entered into a least squares regression of the form: Back = A + Bx + Cy + Dx^2. The medians are weighted by the inverse of the difference between the 50- and 80-percentile points so that histograms which are strongly biased by the presence of images (e.g. bright stars) are weighted less relative to those where most pixels are around the 'true' background. For each row of pixel data this process is repeated and the four coefficients are written to a database. Figure 2 shows how the background varies during an entire night as observations move out of the Galactic plane, as bright stars pass near or into the field of view, and as the moon rises.
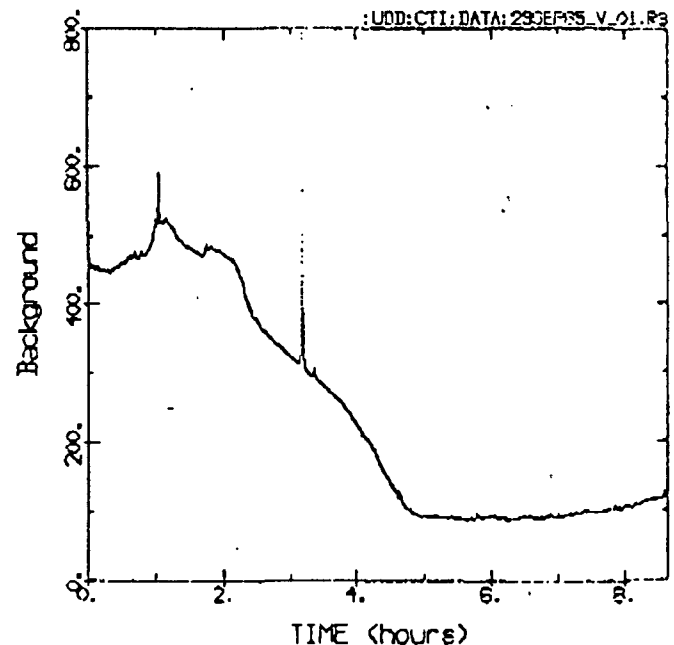

Figure 2. Variation of the background during an entire night.

3.5 Cosmic ray detection and removal

During the same pass through the pixel data in which the background coefficients are calculated, cosmic rays are detected and removed. The detection criterion is that a cosmic ray pixel must be greater than a pre-set threshold above the calculated background while the mean pixel value in the neighbouring 8 pixels be below another pre-set threshold (lower than the first). Cosmic ray pixels in the pixel file are set to the mean of their 8 neighbours.

The positions of all detected cosmic rays, together with the original pixel values and the neighbourhood mean, are entered into a database so that the original pixel data could be re-constructed if necessary, so that the stability of the algorithm can be tested and to allow possible statistical tests on the properties of the cosmic rays and any changes in

the sensitivity of the CCD chip to cosmic rays.

1.6 Analysis of the pixel data

The single most important stage in the analysis of the observations is the reliable detection and parameterisation of each astronomical image represented in the pixel data. This amounts to a transformation of information from pixel representation to attributes of the images stored in a database. A pair of inter-linked algorithms operating on the pixel data (after the fitted background estimates have been subtracted from them) are used for this transformation.

