PROCEEDINGS OF SCIENCE

HAP-Fr, a pipeline of data analysis for the H.E.S.S. II experiment

Bruno Khelifi, for the H.E.S.S. Coll.*
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: khelifi@in2p3.fr

Arache Djannati-Ataï
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: djannati@in2p3.fr

Léa Jouvin
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: ljouvin@apc.univ-paris7.fr

Julien Lefaucheur
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: julien.lefaucheur@apc.univ-paris7.fr

Anne Lemière
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: alemiere@apc.univ-paris7.fr

Santiago Pita
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: pita@apc.univ-paris7.fr

Thomas Tavernier
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: thomas.tavernier@apc.univ-paris7.fr

Régis Terrier
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: terrier@apc.univ-paris7.fr

H.E.S.S. members have developed several pipelines for data analysis, together with two calibration chains and two Monte-Carlo simulations methods, ensuring results robustness. HAP-Fr (HESS Analysis Package-France) is a version of the HESS Analysis Package developed within the collaboration to allow the use of all different calibration outputs and the different simulations data. This chain aims to process raw data in order to reduce them, reconstruct the shower properties with different algorithms for the mono-telescope mode (CT5), stereoscopic mode with 4 12m-telescopes (CT1-4) and with all telescope types (CT1-5), reduce the cosmic-ray background with advanced multivariate analysis and derive high-level products by controlling finely their statistical properties. In this communication, the main algorithms of HAP-Fr are introduced and the analysis performance are given for different instrument configurations in mono for a bright source and stereo for a faint one.

The 34th International Cosmic Ray Conference,
30 July- 6 August, 2015
The Hague, The Netherlands

*Speaker.

PoS(ICRC2015)837

1. Introduction

The H.E.S.S. (High Energy Stereoscopic System) array is a system of five Imaging Atmospheric Cherenkov Telescopes located in the Khomas Highland of Namibia. It measures cosmic gamma-rays of very high energies ($\gtrapprox$20 GeV) using the Earth's atmosphere as a calorimeter. Since the inauguration of the phase II of the system in 2012, the latter consists of a hybrid array of two telescope types, four 12m-class telescopes on corners of a 120m square (CT1-4) and a 28m-class telescope at the center (CT5) [1]. In this context, the data analysis pipelines and algorithms have evolved to take full advantage of information provided by this large telescope.

HAP-Fr is one of the data analysis chains of the H.E.S.S collaboration and is developed in France. It allows to analyse raw data to provide final products, e.g spectra or sky maps. This paper provides a description of the data pipeline in section 2. The algorithms of events reconstruction and γ/background discrimination are presented in section 3, and the main performance of the analysis chain are presented in section 4.

2. Overview of the data pipeline

HAP-Fr is implemented in the historical analysis chain of H.E.S.S., the H.E.S.S. Analysis Package (HAP). Its architecture relies on a custom framework developed for both the on-line and off-line analysis, SASH (Storage & analysis software at H.E.S.S.), allowing a seamless integration of the data analysis code for the on-line data quick view during the data acquisition [2]. Its implementation is made in C++, using the ROOT data analysis framework on which the raw H.E.S.S. data format is based. ROOT offers a serialisation mechanism for the transport and storage of objects and classes. MySQL was chosen to design databases of monitoring, calibration and astrophysical data. The SASH framework and data analysis classes are handled by the C-shell and python scripting languages, allowing to implement the analysis steps for massive datasets.

HAP-Fr is able to process the two chains of H.E.S.S. Monte-Carlo (MC) simulations of air-showers and the telescope array. One uses primarily air-showers simulations made with a custom C++ implementation of Kaskade [3] that also uses the ROOT framework. The detector simulations are computed with a custom package, SMASH, embedded within SASH and ROOT environment. The other simulation pipeline is based on Corsika and Simtelarray [4]. In the same way, HAP-Fr uses different products of the two H.E.S.S. calibration chains. This allows comparisons and evaluations of systematic errors.

