

Needles In Haystacks

Rohan Chabukswar



Engineering Physics
Indian Institute of Technology Bombay

April 29, 2008



Outline

- 1 Introduction
 - Quantum Chromodynamics and Strong Interaction
 - The Problem
 - Wavelet Analysis
 - Discrete Wavelet Transform
 - Preliminary Analysis
- 2 The Search Algorithm
 - Advantages Of The Algorithm
 - Complexity Analysis
 - Simulation Results
- 3 Further Work
 - Limitations Of The Algorithm
 - Further Work
- 4 Bibliography



The Strong Force

- The strong force is used to represent the interactions at the most basic level.
- It is a fundamental force mediated by gluons acting upon quarks, antiquarks and gluons themselves.
- It is detailed by the theory of Quantum Chromodynamics.



Particles Involved

- Quarks are one of the two basic constituents of matter. There are six different flavors of quarks — up, down, charm, strange, top and bottom. Antiquarks are the antiparticles of quarks.
- To uphold Pauli's Exclusion Principle, there has to exist another internal quantum number for the quarks. This quantum number, given the whimsical name "color", is the charge involved in the gauge theory of quantum chromodynamics (QCD).
- The gluon is the gauge boson of QCD. QCD being a non-Abelian gauge theory the force carrying particles are also colored — gluons come in eight colors.



Properties of QCD

Quantum Chromodynamics displays two peculiar properties:

- Confinement, which means that the force between quarks does not diminish as they are separated. Because of this, it would take an infinite amount of energy to separate two quarks. They are forever bound into hadrons such as the proton and the neutron.
- Asymptotic freedom, which means that in very high-energy reactions, quarks and gluons interact very weakly.



Asymptotic Freedom

- Asymptotic Freedom is a direct consequence of *antiscreening*, an effect opposite of the screening found in Quantum Electrodynamics.
- As QCD is non-Abelian, the force carrying particles are themselves colored.
- The net effect of polarization of virtual gluons in the vacuum is not to screen the field, but to augment it.
- Getting closer to the quark diminishes the antiscreening effect of the surrounding virtual gluons, and this weakens the effective charge.
- This implies that within nucleons, quarks move mostly as free, non-interacting particles, which is termed “Asymptotic Freedom”.



Confinement

- A consequence of antiscreening is that the color force experienced by quarks remains constant regardless of their distance from each other (after a certain point).
- When two quarks become separated, as in particle accelerator collisions, at some point it becomes energetically more favorable for a new quark-antiquark pair to be created spontaneously in vacuum, than to allow the quarks to separate further.
- Quarks can never be separated, leading to “Confinement”. In such an interaction, instead of individual quarks, many color-neutral particles are detected.
- This process is called *hadronization*, and is one of the least understood processes in particle physics.



Jet Events

- The tight cone of particles created by the hadronization of a single quark is called a jet.
- These jets must be measured in a particle detector and studied in order to determine the properties of the original quark.
- Figure shows the simulation of a jet event.



Jet Event Depiction

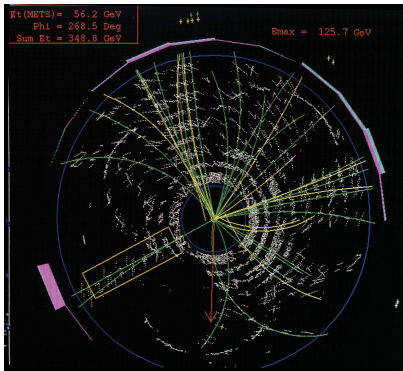


Figure: Top quark and anti-top quark pair decaying into jets visible as collimated collections of particle tracks in CDF detector at Tevatron.



Simulation

- The properties of the quarks and other elementary particles can be accrued by studying the hadronization jets.
- These are simulated using PYTHIA, which is a program for the generation of high-energy physics events, including collisions at high energies between elementary particles.
- It contains theory and models for a number of physics aspects, including hard and soft interactions, parton distributions, initial- and final-state parton showers, multiple interactions, fragmentation and decay.



Data Analysis

- Each event generates particles in thousands, and hundreds of such events need to be analyzed. The volumes of data involved are huge.
- These volumes will increase many times, when actual jet events are observed instead of simulations.
- Analyzing these massive amounts of data, even for something as simple as figuring out direction of the jets, will take years for normal analysis.



Attempt At A Solution

- The project attempts at using Multiresolution Analysis using Wavelets, to find the proverbial *needles in the haystack*.
- This is achieved by implementing a kind of Binary Search using the outputs of a normal Multiresolution Analysis.



What is A Wavelet Transform?

