Multiresolution Jet Reconstruction with FFTJet

Igor Volobouev
Texas Tech University, Department of Physics, Box 41051, Lubbock, Texas, USA 79409
E-mail: i.volobouev@ttu.edu

Abstract. This article introduces a new jet reconstruction method for particle physics. Event energy flow analysis is performed in two stages. Pattern recognition stage is applied first, utilizing multiresolution filtering techniques in the Fourier domain. Jet energy determination follows, conditional upon the choice of signal topology. The resulting algorithm is global, efficient, collinear and infrared safe, and allows the user to identify and avoid the event topology bifurcation points when energy reconstruction is performed. The performance of the new approach is better than that of commonly used jet algorithms such as $k_T$, anti-$k_T$, and seedless cone. An open source implementation is available at http://projects.hepforge.org/fftjet/.

1. Introduction
The necessity of improving pattern recognition capabilities of jet reconstruction algorithms has been recognized both in the context of multijet, high occupancy events and in the cases when massive particles decaying hadronically via electroweak interaction ($W$ and $Z$ bosons, top quarks) are sufficiently boosted so that the energy flow of their decay products can not be easily partitioned into well-separated jets. Due to software availability, LHC-targeted studies of particle physics processes of this kind utilized predominantly sequential recombination techniques [1, 2, 3, 4, 5]. However, sequential recombination algorithms are inherently more sensitive to fluctuations in the event energy flow than global methods, and clustering stability properties of sequential recombination algorithms are not fully understood. Changes in the recombination sequence at early clustering stages are not guaranteed to dissipate and may affect clusters obtained at later stages — this artifact led to the introduction of such concepts as “passive” and “active” jet areas which turn out to be distinct [6]. On the other hand, commonly used global reconstructions algorithms — such as seedless iterative cone and its various seeded variations — suffer from another important deficiency inherent in the cone-based jet reconstruction. This deficiency manifests itself as the pattern recognition ambiguity illustrated in Figure 1. Two energy deposits of similar magnitude separated by a distance larger than $R$ but smaller than $2R$ produce three stable cone centers whose positions are shown with the arrows at the bottom of the figure. In various implementations of cone-based jet reconstruction procedures, this problem is usually addressed by the “split-merge” stage which happens after the stable cone locations are determined. During this stage, jets are merged if the energy which falls into the common region exceeds a predefined fraction of the energy of the jet with smaller magnitude. Even if the search for stable cones is performed in the infrared and collinear safe manner, the outcome of the split-merge stage is often unstable because the decision on whether to merge the two jets depends on the minute details of the energy deposition structure. Seeded
cone algorithms suffer from a variety of additional problems which lead to collinear and infrared instabilities of the reconstructed jet configurations [7].

2. Improved Pattern Recognition

The jet reconstruction method described here is based on a global technique which does not suffer from the problems mentioned in the previous section. This method was originally inspired by Refs. [8] and [9] and initially proposed in [10]. The study by Cheng [8] establishes an important connection between the iterative cone algorithm and kernel density estimation (KDE) [16]. Cheng proves that the locations of stable cone centers correspond to modes (peaks) of the energy density built in the $\eta$-$\varphi$ space using KDE with the Epanechnikov kernel. That is, all such centers can be found by convolving the empirical energy density

$$\rho_{\text{emp}}(\eta, \varphi) = \sum_i \varepsilon_i \delta^2(\eta - \eta_i, \varphi - \varphi_i)$$

with the function

$$\text{Epanechnikov}(\eta, \varphi) = \begin{cases} 1 - (\varphi^2 + \eta^2)/R^2, & \varphi^2 + \eta^2 < R^2 \\ 0, & \varphi^2 + \eta^2 \geq R^2 \end{cases}$$

and then finding all local maxima of the convolution. Here, $\varepsilon$ is an energy variable which can stand for transverse momentum or transverse energy of a particle, calorimeter tower, etc. $\eta$ and $\varphi$ define the direction of the energy deposit as seen from the interaction point (depending on the choice of $\varepsilon$, $\eta$ can stand for either rapidity or pseudorapidity, while $\varphi$ is the azimuthal angle). $\delta^2(\eta, \varphi)$ is the two-dimensional delta function. The sum in the expression for $\rho_{\text{emp}}$ is performed over all energy deposits expected to be jet constituents (typically, leptons and photons produced in the hard scattering process are excluded). $R$ in the Epanechnikov kernel corresponds to the cone radius in the $\eta$-$\varphi$ space.

Following Cheng, the reason for the pattern recognition ambiguity in the cone-based jet reconstruction can be explained as follows: it occurs because the sum of two Epanechnikov kernels placed at the locations of the deposits has a spurious peak in the middle, as shown in Figure 2 (left). Fortunately, there is a variety of kernels which do not suffer from this problem. In particular, the Gaussian kernel produces either two (narrow kernel) or one (wide kernel) peaks, as shown. Even though one still has to address the question of choosing the kernel width, the Gaussian kernel has a very important advantage: the whole split-merge stage is no longer necessary.