The pipeline analyses with the same algorithms the real and simulated data in three major steps. The raw data (Data Level 0, DL0) are first reduced using the calibration products and the first steps of the analysis are made (telescope images analysis, shower pre-reconstruction). The associated products (DL1) are stored in a 'DST' file, with a dedicated HAP format, for each observation run (typically of 28min). Then, the DSTs are processed to first compute intermediate data (DL2) that are the final event parameters and discriminant parameters. Then, on the fly, the events are filtered to increase the signal-to-noise ratio and tagged if their direction is compatible with a source region (the ON region) or a background region with a instrument response similar to the source region (the OFF regions), the whole (DL3) being stored into a file in ROOT or in FITS format. The last analysis step combines the DL3 of individual runs to derive the high level

products (DL4 and DL5), such as statistics of the source region, sky maps, excess morphology and spectrum, light-curves or phasograms.

To derive high level products, custom tools have been developped. The package START [12] allows to reconstruct spectral shape of γ-ray sources. It relies on libraries provided by the ROOT framework, MINUIT2, and the algorithms for numerical integration and interpolation from the GNU Scientific Library (GSL). A 'forward-folding' method is used to adjust an assumed spectral shape on data, taking into account the instrument functions [13]. In order to test the consistency of the fitted spectral parameters, a MC tool was developed within the software, providing a simulated number of 'measured' ON and OFF events in each energy bin for a given observation and a source spectral shape. This allows to test the robustness of the spectral START convergence [14]. The START package can also convert H.E.S.S. spectra and responses in OGIP-compliant format (pha, arf and rmf files). We have developed a python package for the Sherpa environment, 'HSPEC', that allows to perform full spectral analyses for large number of files with a profile likelihood technique [13]. HSPEC is part of GammaPy. A 2D-morphology Maximum-Likelihood fitting package was also developed and has already been used for a number of scientific results (e.g. HESS J1018 [15]).

3. Reconstruction and γ/background discrimination

This section describes the algorithms used to produce the DL3 data. Events are recorded either in the stereoscopic mode, i.e. when any pair of telescopes among 5 triggers the system, or in a monoscopic mode when only the large dish CT5 triggers. The mono mode is designed to lower the energy threshold of the H.E.S.S. II phase down to 20 GeV for specific scientific topics such as γ-ray pulsars [5], γ-ray bursts or high redshift Active Galactic Nuclei (AGN).

The overall scheme for event reconstruction and discrimination against background is similar for the mono and stereo modes, albeit some differences as outlined below.

The shower images are extracted from the calibrated raw data (DL1) after application of a tail-cuts algorithm with two thresholds ((5,7) photo-electrons or p.e) to remove noisy pixels and the Night Sky Background effects. For each telescope, a list of 'clean pixels' (DL2) is derived in order to compute the image moments or 'Hillas parameters' [7] (DL2). After pre-selection cuts over the Hillas parameters (pixel number, image charge and image position relative to camera borders) the images are used for the reconstruction of a first set of shower parameters, either in mono or stereo mode (see below), which in turn are used as starting parameters for a 3D-photosphere Gaussian model fit of the shower, the 'Model3D' fit [8]. The latter provides physical estimates of the shower characteristics, such as its width and maximum height, which are then used as powerful discriminant variables for the γ/background separation. In order to improve the efficiency of the background rejection a multivariate analysis method is employed using the TMVA framework [10]. Two different Boosted Decision Tree (BDT) classifiers are built for stereo and mono mode observations. For the training step of the classifiers, γ-rays are simulated as power law spectra for different observation conditions, and real background events are selected in regions where no γ-ray source is known. Different configurations are defined so as to optimise the analysis for various source classes.

3.1 Stereoscopic mode specificities

The shower parameters and the final event direction (DL3) are reconstructed after DST production thanks to the simple and robust geometric method based on the weighted intersections of major axis of shower images ((5,10) p.e. cleaning) recorded in at least 2 telescopes (see details in [6]). If the image number is larger than three, a cut on the minimal angle between two axis is applied to improve the accuracy of the direction reconstruction.