- A wavelet is a kind of mathematical function used to divide a given function into different frequency components and study each component with a resolution that matches its scale.
- In formal terms, this transform is a wavelet series representation of a square-integrable function with respect to either a complete, orthonormal set of basis functions, or an overcomplete set of frame of a vector space (also known as a Riesz basis), for the Hilbert space of square integrable functions.
- The wavelets are scaled and translated copies (known as “daughter wavelets”) of a finite-length or fast-decaying oscillating waveform (known as the “mother wavelet”).



Advantages over Fourier Transform

Wavelet transforms have advantages over traditional Fourier transforms for

- representing functions that have discontinuities and sharp peaks
- for accurately deconstructing and reconstructing finite, non-periodic and/or non-stationary signals.



Why Wavelets?

- Multiresolution Analysis or multiscale modeling is used in solving physical problems which have important features at multiple scales, particularly multiple spatial or temporal scales, which can only be achieved by using bases which have finite spread in both time and frequency domains.
- This dual localization achievable by wavelet analysis renders many functions and operators sparse to a high accuracy, which makes a large class of computations very fast in the wavelet domain.
- In our example, we do not know the extent or the strengths of the jets a priori.
- There is no reason for choosing a particular wavelet at this stage. The Haar wavelet has been chosen for this analysis because of its conceptual simplicity and easy visualization.



Why Discrete?

- In continuous wavelet transforms, a given signal of finite energy is projected on a continuous family of frequency bands.
- It is computationally impossible to analyze a signal using all wavelet coefficients.
- It is sufficient to pick a discrete subset of the upper halfplane to be able to reconstruct a signal from the corresponding wavelet coefficients.
- One such system is the affine system for some real parameters $a > 1$, $b > 0$.

$$\psi_{m,n}(t) = a^{-m/2} \psi(a^{-m}t - nb). \quad (1)$$



Wavelet Transform by Hardware

- The signal is passed through a quadrature mirror filter.
- The filter outputs are then downsampled by 2.
- This decomposition is repeated, the approximation coefficients decomposed and down-sampled again.

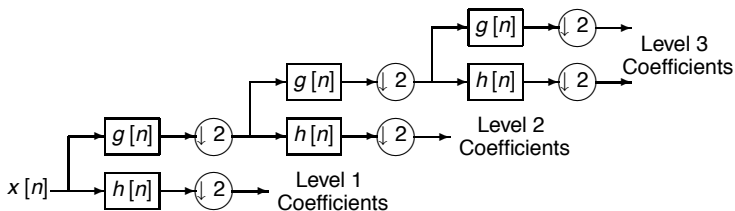


Figure: Level 3 filter bank



DWT Algorithm

- The Discrete Wavelet Transform (DWT) as an algorithm consists of applying a wavelet coefficient matrix hierarchically, first to the full data, then to the smooth part of half the length, then the smooth-smooth part of quarter the length, and so on.
- The Haar wavelet can be considered as replacing every two consecutive data values by their average and difference (neglecting a numerical factor), and then rearranging the resulting data to have all the averages first and the differences later.
- To invert the transform, the process is reversed, with the inverse of the matrix being used.



Wavelet Transform in Higher Dimensions

- For an essentially one-dimensional wavelet, the wavelet transform of a d -dimensional array is obtained by transforming the array sequentially on its first index for all values of its other indices, then on its second, and so on.
- Each transformation corresponds to a multiplication by an orthogonal matrix, hence by matrix associativity, the result is independent of the order of the indices.
- The levels of analysis and the indices can be interchanged without affecting the input, i.e., one particular index can be analyzed to the highest level before the next one.
- While this is not true for essentially multidimensional wavelets, it is not a very great concern for this application, since only one-dimensional wavelets will be used.



2D Wavelet Transform

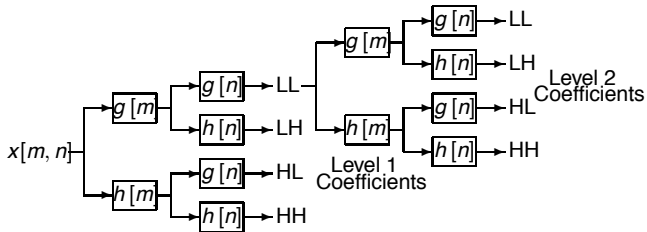


Figure: Two dimensional wavelet transform



Preliminary Analysis

- The first step of the analysis is to choose the optimum resolution at which to search for the jets.
- So, we have to carry out two preliminary steps:
 - 1 Convert the data to spherical (r, θ, ϕ) co-ordinates
 - 2 Bin the data into appropriately sized parts and get a count of the number of particles per unit solid angle in each part, like a two dimensional histogram.



