A pre-identification for electron reconstruction in the CMS particle-flow algorithm.

Michele Pioppi
CERN, CH-1211 Genève 23, Switzerland
E-mail: michele.pioppi@cern.ch

Abstract. In the CMS software, a dedicated electron track reconstruction algorithm, based on a Gaussian Sum Filter (GSF), is used. This algorithm is able to follow an electron along its complete path up to the electromagnetic calorimeter, even in the case of a large amount of Bremsstrahlung emission. Because of the significant CPU consumption of this algorithm, it can, however, only be run on a limited number of candidates. The standard GSF electron track reconstruction is triggered by the presence of high energy isolated electromagnetic clusters, but it is not suited for electrons in jets (usually soft and not isolated). A pre-identification algorithm based on both the tracker and the calorimeter, was therefore recently developed. It allows electron tracks within jets to be efficiently reconstructed even for small electron transverse momentum. This algorithm as well as its performance in terms of efficiency, mis-identification probability are presented.

1. Introduction
The CMS detector is equipped with very accurate electromagnetic calorimeter [1, 2] (ECAL) and silicon tracker [3, 4]. The reconstruction of the electrons in CMS is, nevertheless, a challenging task. Indeed, the amount of tracker material, one and half radiation lengths at a pseudo-rapidity $\eta = 1.5$, induces a large Bremsstrahlung photon emission, which, because of the 4T magnetic field, is significantly spread along the azimuthal direction. The situation becomes even more intricate for electrons in jets, the Bremsstrahlung photons of which are merged within jets. The so-called particle-flow algorithm aims at reconstructing all the individual particles in jets by making an optimal combination of the different sub-detectors. In the case of an electron, it implies reconstructing the track up to the ECAL entrance, associating this track with the electron cluster, and identifying the clusters of the emitted photons.

The Kalman Filter [5] is used by default for the track reconstruction in CMS. It assumes a Gaussian energy loss probability while the Bremsstrahlung emission is highly non Gaussian. The KF usually stops following the track after the emission of a photon because the next hit is often far away from the position expected by the algorithm. Therefore, the GSF [6, 7] which is a non linear generalization of the KF and models both the trajectory and energy loss probabilities by Gaussian mixtures, thus allowing a better momentum measurement, is used for the electron track reconstruction. Contrary to the KF, the GSF algorithm is also able to follow the electron track even after a dramatic Bremsstrahlung photon emission. The GSF is, however, 10 times slower than the KF. It cannot be used for tracking all the charged particles. Moreover, it would not improve the reconstruction significantly for charged hadrons. For all these reasons, a
tool to pre-identify quickly the electron tracks and then re-reconstruct them with the GSF was developed.

This note is organized as follows: in Section 2 the pre-identification strategy is described. In Sections 3 and 4 the tracking procedure and the pre-identification algorithm are respectively discussed. The results are summarized in Section 5, and the conclusions are in Section 6.

2. Pre-identification Strategy
In the balance between electron pre-identification efficiency and pion mis-identification, the emphasis was given to the former. Indeed a final identification, not described in the present paper, based on GSF tracks and calorimetric clusters will be applied and will reduce significantly the hadron contamination. Therefore, the goal of the pre-identification is to be 90-95 % efficient in the whole tracker region acceptance for the electrons that reach the ECAL. This is actually a true challenge because of the highly $\eta$-dependent tracker thickness, which results into a $\eta$-dependent amount of Bremsstrahlung emission. This pre-identification which starts from the list of the KF tracks as an input and selects a few of them. The number of pre-identified electrons, $N_{el}^{\text{pre-id}}$, is thus:

$$N_{el}^{\text{pre-id}} = N_{el} \epsilon^{KF} \epsilon^{\text{pre-id}}$$

where $N_{el}$ is the total number of electrons, $\epsilon^{KF}$ the efficiency of tracking the electrons with the KF and the $\epsilon^{\text{pre-id}}$ the efficiency of pre-identification for KF tracks associated to electrons. As a result, the maximization of the global pre-identification requires both the KF tracking strategy and the pre-identification of KF tracks to be efficient.

3. Iterative Tracking
In the standard KF-based track reconstruction applied in CMS, the tracks are required to be reconstructed with a minimum of eight hits, a maximum of one layer missed along the trajectory and a transverse momentum in excess of 900 MeV/c. These settings are far from being optimal for electrons, because of the high probability to have an energetic photon emitted before eight layers are crossed. If the minimum number of hits is abruptly reduced to recover some efficiency, the fake rate becomes unmanageable. For this reason, an iterative approach was developed. For each iteration the following operations are repeated:

(i) A seeding condition is defined, looser than in the previous iterations;
(ii) Seed tracks are reconstructed;
(iii) Tracks incompatible with coming from one of the primary vertices are rejected;
(iv) Hits used by tracks reconstructed in this iteration. are removed from further considerations.

