A fast contour descriptor algorithm for supernova image classification

Cecilia R. Aragon*[a], David Bradburn Aragon[b]
[a]Lawrence Berkeley National Laboratory, One Cyclotron Rd., Berkeley, CA, USA 94720;
[b]DCA, 1563 Solano Rd. #434, Berkeley, CA USA 94707

ABSTRACT

We describe a fast contour descriptor algorithm and its application to a distributed supernova detection system (the Nearby Supernova Factory) that processes 600,000 candidate objects in 80 GB of image data per night. Our shape-detection algorithm reduced the number of false positives generated by the supernova search pipeline by 41% while producing no measurable impact on running time. Fourier descriptors are an established method of numerically describing the shapes of object contours, but transform-based techniques are ordinarily avoided in this type of application due to their computational cost. We devised a fast contour descriptor implementation for supernova candidates that meets the tight processing budget of the application. Using the lowest-order descriptors (F_1 and F_{-1}) and the total variance in the contour, we obtain one feature representing the eccentricity of the object and another denoting its irregularity. Because the number of Fourier terms to be calculated is fixed and small, the algorithm runs in linear time, rather than the *O(n log n)* time of an FFT. Constraints on object size allow further optimizations so that the total cost of producing the required contour descriptors is about *4n* addition/subtraction operations, where *n* is the length of the contour.

Keywords: image processing, contour descriptors, Fourier descriptors, shape detection, real-time image processing, supernova, astronomy, applications

1. INTRODUCTION

One of the most important scientific discoveries of the last decade is that the universe is expanding at an increasing rate [1, 2]. Type Ia supernovae play a critical role in this discovery because they provide measurements of the universe's expansion history since the time they began emitting light, several billion years ago. Such stellar explosions are very rare, occurring only a few times per millennium in a typical galaxy, and remaining bright enough to detect for only a few weeks (Figure 1). The scientific challenge of finding such supernovae stems from their rare occurrences, short lifespans, and the necessity of detecting them as early as possible after explosion, before maximum brightness is reached. The computational challenge lies in the large volume of data that must be rapidly and accurately analyzed in order to find the one or two supernovae that could appear on a given night out of hundreds of thousands of potential candidates. Additionally, image quality and characteristics may change significantly over time due to weather, the changing phase of the moon, and frequent modifications to the data acquisition procedures (e.g., new telescopes, equipment and camera changes, calibrations, image processing improvements), so there is need for a robust shape-detection algorithm that can produce reproducible results over these changes.

Good candidate supernovae are distinguished from poor candidates partly by their shape. Being stars, supernovae are point sources that should produce round images, with a small-diameter Gaussian intensity distribution dependent on optical and atmospheric conditions ("seeing"). Also present in astronomical images are spiral galaxies which, being viewed at an angle, appear elliptical. The images can contain spurious objects resulting from optical and electronic aberrations such as saturation and flaring, and artifacts left by processing to remove those objects (Figure 2). The process of subtracting images of the supernova's host galaxy to reveal the supernova can also generate artifacts. Finally, the images are captured near the limit of the CCD's light-gathering ability and contain speckle noise, in which there may be groups of connected pixels of comparable size to real celestial objects. While galaxies are elliptical and processing artifacts may be of arbitrary shape, good supernova candidates are consistently circular. Thus, the shape of the outer contour of an object can be used to qualify the object as a candidate supernova.

*aragon@hpcrd.lbl.gov; phone 1 510-486-4106; fax 1 510-486-5812; www.lbl.gov

Figure 1. Hubble Space Telescope image of supernova SN1994D in galaxy NGC4526, observed 1994.

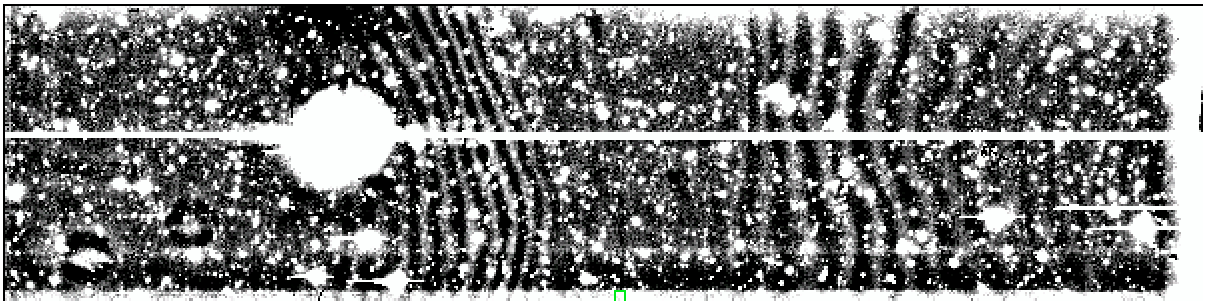


Figure 2. Typical NEAT CCD image used by the Supernova Factory to look for supernovae, containing noise and CCD artifacts such as fringing, saturated pixels, and reflections.

1.1. The supernova image analysis pipeline

The Nearby Supernova Factory (SNfactory) is an international project to obtain spectrophotometry data on a large sample of Type Ia supernovae in order to measure the expansion history of the universe [3]. On a typical night, approximately 30,000 images containing 600,000 potential supernovae (80 GB of data) are received by SNfactory software. These images must be processed and subtracted in 24 hours or less (ideally, in 12 hours) in order that the resulting potential supernovae can be scanned by scientists the next morning, to determine whether they are good candidates for follow-up analysis by spectrography. Time is of the essence, as the supernova's scientific value increases the earlier in its lifecycle it is discovered; the maximum scientific value is gained when it is discovered well before it reaches its peak brightness (approximately two weeks after initial explosion). Therefore it is important that the features which are used to screen the potential supernova candidates be both computationally inexpensive and extremely effective at separating good from bad candidates.