Preliminary Analysis

- The initial resolution of the histogram is arbitrarily taken to be 1024 in both θ and ϕ .
- This does not divide the space into areas subtending equal solid angles
- While creating the histogram, the actual number of particles in that area has to be divided by the solid angle it subtends at the center.
- If the bin is between θ_1 and θ_2 and has a width of $\Delta\phi$, the solid angle subtended is

$$\Delta\Omega = (\cos\theta_1 - \cos\theta_2) \cdot \Delta\phi. \quad (2)$$



Preliminary Analysis

- The optimum resolution can be taken to be the one where the histogram shows the most uneven distribution, or the most entropy.
- First we carry out a two-dimensional wavelet transform on the histogram, down to the lowest level.
- This will give us, for a resolution of 1024 in each direction, 1 average component (LL), and 9 different levels with three detail components (LH , HL and HH) each.
- The energy is defined as the average of the squares of all the detail components.
- The energy is calculated for each component, and the level with the maximum energy is outputted for each event.



Results of Preliminary Analysis

- 96% of the events showed a maximum energy at level 2, i.e., at a resolution of 4 in each direction, which implies that an optimal resolution of 4 should be used at each step of the analysis.
- Once the area with the maximum particles is identified, the next step can again be undertaken with a further resolution of 4, *in that area*.
- This recursion would have to be implemented at least 4 times, to get the position of the jets accurate to degrees.



Results of Preliminary Analysis

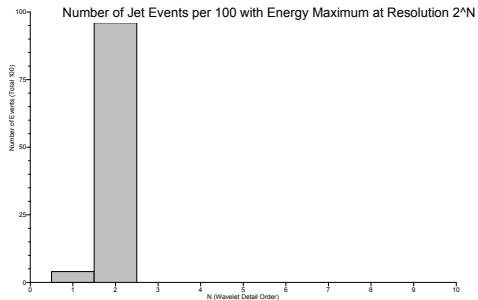


Figure: Optimal Resolutions in 100 Events.



The Problem

- The histogram of the data can be considered as an image, where the gray value of each pixel is either the number of particles observed near those co-ordinates, or the energy content near those co-ordinates
- The random white gaussian noise can have widely distributed intensities.
- This noise represents itself as randomly occurring white and black pixels, commonly referred to as intensity spikes, speckle or salt and pepper noise.
- Usual and effective noise reduction method for this type of noise involves the usage of a 2-dimensional median filter, or alpha-quantile averaging.
- Such filters have a significant time complexity, which in fact will govern the time complexity of the entire algorithm.



The Idea

- The key idea here is to use the the outputs of the filters themselves to implement a sort of binary search on the data.
- A binary search algorithm (or binary chop) is a technique which narrows the search by a factor of two each time, and finds the target value — a dichotomic Divide and Conquer.
- It is usually used for finding a particular value in a sorted list, its combination with a multiresolution analysis can be used to detect peaks in the image.



The Algorithm

- The main difference between the algorithm and a normal MRA is that instead of keeping the differences at each tier of the analysis, the difference coefficients are discarded but the smooth components are saved instead.
- The output of the multiresolution analysis, which is the first step in the algorithm, can be represented as a tree — each pixel in a lower tier corresponds to four pixels in the tier immediately above it. Thus the tree is a collection of successively poorer resolutions of the original image.
- It is necessary that the resolution in each direction be a power of two. If the resolution cannot be chosen for any reason, it should be appropriately padded by zeros upto the next power of two.



The Algorithm

- Since we are considering only two-jet events, theoretically for conservation of momentum, the two jets have to be antipodal. If one jet is assumed to be at (θ, ϕ) , the antipodal jet will occur at $(\pi - \theta, \phi \pm \pi)$.
- For a resolution of 2×2 , where the (θ, ϕ) values can only be $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$, the antipodal pixel is given by $(1 - \theta, 1 - \phi)$.



The Algorithm (cont.)

- Thus for the first step the pixel with the highest value is chosen, along with its antipodal pixel. The further steps are run iteratively on each of the two pixels independently.
- For each step in the further analysis, chosen pixel is taken, and the four pixels it corresponds to in the next higher tier are considered. The pixel with the highest gray value in this set is the chosen pixel for the next iteration.
- The resolution of the location of the peak is successively doubled, down to the resolution of the original image.