This approach is intended to reduce the combinatorial reconstruction after each step and allows all the seeding and tracking thresholds to be gradually reduced. For this study, three iterations were used. The settings for each step are summarized in Table 1.

| Iter | Min Hits | Max Missed Hit | $P_T$ threshold (MeV/c) |
|------|----------|---------------|------------------------|
| 1    | 8        | 1             | 900                    |
| 2    | 5        | 1             | 500                    |
| 3    | 3        | 1             | 500                    |
The increase in the electron detection efficiency as shown by figure 1 is obvious for low transverse momentum and for $\eta$ region where the crossed material is larger in terms of radiation length. The tracking KF efficiency for electrons is in excess of 95% with a very low dependence on the pseudo-rapidity of the track. This approach also improves significantly the efficiency for pions and, in general, for the charged particles inside jets, without increasing the fake rate [8].

4. Electron Track pre-identification
The KF tracks that are obtained with the iterative tracking must then be filtered to select a few electron track candidates, to give seeds for the final GSF reconstruction. Since they are the most common hadrons produced in proton-proton collisions, the pions represent the background to reject. Several discriminating variables were developed and can be put in two categories:

- Track-cluster matching variables
  These variables measure the quality of the association (Section 4.1) between the KF track and the clusters in the electromagnetic calorimeter. For the tracks in the endcaps the clusters built in the preshower are also considered. This matching is mostly efficient for the electrons which do not lose much energy in the tracker, for which the corresponding KF candidate contains a large number of hits.

- Tracker-only-based variables
  The different behaviour between the electrons (especially those which emit Bremsstrahlung photons) and the pions in the tracker allows the short KF electron tracks to be selected and the pion rejection to be improved.

In order to maximize the electron pre-identification efficiency and the hadron rejection, and to take into account the correlation between all the variables, a multivariate approach is followed. The so-called Boosted Decision Trees (BDT) method was chosen. It is widely used in social sciences, and has been recently introduced in High Energy Physics [9] for several purposes, from particle identification to the detection of rare processes.
4.1. Track-cluster matching

In order to associate the track and the calorimetric cluster of an electron, each track is propagated to the inner surface of the ECAL. Then a straight line extrapolation to reach the expected maximum of the electron shower is applied. All the clusters with an energy greater than 900 MeV are considered and the closest one is associated with the track according to the normalized distance

\[ \chi^2_{\text{geom}} = (\Delta x)^2 + (\Delta y)^2 \]

where \( \Delta x \) is the distance in \( x \) between the position of the cluster and the one of the propagated track, and \( \sigma_{x(\eta)} \) is the intrinsic cluster position resolution. The ratio between the cluster energy and magnitude of the momentum of the track in the outermost layer \( E_{\text{cl}}/|P_{\text{tk}}| \) is also used to select electrons.

After finding the best matching, the \( \chi^2_{\text{geom}} \) and the ratio between the cluster energy and the track momentum, \( E_{\text{cl}}/|P_{\text{tk}}| \), are used to select the electrons. The \( E_{\text{cl}}/|P_{\text{tk}}| \) ratio is expected to be close to unity for electrons, when the track points precisely towards the electron cluster, which is usually the case when the electron does not lose much energy in the tracker.

In Figure 2, the geometrical and energetic compatibility is shown for electrons and pions with \(|\eta|<0.8\) and \(P_T>15\text{ GeV}/c\).

![Geometrical compatibility](image)

![Energetic compatibility](image)

**Figure 2.** Distribution of \( \chi^2_{\text{geom}} \) (left) and \( E_{\text{cl}}/|P_{\text{tk}}| \) for electrons(solid line) and pions(dotted line).

In the endcaps region \((1.65<|\eta|<2.6)\) CMS is equipped with a preshower detector located in front of the ECAL which allows the pion rejection to be improved. A clustering algorithm is applied in the two preshower layers and a geometrical matching similar to that described above is applied.