The Supernova Factory search software computes photometric and geometric features of subimages thought to contain potential supernovae, thresholds each of these features, and sends those subimages that satisfy all thresholds to human scanners, who reject or accept them for follow-up studies. Due to high and variable image noise levels and image processing artifacts, the great majority of subimages that pass the thresholds are false positives that humans must manually reject. The version of the software under development applies machine learning techniques (including support vector machines and boosted decision trees) to these features in order to improve the ratio of good to bad candidates sent to humans for manual scanning [4]. The effectiveness of such machine learning algorithms depends on the ability of the input features (describing the astronomical images) to accurately separate the classes.

1.2. Fourier contour descriptors and their application to supernova detection

Fourier descriptors are an established method of numerically describing the shapes of object contours [5]. However, Fourier transform-based methods were initially considered infeasible, due to their computational complexity, for the high throughput requirements of the Supernova Factory. We devised an extremely fast contour descriptor implementation that meets the processing budget of the application. Noting that the lowest-order descriptors (F_1 and F_{-1}) convey the circular or elliptical aspect of the contour, our features are based on those two terms and the total variance or power in the contour. The features are thus equipped to detect eccentric or irregular objects, classes to which the poor candidates generally belong.

The remainder of this paper is structured as follows: Section 2 describes the scientific mission of supernova search, the software that generates the image data in our study, and the features extracted from subimages of celestial objects. Section 3 defines the linear-time contour descriptor algorithm we devised, and describes empirical studies to demonstrate the validity of the Fourier contour descriptors in this astronomical application. Finally, Section 4 discusses conclusions and the current impact of this work on a large-scale supernova survey, and future impact on the fields of astronomy and cosmology as many more such large-scale surveys are planned for the future as physicists attempt to unravel the mystery of dark energy.

2. THE SEARCH FOR SUPERNOVAE

The Nearby Supernova Factory project is intended to obtain spectrophotometry data on a large sample of Type Ia supernovae in a nearby redshift range ($0.03 < z < 0.08$, or about 0.4 to 1.1 billion light years away) in order to measure the expansion history of the universe [3]. Previous studies of Type Ia supernovae led to the discovery of the mysterious "dark energy" that is causing the universe to expand at an accelerating rate [1, 2].

Exploring the implications of these discoveries requires reducing the statistical uncertainties in previous experimental data. For that purpose, extensive spectral and photometric monitoring of a larger number of Type Ia supernovae is essential. The SNfactory collaboration has built an automated system consisting of specialized software and custom-built hardware that systematically searches the sky for new supernovae, screens potential candidates, then performs multiple spectral and photometric observations on each supernova. These observations will be stored in a database to be made available to researchers worldwide for further study and analysis.

Each night, the SNfactory receives about 80 GB of wide field CCD imaging data taken by the Quest-II camera on the Oschin 1.2-m telescope on Mt. Palomar for the Near Earth Asteroid Tracking (NEAT) project at JPL. Images are transferred from Mt. Palomar via the High Performance Wireless Research and Education Network (HPWREN) to the High Performance Storage System (HPSS) at the National Energy Research Scientific Computing Center (NERSC) in Oakland, California. Each morning, SNfactory search software running on NERSC's 700-node computing cluster, the Parallel Distributed Systems Facility (PDSF), matches images of the same area of the sky, processes them to remove noise and CCD artifacts, then performs an image subtraction from previously observed reference images on each set of matched images (Figure 3)[6]. Supernovae are detected by subtracting a reference image from a new one to reveal the newly luminous object. Each such object becomes a supernova candidate. Figure 3 below depicts a typical subtraction; the leftmost subimage is the "reference" or original subimage depicting a galaxy before the supernova explosion; the middle subimage is the "new" subimage taken after the supernova explosion; and the rightmost subimage is the resulting subtraction of the two, revealing the presence of the supernova.

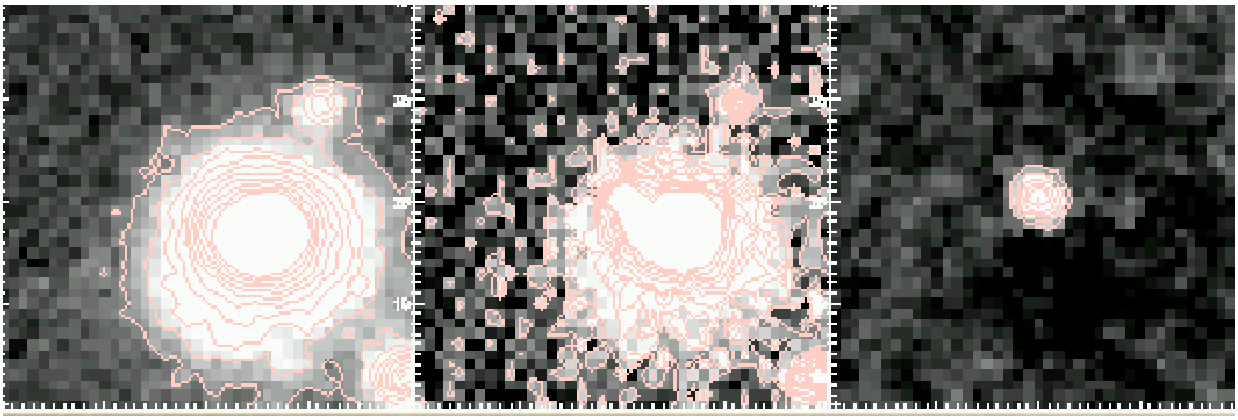


Figure 3. Example supernova images, from L to R: reference image of galaxy, new image of galaxy plus new bright area, subtracted image reveals new supernova.