Flowchart

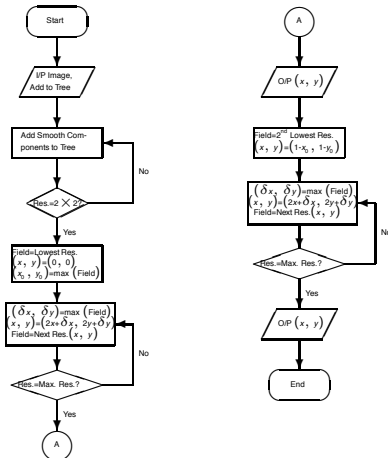


Figure: Algorithm Flowchart



Advantages Of The Algorithm

- The number of iterations is fixed for a particular maximum resolution. For an image of size $2^N \times 2^N$, the number of iterations is $(N - 1)$ for each jet.
- In each iteration, there are only a few comparisons which need to be undertaken, to find which the pixel with the maximum gray value is.
- As the wavelet transform itself involves a low pass filter, the noise is automatically taken care of at each resolution.
- The algorithm was found to be extremely resistant to noise with signal-to-noise ratio (SNR) as low as 4.



Complexity Analysis

- The transform implemented in the algorithm is runs with a complexity same as that of a 2-dimensional wavelet transform, $\mathcal{O}(N^2)$.
- The complexity of the search algorithm does not depend upon the number of particles observed, but on the resolution required.
- The search part of the algorithm implements a constant number of comparisons in each iteration, $2(N - 1)$ for an image of size $2^N \times 2^N$.
- Thus the complexity of the search is $\mathcal{O}(\ln N)$.



Complexity Analysis

- The complexity of the k_t jet search algorithm is $\mathcal{O}(n^3)$, where n is the number of particles.
- Using two-dimensional nearest neighbor location, the complexity can be reduced to $\mathcal{O}(n \ln n)$.
- The total complexity of the proposed algorithm will be governed by the wavelet transform and will go as $\mathcal{O}(N^2)$. In situations where the wavelet transform can be computed separately prior to analysis the search algorithm will only take $\mathcal{O}(\ln N)$ computations.



Simulation

- The algorithm could not be tested out on actual PYTHIA data, as no results from other algorithms on the *same* data were available for comparison of results or time complexity.
- To test whether the jets were actually found, a grayscale image was generated which resembles two antipodal jets with speckle noise, and run the algorithm on this image.
- The results of the algorithm were then compared with the parameters which generated the image.
- This gives us an advantage in the sense that we can test the limits of the algorithm in terms of noise resistance and accuracy, independently of whether that noise would ever be present in a real image.



Image Generation

- The data was generated as a 1024×1024 image. Each pixel can have a gray value ranging from 0 through 255.
- The jet is a gaussian, $192 \pm (64 \cdot \text{random})$ in height, $7.5 \pm (2.5 \cdot \text{random})$ in width, located at a random point $(\theta_1, \phi_1) = (1024 \cdot \text{random}, 1024 \cdot \text{random})$
- Though the other jet should theoretically be antipodal, practically, it might deviate from the exact antipode due to instrumental and other errors.
- Thus in the simulation, the antipodal jet is generated at the approximately antipodal point $(\theta_2, \phi_2) = (1024 - \theta_1 \pm 64 \cdot \text{random}, 512 + \phi_1 \pm 64 \cdot \text{random})$.
- There is also random white gaussian speckle noise of signal-to-noise ratio around 4, i.e. maximum strength is 25% of the average strength 192, or 48.



Sample Image



Figure: An Example Simulation Image Generated, 1024×1024 .



Simulation Results

- After a negligible time interval, the algorithm calculated the jet locations to be (623, 614) and (373, 88), which is very accurate for such a low SNR.
- These points were the highest in the jet, deviating from the actual center due to the random noise.
- This simulation was run for many such images, and all of the results were within ± 3 of the actual center, and a large fraction were within ± 2 .



Sample Image Tree

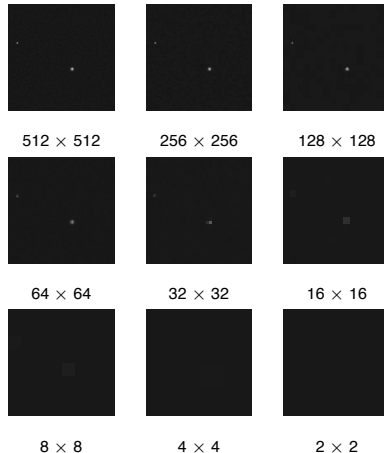


Table: The Multiresolution Tree Generated By The Search Algorithm.



