Nuclear Instruments and Methods in Physics Research A 697 (2013) 64–76

Contents lists available at SciVerse ScienceDirect

Nuclear Instruments and Methods in Physics Research A

journal homepage: www.elsevier.com/locate/nima

Technical Notes

Introduction to HOBIT, a b-jet identification tagger at the CDF experiment optimized for light Higgs boson searches ☆

J. Freeman [a], T. Junk [a], M. Kirby [a,*], Y. Oksuzian [b], T.J. Phillips [c], F.D. Snider [a], M. Trovato [e], J. Vizan [f], W.M. Yao [d]

[a] Fermi National Accelerator Laboratory, Batavia, IL 60510, USA
[b] University of Virginia, Charlottesville, VA 22906, USA
[c] Duke University, Durham, NC 27708, USA
[d] Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[e] Isituto Nazionale di Fisica Nucleare Pisa, Scuola Normale Superiore, I-56127 Pisa, Italy
[f] Université catholique de Louvain, Louvain la Neuve, B-1348, Belgium

ARTICLE INFO

Article history:
Received 9 May 2012
Received in revised form
10 September 2012
Accepted 14 September 2012
Available online 23 September 2012

Keywords:
b-Jet identification
b-Tagging
Standard model Higgs boson
CDF
Tevatron

ABSTRACT

We present the development and validation of the Higgs Optimized b Identification Tagger (HOBIT), a multivariate b-jet identification algorithm optimized for Higgs boson searches at the CDF experiment at the Fermilab Tevatron. At collider experiments, b taggers allow one to distinguish particle jets containing B hadrons from other jets; these algorithms have been used for many years with great success at CDF. HOBIT has been designed specifically for use in searches for light Higgs bosons decaying via $H \rightarrow b\bar{b}$. This fact combined with the extent to which HOBIT synthesizes and extends the best ideas of previous taggers makes HOBIT unique among CDF b-tagging algorithms. Employing feed-forward neural network architectures, HOBIT provides an output value ranging from approximately -1 ("light-jet like") to 1 ("b-jet like"); this continuous output value has been tuned to provide maximum sensitivity in light Higgs boson search analyses. When tuned to the equivalent light jet rejection rate, HOBIT tags 54% of b jets in Monte Carlo simulated Higgs boson events ($m_H = 120\ \mathrm{GeV}/c^2$) compared to 39% for SecVtx, the most commonly used b tagger at CDF. We present features of the tagger as well as its characterization in the form of b-jet finding efficiencies and false (light-jet) tag rates.

1. Introduction

At CDF, the search for a light Higgs boson has been a subject of increasing interest and focus in recent years. While there have been numerous successful b-jet identification algorithms (commonly referred to as "b taggers") over the years, most have been intended for use in analyses other than searches for $H \rightarrow b\bar{b}$. Aspects of a given analysis, however, such as the optimal signal-to-background ratio, or the relative rate of non-b jets originating from gluons in the data sample before tagging, can influence whether a tagger is optimal for the analysis in question. Traditional taggers have tended toward a higher purity and lower efficiency than would be ideal for Higgs boson searches given the relatively low cross-section of Higgs boson production at Tevatron energies. While this problem has been circumvented somewhat by taking the logical OR of several taggers, a more elegant and flexible solution can be found in the continuous output of a neural network, tunable for each analysis application.

In this paper, we describe the Higgs Optimized b Identification Tagger (HOBIT). The strategy used in developing HOBIT is to build upon the strengths of previous CDF b taggers, address their weaknesses, and construct a new tagger that is highly optimized specifically for finding light Higgs boson decays. HOBIT produces a continuous output variable, allowing efficiency and background rejection to be tuned to meet the requirements of a given search. In the next section, we review some of the general features of b quark decays used by HOBIT to distinguish jets containing B hadrons from jets produced by gluons or light quarks (up, down, or strange). Section 3 then describes some of the previous b-tagging algorithms used by CDF upon which HOBIT is built. We then discuss some features of the CDF detector in Section 4, followed by a detailed description of the HOBIT algorithm and training regimen. The performance of HOBIT as characterized by the b-jet tagging efficiency and background rejection rates in data and Monte Carlo (MC) simulation is presented in Section 6. We conclude in Section 7.

2. Physics of b's from Higgs boson decay

Jets containing high-E_T B hadrons are created in a light Higgs boson decay possess several features that distinguish them from

☆ FERMILAB-PUB-12-118-PPD.

* Corresponding author. Tel.: +1 630 840 4326.
E-mail address: kirby@fnal.gov (M. Kirby).

0168-9002/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.nima.2012.09.021

jets produced by light quarks or gluons. The most important of these is the relatively long lifetime of a B hadron, augmented in the lab frame by its relativistic boost, which allows it to travel a distance on the order of a millimeter.[1] The B hadron's travel across these macroscopic distances results in a displacement between the location of the $p\overline{p}$ collision (the "primary" vertex) and the B hadron decay (the "secondary", or "displaced" vertex). These displacements are resolvable by the CDF tracking system, and in particular by its silicon detector. Almost all information as to whether or not a given jet originates from b-quark production is carried in the tracks reconstructed from detector signals left by the jet's charged particles. Specifically, it is possible to identify the decay of a B hadron through the displacement from the primary vertex of the individual tracks it leaves in the detector, and also through the displacement of a B-hadron decay vertex formed by combining multiple displaced tracks in a fit.