A set of 19 features (Table 1) is computed on subimages of each subtraction image in which bright, point-source-like objects are detected. (The last two features, R1 and R2, are the new roundness descriptors detailed in Section 3.) In the existing software, each feature is thresholded from above and/or below, and subimages that satisfy all thresholds are identified as containing a potential supernova candidate and saved to a database. Several hundred images containing one or more candidate subimages are saved each night.

Feature Name	**Feature Definition**
apsig	signal-to-noise ratio in candidate aperture
perinc	% flux increase in aperture from REF to NEW
pcygsig	difference of flux in 2*FWHM of aperture and 0.7*FWHM; detects misaligned REF and NEW images)
mxy	x-y moment of candidate
fwx	FWHM of candidate in x
fwy	FWHM of candidate in y
neighbordist	distance in pixels to the nearest object in REF
new1sig	signal-to-noise of candidate in NEW1
new2sig	signal-to-noise of candidate in NEW2
sub1sig	signal-to-noise of candidate in SUB1
sub2sig	signal-to-noise of candidate in SUB2
sub2minsub1	weighted signal-to-noise difference between SUB1 and SUB2
dsub1sub2	difference in pixel coordinates between SUB1 and SUB2 (motion measurement)
holeinref	measure of negative pixels on REF in region of candidate
bigapratio	ratio of sum of positive pixels to sum of negative pixels within aperture
relfwx	REF image FWHM in x divided by NEW image FWHM in x
relfwy	REF image FWHM in y divided by NEW image FWHM in y
R1	object contour eccentricity; ratio of powers in lowest order negative and positive Fourier contour descriptors
R2	object contour irregularity; power in higher order Fourier contour descriptors divided by total power

Table 1: Definitions of features computed from image regions containing potential supernovae. Additional definitions: aperture = subarea of image within which features are computed; REF = reference image of a piece of sky composited from multiple nights; NEW1 (NEW2) = new images of same piece of sky; SUB1 (SUB2) = subtraction of NEW1 (NEW2) and REF images; FWHM = full-width half-max; flux = integrated photon levels over image subarea, i.e., sum of all pixel values within aperture. The new contour-based features are R1 and R2.

These subimages are then visually scanned by humans, a process that takes approximately 16 person-hours for each night's data. Figures 4, 5, and 6 below show examples of the subimages viewed during this process. Each figure depicts

a triplet of reference/new/subtraction images (the reference image is from before the supernova explosion, the new image is the current one including the supernova, and the subtraction depicts the new bright spot alone). All of these subimages were "passed" by the search software before our new roundness features were added. Figure 4 depicts a good candidate (note the compact circular form), but figures 5 and 6 display bad subtractions which should have been culled out. After our algorithm had been implemented (features R1 and R2 added to the thresholded criteria), Figure 5 failed the R1 cut and would not have been passed for human scanning. Figure 6 failed the R2 cut and likewise would not have been passed.

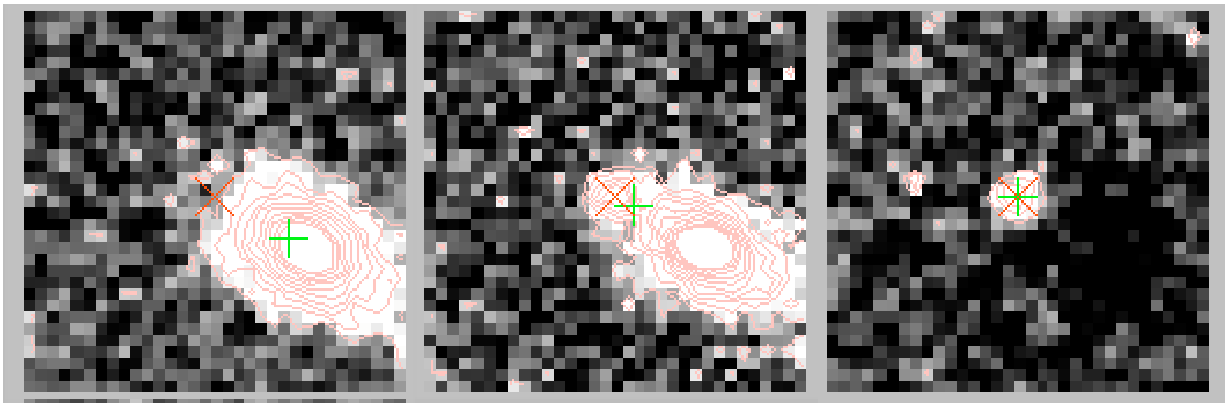


Figure 4. Supernova found in Nov 2006 by the Nearby Supernova Factory: SNF20061117-000. Images from L to R: reference image of galaxy, new image of galaxy plus bright spot, subtraction revealing new supernova.

