TOWARDS BETTER SEGMENTATION OF LARGE FLOATING POINT 3D ASTRONOMICAL DATA SETS: FIRST RESULTS

U. Moschini, P. Teeninga, M.H.F. Wilkinson
Johann Bernoulli Institute
University of Groningen
The Netherlands

N. Giese, D. Punzo, J.M. van der Hulst, S.C. Trager
Kapteyn Astronomical Institute
University of Groningen
The Netherlands

ABSTRACT

In any image segmentation task, noise must be separated from the actual information and the relevant pixels grouped into objects of interest, on which measures can later be applied. This should be done efficiently on large astronomical surveys with floating point datasets with resolution of the order of Gigapixels. We illustrate in this paper how the combination of two techniques presented in previous works can help in this task. We summarise the benefits and initial outcomes of combining a parallel algorithm to build max-trees of floating point data sets and a connected attribute filter that uses a statistical approach to identify structures due to noise and to perform segmentation on 3D radio cubes.

Index Terms— radio astronomy, attribute filter, max-tree

1. INTRODUCTION

Big data from space spans many different application fields. Examples are single-band or multi-band astronomical surveys of regions of the sky, astronomical radio surveys that often produce three-dimensional data volumes, and remote sensing images from satellites. Such data are considered big with a double meaning: their resolution (or simply the number of separate observations) is high and so is the bit depth of the data type they carry. The focus of this work is on the processing of radio astronomical spectral line data. Radio astronomy studies the radio emission from astronomical objects, which is not absorbed by dust clouds in galaxies nor affected by the Earth’s atmosphere. In particular radio spectral line emission of galaxies is captured as 3D volumes and contains important information for investigating the distribution and kinematics of gas in galaxies. Such radio cubes carry floating point values and have high resolution in the order of Gigavoxels.

A better segmentation of the objects means that better and more meaningful measures and statistics can be computed. With upcoming large surveys of the sky, the size of a 3D data cube will increase to the order of Terapixels. A manual extraction of possible sources by hand is not feasible any more and automatic segmentation methods are needed. Max-trees [2] are a powerful image representation that can help in this task. A max-tree is a tree structure that represents the hierarchy of the connected sets (components) of any image or volume. Each node corresponds to a connected set. Several attributes can be computed efficiently for every node and many filtering strategies based on them can be applied on the tree to perform segmentation of the objects of interest.

In the next section, to separate efficiently objects from noise in big volumes, the combination of a new parallel algorithm to build max-tree of floating point 3D volumes [3] with a connected statistical attribute filter [1] is introduced. An example from [1] is reported to illustrate the filter in the case of a two-dimensional astronomical image taken from the Sloan Digital Sky Survey DR7 [4]. In the other sections, the extension of the filter to 3D radio cubes is discussed together with the results compared with the output from SoFiA [5], a source finder used with this kind of data.

---

1Part of this work was funded by the Netherlands Organisation for Scientific Research (NWO) under project number 612.001.110 and by the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement nr. 291531.

---

Fig. 1: (a) Two merging galaxies. The segmentation performed (a) by SExtractor; (b) by the method in [1] with the background estimated by SExtractor and (c) by the method in [1] with our background estimate. The filament is identified.
2. BUILDING MAX-TREES OF FLOATING POINT DATA SETS IN PARALLEL

Among the state-of-the-art algorithms to build max-trees of floating point images, no shared memory parallel solution existed before the method proposed by Moschini, Meijster and Wilkinson [3]. The Moschini et al. method is used for building in parallel max-trees of images with high bit depths. It exploits both a bottom-up flooding and a top-down merging approach in two stages. For this reason, it was named the diplomatic algorithm. First, a pilot max-tree of a quantized version of the image is built in parallel by the flooding approach in [6], based on spatial partitioning. This tree is used as a support data structure in the second parallel stage, based on the top-down approach in [7], that builds the tree of the original image. The pixels are partitioned such that every thread works with all the pixel values that were mapped on a quantization level: the pilot max-tree allows for a correct node and attribute merging among the partitions handled by the threads. Note that the final tree represents all the original pixel values and no information is lost: the quantization is used just to support the second parallel stage.