The energy of each event is estimated for each telescope as a function of the corresponding image amplitude and impact parameter using a lookup table. The final energy is derived from a weighted average of the individual telescope energies [9].

A total of 8 discriminant parameters are used. Two are based on the averaged and scaled Hillas parameters for the width and length of the images, and three are directly provided by the Model3D fit: the physical shower width, its associated error and the shower maximum depth in the atmosphere [8]. In order to exploit the azimuthal symmetry of the electromagnetic showers – as compared to more irregular background ones – for sake of better background rejection, two more parameters are calculated as following : the fitted Model3D shower is projected into the cameras of all telescopes, then the Hillas parameters of these predicted images are used to compute the shower parameters and direction as for the real images, but this time for predicted shower. The parameter labelled Ω compares the two reconstructed directions, while $R_{\rm E}$ measures the difference between the reconstructed energies [9]. The use of these parameters avoids the necessity of goodness-of-fit type variables and helps to preserve as much as possible the robustness of the analysis against systematic effects. The last parameter ΔQ is based on the differences between the measured and expected charge for each telescope given the reconstructed energy. All these steps producing DL3 information are made after the DST production.

With these discriminant parameters, a first set of events is used to generate a forest of decision trees with the boosted decision trees method. Compared to [9], a large number of short trees are used and no pruning is made [12], in order to fully exploit the AdaBoost algorithm [11]. For each event, the forest gives a rate, called ζ, according to its γ-like nature (close to +1 for γ-like and close to -1 for background). A second set of independent events allows us to determine the performance of classifiers such as the γ-ray and background reconstruction efficiencies as a function of ζ.

The operating point [12] is determined as the value $\zeta^{\star}$ which gives the best significance for a given source, as defined by a spectral index and a level of flux at a given energy. However these depend upon the observation conditions, e.g. zenith angle, wobble offset but also characteristics of the incident particles such as their energy. In order to increase the γ/background separation power, the parameter space is then divided as a function of the cosine of zenith angle, offset and the logarithm of the number of photons in the photosphere, $\log N_{\rm C}$. A BDT classification is generated for each interval of the parameter space.

3.2 Monoscopic event specificities

For the events triggered by CT5 only, direction, impact parameter and energy are estimated using a Neural Network (NN) regression method. Due to the limited information available for monoscopic events, the Model3D adjustment of the shower uses the above estimates of direction

and impact parameters as fixed inputs in order to fit the physical length, width and depth of the shower maximum, as well as the number of photons in the photosphere, N_C.

The reconstruction of the event direction by this NN assumes that the direction projected on the camera plane lies on the major axis of the image and its angular distance d from the barycenter is a function of its length, λ, width, ω and charge, Q. As the first foreseen application of the mono reconstruction is dedicated to pulsar studies [5], which have a known sky position, it is further assumed that the source lies towards the inner part of the field of view. This provides a better angular resolution at the lowest energies, but at the expense of a higher background. With this method, a Gaussian distribution of reconstructed d is achieved with a $\sigma = 0.22^\circ$ (see Fig. 1).

For the impact distance (ρ) estimation, the value of d calculated using the true source position in the field of view is used as an input during the training process in addition to the above image parameters.

The energy estimator relies on five parameters, λ, ω, Q, d, plus ρ. During the training phase, the true value ρ_{true} is used in order to avoid the contribution of the impact parameter error to the NN weight tables.