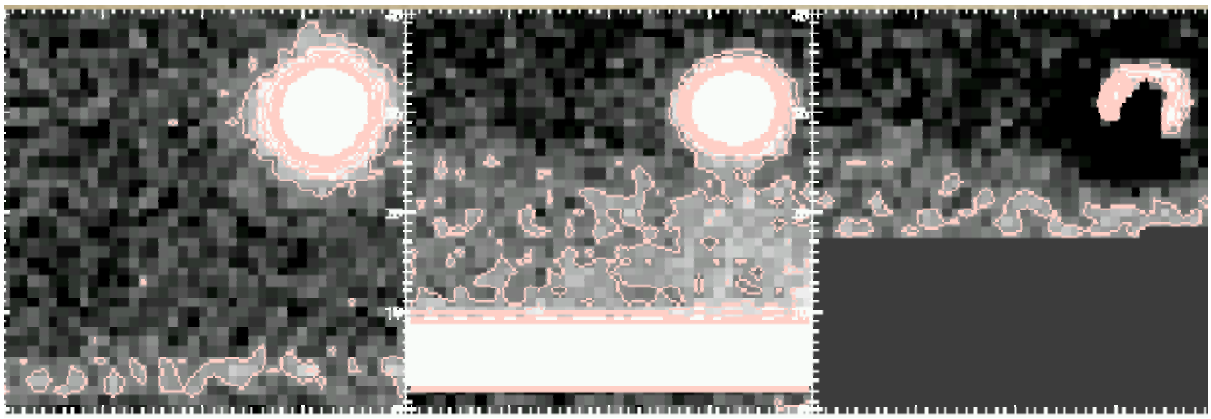


Figure 5. Example of a bad subtraction. Images from L to R: reference image, new image, subtraction (the images were misaligned leaving the subtraction with a meniscus-shaped object). This object is screened out by R1.

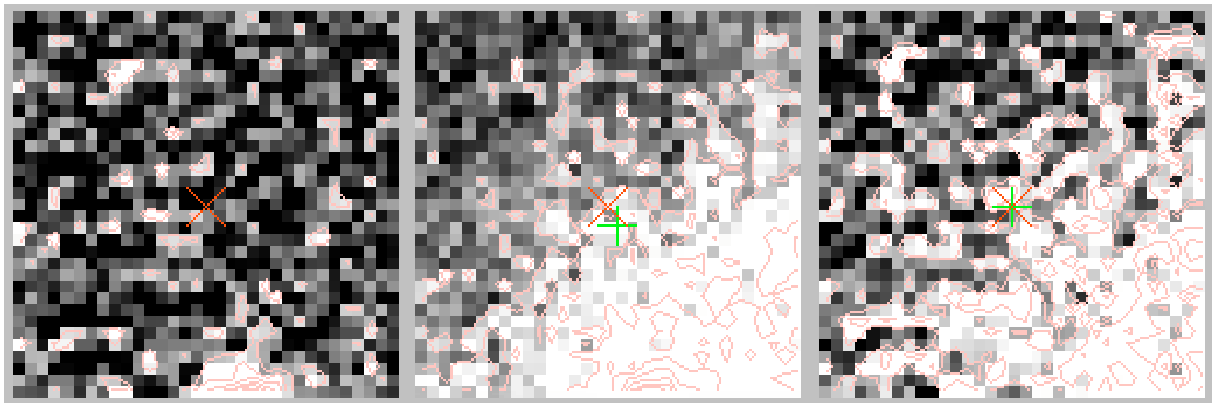


Figure 6. Example of a bad subtraction. Images from L to R: reference image of empty space, new image with edge of bright object present, subtraction with an object that has an irregular contour. This object is screened out by R2.

Of these visually scanned subimages (about 1000 per night prior to our implementation of the contour descriptors), about 1% are flagged as containing potential supernova candidates. These candidates (about 5 to 30 per night, or about .001% -.01% of the objects originally imaged) are sent to the Supernova Integral Field Spectrograph (SNIFS) on the University of Hawaii 2.2-m telescope on Mauna Kea for spectrophotometric screening and follow-up. The resulting image and spectral data are stored in a database for scientific study. Of these, only a few are actual supernovae and the rest are other transient objects that appear in the image subtractions, e.g. asteroids, variable stars, and AGNs, or imaging artifacts that resemble transient objects, both in the feature space and when visually scanned. Typically from one to three of the objects examined by SNIFS will finally be determined to be supernovae, of which (on average) about one per night is the sought-for Type Ia supernova. Thus of the objects originally imaged, somewhere near .0002% prove to be Type Ia supernovae.

The bottleneck in this process is the large amount of manual labor required to find good supernova candidates. Feature thresholds are often adjusted manually (known as "tightening the cuts") in order to reduce the number of false positives (frequently due to bad subtractions) sent to human scanners. However, this tuning results in the occasional loss of a good supernova candidate, and due to their rarity, it is important to miss as few candidates as possible.

3. FAST CONTOUR DESCRIPTOR ALGORITHM

Fourier descriptors [5, 7, 8] are a well-known method for parameterizing the outer contour of an object by decomposing it in terms of complex exponential functions. Although long applied in other areas such as OCR (optical character recognition) [9], medical imaging [10] and industrial defect detection [11], previous applications to shape analysis of celestial objects are not applicable to supernova search [12, 13].