Other features also distinguish the b jet from other jets. Due to the large mass of the b quark, the collective invariant mass of the decay products of B hadrons will be larger than those from the decay products of hadrons not containing b quarks. Furthermore, the large relativistic boost typical of a B hadron will result in decay products which tend to be more energetic and collimated within a jet cone than other particles. Finally, particle multiplicities tend to be different for jets containing B hadron decays compared to other jets; in particular, muons or electrons appear in approximately 20% of jets containing a B hadron, either directly via semileptonic decay of the B or indirectly through the semileptonic decay of charm hadrons resulting from a B decay.

3. b-Tagging algorithms

As a tremendous amount of effort has gone into the construction of b taggers at CDF and other experiments [1–3], we build upon previous experience when constructing HOBIT. In particular, HOBIT explicitly uses as inputs the output of the SecVtx algorithm set to its "loose" operating point [4], the output of CDF's soft muon tagger [5], and inputs to the earlier RomaNN [6,7] and Bness [8] multivariate taggers. Consequently, it is useful to describe these taggers.

3.1. SecVtx

SecVtx is a displaced vertex tagger and the most commonly used b tagger at CDF. SecVtx only uses tracks which are significantly displaced from the primary vertex, accepted by quality requirements, and within a distance of $\Delta R < 0.4$ of the jet axis. Here, $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$, where ϕ is the azimuthal angle of the track around the beam axis, and η is its pseudorapidity defined as $\eta = -\log(\tan(\theta/2))$, with θ the polar angle of the track with respect to the beam axis. With these tracks, SecVtx uses an iterative method to fit a displaced vertex within the jet, where the χ^2 of the vertex fit is employed to guide the process. Assuming that this displacement is due to the long lifetime of the B hadron, the significance of the two-dimensional decay length L_{xy} in the plane perpendicular to the beampipe axis is used to select b-jet candidates. The algorithm is utilized with different track requirements and threshold values in order to achieve different efficiencies and purity rates. In practice, three operating points are used, referred to as "loose", "tight", and "ultra-tight". The loose SecVtx operating point decision is used as an input to both the RomaNN and HOBIT tagger. One drawback of the SecVtx tagger is that it is unable to fit a vertex in every b jet. In the Pythia [9] Higgs boson Monte Carlo (MC) simulation ($m_H = 120$ GeV/c^2) whose b jets are used to train HOBIT, SecVtx operating at its "loose" setting fails to find a vertex in 44.3% of these jets.

3.2. Soft lepton taggers

Soft lepton taggers [5] (SLT) take a different approach to b tagging. Rather than focusing on tracks within a jet, they select B hadron decays by identifying charged leptons inside a cone around the jet axis. Since the b semileptonic branching ratio is approximately 10% per lepton flavor, this class of tagger is not competitive with SecVtx or the other taggers described below if used alone. However, because a soft lepton tagger does not rely on the presence of displaced tracks or vertices, it has a chance to identify b jets that the other methods cannot. In practice, CDF uses only a soft muon tagger since high-purity electron or τ identification within jets is difficult. HOBIT uses as inputs the number of soft muon tags within a jet as well as the momentum transverse to the jet axis of the muon with the highest-likelihood tag.

3.3. The RomaNN tagger

The "RomaNN tagger" has been used at CDF in light Higgs boson searches [6,7] and employs neural network architectures. Neural networks (NNs) can use as many flavor-discriminating observables as is computationally feasible; hence the efficiency of NN taggers is equal to or greater than that of conventional taggers for a given purity. While the SecVtx tagger attempts to find exactly one displaced vertex in a jet, the RomaNN tagger uses a vertexing algorithm that can find multiple vertices, as may be the case when multiple hadrons decay within the same jet cone (for example, in a $B \to D$ decay). The RomaNN tagger uses several types of NNs: one to distinguish vertices which come from a heavy flavor (B or charm) hadron from false vertices or vertices coming from other hadrons; another to identify unvertexed tracks which come from a heavy flavor hadron; and then another NN which takes as inputs the output of the first NNs along with other inputs, including the loose SecVtx tag status, the number of SLT-identified muons, and the vertex displacement and mass information. Distinct versions of this third NN are trained to separate b jets from light jets, charm jets from light jets, and b jets from charm jets; the outputs of these three flavor-separating NNs are then used to train a final NN whose output is the RomaNN discrimination variable. The RomaNN tagger not only has superior performance to that of SecVtx at equivalent purities (see Fig. 7), but also allows for an "ultra-loose" operating point yielding greater efficiency, particularly useful in light Higgs boson searches.

However, the RomaNN tagger is not guaranteed to fit a vertex or to have sufficient input information to reliably tag a jet. In the event that the RomaNN tagger fails to receive sufficient information from its inputs, it is unable to assign an output value to that jet. This is the case with 20.6% of the b jets in the aforementioned light Higgs boson MC simulation sample. Regardless, due to the usefulness of the RomaNN inputs, a majority of them are employed as inputs into the HOBIT tagger, which allows HOBIT to take advantage of the same extensive vertex information that the RomaNN tagger uses.

3.4. The Bness tagger

While the RomaNN tagger focuses on the vertices it finds within a jet, in the event that it is unable to fit any vertices, it is unable to distinguish b jets from light jets. However, a significant proportion of b jets (approximately 20% in Higgs boson candidate events) do not contain a sufficient number of well-reconstructed

[1] This distance is achieved due to the fact that c times the rest frame lifetime of a B^0 ($B^{\pm}$, B_s, Λ_b) hadron is 460 μm (501 μm, 441 μm, 367 μm).