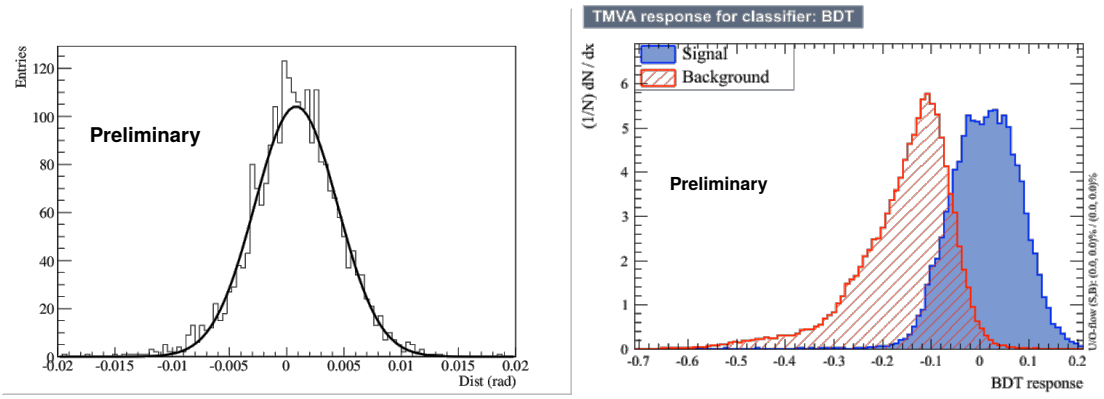

Figure 1: *Left* : Distribution of the Mono reconstruction of the distance parameter d at 20° zenith for a power-law distribution between 5 and 120 GeV with index $\gamma = -2$; the estimator has a Gaussian distribution with $\sigma = 0.22^\circ$. *Right* : Distribution of the multivariate discrimination parameter ζ, for signal and background events; a conservative cut value at $\zeta^\star = -0.1$ provides for a rejection of 70% of background events, while keeping 95% of the signal.

In the mono mode, the discrimination against background events using a DBT method is based on 7 variables, including λ, ω of the image (as for the stereo mode), the physical length and width of the fitted Model3D shower, Model3D-λ and Model3D-ω, the shower maximum depth and the error of the fitted Model3D-ω. The number of Cherenkov photons, N_C, is also included in this case as it adds some discrimination power.

The ζ distributions of real background and simulated γ-rays are compared in Fig. 1. Although the very small size of the shower images at the lowest energies (5-120 GeV) makes discrimination difficult, one can notice that the BDT method provides still some power. Here, a conservative value of $\zeta^\star = -0.1$ allows enough discrimination for a strong source like the Vela pulsar. Studies are under-way to improve the γ/background separation for fainter point sources.

4. Performance and examples

4.1 Monoscopic mode

The performance of the mono reconstruction algorithms, as obtained between 5 and 120 GeV for a power-law of index $\gamma = -2$, 20° zenith angle, and an image amplitude cut of 30 p.e., summarises as follows.

The angular resolution yields $R_{68\%} = 0.30°$, the impact parameter is determined with an accuracy of $\sigma = 33$ m and a very small bias (4 m), while the integrated energy distribution $(E - E_{true})/E_{true}$ shows a dispersion of 32% and a bias of -3%. As shown on Fig. 2 the energy bias rises steeply towards the threshold where it reaches values $> 50\%$. However it allows to reconstruct satisfactorily the spectra, even steep ones such as of the Vela pulsar with a differential index ~ -4 (details are given in [5]).

The collection area and the performance of the energy estimator are plotted in Fig. 2. One should remind that in these figures mono and stereo analyses are optimised for different source flux levels and comparisons should be made with care. As this analysis has not been optimised to achieved the greatest CT5 sensitivity, no sensitivity plot with this configuration is presented here.

