An Image Analysis Approach To Analyzing Milagro Data
I. Monte Carlo Study of γ/hadron-Separation

Richard S. Miller
Los Alamos National Laboratory
15 March 1999

1 Introduction

Imaging has played an important role in improving the sensitivity of many astrophysics experiments. The ability to analyze sky exposure images (ROSAT, COMPTEL, etc.) or individual events (Whipple) has led to significant improvements in background rejection capabilities and hence improved source sensitivity. Here I describe a new analysis method, based on an image analysis approach, that should provide Milagro with γ/hadron-separation capabilities. The study described here is based solely on Monte Carlo simulations; comparison with actual data will be discussed in a future memo.

2 Motivation for an Imaging Analysis

Throughout this memo the term "imaging" will be used. According to Webster this is defined as follows: ***image: to create a representation of; to represent symbolically***. The goal here is to develop a technique that enhances Milagro's imaging capability such that one can create a representation of air-showers and identify the key features that differentiate signal γ- from the background hadronic-showers. An analogous imaging capability was critical to the success (ie. background rejection) of the Whipple ACT - the instrument that opened the field of TeV γ-ray astronomy.

Milagro is more than a light bucket. Although not originally designed as an instrument with imaging capabilities (at least not in the purest sense), Milagro can be thought of as a camera. Here the camera has n pixels (on two layers each of which has a different "exposure") that record the distribution of shower particles indirectly via Cherenkov radiation - an air-shower snapshot.

About 2 years ago at a collaboration meeting I presented an initial attempt at γ/hadron-separation in Milagrito. That presentation discussed some of the characteristic differences that exist between γ- and hadronic-showers (via Monte Carlo). Specifically, I found that in the raw particle distributions that reach the ground there were more clumps in hadronic-showers than in electromagnetic-showers - thus potential for significant γ/hadron-separation. However, when the detector response (ie. detector Monte Carlo) to the same showers was analyzed in an identical way there was little or no difference found between the two shower types, thus the technique was not used in Milagrito analysis. The expectation of enhanced clumpiness in raw (albeit Monte Carlo) hadronic-showers is certainly not a new result; however, to my knowledge this was the first attempt to systematically quantify the level of clumpiness or non-uniformity in air-showers.

I have revisited, and improved, this analysis using Milagro Monte Carlo with the thought that its deeper muon layer may be beneficial. The result can be summarized here: **there appears to be a significant γ/hadron-separation capability in Milagro**. Details can be found in the sections that follow.

3 γ/hadron-Separation

It is well known that the muon component of hadronic-showers is larger than γ-showers; however, the absolute number of muons hitting a ground-based detector (for triggered showers) is small and the fluctuations in the number of muons is not Poissonnian (see Figure 1). Thus relying solely on muon identification may not be the most reliable method for background rejection and is not likely to provide the large background rejection factors desired in Milagro.

The Monte Carlo studies performed to date on Milagro's γ/hadron-separation capabilities have hinted that some differences may exist between γ- and hadronic-showers. Yodh & Atkins and Westerhoff (internal Milagro memos) have shown that the information content of showers (number of PEs, number of PMTs, etc.) may provide an additional level of γ/hadron-separation beyond what one might expect from simply

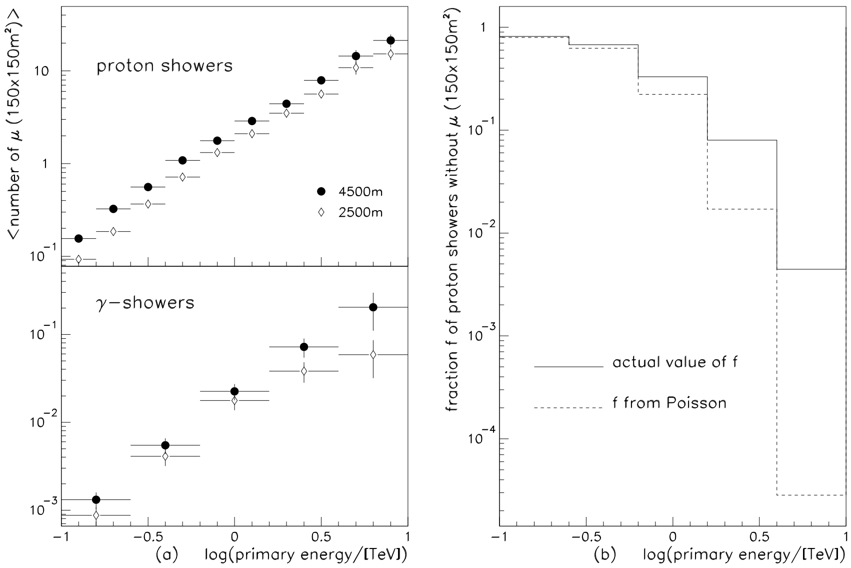

Figure 1: For a $150 \times 150\,\mathrm{m}^2$ detector area and cores randomly distributed over the detector area, (a) shows the mean number of $\mu^{\pm}$ in proton-induced showers as a function of the energy of the primary particle, and (b) shows the fraction f of proton showers without $\mu^{\pm}$ as a function of the energy of the primary particle. The solid line is the actual value of f, the dashed line is the expected value assuming the number of $\mu^{\pm}$ follows a Poisson distribution. From Miller & Westerhoff (1999).

counting muons. In the following sections I present another analysis technique which also shows a useful γ/hadron-separation, and represents a true image analysis approach to analyzing Milagro shower data.

4 Wavelet Image Analysis

The wavelet transform allows one to analyze an image in both *direct-space* and *Fourier-space*, and thus has the ability not only to analyze image features on various spatial scales (Fourier), but to identify the locations and morphologies (eg. edges, moments, etc.) of these features (direct).