Limitations Of The Algorithm

- One serious limitation of the algorithm is the fact that it does not delineate the jet completely or count the number of particles in the jet or their energy. This is rendered very difficult by the use of polar co-ordinates (θ, ϕ) .
- We can assume that the exact equation of the jet is given by gray value = $e^{-r^2} \cdot \sin r dr d\eta$.
- Using standard equations of spherical trigonometry, the circle $r = 15^\circ$ can be transformed into (θ, ϕ) coordinates.
- These curves will be circles when viewed on a sphere, but in (θ, ϕ) they do not resemble circles, except perhaps when $\theta_0 = 90^\circ$. Near $\theta_0 = 0^\circ$, they do not even close.
- This means that the boundaries cannot be defined easily for an arbitrary jet, which excludes the usual boundary finding algorithms like the Hough Transform.



Circles On A Sphere

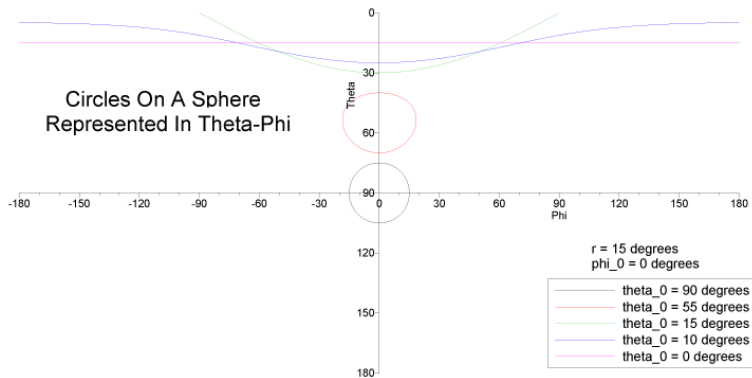


Figure: Circles On A Sphere Represented In (θ, ϕ) .



A Rough Approach

- As long as the jet extent is less than the current pixel size, the value of the maximum pixel, which is the average of all the (maximum resolution) pixels containing the jet, will continue to increase by an approximate factor of 4.
- As soon as the jet extent exceeds one pixel, the average will remain approximately constant.
- This will give the extent of the jet to the nearest power of two.
- Further analysis can be carried out at a finer resolution to find the jet boundary.



Simulation Results

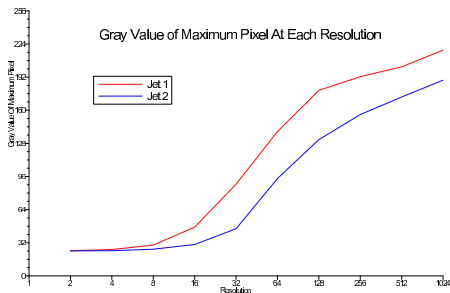


Figure: Gray Values Of Maximum-Valued Pixels At Each Resolution.






A Rough Approach

- Currently a brute force algorithm to find the boundary, which finds the pixels with a specific strength nearest to the peak, has not been implemented.
- This algorithm will be prone to noise, but the native resolution found out in stage one can be used here to apply a low pass filter of size 4×4 .
- Further work will hopefully come up with a good algorithm to find the boundaries of the jets. It is then easy to find the number of particles in each jet, along with the total energy.
- Another direction for work will be using a wavelet other than the one used, say DAUB4 or DAUB20, or a different family like Morlet. The use of Biorthogonal and Continuous Wavelet Transforms might also be explored, though they will almost certainly require extensive computation.



References

-  Graps, Amara,
An Introduction to Wavelets,
IEEE Computational Science and Engineering, vol. 2,
num. 2, 1995.
-  *Numerical Recipes in C++ - The Art of Scientific
Computing*, 2nd Ed.,
William H. Press, Saul A Teukolsky, William T. Vatterling,
Brian P. Flannery,
Cambridge University Press, 2002.
-  HyperPhysics Concepts
[http://hyperphysics.phy-astr.gsu.edu/
hbase/particles/parcon.html](http://hyperphysics.phy-astr.gsu.edu/hbase/particles/parcon.html).



References (cont.)



Wikipedia

[http://en.wikipedia.org/.](http://en.wikipedia.org/)



PYTHIA Website

<http://www.thep.lu.se/~torbjorn/Pythia.html>



I would like to thank Prof. Raghav Varma and Prof. Vikram M. Gadre for giving me the opportunity to undertake my B. Tech. Project under their esteemed guidance and in a subject concurrent with my academic interests. I also thank Prof. Basanta Kumar Nandi for acquiring the data and explaining its format in detail. This project was extremely stimulating, and the knowledge and experience gained by undertaking it will be useful in my future life.