3. ATTRIBUTE FILTER FOR ASTRONOMICAL OBJECT SEGMENTATION

The bivariate statistical filter [1] performs connected attribute filtering on the max-tree. It separates nodes likely to be noise from objects and assigns identifiers to group nodes that are considered as belonging to the same object. The attribute used is the ratio of the integrated power (flux) of a component over its local background. A statistical test is performed on this attribute, based on its distribution due to the estimated noise, as a function of the area of the component. This measure follows a $\chi^2$ distribution, with degrees of freedom equal the area of the component. Such distribution holds under the conditions that the noise is Gaussian, the pixel values are independent, and the intensity of every parent node has the same value as the noiseless image signal. The conditions are verified when the parent node is the root and the background is flat. Otherwise, the $\chi^2$ test is a worst case test. The local background of a node is assumed to be the parent node and its value is computed scaling the noise variance with the parent node intensity, since noise increases at higher intensities. The $\chi^2$ inverse cumulative distribution function defines a rejection boundary to identify noise nodes, for a given significance level and area. A more mathematical description is given in [1].

The filter was tested on images taken from the Sloan Digital Sky Survey DR7 [4]. On such images, the (noise) background and its variance are estimated looking for regions devoid of objects. The background value is subtracted from the image and values below zero are truncated. The diplomatic parallel algorithm builds the max-tree of the truncated image, nodes are marked as significant and objects are identified. Object identifiers are then moved up in the tree hierarchy to reduce the probability of noise pixels to be included in the segmentation. The image in Fig. 1 shows an example of the improved segmentation that was achieved in [1] with respect to SExtractor [8], the state-of-the-art application to segment astronomical objects in two-dimensional images. We refer to the work in [1] for considerations about the rejection boundary when smoothing is applied.

4. SEGMENTATION OF 3D RADIO CUBES

Radio spectral line data have very different characteristics from the images in the Sloan catalogue. The bivariate statistical filter was customized with settings more suitable to this kind of volumes. For these data, the background value is equal to 0. All negative real values of the cube are due to noise, whereas sources have positive values. The noise is Gaussian, and therefore it follows a $\chi^2$ distribution. Its standard deviation can be estimated using the Median Absolute Deviation. Since in radio data the noise is independent of the local background, the background variance is not adapted to the level of the parent node when the power attribute is used in the filtering process. An optimal factor to move the object identifiers up in the tree hierarchy was worked out after trial and error. The parallel max-tree algorithm detailed in the previous section was adapted to support three-dimensional volumes, and the bivariate filter was also parallelised wherever it was possible.
The volumes that are examined in the rest of the paper are data cubes of the 21 cm neutral hydrogen (HI) emission of modelled galaxies as they would be observed by the WSRT (Westerbork Synthesis Radio Telescope, courtesy of P. Serra). The cubes show velocities versus projected spatial dimensions on the sky. The output of the segmentation process is a volume with object identifiers, commonly referred to as mask by astronomers.

Many of the object sources in such radio cubes are just barely above the noise level: those are difficult targets for state-of-the-art source finder applications. We compare our new algorithm to SoFiA [5], a source finder that combines different 3D source finding algorithms in one application. In SoFiA we chose to use the Smooth and Clip method, which allows us to search for sources on different spatial and velocity scales by smoothing the cube with 3D kernels provided by the user. At each resolution, volume pixels with flux above a fixed sigma-to-noise threshold are identified as signal and added to the mask. This method corresponds to the most common technique used in HI source finding.