Isophotal Images

blended image

Isophotal threshold

Segments

segment 1
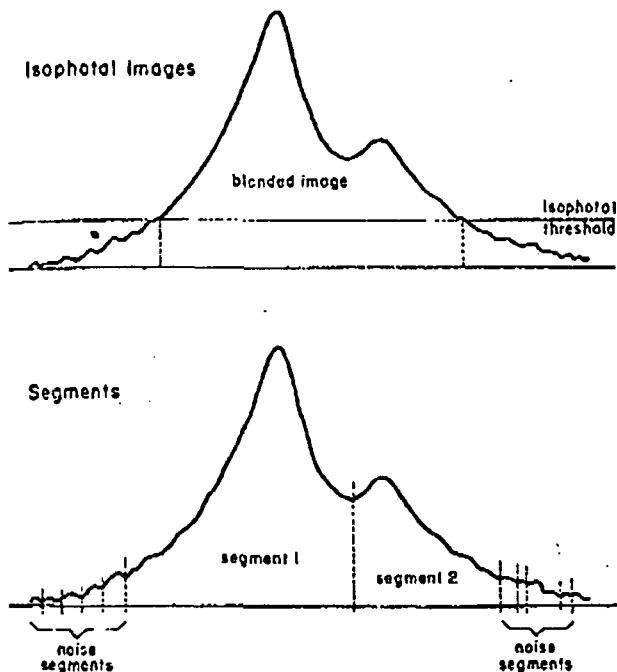
segment 2

noise segments

noise segments

Figure 3. Segmentation of an extended image.

In the first algorithm, all pixels below a pre-set isophotal threshold are ignored. Those above the threshold form groups which are connected either adjacently or diagonally. All pixels in each group are considered to be part of the same 'image'. The properties of the 'image' are calculated from the first and second moments of the group of pixels of which it is composed. These moments provide the integrated luminosity, centroid, radius of gyration, ellipticity amd position angle of the image. The peak pixel value and its position and the local background are also stored for each image.

In the second algorithm, all pixels are allowed to migrate to the highest neighbouring pixel (adjacent or diagonal) and from there upwards until a peak is reached. In this manner the pixel data is segmented into single pixels at the location of each peak and separated by valleys and saddle points in the data. During the migration, first and second moments are summed for all pixels which arrive at each peak. This results in 'segments' for which the same parameters are available as for 'images'.

The two algorithms are inter-linked by a third which cross-references the segments and images. If any pixel is common to both an image and a segment, the segment is made to refer to the image and the image updates a list of included segments (the definition of a segment makes it impossible for a segment to refer to more than one image). Segments may have a 'parent' image and images may have several segments, however segments do not necessarily have a parent image if all pixels in the segment are below the isophotal threshold for images.

The three algorithms have been designed in such a way that they can be executed together in a single linear pass through the data with each pixel being considered only once.

There are two principal advantages of 'segments' over 'images': no threshold is required to segment the image data so the luminosity contribution to an image below the isophotal threshold is not lost; and blends are detected as such so long as each object has its own distinct peak. The only disadvantage of segments is that every peak in the data is guaranteed to produce a segment and noise ensures that about one pixel in nine all over the background is a peak (if adjacent connectivity only were allowed then about one pixel in five would be a peak, which accounts for the use of diagonal connectivity in the algorithms).

However, in practise, segments and images are thrown away if their integrated value over all pixels (their luminosity) is below some detection threshold. Furthermore, an isophotal threshold of zero above background is usually used for segments to prevent their properties from being biased by a 'moat' of negative luminosities around each segment migrating to their peaks along with the central positive pixels.

These steps virtually eliminate 'background segments' but extended objects such as galaxies are also segmented at a radius where the gradient of their profiles becomes less than the pixel-to-pixel noise inherent in the data (see Figure 3). It has been found that a simple 3x3 Hanning smooth of the data prior to analysis reduces such segmentation and allows more concise separation of noise from 'real' segments at the next stage. The smoothing acts much like a matched filter at improving detection reliability and is performed as the data are read into the analysis program without editing the raw pixel file on disk.

3.7 Selection of 'real' segments

As indicated in the previous section, the segmentation algorithm can produce spurious detections due to local peaks in the data caused by noise. In order to remove such effects and to assign a 'probability value' to each detection the concept of a contrast filter is used.