Because transform-based image analysis methods are often avoided in high-throughput applications due to their computational cost, it is important to distinguish the contour Fourier descriptor from other Fourier-based methods. For a two-dimensional image of diameter D pixels, the processing time for a Fourier transform of the image can scale as severely as D^4. In fortunate cases where the Cooley-Tukey FFT can be used, the processing still scales as $D^2 \log(D^2)$. The contour Fourier descriptor operates not on the object image but on its outer contour, which is essentially a one-dimensional (albeit complex-valued) sequence, of length proportional to object diameter D. Since we calculate a fixed number of terms rather than requiring a full DFT, the calculation time is actually proportional to D.

3.1. Algorithm overview

We compute two new features (or "scores") on each subimage containing an object, and add them to the set of features already calculated in the supernova search pipeline. These two features are: R1, which measures the "roundness" or eccentricity of the candidate object, and R2, which measures the amount of irregularity in the object contour.

Our approach inserts the following steps into the processing pipeline for candidate supernova objects:

1. Find the outer contour of the object and store it as a sequence of (x, y) points.
2. Calculate the lowest Fourier descriptor terms F_1 and F_{-1} on the contour point sequence.
3. Form two roundness-related features: R1 measures the eccentricity of the object (0 is a circle, 1.0 is a line and intermediate values are ellipses). R2 measures the amount of irregularity in the object contour.

$$\mathrm{R1} = \frac{\left\|F_1\right\|^2}{\left\|F_{-1}\right\|^2}$$

$$\mathrm{R2} = \frac{\sum_{|k|>1} \left\|F_k\right\|^2}{\sum_{k\neq 0} \left\|F_k\right\|^2}$$

The features R1 and R2 are added to the list of thresholds or "cuts" applied to image objects, so that objects with too high R1 or too high R2 are rejected as supernova candidates, and not sent to human scanning.

3.2. Contour discovery

The input data for the image object includes an intensity threshold, above which a connected pixel is to be considered part of the object. Typically this is the "half-max" value, i.e. half the intensity of the highest-intensity central pixel in the object (although our algorithm does not enforce this convention but merely accepts the threshold as input). The input also includes coordinates of a pixel known to belong to the object; due to the operation of previous pipeline stages this seed pixel will often be in the center of the object, but it need not be. (Although for stars the center pixel will be part of the object, this may not be true for a pathologically-shaped object resulting from a bad subtraction). The algorithm finds an object edge by moving rightward from the seed pixel, until a pixel below the intensity threshold is found.

Having found an edge, the object contour is found using a chain-code generator of the "crack code" style (i.e. similar to a Freeman code but using four directions instead of eight) [14, 15]. A 2x2 pixel window is placed over the detected edge, so that the window contains at least one pixel belonging to the object and at least one that does not. The center of the window thus lies on an edge or corner of the object. The memberships of the four pixels under the window form a 4-bit address into a lookup table that gives the direction in which the window is to be moved next, so as to follow the object contour counterclockwise.

The output of the contour follower is a list of the x,y coordinates of the corners of the object's pixels. The origin of the coordinate system is not the same as for the original image, first because the chain code starts from the detected object edge rather than the image origin, and second because the coordinates describe the corners of pixels rather than their centers. However, no correction is necessary because the subsequent processing is invariant under translation.

3.3. Fourier descriptor calculation

The discrete Fourier transform of a periodic sequence of complex numbers z_k is a periodic sequence of complex values F_n where:

$$F_n = \frac{1}{N}\sum_{k=0}^{N-1} z_k e^{-j2\pi nk/N}$$

In this case, we use the (x,y) coordinates of the object contour to form the sequence of complex $z_k = (x_k + jy_k)$. Then expanding the complex exponential as a sum of circular trigonometric functions we have

$$F_n = \frac{1}{N}\sum_{k=0}^{N-1} (x_k + jy_k)(\cos(\tfrac{2\pi nk}{N}) - j\sin(\tfrac{2\pi nk}{N}))$$

and expanding and then collecting the real and imaginary results gives

$$F_n = \frac{1}{N}\sum_{k=0}^{N-1} (x_k \cos(\tfrac{2\pi nk}{N}) + y_k \sin(\tfrac{2\pi nk}{N})) + j\sum_{k=0}^{N-1} (y_k \cos(\tfrac{2\pi nk}{N}) - x_k \sin(\tfrac{2\pi nk}{N}))$$

where the complete Fourier transform consists of the entire set of F_n for $n \in \{0, 1, \ldots, N-1\}$. The periodicity of the functions allows us to number the F_n alternatively using $n \in \{-\frac{N}{2}, \ldots, -1, 0, 1, \ldots, \frac{N}{2}\}$, where the positive and negative indices correspond to traversing the same contour in opposite directions.

However, of all these F_n we are only interested in a specific few. The contour of an ideal supernova candidate is circular, so that ideally the only nonzero term would be F_1, which corresponds to a circular contour traced counterclockwise.

The closely related term F_{-1} corresponds to a circular contour traced in the opposite direction from F_1. Increasing magnitude of F_{-1} corresponds to increasing eccentricity of the object, changing it from a circle to an ellipse and finally to a line when $\|F_1\| = \|F_{-1}\|$. Eccentric, elliptical objects are common among the poor candidates we wish to detect and eliminate. Therefore, we define the first roundness feature as

$$R1 = \frac{\|F_1\|^2}{\|F_{-1}\|^2}$$

Higher-order Fourier terms encode more complicated curves. For example, meniscus-shaped objects (such as may result from subtracting slightly mis-registered images) usually have significant energy in F_2 and/or F_{-2}. We define our second roundness discriminant R2 as the combined power (squared magnitude) at all frequencies higher than the circular terms, divided by the total power in the circular and higher terms together:

$$R2 = \frac{\sum_{|k|>1} \|F_k\|^2}{\sum_{k\neq 0} \|F_k\|^2}$$

As we will show below, once the values necessary for R1 have been obtained, R2 can be calculated without any further trigonometric operations.