The results of the proposed method have been promising since the early tests. By simply giving the background value and the standard deviation of the noise as input to the bivariate filter, our method was able to identify objects. Fig. 2a shows a section of a larger radio volume containing a ring-structured galaxy in a 60x60x70 floating-point cube, with velocity as the third dimension. In Fig. 2b, noise has been added and segmentation performed on the image. Fig. 2c shows the segmentation computed by SoFiA. Fig. 2d shows the segmentation of the bivariate attribute filter: the outer boundary of the ring adheres more to the original structure, while some extremely faint structures within the centre of the ring are wrongly considered not part of the object. Fig. 3 shows six selected frames of the cube in Fig. 2. The mask obtained with the proposed method is contained in the galactic structure, whereas the SoFiA mask includes the fainter regions as well. The smoothing applied by SoFiA has the positive effect of helping to find the faint outer parts of the objects. The absence of faint details in the mask of the proposed method is a problem that must be addressed. However, there already are methods used by astronomers that allow the mask to grow to circumvent the possible loss of object details. These methods could be implemented to improve the segmentation. The drawback of growing the mask is the possibility of including noise voxels in the segmentation. This however should be preferred over losing information on the faint parts of the galaxy. On simulated data, it is possible to study the features of the parts of objects missed by the algorithm. Extra information can be included in the max-tree structure: the mask can be grown accessing the parent nodes of the identified structure and make them part of the object according to relevant attributes, such as volume, total flux, or maximum flux.

Fig. 4 represents all the sources found in a large cube, henceforth called the WSRT cube, with 360x360x1464 resolution. The image is a moment-0 image of the identified sources. It is computed by adding up all the flux in the detected objects along the velocity axis. The comparison between SoFiA and the proposed method again shows the above-mentioned problem, in which the fainter parts of objects are not captured by the proposed method. However, there are only few sources in the SoFiA mask that the proposed method does not detect. Interestingly, there is one point-like source that is only identified by the new method (see red circle in Fig. 4). Fainter parts are probably detected by SoFiA due to the fact that method first smooths the image. Smoothing has the positive effect of helping to find fainter structures, with the drawback of removing some actual, compact objects. Results from [1] show that use of...
smoothing increases the sensitivity of the connected filter method significantly. This suggests that there is potential for developing the proposed method further. Using the flexibility given by the max-tree structure, we believe that the segmentation can improve significantly by studying the features of these sources and computing them in the max-tree.

5. PERFORMANCE TESTING

The method described in this paper was implemented in C with POSIX Threads and OpenMP API. Timings were performed on a shared-memory Dell R815 Rack Server with four 16-core AMD Opteron processors and RAM memory of 512 GB. It has 32 floating point units, each shared by a pair of cores. Previous experiments [3] on a 1024 x 1024 (spatial) x 1080 (temporal) cube of a field of radio sources taken from the Low-Frequency Array (LOFAR) telescope in The Netherlands showed that its max-tree is built in 3 minutes and 30 seconds on 64 threads compared to 1 hour and 22 minutes with the fastest sequential method for floating point volumes [7]. The algorithm goes from 11 minutes on one thread to 2 minutes and 10 seconds on 16 threads on the WSRT cube. The minimum timing of a series of ten runs of the algorithm was taken. No improvement appears after 16 threads. That is due to both sub-optimal load balance (the first thread must handle all the 0-value voxels that are 50% of all the values after background subtraction) and some sequential parts of the method. The time spent for the bivariate filter is 6% and 28% of the overall execution time with one and sixteen threads, respectively. The sequential program SoFiA takes about 33 minutes. This value varies with the number of smoothing steps at different standard deviations that SoFiA performs.

The maximum memory occupied by the method presented is about 25 times the size of the cube: that includes the pilot max-tree and the final tree structures. On the WSRT cube (725MB), this is about 18GB. Afterwards the pilot tree memory is deallocated, but 5GB are reserved for the bivariate filter structures. The maximum memory needed by SoFiA is four times smaller: it is equal to about six times the size of the cube (4GB for the WSRT cube), which includes the size of the cube, the current smoothed cube and the mask holding the results.

6. CONCLUSIONS

In this paper, the first limited but promising results of object segmentation in large radio cubes are presented. These results show that the use of max-trees and the bivariate filter should be explored further. A better understanding of the segmentation process and the noise model is needed. The effect of smoothing the cube should be tested as well, together with voxel connectivity. Checking the features of the radio sources missed and including in the max-tree the attributes that describe them will be important to improve the output of the bivariate filter. The ultimate goal is to segment large cubes directly without splitting them or the objects in them into smaller sections. We are also addressing the load-balance issue by altering the max-tree building algorithm for these highly skewed intensity distributions. Better segmentation should lead also to a better classification of radio sources that could be achieved by computing a pattern spectrum of meaningful attributes relative to every correctly segmented object.

7. REFERENCES