In order to apply the wavelet analysis technique described below to Milagro the pulse-height information from the pond is used - no timing information has yet been incorporated. The model of the Milagro used is one of independent pixels, each 3m $\times$ 3m in area centered on the true PMT locations. Each pixel is given an amplitude equal to the number of photoelectrons measured by the corresponding PMT. It is important to note that in Milagro the “pixels” are not independent (ie. optically isolated from each other) and are therefore subject to cross-talk over a length scale corresponding to the attenuation length of Cherenkov radiation in water, combined with the characteristic scale size of clumps; however, the assumption of independent pixels simplifies the analysis.

4.1 The Wavelet Transform

Given a signal, or image, $s(x,y)$ and an analyzing wavelet function $g(x,y)$, the wavelet transform is simply a correlation

$$w(x,y,a) = s(x,y) \otimes \frac{1}{a} g\left(\frac{x}{a}, \frac{y}{a}\right), \tag{1}$$

where x and y are the coordinates of a given "pixel" (measured from some origin) and a represents a characteristic scale size being analyzed[1]. There are many choices for the analyzing wavelet $g(x, y)$. For the application described here the Urban Sombrero

$$g\left(\frac{x}{a}, \frac{y}{a}\right) = \left(2 - \frac{x^2 + y^2}{a^2}\right) e^{-\frac{x^2+y^2}{2a^2}} \tag{2}$$

is particularly well suited since it is a localized function having zero mean; therefore, image features analyzed at multiple scales a will maintain their location in image space. The Urban Sombrero has the same functionality as the second derivative of a Gaussian (same functionality as the 2nd Difference Method I described at the Feb '99 collaboration meeting), and thus is optimized to "pick out" peaks in an image while at the same time reducing flat components and gradients within images. This last feature is not critical to γ/hadron-separation but will be useful in making skymaps and source searches (forthcoming memo).

The correlation of Equation 1 can be discretized for ease of computation using

$$w(m, n, a) = \frac{1}{a} \sum_{i,j} c_{ij} g\left(\frac{x_i - x_m}{a}, \frac{y_j - y_n}{a}\right) \tag{3}$$

where the c_{ij} are the amplitudes of individual pixels. Of course, even for a modest number of pixels ($\mathcal{O}(10^3)$) this correlation represents a significant number of operations. To speed up the convolution I therefore use FFTs[2]. The use of FFTs reduces the number of operations from $\mathcal{O}(\mathcal{N}^2)$ to $\mathcal{O}(\mathcal{N}\log\mathcal{N})$, where $\mathcal{N}$ is the number of pixels.

5 Feature Identification

Once the image has been transformed via convolution with the analyzing wavelet characteristic features must now be identified. The features of interest here are "clusters"; as discussed above it is generally believed that hadronic showers should be clumpier than gamma showers. Assuming this to be correct the goal is to quantify the degree of clumpiness in a systematic way.

In the transformed image each pixel has an amplitude, or wavelet power. This power is proportional not only to the number of photoelectrons originally in that pixel, but also proportional to the number of photoelectrons in surrounding pixels out to a scale size of approximately a. Pixels can have both positive and negative amplitudes, including zero. Identifying clusters is a two step process: 1) identifying localized peaks in the transfromed image, and 2) "chaining" localized peaks together into clusters. These two steps are discussed below.

5.1 Identifying Localized Peaks

The first step in identifying and locating clusters is to identify what I call *localized peaks*. Not all tranformed pixel amplitudes w are independent, since the analyzing function has a characteristic scale size. Because the zero-point of the Urban Sombrero function (ie. the point where the amplitude crosses zero) is at $\sqrt{2}a$ any pixels that are within this distance of each other are related.

Here only pixels whose amplitude satisfies the condition $w > 0$ are used. To begin, this subset of pixels is sorted in descending order in w. Then treating the position of the largest amplitude pixel as an origin, all related pixels within a distance $r = (\Delta x^2 + \Delta y^2)^{\frac{1}{2}} \leq \sqrt{2}a$ are identified. This procedure continues until all positive amplitude pixels have been allocated to a peak; localized peaks can contain anywhere from 1 to all pixels, depending on the event and the analyzing scale size used. Pixels that have been included in one localized peak are not included in others.

[1] The scale size a is quoted as a multiple of the pixel separation. For Milagro, a scale size $a = 1$ is equivalent to 3 meters, a scale size $a = 2$ is equivalent to 6 meters, etc.

[2] Fourier transform the image and the analyzing wavelet, multiply the spectral components, and take the inverse transform. This is equivalent to the brute force convolution.

5.2 Cluster Formation

Just as individual pixels may be related, some of the localized peaks identified above may also be related. Although the analyzing function (Urban Sombrero wavelet) has a zero-point of $\sqrt{2}a$, the shape of the cluster is not guranteed to be circular. One must infer the shape of the actual cluster by chaining together many related localized peaks, ie. forming clusters by constructing a mosaic of related pixels.

The cluster formation procedure is quite simple: if the origins of two localized peaks are within a distance $d = 2\sqrt{2}a$ then they are related and belong in the same cluster. This procedure is continued until all localized peaks are identified with a cluster. Again, a cluster can contain anywhere from 1 localized peak to all the localized peaks. The morphology of each cluster can assume any shape; however, as one moves to larger scale sizes a the clusters approach circular shapes.

The position of each identified cluster is taken to be the position of the pixel with the largest amplitude w. A position estimate based on amplitude weighting could also be used, but does not impact the analysis presented here. The amplitude of each identified cluster is simply:

$$A_c = \sum_i w_i \tag{4}$$

where the sum is over all related pixels. Once the cluster formation procedure is complete a list of identified clusters, their positions, and their amplitudes exists. In practice the localized peak identification and cluster formation are performed in a single step.