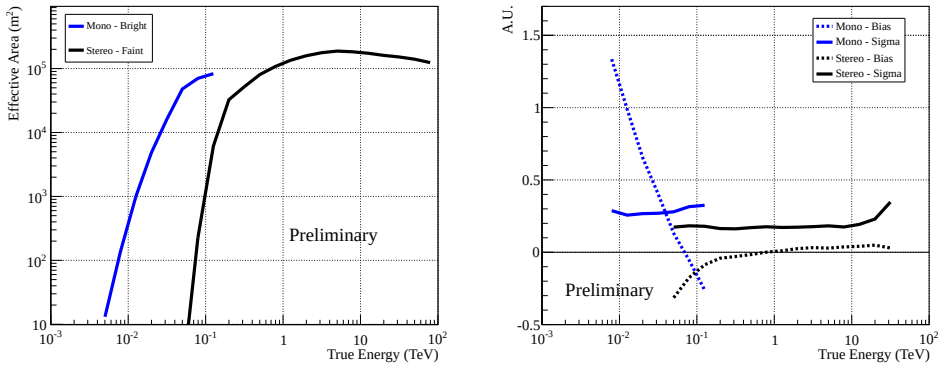

Figure 2: *Left* : Effective area for the mono (CT5-only) and stereo (CT1-CT5, two telescopes at least) modes; the former is optimized for a bright source (e.g. the Vela pulsar in the 10-30 GeV range), while the latter is defined for a *Faint* source in the 100 GeV range (2% of the Crab Nebula flux at 300 GeV with a -3.3 spectral index). *Right* : Energy resolution and bias for the same configurations derived from $(E - E_{true})/E_{true}$.

4.2 Stereoscopic mode

As mentioned above, the stereo analysis can be optimised for different types of signal (spectral shape and flux). In this section, its performance is given for an analysis dedicated to 'Faint' source, like distant AGN. The BDT cuts, $\zeta^{\star}$, are optimised to achieve the best significance for a signal of 2% of the Crab Nebula flux at 300 GeV with a $\Gamma = -3.3$ spectral index. For this configuration which the main performances are listed below, the image amplitude cut is 40 p.e and the θ^2 cut is $0.0125^{\circ 2}$.

The energy bias and resolution (see Fig. 2) are of the order of less than 10% and 15% respectively. The shower core distance error has a small bias ($< 1\%$ at 1 TeV) for an RMS smaller than 20%. The angular resolution is about 0.14° at 100 GeV, 0.07° at 1 TeV and $< 0.05°$ above 10 TeV. The collection area is given in Fig. 2, as well as the energy estimator performance.

The corresponding differential sensitivity curve (see Fig. 3) is computed for 50 hours of observation with at least a significance of 5 σ in each energy bin, and with 5% of background systematics and 8 background regions (i.e. the area ratio between the ON and OFF regions is thus $\alpha = 0.0125$). A sensitivity of around 1% C.U.[1] is effectively reached after training, the systematics errors being under investigation.

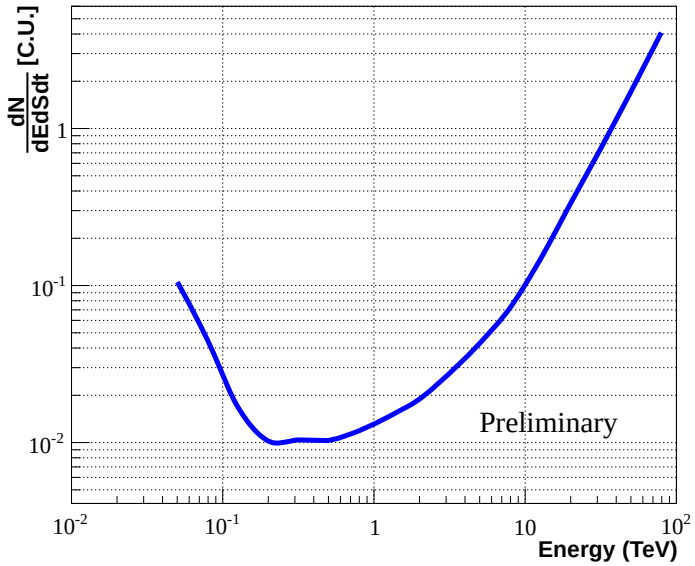

Figure 3: Stereo differential sensitivity in Crab units for the stereo mode and the *Faint* source configuration, computed at $Zenith = 18^\circ$ and $Offset = 0.5^\circ$ for 50 hours of observation.