4.2. GSF light refitter

To strengthen the discriminating power of the tracker-based variables, all the tracks selected using the iterative tracking are refit with the electron hypothesis and using the GSF algorithm. The standard configuration for the GSF algorithm in the reconstruction software uses a mixture of twelve Gaussians for the trajectory state and six for the energy loss making it really time consuming. A factor six in CPU time is gained by reducing the number of components to four for the trajectory and three for the energy loss. The choice of the number of components results from a balance between the CPU time and the discriminating power of some tracking variables like the track \( \chi^2 \) or the difference between the reconstructed transverse momentum calculated in the outermost and innermost layer \(|P_T^{\text{out}} - P_T^{\text{in}}|\) between electrons and hadrons.
4.3. Boosted Decision Trees

The Boosted Decision Trees is interfaced in CMSSW using the package TMVA [10]. The whole list of variables used in the BDT is:

- $E_{\text{ECAL}}/P_{\text{track}}$
- $\chi^2_{\text{ECAL}}$
- $E_{\text{PS1}}$ (Only in the endcaps)
- $\chi^2_{\text{PS1}}$ (Only in the endcaps)
- $E_{\text{PS2}}$ (Only in the endcaps)
- $\chi^2_{\text{PS2}}$ (Only in the endcaps)
- Number of hits per track
- $|P_{\text{Tout}} - P_{\text{Tin}}|/P_{\text{T}}$
- $\chi_{\text{GSF}}^2$
- $\chi_{\text{KF}}^2/\chi_{\text{GSF}}^2$
- $P_{\text{T}}$
- $\eta$

The transverse momentum and the pseudo-rapidity are not directly discriminating like their correlations with all the other variables. The training was done using two samples (100K events each) of single electrons and single pions with a flat distribution on $p_T$ and $\eta$. Since the set of variables used are different for endcaps and barrel, two different files of weights are produced. In Figure 3 the output of the boosted decision trees is shown for single particles and for particles within jets. The shapes of the distribution are similar, and the remaining differences are related to the different transverse momentum spectrum.

Independently of the value of transverse momentum and pseudo-rapidity, the cut to discriminate between electrons and pions is set to 0.

5. Results

The samples used for this work are

- one million single electrons and one million single pions within the tracker acceptance ($|\eta| < 2.4$) and with a transverse momentum, $P_T$, in the $[1 - 50]$ GeV/c range
- 70000 of $b\bar{b}$ events generated with a $P_T$ in the hard process between 50 and 120 GeV/c

These events were simulated and processed with the CMS software (CMSSW). The optimization of the pre-identification algorithm was done with single particles, while the performance were evaluated both on single particles and on tracks inside jets.

The pre-identification efficiency for single electrons is always greater than 95% with a pion mis-identification below 10 %. The agreement between single particles and particles within jets as visible in Figure 4 is good especially in the barrel region, with a residual difference below 5%.

6. Conclusions

An electron pre-identification within the CMS particle-flow algorithm has been presented. An iterative KF tracking procedure has been applied to improve the KF tracking efficiency. An efficiency in excess of 95% almost independent of the transverse momentum, the pseudorapidity and the fraction of energy loss by Bremsstrahlung has been obtained for the electrons. In addition, this iterative procedure allows the overall tracking efficiency to be improved for all sorts of particles and all environments.

An electron pre-identification algorithm has then been applied to select a subset of tracks which are reconstructed with the GSF algorithm. Several discriminating variables based on the
Figure 3. BDT output normalized distribution for electrons (solid line) and pions (dotted line) in barrel (top), in endcaps (bottom), for single particles (left), for particles in jets (right).

Figure 4. $\epsilon^{\text{pre-id}}$ for single particles $-$ $\epsilon^{\text{pre-id}}$ for particles in jets as a function of transverse momentum for electrons (solid line) and pions (dotted line).
tracker-only have been determined. They have been combined in a BDT together with track electromagnetic cluster matching variables. A 95% efficiency for electrons almost independently of their isolation has been achieved, with a pion mis-identification at the level of 10%.

References
1. CERN-LHCC-97-33
2. CERN-LHCC-2002-027
3. CERN-LHCC-98-06
4. CERN-LHCC-2000-016
5. Fruhwirth R 1987 *Nucl. Instrum. Meth.* **A262** 444–450
6. Fruhwirth R 2003 *Comput. Phys. Commun.* **154** 131–142
7. Adam W, Fruhwirth R, Strandlie A and Todorov T CERN-CMS-NOTE-2005-001
8. The particle-flow algorithm in CMS. CERN-CMS-NOTE in preparation
9. Yang H J, Roe B P and Zhu J 2005 *Nucl. Instrum. Meth.* **A555** 370–385 (*Preprint* physics/0508045)
10. Hocker A et al. 2007 (*Preprint* physics/0703039)