When visually inspecting the image of a bright galaxy the eye is able to distinguish undulations in the surface brightness of the galaxy due to large-scale structure, such as spiral arms, from superposed images like stars or H-II regions. It seems clear that this ability is based on a comparison of the 'peakedness' of the region in question with a measure of the mean surface brightness of the surrounding environment. In other words the eye picks out regions centered on peaks on the basis of contrast from their surroundings. In an exactly parallel manner segments may be plotted on a scattergram with peak surface brightness plotted against mean surface brightness (i.e, total luminosity divided by area or number of pixels in the segment). Such a scattergram is shown in Figure 4a. Clearly no segments can lie below the y=x line (zero contrast) but segments caused by local noise spikes in the data superposed on an enhanced galaxy background appear very close to the line and increasingly so at higher surface brightnesses. This suggests rotating the axes of the plot through 45 degrees and using an exponential asymptote to filter segments close to the 'zero contrast' line. Figure 4b shows the result of this transformation where circles have been used to represent images which were selected by careful inspection by eye. A logarithmic scale has been used for the y-axis so the line represents the desired exponential asymptote.
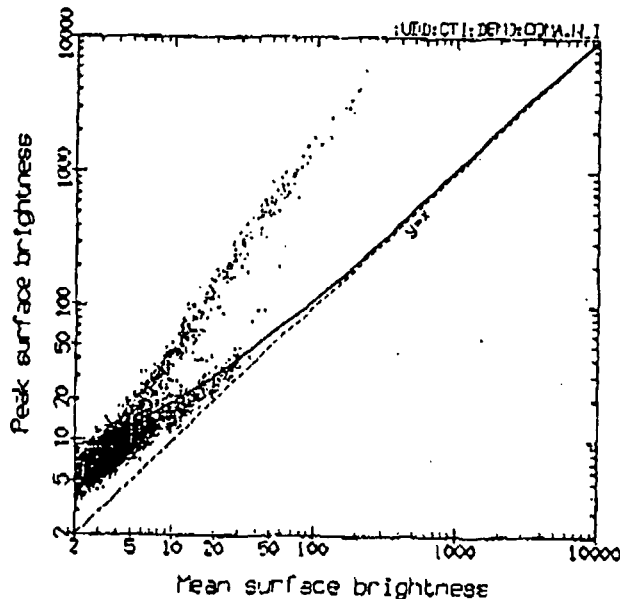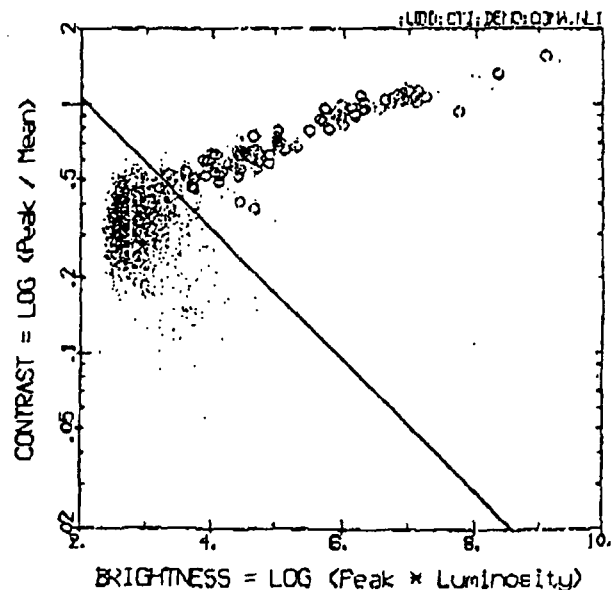

Figure 4a

Figure 4b

Figure 4.   Contrast filter to select 'real' detections.

The line rather successfully divides 'noise' segments below from 'real' segments above. If we arbitarily define the 'probability-value' of segments below the line as zero then the distance of a segment above the line can represent a measure of our confidence that the segment is a 'real' detection. We call this measure the 'probability-value' (or p-value) of the segment because it monotonically increases with genuine detection probability although its zero point is arbitrarily defined and it is not normalized. However we may choose to consider as 'definite' detections all segments which have a p-value above some threshold. Furthermore, if we detect a segment at the same x-y position on different nights we may combine the p-values as though they were genuine probabilities to decide whether the double uncertain detection warrents the status of a 'definite' detection. Thus the system maintains the ability to detect obvious changes within a single day whilst requiring repeated confirmations for less certain objects (this capability is referred to as 'fuzzy logic').