1 Similar clustering and pattern recognition schemes have been introduced in various disciplines [11, 12, 13, 14].

2 The iterative cone algorithm is known as the “mean shift” algorithm [15] in the pattern recognition literature.
3. Multiresolution Description of the Event Energy Flow

An intelligent choice of the kernel width (or the \( R \) parameter in the \( k_T \) and cone algorithms) can not be performed until some assumptions are made about the expected jet shapes. In fact, optimal choice will be different for different signals. For example, a data analysis which searches for high energy dijet events with two well-separated jets is likely to make very different assumptions about jets from a data analysis which looks for \( t\bar{t} \) events in the all-hadronic, 6-jet mode. Moreover, the optimal width is not necessarily the same for every jet in an event, as low momentum jets tend to have wider angular profiles, especially in the presence of magnetic field. Because of this, it is instructive to look at the jet structure of an event using a variety of kernel width, cone radius, etc. choices — all such parameters will be referred to by a general term “resolution scale” for the remainder of this paper. In the limit of continuous resolution scale one arrives at the description of event energy flow known as “mode tree” in the nonparametric statistics literature or “scale space image representation” in the computer vision theory.

Energy flow peaks found at different resolution scales in the \( \eta-\phi \) space are used as “preclusters” which determine possible initial jet locations. In order to permit an effective use of the information contained in the multiresolution energy flow picture, preclusters are combined into a single hierarchical structure called “clustering tree”. This tree is built using a distance function\(^3\). A precluster found at some resolution scale \( s_i \) is assigned a parent from the previous (larger) resolution scale \( s_{i-1} \) as follows: the distance between the precluster at the scale \( s_i \) is calculated to all preclusters at the scale \( s_{i-1} \). The precluster at the scale \( s_{i-1} \) with the smallest such distance becomes the parent. This agglomeration strategy follows the approach of Ref. [9] and has the advantage that the obtained tree structure can also be utilized as a balltree [17]. A simple tree-like structure is obtained if the pattern recognition kernel is chosen in such a manner that the number of preclusters does not increase when the resolution scale increases, no matter how the event energy flow looks like. Although the 2-d Gaussian kernel is known to violate this condition on some rare occasions [18], it appears to perform very well for all practical purposes. An example clustering tree image can be found in Ref. [19].

Once the parent/daughter relationships are established between preclusters found at different resolution scales, dependence of various precluster characteristics on the scale parameter can be analyzed. Perhaps, the most useful such characteristic is the speed with which the precluster \( \eta-\phi \) location changes as a function of scale (scale space drift). The drift speed becomes relatively high when two preclusters are about to merge. This indicates that the resolution scale is close to a bifurcation point — the scale at which two smaller preclusters form a bigger one. Near the bifurcation point the locations of the affected preclusters become very sensitive to small changes

\(^3\) A simple distance is given by \( d = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2} \), with some constant parameters \( h_\eta \) and \( h_\phi \). More complicated functions which take into account resolution scales and peak magnitudes can be constructed as well.
in the event energy flow. This problem results in an increased uncertainty of jet energy and direction determination. The bifurcation points can be avoided by detecting them in the scale space with the clustering tree, and by using slightly modified resolution scales in case such points are found.

4. Jet Reconstruction Strategies
The information contained in the multiresolution energy flow description permits multiple strategies for jet reconstruction. The common technique of choosing a single best overall resolution scale according to some optimization criterion can be easily improved upon by augmenting it with the bifurcation point detection method. A number of less traditional techniques can be employed as well:

- The jet configuration can be selected according to a clustering stability criterion. In the multiresolution picture, it is natural to assume that jet configuration which persists through a wide range of pattern recognition scales represents a salient feature of the event energy flow. Therefore, clustering stability can be expressed in terms of the configuration “lifetime” in the scale space. A reasonable lifetime definition is \( J^\alpha \log(s_{\text{max}}/s_{\text{min}}) \). Here, \( J \) is the number of preclusters (perhaps, satisfying some imposed selection criteria) present in the clustering tree at each resolution scale \( s_{\text{min}} \leq s \leq s_{\text{max}} \). Parameter \( \alpha \) can be chosen empirically depending on the process under study (normally \( 0 < \alpha < 1 \)).

- A global resolution scale can be chosen on event-by-event basis in such a manner that the number of preclusters found in the clustering tree for that scale corresponds to the number of jets expected in the signal. The scale space lifetime of the expected jet configuration should be checked and events with unsatisfactory lifetimes should be discarded.