3.4. Computational efficiency

Now we examine the computation required to obtain R1 from the image contour. Substituting the indices 1 and -1 into the above definition of the F_n and using the identities $\sin(-\theta) = -\sin(\theta)$, $\cos(-\theta) = \cos(\theta)$ gives:

$$F_1 = \frac{1}{N}\left(\sum_{k=0}^{N-1}(x_k \cos(\tfrac{2\pi k}{N}) + y_k \sin(\tfrac{2\pi k}{N})) + j\sum_{k=0}^{N-1}(y_k \cos(\tfrac{2\pi k}{N}) - x_k \sin(\tfrac{2\pi k}{N}))\right)$$

$$F_{-1} = \frac{1}{N}\left(\sum_{k=0}^{N-1}(x_k \cos(\tfrac{2\pi k}{N}) - y_k \sin(\tfrac{2\pi k}{N})) + j\sum_{k=0}^{N-1}(y_k \cos(\tfrac{2\pi k}{N}) + x_k \sin(\tfrac{2\pi k}{N}))\right)$$

In computing these sums, at each value of k both descriptors use the same four partial results, namely each of (x_k, y_k) multiplied by each of the two coefficients $\cos(\theta)$, $\sin(\theta)$. Denoting the partial results by $q_k = x_k \cos(\frac{2\pi k}{N})$, $r_k = y_k \sin(\frac{2\pi k}{N})$, $s_k = y_k \cos(\frac{2\pi k}{N})$, $t_k = x_k \sin(\frac{2\pi k}{N})$ shows the commonality in the calculations of the two descriptors:

$$F_1 = \frac{1}{N}\left(\sum_{k=0}^{N-1}(q_k + r_k) + j\sum_{k=0}^{N-1}(s_k - t_k)\right) = \frac{1}{N}\left(\sum_{k=0}^{N-1} q_k + \sum_{k=0}^{N-1} r_k + j\left(\sum_{k=0}^{N-1} s_k - \sum_{k=0}^{N-1} t_k\right)\right)$$

$$F_{-1} = \frac{1}{N}\left(\sum_{k=0}^{N-1}(q_k - r_k) + j\sum_{k=0}^{N-1}(s_k + t_k)\right) = \frac{1}{N}\left(\sum_{k=0}^{N-1} q_k + \sum_{k=0}^{N-1} r_k + j\left(\sum_{k=0}^{N-1} s_k - \sum_{k=0}^{N-1} t_k\right)\right)$$

The cost of calculating R1 is thus linear in the contour length. At each contour point, two trigonometric operations and four multiplications are performed to obtain q_k, r_k, s_k, t_k. We now proceed to eliminate first the trigonometric calculations and then the multiplications.

The optimization depends on the fact that N is small for candidate supernova objects – never greater than 64 points. (Recall that the candidates should be circular and are typically about 6 pixels in diameter.) This is small enough to consider the use of lookup tables. Due to the manner in which the chain code is generated, the contour lengths are always even, so the symmetry of the trigonometric functions can be used to eliminate half the calculation. In fact, we actually generate the tables only for radices r = 4n where n = 9, 10, 11, …16, which removes the need for separate sine and cosine tables since those functions differ by a quarter-circle offset. For radices below 32, the lookup table entries are found by consulting the table for a multiple of the chosen radix. The lookups are therefore exact for radices of 32 or less, i.e. for round supernova candidates of diameter about 10 pixels or less, and are approximate (but quite a good approximation) for larger objects. After these various optimizations, the total number of cosines stored in the table is

under 100, and no further compacting of the table was judged necessary. The calculation of trigonometric coefficients at run-time is thereby eliminated.

We then note that the (x, y) coordinates on the contour are integers, and usually very small ones (<10 in absolute value). This allows us to eliminate the multiplications, by using the contour coordinates as array indices to locate cosine values already multiplied by the appropriate small integer. Although with this modification the table still has fewer than 1000 entries, it allows us to obtain q_k, r_k, s_k, t_k with no floating-point operations at all. These four values are then accumulated over the contour, and the descriptors F_1 and F_{-1} are obtained as shown above.

The real and imaginary parts of the two descriptors are separately squared to obtain the required squared magnitudes for R1. Using the abbreviated notation $q = \sum_k q_k$ and similarly for *r, s, t,* the roundness discriminant R1 is obtained as:

$$\mathrm{R1} = \frac{\left\| \mathrm{F}_1 \right\|^2}{\left\| \mathrm{F}_{-1} \right\|^2} = \frac{(q+r)^2 + (s-t)^2}{(q-r)^2 + (s+t)^2}$$

Because the q_k, r_k, s_k, t_k are formed by table lookup, the major computational cost in forming R1 is the need for 4N additions to accumulate the q, r, s, t. (There is also a constant cost of six addition/subtractions and four multiplications to form the squared magnitudes $\left\|\mathrm{F}_1\right\|^2$ and $\left\|\mathrm{F}_{-1}\right\|^2$, and a single division to form R1 from those squared magnitudes.)