6 Image Based γ/hadron-Separation

Using the procedures outined above one can now identify clusters in images using an unbiased technique. Figure 2 shows the application of the image analysis procedures to a shower. The goal now is to develop a metric with which one can quantify the degree of "gamma-like" or "proton-like" in a detected Milagro shower. A relatively simple measure is the likelihood ratio. The likelihood ratio is defined as

$$\lambda = -2\log\frac{\mathcal{L}_\gamma}{\mathcal{L}_\mathrm{p}} \tag{5}$$

where $\mathcal{L}_\gamma$ and $\mathcal{L}_\mathrm{p}$ are the likelihoods for an event being a γ- or hadronic-shower, respectively[3]. These likelihoods are a product of probabilties

$$\mathcal{L}_\mathrm{x} = \prod_{\mathrm{clusters}} \mathrm{p_x(A_c;N_c)} \tag{6}$$

where $(x = \gamma, p)$. The $p_x(A_c; N_c)$ represent the probability of observing a cluster with amplitude A_c when there are N_c clusters identified in the event. Probability distributions $p_x(A_c; N_c)$ $(x = \gamma, p)$ have been generated using Monte Carlo events ($5\times10^4\gamma$, 2×10^4 protons). Figure 3 shows examples of these probability distributions for specific values of N_c. There are clear differences in the distributions for the two shower types, particularly when using the μ-layer of Milagro.

The complete analysis procedure is summarized below:

1. Convolution of pond image and analyzing wavelet
2. Sort peaks in transformed image in descending order in w
3. Form clusters
4. For each identified cluster:
   - Lookup $\mathrm{p}_\gamma(A_c; N_c)$
   - Lookup $\mathrm{p}_p(A_c; N_c)$

[3]The parameter $\log\frac{\mathcal{L}_\gamma}{\mathcal{L}_\mathrm{p}}$ could also be used. The form of λ used here was chosen simply for historical reasons.

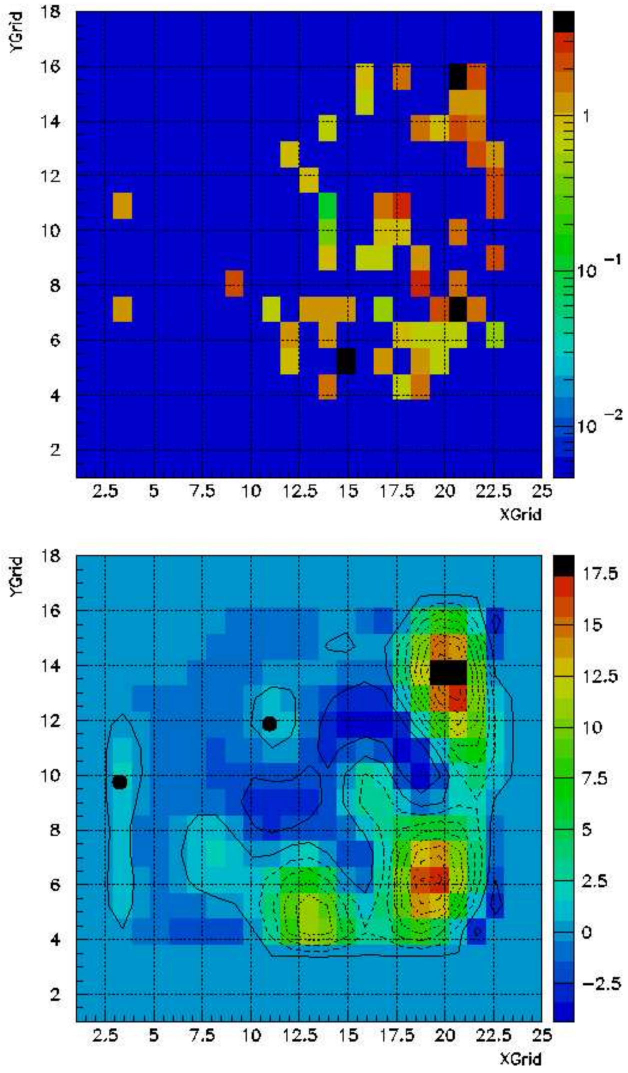

Figure 2: Example of wavelet based image analyis with a=2. (top) Milagro muon-layer image of a γ-shower. The amplitude (z-axis) is photoelectrons and the squares represent individual PMTs. (bottom) Same event after application of wavelet analysis. The amplitude is now wavelet power. The black circles mark identified clusters. For $a = 1$ the two additional "clusters" in the lower-right were identified; the $a = 2$ scale size shown here simply demonstrates the smoothing and peak identification concepts more easily.