This analysis configuration has been applied to data taken with the full array (CT1-5) (see Table 1). On the sources PKS 2155-304 and PG 1553+113, the measured rates of background BDT rejection ($7.5 \pm 0.9\%$ and $5.1 \pm 0.4\%$ respectively) are compatible within 20% of the rejection rates expected from the BDT analysis (7.9% and 4.4% respectively), showing that the cuts efficiencies are under control. For PG 1553+113, the analysis gives about $5.5\sigma/\sqrt{h}$, confirming the expected performance derived by simulations. While satisfactory, the results on PKS 2155-304 could be further improved by using an analysis configuration dedicated to brighter sources. For these datasets, the spectra measured with START give very satisfactory results and will be presented in future communications.

Source	T_{Obs}	$< Offset >$	$< Zen >$	ON	OFF	α	$Excess$	$Sigma$
PKS 2155-304	4.15	0.66	15.87	410	2304	0.067	254.3	16.1
PG 1553+113	11.83	0.5	35.25	680	3143	0.091	394.3	18.6

Table 1: Exemples of application of the stereo analysis for observations with the full array (CT1-5). T_{Obs} is the livetime in hours, $< Zen >$ and $< Offset >$ are respectively the mean zenith angle and mean offset in degrees, ON and OFF are the numbers of events in the test and background regions respectively, α is the area ratio between these regions, $Excess$ is the corresponding event excess, $Sigma$ is the Li&Ma significance.

[1] The used Crab Nebula spectrum is $2.83 \times 10^{-11} \mathrm{E_{TeV}}^{-2.62}$ cm^{-1}s^{-1}TeV^{-1}.

5. Conclusions

HAP-Fr is a software pipeline allowing to analyse the H.E.S.S. data taken with the full array composed of 5 telescopes. It processes the raw data by using any H.E.S.S. calibration or simulated products to derive high level products, such as spectra or phasograms. This pipeline has been adapted to account for the running of our array of different telescope types and new algorithms have been set up to analysis data in mono mode (with CT5 only) and in stereo (CT1-5).

A special care has been carried on the γ/background discrimination. A multi-variate analysis using boosted decision trees is used together with robust discriminant parameters built with the usual Hillas analysis and Model3D fit of the shower. One forest of decision trees is trained for each interval of the parameter space of discriminant variables and observation parameters. Their operating points are derived by optimising the signal significance for a fixed γ-ray flux and spectral shape.

The performances of the stereo analysis for a *Faint* source configuration reach the expected ones by comparing them with the preliminary results on AGN datasets. The mono analysis is also quite promising despite modest performances of the energy estimator and the difficulty of the γ/background separation. It allows to treat robustly data on strong and very soft sources, such as the Vela pulsar.

References

[1] M. Punch *et al.*, in: Towards a Network of Atmospheric Cherenkov Detector VII, 2005, p.379

[2] K. Mauritz, C. Borgmeier and C. Stegmann for the H.E.S.S. Coll., in proc.of the ICRC 2001, p.2896

[3] M.P. Kertzmann and G.H. Sembroski, NIM 343 (1994), p.629

[4] K. Bernlöhr, Astroparticle Physics 30 (2008), Issue 3, p.149

[5] Michael Gadjus *et al.* for the H.E.S.S. Coll., *'Pulsations from the Vela pulsar down to 20 GeV with H.E.S.S. II'*, in this proc.

[6] F. Aharonian *et al.*, A&A, 457 (2006), p.899

[7] A.M. Hillas, in proc. of the ICRC 1985, Vol.3, p.445

[8] M. Lemoine-Gourmard *et al.*, Astroparticle Physics, 25 (2006), p.195

[9] Y. Becherini *et al.*, Astroparticle Physics, 34 (2011), p.858B

[10] http://tmva.sourceforge.net

[11] Y. Freund and R.E. Shapire, in proc. of *Machine Learning: Proceedings of the* 13^{th} *International Conference* (1996)

[12] PhD Thesis of Julien Lefaucheur, in French: https://tel.archives-ouvertes.fr/tel-01128429

[13] F. Piron *et al.*, A&A, 374 (2001), p.895

[14] L. Jouvin *et al.* for the H.E.S.S. Coll., *'Statistical biases of spectral analysis with the ON-OFF likelihood statistic'*, in this proc.

[15] A. Abramoswki *et al.*, A&A 577 (2015), A131