3.5. Calculation of roundness discriminant R2

Feature R2 measures the total impact of the higher-order Fourier terms that cause the object contour to be "out of round" in various ways. To form R2, it is not necessary that we calculate any higher-order terms individually. Rather, we calculate the total power and subtract the contributions of the already-calculated circular terms F_1 and F_{-1} and the DC component F_0:

$$\sum_{|\mathrm{n}|>1} \left\|F_n\right\|^2 = \sum_n \left\|F_n\right\|^2 - \left\|F_0\right\|^2 - \left\|F_1\right\|^2 - \left\|F_{-1}\right\|^2$$

The power in a periodic signal can be expressed either as the mean squared magnitude of the signal itself, or as the mean squared magnitude of the terms of its Fourier series:

$$\tfrac{1}{N}\sum_{k=0}^{N} \left\| z_k \right\|^2 = \tfrac{1}{N}\sum_{k=0}^{N} \left({x_k}^2 + {y_k}^2 \right) = \tfrac{1}{N}\sum_k \left\| F_k \right\|^2$$

Closely related to power is the variance σ^2 of the signal, which is the power in its non-DC terms. Variance can be computed using the "sum of the squares minus the square of the sum":

$$\sigma^2 = \tfrac{1}{N}\sum_k \left\| z_k \right\|^2 - \left\| \tfrac{1}{N}\sum_k z_k \right\|^2$$

But the first part of the above expression is also the total power in all the F_n, and the second part is also the squared magnitude of the DC term F_0. Therefore we can obtain the combined power in all the Fourier terms higher than F_0 without any trigonometric operations, simply accumulating as we read the coordinates around the object contour and finally subtracting the squared sums, as follows:

$$\sigma^2 = \tfrac{1}{N}\sum_{k\neq 0} \left\|F_k\right\|^2 = \tfrac{1}{N}\sum_{k=0}^{N} \left({x_k}^2 + {y_k}^2 \right) - \left(\left(\tfrac{1}{N}\sum_{k=0}^{N} x_k \right)^2 + \left(\tfrac{1}{N}\sum_{k=0}^{N} y_k \right)^2 \right)$$

(Note that these are all integer operations, except for the division by N.) We can then form a normalized R2 from the variance and the values we already obtained for forming R1:

$$\mathrm{R2} = \frac{\sum_n \|F_n\|^2 - \|F_0\|^2 - \|F_1\|^2 - \|F_{-1}\|^2}{\sigma^2}$$

$$= 1.0 - \frac{\|F_1\|^2 + \|F_{-1}\|^2}{\sigma^2}$$

R2 entails calculation proportional to the contour length, as does R1. But since we have already formed $\|F_1\|^2$ and $\|F_{-1}\|^2$, R2 is computationally even lighter than R1 because the (x, y) coordinates are small integers. R2 requires only a constant three floating-point operations: adding the squared magnitudes of F_1 and F_{-1}, dividing by the variance, and subtracting from 1.0. The divisions by N shown in several preceding steps can be dropped by noting that R1 and R2 are both quotients in which those factors cancel.

Finally, we note that if desired, R2 could have been formed without explicitly forming F_1 and F_{-1}, using the partial results q, r, s, t that were formed originally for R1:

$$\mathrm{R2} = 1.0 - \frac{2(q^2 + r^2 + s^2 + t^2)}{\sigma^2}$$

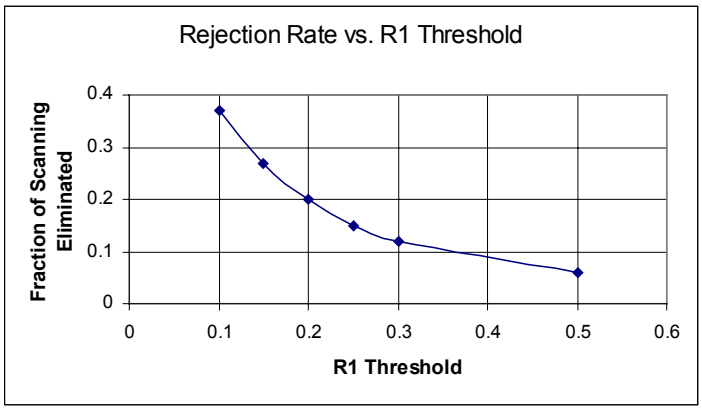
The outcome is that we form both features R1 and R2 at a cost of approximately 4N floating-point additions. Thus, we obtain well-motivated contour features with considerably better scaling behavior than is usually expected from transform-based image features.

The next concern then becomes to measure the effectiveness of those features in separating good from bad supernova candidates.

3.6. Results

For testing, the fast contour descriptor algorithm was implemented and incorporated into the existing supernova search pipeline software. The new scores R1 and R2 were added to the existing image feature set, without yet setting thresholds, but with diagnostic output to permit analyzing the distribution of scores. First, we examined the scores within the class of good candidates, using a validation set of 90 supernovae. For the first roundness feature, all of the supernovae in the validation set yielded R1 ≤ 0.08 (and frequently much closer to zero). For the second feature, nearly all (89 of 90) gave R2 < 0.1.

A second evaluation was performed on a much larger data set representing both classes (good and bad candidates), to determine appropriate threshold values for R1 and R2. The goal was to reduce the scanning load (remove the bad candidates) so far as possible without sacrificing any good candidates (possible supernovae) at all. The algorithm was run on a test set of approximately 35,000 subtracted images, of which 1094 would have been sent to a human for evaluation under the old parameters. The fraction of those 1094 which would be rejected (i.e. not sent for human scanning) under various settings of R1 was calculated; the results are shown below in Figure 7.