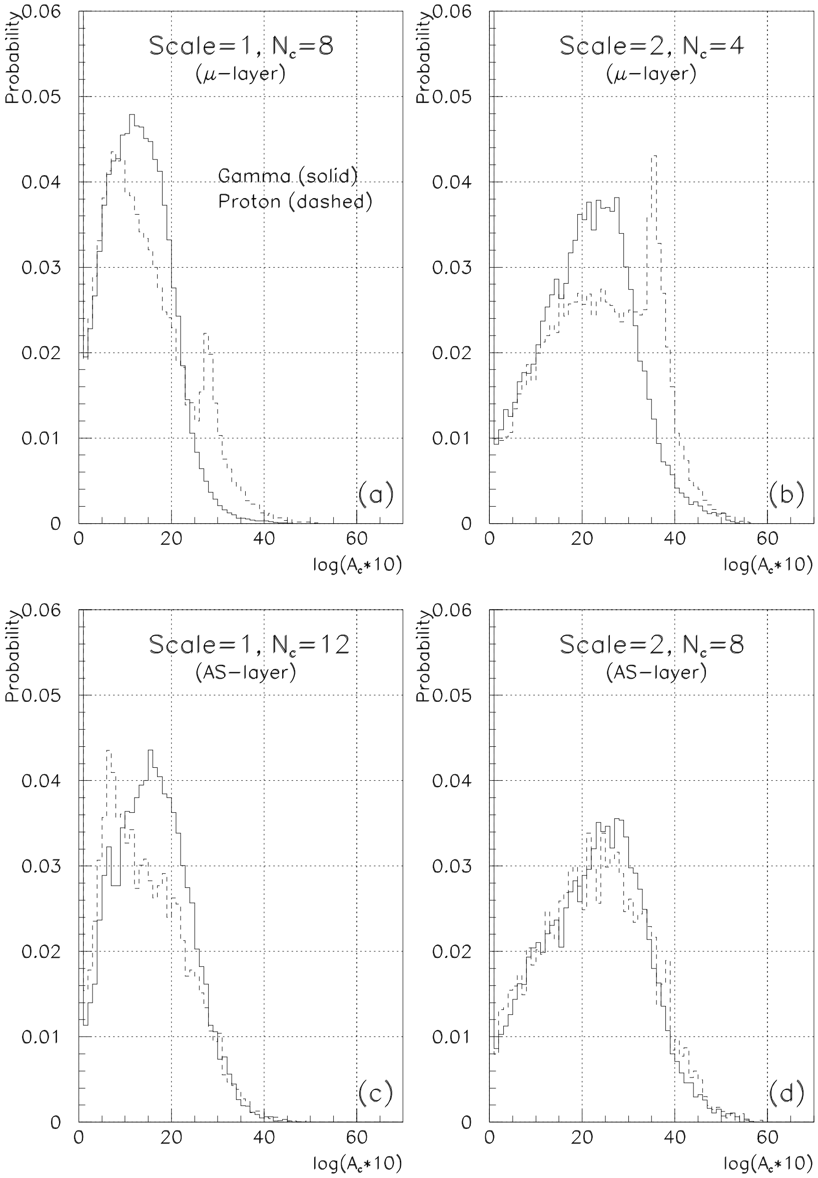

Figure 3: Examples of the amplitude probability distributions: (a) scale size $a = 1$ (3 m) and $N_c = 8$ using μ-layer, (b) scale size $a = 2$ (6 m) and $N_c = 4$ using μ-layer, (c) scale size $a = 1$ (3 m) and $N_c = 12$ using air shower-layer, (d) scale size $a = 2$ (6 m) and $N_c = 8$ using air shower-layer.

5. Compute Likelihoods:
   $\mathcal{L}_\gamma = \prod \mathrm{p}_\gamma(\mathrm{A_c}; \mathrm{N_c})$
   $\mathcal{L}_\mathrm{p} = \prod \mathrm{p_p}(\mathrm{A_c}; \mathrm{N_c})$
6. Compute $\lambda = -2 \log \frac{\mathcal{L}_\gamma}{\mathcal{L}_\mathrm{p}}$

Given a Milagro shower event one can now quantify how "gamma-like" or "proton-like" the event is through the use of the parameter λ:

$$\lambda = \begin{cases} \ll 0 & \text{"}\gamma\text{-like"} \\ \gg 0 & \text{"hadron-like"} \end{cases}$$

For purposes of clarity in this memo cuts on λ will be used to determine quality factors. However, a more sensitive use of this parameter would be to use it (actually a modification of λ) as a *weight*, thus making no hard cuts and using all available data in our analysis. Figure 4 shows an example of how the optimum value of λ_{cut} is chosen; the value of λ_{cut} is chosen such that the quality factor $Q = \epsilon_\gamma / \sqrt{1 - \epsilon_p}$ is maximized. As usual ϵ_γ and ϵ_p are the efficiencies for identifying γ and protons, respectively.

7 Results

Unlike other γ/hadron-separation analyses that determine quality factors for an ensemble of events or based on the number of hit PMTs, for the imaging analysis events *with the same number of clusters* are used to compute quality factors. In other words, γ-showers with 5 clusters are compared to proton-showers with 5 clusters. This aspect of the analysis was critical to achieving the highest effective quality factors. Why? Because it is not only the number of clusters that is important but also their amplitude. Thus, although a γ- and proton-shower might each have 5 identified clusters, the probability distribution of cluster amplitudes is considerably different (see for example Figure 3).

Before jumping to quality factors let me show a summary of how the likelihood parameter λ varies with the number of identified clusters; it is actually from this distribution that any γ/hadron-separation is extracted. Figure 5 shows profile plots for a scale size of $a = 1$ for both air shower- and μ-layers. It is interesting to note that although the likelihood parameter variation for both γ- and proton-showers flattens on the air-shower layer, the ability to differentiate the two shower types improves when using the μ-layer. The apparent divergence for small N_c is primarily due to the fact that there are very few proton-showers with a small number of identified clusters. This may or may not be an artifact of the Monte Carlo statistics.

A number of different characteristic scale sizes a have been studied to optimize the γ/hadron-separation, however only the smallest scale sizes show any sort of separation capability; shower differentiation degrades as the scale size increases since features are being smoothed out on the larger scales. This is partly due to the coarseness of the Milagro PMT spacing (ie. limited spatial resolution), the size of the pond, and the cross-talk between pixels.