- The resolution scale can be adapted for each jet separately, in a manner that is consistent with the expected event topology. For example, if the process under study is expected to produce \( M \) jets at the leading order perturbation theory, the jet configuration can be chosen according to the principle by which the minimum transverse energy among \( M \) hardest jets is maximized over all possible cluster sets logically permitted by the clustering tree structure (i.e., over all sets of \( M \) preclusters formed in such a way that no set member can be a clustering tree ancestor of another set member).

- Scale space differential blob detectors [23] can potentially be used for jet identification.

- Nontrivial clustering patterns can be identified in the signal, and similar patterns can be searched for in the clustering tree. For example, boosted resonances, such as \( W \) bosons or top quarks, are expected to produce one wide jet at higher resolution scales which has a prominent substructure at lower scales.

The best method will be process and detector-dependent. Perhaps, simple recipes for choosing an optimal jet reconstruction strategy will emerge as more experience with these techniques is accumulated.

5. The FFTJet Package
The FFTJet software package [19] implements the multiresolution pattern recognition ideas outlined in the previous sections. Package users are expected to reconstruct jets using the following sequence of steps:

---

4 Bifurcation points are present in every jet reconstruction algorithm, including those which are normally considered to be infrared and collinear safe. Most algorithms do not have the capability to detect such points.

5 For example, the fraction of events in which the number of reconstructed jets above certain \( p_T \) threshold equals the number of partons produced at the leading order perturbation theory can be maximized for the signal of interest.

6 More powerful but significantly more complicated definition of the lifetime in the scale space is proposed in [22].
(i) The event energy flow is discretized using an equidistant grid in the $\eta - \varphi$ space. The grid cell size is normally chosen in a manner consistent with the granularity of the detector calorimeter.

(ii) The discretized energy distribution is convolved with a kernel function $K(\eta, \varphi, s)$. The resolution scale parameter $s$ determines the width and, possibly, the shape of the kernel. Many standard kernel functions are included in the FFTJet package, and user-defined kernels can be seamlessly added as well. The convolutions are efficiently performed by Discrete Fast Fourier Transform (DFFT).

(iii) The peaks of the convolved energy distributions are found. These are potential preclusters.

(iv) Preclusters with small magnitudes are eliminated in order to suppress the calorimeter noise.

(v) Steps ii through iv are repeated as many times as necessary using different values of $s$. The resulting preclusters are arranged in the clustering tree structure. Facilities for clustering tree sparsification and bifurcation point avoidance are provided.

(vi) Using the clustering tree information and assumptions about the signal spectrum, a decision is made about the event topology by choosing a set of preclusters. These preclusters are passed to the jet energy reconstruction stage.

(vii) Jet constituents are determined as follows. The event is viewed as a collection of energy deposits characterized by their direction $(\eta, \varphi)$ and energy variable $\varepsilon$. Depending on the environment in which the code is used, these deposits can originate from detector calorimeter cells, reconstructed tracks, Monte Carlo particles, etc. A cluster membership function $M_j(\eta - \eta_j, \varphi - \varphi_j, \varepsilon, s_j)$ is associated with each precluster $j$ at angular coordinates $(\eta_j, \varphi_j)$ and scale $s_j$. There is also a membership function for the unclustered energy/underlying event. The cluster membership functions are evaluated for every energy deposit in the event. In the “crisp” clustering scenario, an energy deposit is assigned to the jet whose membership function for this deposit is the largest. In the “fuzzy” scenario, the deposit is split between all jets with weights proportional to their respective membership function values (the sum of all weights is normalized to 1 to ensure energy conservation).

(viii) Jet energies and directions are calculated according to one of the standard energy recombination schemes using weights determined in the previous step.

In order to further improve the jet energy resolution, the last two steps of the algorithm can be applied iteratively. In such a procedure, the jet directions $(\eta_j, \varphi_j)$ and the membership function scales $s_j$ are updated at each iteration using reconstructed jets from the previous iteration until some convergence criterion is satisfied.$^7$

The computational complexity of the pattern recognition stage (steps i through v) is $O(SN \ln N)$, where $N$ is on the order of the number of towers in the detector calorimeter and $S$ is the user-selectable number of angular resolution scales (cone and $k_T$ algorithms use only one resolution scale). This complexity is independent from the detector occupancy and therefore allows for predictable execution times which can be important for online use. The computational complexity of the jet energy reconstruction stage is $O(IMI)$, where $J$ is the number of jets found, $M$ is the number of objects (4-vectors) used to describe the event energy flow ($M \leq N$), and $I$ is the number of times the last two steps of the algorithm are iterated (non-iterative version with $I = 1$ is often sufficient).