R1	Fraction of images rejected
< 0.10	37 %
< 0.15	27 %
< 0.20	20 %
< 0.25	15 %
< 0.30	12 %
< 0.50	6 %

Figure 7. Graph and table of image rejection rate vs. R1 threshold

A similar test was then performed for R2. On the basis of the combined results, thresholds for R1 and R2 were chosen at 0.1 and 0.15 respectively. At those settings, the number of images passing the thresholds and sent to human scanners decreased to 647 (vs. 1094 under the old criteria). This is a scanning load reduction of 41%, representing a savings of approximately 6 hours of manual scanning labor each day.

It was not immediately clear how well the calculated tradeoff would generalize, particularly with respect to good candidates. The new algorithm has now been incorporated into the production version of the supernova search pipeline and has been running for six months, during which the yield of actual supernovae has remained at the previous level (suggesting that the thresholds are correctly passing the good candidates), while the reduction in scanning load has remained approximately as anticipated.

4. CONCLUSIONS

In an image analysis pipeline for identifying supernovae in wide-field CCD images, we introduced two discriminants based on Fourier descriptors to eliminate objects with severely non-circular outer contours. The improved system has a dramatically (41%) reduced requirement for human inspection of images containing poor candidates, allowing astrophysicists to more readily discover actual supernovae.

Digital sky surveys can now generate very large amounts of image data that can support needle-in-a-haystack search applications such as the Nearby Supernova Factory. The amount of data to be sifted mandates that image analysis be highly automated, but also constrains the permissible algorithmic complexity, so that classical image analysis techniques can be applied only with great care.

We would expect the Fourier transform, being a decomposition in terms of circular functions, to be especially suited to the task of recognizing circular celestial objects, and this intuition is borne out by the improved discriminating power of the supernova feature set with the new Fourier-based features included. To meet the tight processing constraints of the high-volume application, we developed contour descriptors that can be calculated in linear time (proportional to the object circumference) rather than the $O(n \log n)$ time of a Fast Fourier Transform. We exploited the symmetry of the descriptors, the size of the objects, and the fact that pixel coordinates are integers to reduce the proportionality factor to about 4 floating point additions. The successful implementation of an image analysis technique ordinarily avoided in this type of application demonstrates the impact that an appropriate choice of parameters and algorithms may have on future sky surveys.

ACKNOWLEDGMENTS

This work was supported in part by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098, and by the Director, Office of Science, Office of High Energy Physics, of the U.S. Department of Energy under Contract No. DE-FG02-92ER40704, and by a grant from the Gordon & Betty Moore Foundation. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

1. S. Perlmutter, G. Aldering, G. Goldhaber, et al., "Measurements of Omega and Lambda from 42 High-Redshift Supernovae," *Astrophysical Journal* 517, 565-586 (1999).
2. A. G. Riess, A. V. Filippenko, et al., "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant," *Astrophysical Journal*, 1009-1038 (1998).
3. G. Aldering, G. Adam, P. Antilogus, P. Astier, R. Bacon, et al., "Overview of the Nearby Supernova Factory," in J. A. Tyson and S. Wolff, editors, Survey and Other Telescope Technologies and Discoveries, *Proceedings of the SPIE*, vol. 4836, 61-72 (Dec. 2002).
4. R. Romano, C. Aragon, and C. Ding, "Supernova Recognition Using Support Vector Machines," *Proceedings of the 5th International Conference of Machine Learning Applications*, Orlando, FL, 2006.
5. C. T. Zahn and R. Z. Roskies, "Fourier descriptors for plane closed curves," *IEEE Trans. Computers*, vol. 21, 269-281 (1972).
6. W. M. Wood-Vasey, "Rates and Progenitors of Type Ia Supernovae," PhD dissertation, University of California, Berkeley, 2004.
7. E. Persoon and K. Fu, "Shape discrimination using Fourier descriptors," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 7, 170-179 (1977).
8. R. Chellappa and R. Bagdazian, "Fourier coding of image boundaries," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, 102-105 (1984).
9. G. Granlund, "Fourier preprocessing for hand print character recognition," *IEEE Trans. Computers*, vol. 21, 195-201 (1972).
10. D. J. Lee, S. Antani, and L. R. Long, "Similarity Measurement Using Polygon Curve Representation and Fourier Descriptors for Shape-based Vertebral Image Retrieval," *Proceedings of the SPIE* (2003).
11. I. Kunttu, L. Lepisto, and A. Visa, "Efficient Fourier shape descriptor for industrial defect images using wavelets," *Optical Engineering*, vol. 44 (2005).
12. A. Fenske and H. Burkhardt, "Affine-invariant recognition of gray-scale objects by Fourier descriptors," *Proceedings of the SPIE -- Applications of Digital Image Processing XIV* (1991).
13. M. Syrjasuo and E. Donovan, "Using Relevance Feedback in Retrieving Auroral Images," *Proceedings of the 4th IASTED International Conference on Computational Intelligence*, Calgary, Alberta, Canada, 2005.
14. H. Freeman, "Computer processing of line drawing images," *Computer Surveys*, 57-97 (1974).
15. L. S. Davis, "Two-Dimensional Shape Representation," in *Handbook of Pattern Recognition and Image Processing*, T. Y. Young and K.-S. Fu, Eds. San Diego: Academic Press, pp. 233-245, 1986.