Using the predetermined λ_{cut} parameter for each value of N_c one can now define efficiencies and hence a quality factor. The results are shown in Figures 6 and 7 for the best scale size ($a = 1$ or 3m) for both the AS-layer and μ-layer. It is clear that the air shower-layer provides only a moderate degree of γ/hadron-separation and its response is approximately independent of the number of hit PMTs on that layer. In contrast the μ-layer shows a much more significant rejection capability which *improves* as the number of hit μ-layer PMTs increases; the efficiency for keeping γ-showers flattens for large μ-layer events, while the efficiency for proton-showers continues to drop. If this separation can be performed with actual detector data, then the high Q μ-layer events would represent a *gold-plated sample* of γ-like showers with a high signal-to-noise ratio. These large events make up a few percent of all >50 PMT γ triggers; Figure 8 shows frequency histograms of all events meeting this trigger criterion. To ease comparison with other γ/hadron-separation algorithms the efficiencies and quality factors obtaioned using μ-layer data only (ie. Figure 7) are plotted versus the number of AS-layer PMTs in Figure 9.

8 Discussion

It is logical to assume that the clusters being identified are simply the result of muons - however, this would be incorrect. The dashed histograms in Figures 6 and 7 show the quality factors that would be obtained if

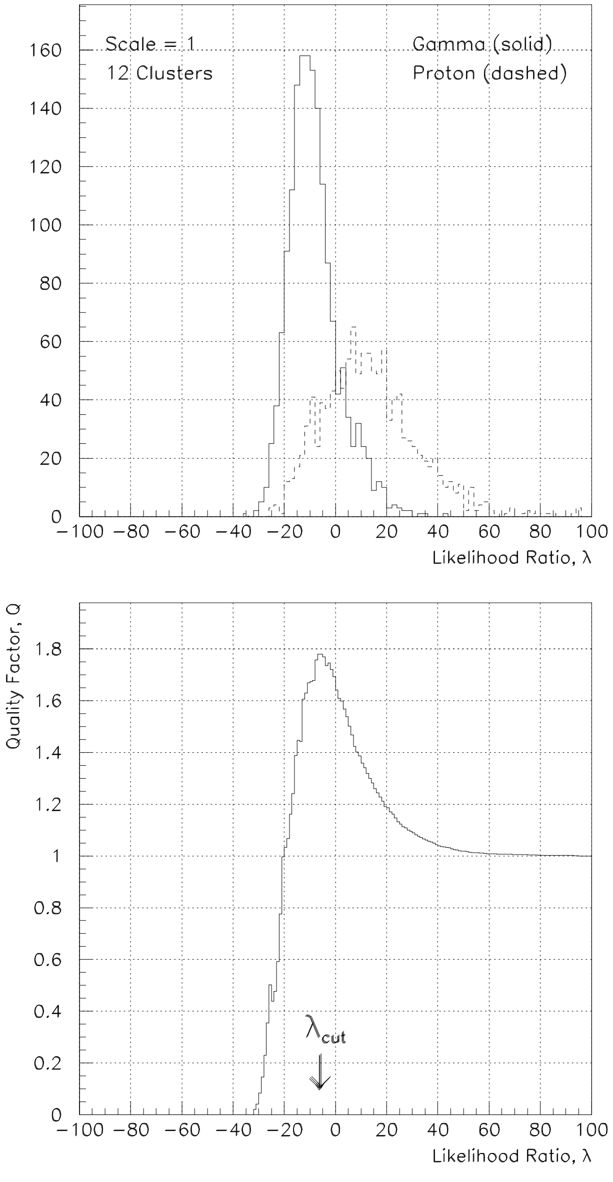

Figure 4: Example of how to determine the value of λ_{cut}. (top) Histogram of λ values for all showers with $N_c = 12$. (bottom) Quality factor vs. λ for $N_c = 12$. This plot is derived by calculating ϵ_γ and ϵ_p for each value of λ and computing the quantity $Q = \epsilon_\gamma/\sqrt{1-\epsilon_p}$. λ_{cut} is the value of λ the maximizes Q.

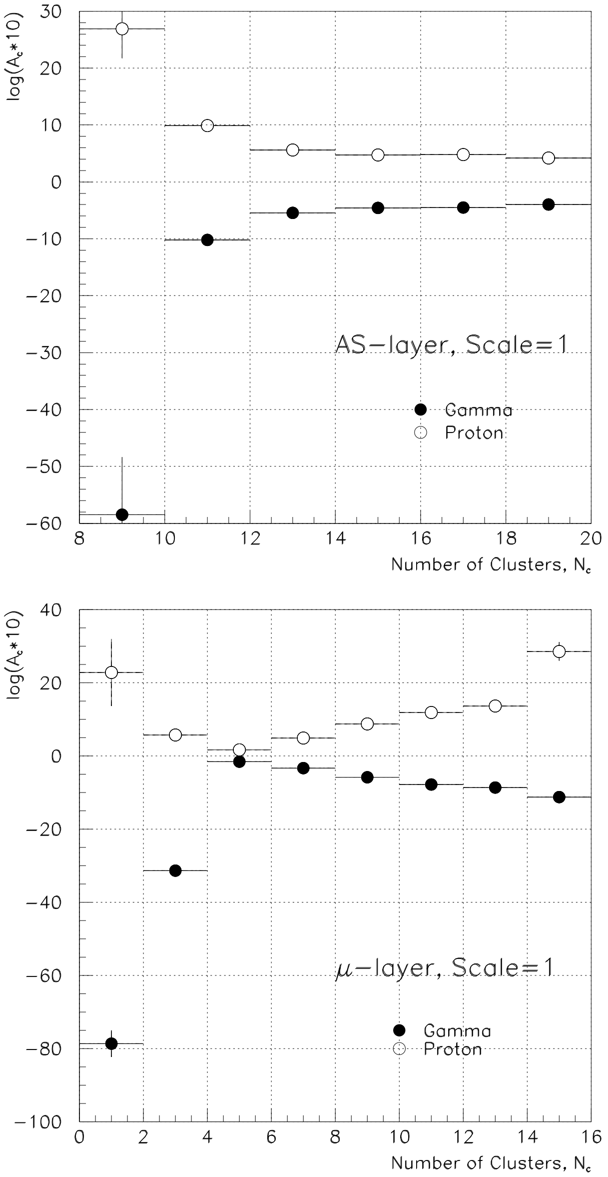

Figure 5: Profile plot showing the variation of the likelihood parameter λ as a function of the number of identified clusters N_c. (top) λ obtained from AS-layer data only, (bottom) λ obtained from μ-layer data only. Both analyses used a scale size of $a = 1$.