FFTJet permits arbitrary pattern recognition kernels and jet membership functions. The ability to use non-circular jet shapes translates into significant advantages in jet energy resolution (for the same expected instrumental noise and pile-up) when the method is used with calorimetric energy information provided by particle detectors which employ strong magnetic fields in their tracking systems. In such detectors, low-$p_T$ jets experience significant widening in $\varphi$ by the time

$^7$ This iterative procedure is known as “generalized mean shift” or “expectation maximization” algorithm [20, 21].
particles reach the calorimeter. As it was shown in [19], a simple replacement of Epanechnikov pattern recognition kernel by 2-d Gaussian and use of elliptical instead of circular cone results in a ≈ 30% reduction of algorithm-related jet energy reconstruction uncertainty (compared to seedless cone algorithm) for a typical cone size of $R = 0.5$ in a detector with 3.8 T magnetic field. The gains over $k_T$ and anti-$k_T$ algorithms are similar. It is also apparent that development of more sophisticated cluster membership functions (which can, for example, mimic expected transverse energy profile of a jet) will result in further improvements in the method efficiency and jet energy resolution.

6. Acknowledgements
The author thanks the Jet Algorithms Group of the CMS collaboration for comments and discussions. The development of the FFTJet package was supported in part by the US Department of Energy grant DE-FG02-95ER40938.

References
[1] G.P. Salam, “Towards Jetography”, The European Physical Journal C, Vol. 67, pp. 637-686, 2010.
[2] J. M. Butterworth, A.R. Davison, M. Rubin, and G.P. Salam, “Jet substructure as a new Higgs search channel at the LHC”, Phys. Rev. Lett., Vol. 100, 242001 (2008).
[3] D. Krohn, J. Thaler and L. T. Wang, “Jet Trimming”, J. High Energy Phys. JHEP 1002, 084 (2010).
[4] S.D. Ellis, C.K. Vermilion, and J.R. Walsh, “Techniques for improved heavy particle search with jet substructure”, Phys. Rev. D, Vol. 80, 051501 (2009).
[5] S.D. Ellis, C.K. Vermilion, and J.R. Walsh, “Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches”, Phys. Rev. D, Vol. 81, 094023 (2010).
[6] M. Cacciari, G.P. Salam, G. Soyez, “The catchment area of jets”, J. High Energy Phys. JHEP04 (2008) 005.
[7] S.D. Ellis et al., “Jets in hadron-hadron collisions”, Progress in Particle and Nuclear Physics, Vol. 60, pp. 484-551, 2008.
[8] Y. Cheng, “Mean Shift, Mode Seeking, and Clustering”, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 17, pp. 790-799, 1995.
[9] M.C. Minnotte, D.W. Scott, “The Mode Tree: a Tool for Visualization of Nonparametric Density Features”, J. Comp. Graph. Statist., Vol. 2, pp. 51-68, 1993.
[10] I. Volobouev, “Density-Based Clustering and Jet Reconstruction”, presentation at the MC4LHC Workshop, CERN, July 2006.
[11] Y. Wong, “Clustering Data by Melting”, Neural Computation, Vol. 5, pp. 89-104, 1993.
[12] S.J. Roberts, “Parametric and non-parametric unsupervised cluster analysis”, Pattern Recognition, Vol. 30, pp. 261-272, 1997.
[13] E. Nakamura, N. Kehtarvaz, “Determining number of clusters and prototype locations via multi-scale clustering”, Patt. Rec. Lett., Vol. 19, pp. 1265-1283, 1998.
[14] Y. Leung, J.-S. Zhang, Z.-B. Xu, “Clustering by Scale-Space Filtering”, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, pp. 1396-1410, 2000.
[15] K. Fukunaga, L.D. Hostetler, “The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition”, IEEE Trans. Information Theory, Vol. 21, pp. 32-40, 1975.
[16] D.W. Scott, “Multivariate Density Estimation: Theory, Practice, and Visualization”, Wiley, 1992.
[17] S.M. Omohundro, “Five Balltree Construction Algorithms”, ICSI Technical Report TR-89-063, 1989.
[18] M.A. Carreira-Perpinan, C.K.I. Williams, “An isotropic Gaussian mixture can have more modes than components”, Technical Report EDI-INF-RR-0185, School of Informatics, University of Edinburgh, UK, 2003.
[19] I. Volobouev, “FFTJet: A Package for Multiresolution Particle Jet Reconstruction in the Fourier Domain”, arXiv:0907.0270v1, 2009.
[20] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, “Numerical Recipes: the Art of Scientific Computing”, 3rd ed., Cambridge University Press, 2007.
[21] M.A Carreira-Perpinan, “Gaussian Mean-Shift Is an EM Algorithm”, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 29, pp. 767-776, 2007.
[22] T. Lindeberg, “Effective Scale: A Natural Unit for Measuring Scale-Space Lifetime”, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 15, pp. 1068-1074, 1993.
[23] T. Lindeberg, “Feature Detection with Automatic Scale Selection”, Int. J. of Computer Vision, Vol. 30, pp. 79-116, 1998.