3.8 Accurate photometric and geometric calibration

Having thrown away all segments whose properties indicate that they are likely to be caused by noise, the remaining segments are then calibrated accurately in position and brightness.

Luminosity calibration is performed using photometric secondary standards within the field of view of the telescope which were obtained from a series of photometric observations with the Steward 1.55m telescope on Mount Bigalow.

Initially, for the positional calibration, AGK-3 stars are used but because most AGK-3 stars are heavily saturated in the CTI data their centroids are not sufficiently reliable (especially as charge bleeding in the CCD occurs under such conditions of saturation). Nevertheless, the AGK-3 stars over many nights' observations provide a statistical calibration from which many thousands of unsaturated secondary positional standards may be derived. These same stars are being calibrated photometrically using CTI data itself from observations made when self-consistency between the photometric secondary standards was extremely good.

All coordinates are calibrated to the 1987.5 epoch (J2000 system) because such a date is in the range of the initial period of operation of the CTI telescope. Although transformations to this system are rigourously calculated for each object during the calibration stage, the dominant terms to convert any single observation to this epoch are linear so that good approximations may be obtained in a very simple manner when extreme positional accuracy is not essential.

3.9 Definition of the CTI knowledge base

The end result of the analysis of each night's observations is a calibrated list of all objects which were observed. These lists are then used to create a knowledge base which summarises the total information extracted from all the observations to date. This knowledge base is maintained in the form of a pair of databases -- the Master List and the History List.

The Master List contains the time-averaged properties of each object ever definitely detected with the CTI telescope, plus various flags and pointers. The Master List is kept sorted strictly by right ascension and in detail each object has an entry of the form:

* right ascension and declination of the object;
* integrated luminosity in each of 5 bands (U, B, V, R and I) and two polarisations;
* the number of observations used to form the 7 average integrated luminosities;
* the area, radius of gyration, ellipticity, position angle and peak surface brightness of the object as measured from the observations of best seeing;
* the best seeing yet encountered during observations of the object;
* flags to indicate that the object is a reference star (for position and luminosity calibration), an extended object (galaxy), a variable object or an object in a confused region (ie. sometimes blended with another object);
* a pointer into an Identification list if the object has a name or other information referring to it;
* a pointer into the History List.

The History List records each individual measurement of the V-magnitude of each object ever detected with the CTI telescope and is stored in the order that each object was first detected. In detail the History List has an entry for each object of the form:

* right ascension and declination of the object (to provide easy access into the Master List);
* integrated V-magnitudes each time the object was detected;

• dates on which each detection was made (to enable a light curve to be drawn).

The user interface to the CTI knowledge base is by means of a powerful database interrogation program with graphical I/O capability. Users may extract all objects obeying complex sceintific, relational and/or logical tests on their properties and these tests may optionally be specified in graphical or functional form (see the accompanying paper by Cawson et al.).

3.10 Merging new observations into the CTI knowledge base

In order to create and maintain the CTI knowledge base new observations are merged into the Master and History Lists on a daily basis. This stage is conceptually extremely simple consisting of a search through the current Lists for each object just detected and either updating the entries or inserting a new one (the Master List is sorted by RA and observations are made in that order so very little searching is required). However, the problem is complicated dramatically by the possibility that two distinct detections on a night of good seeing may be detected as a single blended object on a night of inferior seeing. For this reason unmatched detections must first be compared with possible blends of two or more objects before a new entry is made. Fortunately, the nature of the segmentation algorithm makes it extremely unlikely that complex blends of many detections will be needed to match a single detection from another night (this would not be true for the isophotal threshold algorithm alone). Searching for blends is therefore limited to individuals, pairs or triples from either list matching one another.