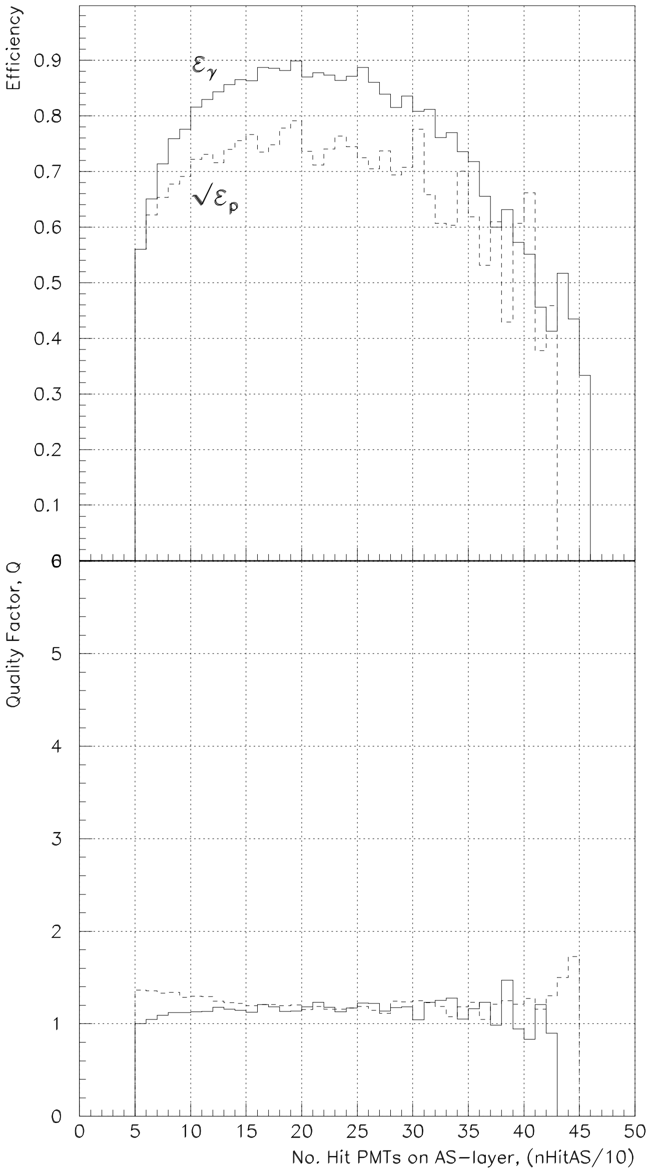

Figure 6: Efficiencies and quality factors obtained via wavelet image analysis technique using **AS-layer data only**. (top) Efficiencies ϵ_γ and $\sqrt{\epsilon_p}$ for keeping γ- and proton-showers respectively. (bottom) Quality factors Q obtained by dividing ϵ_γ by $\sqrt{\epsilon_p}$.

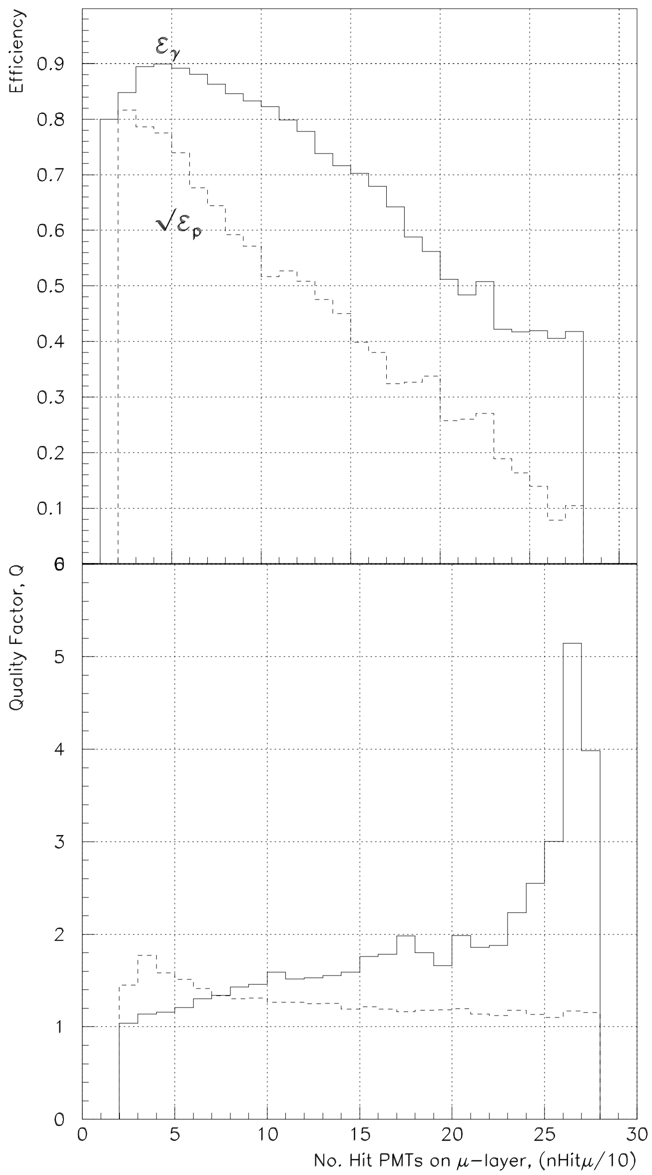

Figure 7: Efficiencies and quality factors obtained via wavelet image analysis technique using **μ-layer data only**. (top) Efficiencies ϵ_γ and $\sqrt{\epsilon_p}$ for keeping γ- and proton-showers respectively. (bottom) Quality factors Q obtained by dividing ϵ_γ by $\sqrt{\epsilon_p}$.

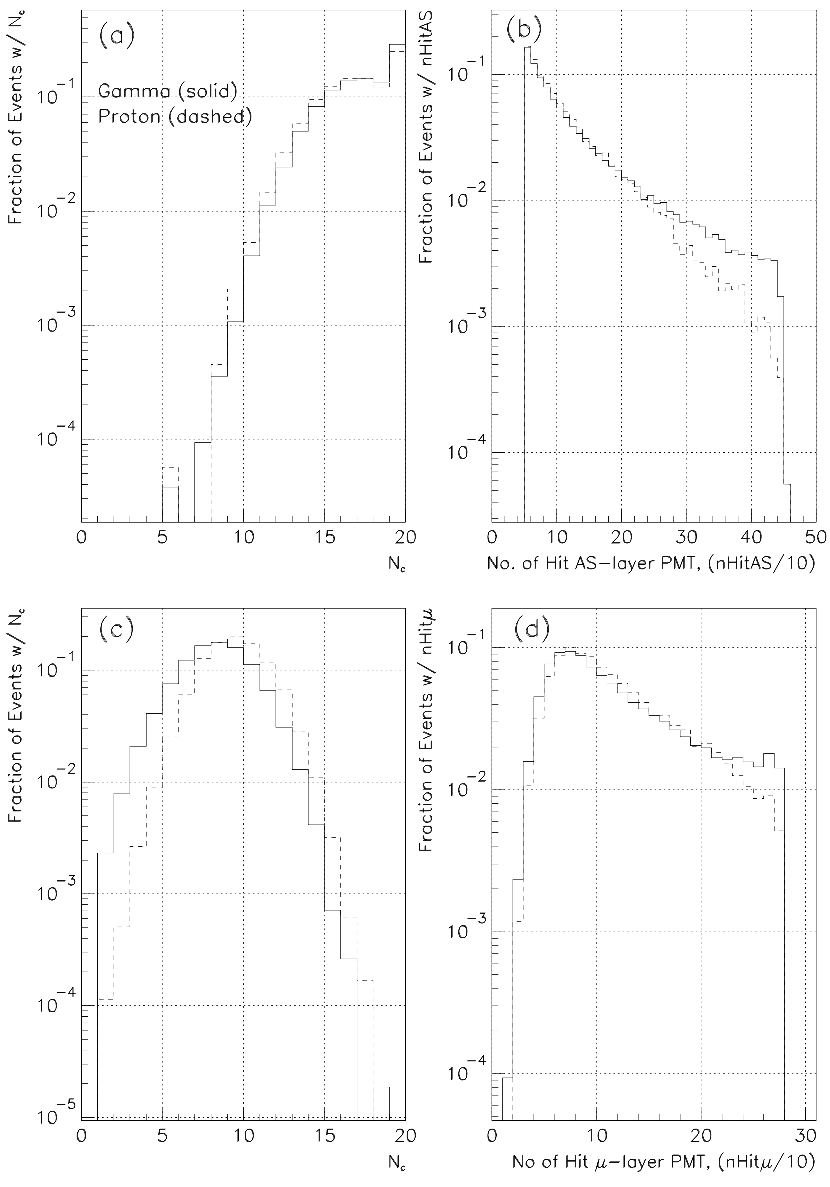

Figure 8: Frequency histograms of triggered events. (a) fraction of triggered events vs. N_c for AS-layer only; (b) fraction of triggered events vs. the number of hit PMTs in AS-layer; (c) fraction of triggered events vs. N_c for μ-layer only; (d) fraction of triggered events vs. the number of hit PMTs in μ-layer. The trigger condition for all events was >50 AS-layer PMTs.