A second problem results from the fact that the telescope appears to have very small time-variations in its optical axis. These variations amount to a few arc-seconds over time-scales of a few minutes and are almost certainly caused by flexure of the telescope due to horizontal thermal gradients across the structure. Because of this flexure it is necessary to allow for small variable x-y offsets between new observations and the Master List. These offsets are locked to low values (around zero) by the previous calibration technique. However, a log of the x-y offsets which are found to be necessary to pull the positions of objects into exact conicidence as a function of time is put into a database for engineering analysis and for use when aligning multiple files of pixel data for coaddition (see Section 5.1 below).

There is one further complication in the maintenence of the CTI Master and History Lists which concerns the threshold at which detections are placed into the Lists. In any noisy system with a sharp decision threshold there will be objects which pass the threshold due to noise adding to their 'true' properties and similarly other objects which are wrongly rejected. Statistically these effects may cancel one another but that is not sufficient when two lists are to be compared and matched on an object-by-object basis. Objects which fall close to the threshold may actually fall just below the threshold in one list (i.e. be absent from the list) and above in the other list -- the result is that a new object is found or an existing one lost and this may trigger some unwarrented action (for instance alerting other astronomers).

To overcome this problem it is necessary to use a 'fuzzy' threshold consisting of two thresholds spaced sufficiently far apart that an object crossing from below the lower threshold to above the higher one would be sufficiently remarkable to warrent triggering the action. This is achieved by storing all objects between the two thresholds in a 'pool'. In practise each object in the pool has a detection probability (p-value) so that objects which continuously fall between the thresholds may reinforce one another and migrate towards and ultimately cross the upper threshold for inclusion in the Master and History Lists. This scheme remains alert to new objects of sufficient certainty of detection in a single observation while using the results of several observations for fainter objects.

3.11 Automatic execution of key projects

After the CTI Master and History Lists have been updated with the results of a night's observations, the observations are compared with the Master and History Lists to search for evidence of newly detected interesting objects. For each current CTI project whose goal is the detection of objects of a particular type, the first step is to interrogate the CTI knowledge base interactively in order to define and refine the logical tests which are optimal in selecting the objects of interest and descriminating against unwanted ones. Once this has been done, the definitions of the logical tests can be entered as 'macros' for automatic execution on a daily basis. In this way key projects are performed on all the data which the telescope collects and 'search lists' of selected objects are automatically compiled for each project.

There is a parallel effort to use AI techniques to analyze and search the CTI knowledge base with the aim of producing an expert system to classify objects and to find unusual

ones (see the accompanying paper by Bernat *et al.*).

4. Archival of the pixel data

4.1 The CTI archival system on optical disks

Each clear night can produce as much as 450 Mbytes of pixel data so it is imperative to flush this data from active computer memory (magnetic disk etc.) as quickly as possible. After the analysis stage the pixel data are not explicitly required by any further stages although purging is usually initialised only after the analyzed databases have been calibrated and successfully merged with the Master and History Lists. This is so that accurate calibration is available for the data when it is archived, so that any problems encountered may be investigated and corrected immediately, and so that raw pixel data is immediately available for regions around any newly detected objects of interest.

When these requirements have been met the pixel data is permanently archived on optical disk. This medium offers a compact, random-access archival system at low cost and is the subject of an accompanying paper (Keane *et al.*). Approximately two weeks of observations can be stored on a single optical disk so physical volume becomes an added advantage over a tape-based archival system.

Due to the large data volume and the fact that astronomical images are characterised by a majority of blank sky background of approximately constant pixel values, there is both the incentive and the opportunity to compact the pixel data without any loss of information. The scheme which is used involves storing the first pixel as its absolute value (using 2 bytes) and then switching to one-byte differences unless two adjacent pixels differ in value by more than 127. When this happens the special value -128 is inserted into the archive and the scheme repeats. This data compression is referred to as 'byte compaction' and yields a typical data volume reduction of 45 percent.