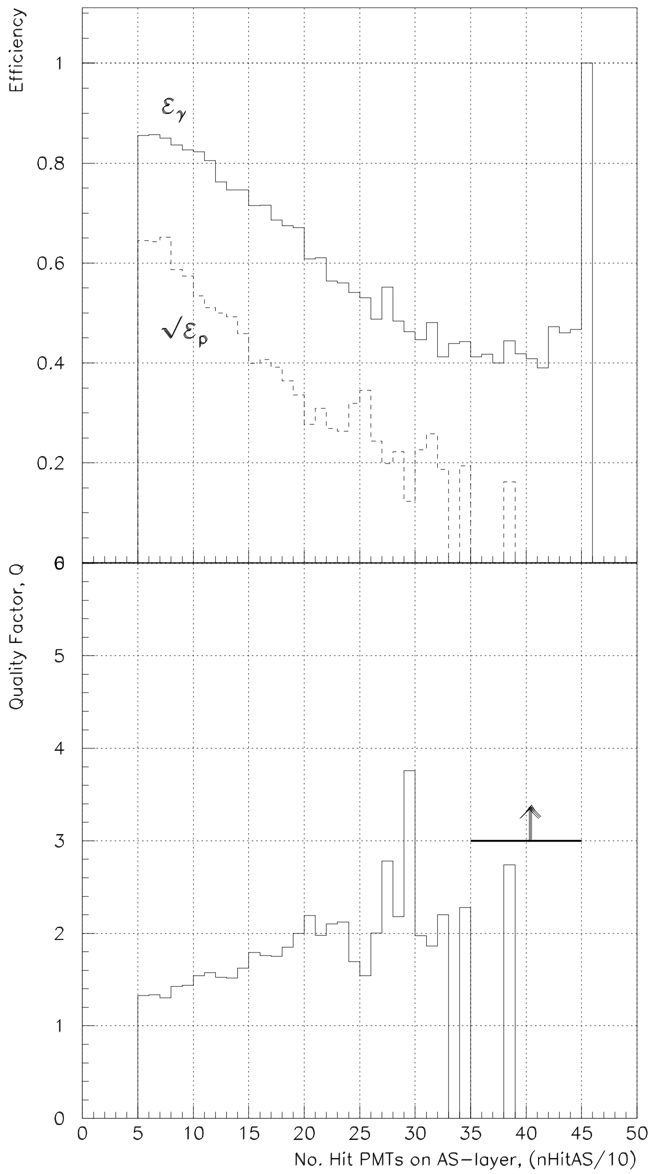

Figure 9: Efficiencies and quality factors obtained via wavelet image analysis technique using μ**-layer data only** vs. the number of hit AS-layer PMTs. (top) Efficiencies ϵ_γ and $\sqrt{\epsilon_p}$ for keeping γ- and proton-showers respectively. (bottom) Quality factors Q obtained by dividing ϵ_γ by $\sqrt{\epsilon_p}$. Note: large nHitAS(> 350) events have very large Q but are not plotted due to artifact of histogram generation. The efficiencies and quality factors are the same as in Figure 7, but have been replotted vs. the number of AS-layer PMT hits for comparison purposes. The arrow indicates the sub-sample of events with very very high signal-to-noise.

only muons were counted. To obtain these quality factors events were rejected that had at least one muon hitting the air-shower layer - no muon identification algorithm was used, only a muon identification flag from the Monte Carlo. This represents the most optimisitc estimate. The interpretation is that the image analysis is identifying characteristic features beyond what is expected from muons. The physics responsible for producing these features requires further investigation, and will be presented in a future memo. The expectation is that these features are the result of local particle density enhancements (ie. sub-showers) which predominate is hadronic-showers.

In the absence of better knowledge about the origin of these features, the results of the image analysis technique can be summarized as follows:

- The wavelet analysis is an unbiased algorithm for cluster or feature identification - ie. no assumption on amplitudes or shapes of features are made except that they represent pulse-height enhancements.
- The number of clusters alone is a *weak* separation parameter.
- The amplitude of a cluster is a *moderately good* separation parameter.
- The combination of the number of clusters *and* their amplitudes represents a *powerful* set of separation parameters.
- The muon-layer is significantly better for finding clusters, and hence providing background rejection information, than the air-shower layer.
- The information content of showers, combined with image analysis, produces a γ/hadron-separation capability that exceeds what is expected from analyzing the muonic component of showers alone.

9 Future Work

In principle there are a number of additional parameters that could be extracted from this image analysis technique and used to further enhance the γ/hadron-separation capabilities:

- *Spatial Distribution of Clusters*: Clusters in proton-showers might be distributed further away from the shower core due to large transverse momentum, than in γ-showers. For events with $N_c > 1$ the spatial distribution of clusters may be a diagnostic.
- *Cluster Morphology*: The shape of clusters, particularly the core, may also be important. In the raw shower particle distributions (Monte Carlo) that reach the ground I have found that the core of γ-showers tends to be quite symmetric in shape, while proton-showers may be more elliptical and/or chaotic. This is similar to the image characteristics identified in air-Crenekov shower images.
- *Cluster Timing*: In principle, it should be possible to extract an arrival time, relative the some global shower front measure, for each identified cluster. If delayed muons (or sub-showers) are present the cluster timing may be an additional hadronic background identifier.

Preliminary investigation of these parameters indicates that they will not significantly improve the background rejection capabilities of the current Milagro configuration. The primary reason is that the coarseness of our "pixellation", the small size of the pond, and the pixel cross-talk set limits on possible cluster scale sizes and determination of their shapes. How these additional parameters improve with a denser array of pond PMTs is currently under study.

It is true that the development of this image analysis technique and its application to Milagro has relied on Monte Carlo simulations. However one aspect of the image analysis can be tested: the probability distributions used to compute the likelihood (ie. cluster amplitude probability vs. number of clusters) *can be tested and compared against actual Milagro data*. Although only the proton-shower distributions can be compared and tested (since we currently have no statistically significant tagged photon source), it is a critical test. In fact actual shower data can be used to compute the proton probability distributions. In this case Monte Carlo simulations would then only be used to supply a guesstimate of the γ-shower probability distributions. Of course, once Milagro achieves high statistical significance on a source direct comparisons or tuning of the γ distributions can occur. Detailed comparisons of the proton probability distributions will follow in another memo.

10 Summary

Milagro has a degree of imaging capability which, if extracted and used properly, may provide discrimination between γ- and hadronic-showers leading to improved source sensitivity. I have described a technique that efficiently extracts image features from the spatial distribution of PMT pulse-heights. This technique, based on the wavelet transform, leads to quality factors conservatively extending beyond $Q = 3$. The Monte Carlo simulations suggest that the ability to distinguish γ- from proton-showers improves as the number of hit PMTs increases.