The disadvantage is that it is no longer possible to jump into the pixel data randomly at any location. In fact a region of pixel data near the end of a strip 8 or 10 hours long could only be regenerated by unpacking the data sequentially from its start. This problem is overcome by re-starting the byte compaction at pre-defined intervals in the data (e.g. every 512 lines) and recording in a directory the location in the archive of the first pixel of each interval. The same directory is used to store the calibration coefficients for the observation and all other information which might be relevent when data is to be extracted from the archive (for example the seeing and mean background values, date and filter, etc.). A database is used for the directory to ease the interrogation of the pixel archive and to allow the archival process itself to be monitored, for instance to check the efficiency of the byte compaction algorithm.

4.2 Extraction of image data from the archive

There are many situations which may be relevent when a user wishes to extract data from the pixel archive. For example he may have a list of coordinates and want a small region of image data around each one in a specified colour, or he may want all observations of a particular named object when the seeing was better than some limit. Thus a general-purpose program to interrogate the pixel archive directory is required which may create a list of the locations of the desired sub-set of pixel data for subsequent extraction in batch mode. At present the user specifies exactly which data he requires but a command language driven system is planned where requirements may be entered and the system will select which data is suitable to be extracted.

5. Coaddition of the pixel data

5.1 Creation of coadded pixel data

Periodically, all the observations taken in a given colour band are extracted from the pixel archive, registered (using the databases of x-y offsets created during the merging of the observations into the the CTI knowledge base), and coadded. This procedure produces a seamless image of the entire CTI strip to fainter limiting magnitude than a single observation. The coadded image is stored in the form of a 4-byte field for each pixel -- three bytes for the simple sum of the pixels from each observation and the fourth byte for the number of values entered into the sum. In this way up to 255 observations can be coadded (many years of data) and individual pixels whose values are in any way suspect in any observation (e.g. cosmic ray pixels) need not be entered into the coadded image.

During the creation of the coadded image all other use of the computer has to be suspended and all the magnetic disks cleared for the sole purpose of making room for the coadded data (1.1 Gbytes per colour). The resulting coadded image is immediatly archived to optical disk at the end of the task and may be 'extracted' in the same manner as nightly

observations.

5.2 Analysis of the coadded pixel data

When the coadded data has reached a sufficiently faint limiting magnitude it will be normalised (by dividing the three-byte sum by the number of entries for each pixel) and fed into the input of the CTI data analysis system to create a single list analogous to the Master List. This 'Super List' is expected to have approximately 20 million entries to a limiting V-magnitude of about 25.

6. Conclusions

The CTI telescope has been in part-time operation for just over one year. During this time engineering and scientific data has been collected and run through the CTI data analysis system. The experiences with this data have resulted in the analysis system described in this paper. Although some sections of the analysis system have not yet been completed it is unlikely that the overall structure of the system will change.

The use of a database approach in all aspects of the project has dramatically eased the design of the system and has speeded up its implementation and the analysis of problems as they have been found. So far the entire system has required just two man-years of software effort which has included the design and implementation of the database definition and interrogation system.

7. Acknowledgements

8. References

1. Bernat, A. P., McGraw, J. T., Proc. SPIE Conference, Instrumentation in Astronomy, 1986.

2. Cawson, M. G. M., McGraw, J. T., Keane, M. J., Proc. SPIE Conference, Instrumentation in Astronomy, 1986.

3. Keane, M. J., McGraw, J. T., Cawson, M. G. M., Proc. SPIE Conference, Instrumentation in Astronomy, 1986.

4. McGraw, J. T., Cawson, M. G. M., Keane, M. J., Proc. SPIE Conference, Instrumentation in Astronomy, 1986.