Identifying the Higgs Boson

Guy Amdur[1], Anton Apostolatos[1], Leonard Bronner[2]
{gamdur, antonaf, lbronner}@stanford.edu

1 Introduction

This paper discusses the identification of the Higgs boson subatomic particle from jet pull energy colorflow images of the particles' decay, as modeled by the ATLAS Experiment at the Large Hadron Collider in CERN [1].

The Higgs field is an hypothesized energy field thought to permeate the entire universe. Without it, the Standard Model of particle physics would break down, as atomic particles would not have the required mass to attract each other, leading them to simply float around in the universe at the speed of light [20]. Its proof would completely alter our understanding of mass as a physical property, making the discovery of the Higgs field the fundamental unanswered question in particle physics in the last half-century [13].

The Standard Model suggests that if the Higgs field were to exist, then its quantum excitation, a particle referred to as the Higgs boson, would also have to exist [20].

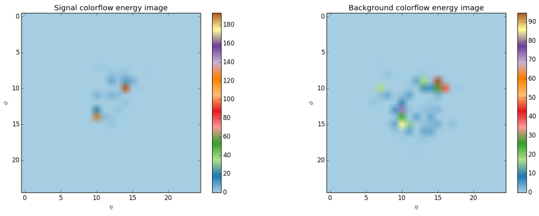
The Large Hadron Collider (LHC), tasked with finding this particle, consisting of multiple super-powered electromagnets, collides charged particles traveling at near lightspeed. These collisions deform space upon impact, breaking the charged particles into subatomic constituents. It has been hypothesized that with a proton collision at high enough energy, the Higgs boson would decay in observable ways.

The ATLAS detector at the LHC records 40 million proton collisions a second, making human curation of these events unfeasible [14]. An accurate classification system that would label the most promising observations as Higgs boson particle decays is, therefore, required.

In collaboration with SLAC and the ATLAS Experiment at CERN we were given access to energy images for both Higgs boson and gluon decay, referred to as signal and background respectively, seen in Figures 1 and 2.

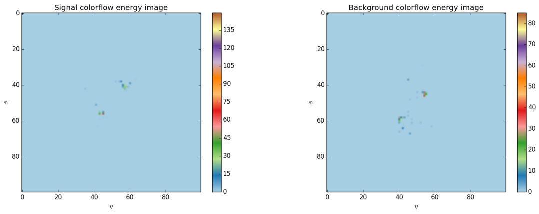
The purpose of this project was to build a binary classifier which, given the colorflow energy image of the decay of an unknown particle, would accurately distinguish whether or not that particle was a Higgs boson.

While we tested a wide array of supervised learning models, we focused primarily on ensemble methods. In particular we developed and fine-tuned an Adaptive Boosting classifier (AdaBoost) with a Random Forest classifier as its base estimator [8, 10]. We also utilized a number of image feature extraction mechanisms including the Laplacian Operator for edge detection and Features from Accelerated Segment Test (FAST) for corner detection [21, 22].



(a) Signal image (b) Background image

Figure 1: 25 × 25 pixel colorflow energy images



(a) Signal image (b) Background image

Figure 2: 100 × 100 pixel colorflow energy images

2 Related Work

Current approaches to this problem involve the use of jet pull features, which is a class of features used to characterize the superstructure of a particle decay event [9]. Jet pull information provides insight as to whether an event was initiated by a quark or a gluon, or if it came from a single object's decay, as would be the case for a Higgs boson particle decay, by describing the angle between energy decay patterns [18, 11]. The current state of the art model leverages this feature-set, while using Fisher Discriminant Analysis (FDA) for classification.

The approach extracts discriminating information between different classes of jets, similar to techniques used in computer vision [17]. The algorithm uses a representation of jets as images, applies preprocessing techniques to construct a consistent set of jet-images, and applies a linear discriminant, which has been trained on a collection of example jets [6].

FDA identifies the plane in the high dimensional feature space which maximizes the separation between the jet classes and simultaneously minimizes the scatter within each jet class. Since FDA uses knowledge of the within-class variations, it is not significantly influenced by variations present in both classes [6].

This method achieved an AUROC of 0.686 for the Higgs Boson classification [6, 17]. We will be comparing

[1] Stanford Computer Science Department

[2] Stanford Statistics Department

[0] Special thanks to Ariel Schwartzman at SLAC, Benjamin Nachman and Michael Kagan at CERN for their guidance, advice and access to data.

our results with the results achieved by this model.

3 Dataset and Features

We were given access to energy images for both Higgs boson and gluon decay. The two dimensions of these images corresponded to the spherical coordinates called η and ϕ, where ϕ is the azimuthal angle in the x-y plane perpendicular to the beam direction and η is the angle in x-z. These were preprocessed to center the jet, with resonance being kept constant in every sampled data point. As is evident in Figure 1 there is a stark difference between signal and background colorflow images, so automatic classification seems feasible.

We had two datasets of images of different sizes – 10,000 images composed of 625 floats (25 $\times$ 25 images) and 30,000 images of 10,000 floats (100 $\times$ 100 images). While these larger images provide higher resolution, the nature of the electromagnetic mechanism of jet pull energy observation leads these to less accurately detect charged particles. Thus, there is an inherent balance between higher granularity and lower quality of information.

Our baseline consisted of vectorizing these images in $\mathbb{R}^{625}$ and $\mathbb{R}^{10,000}$ vector space, respectively, and feeding those to our classifier. While for more advanced features we experimented with different image processing tools.

For low-level image processing, one-dimensional edge detection and multidimensional corner detection have been proven to be very successful image processing techniques [19]. Thus, the first featureset we developed used a Laplacian of the image. We were expecting to see an increase in accuracy as the Laplacian provides us with the edges of the color flow images, filtering any irrelevant noise [3]. Corners, on the other hand, can be found at the regions which have maximum variation when moved in all regions around it. We settled on the Features from Accelerated Segment Test (FAST), a modern corner detection method [3].

We also tested image histograms as features. The motivation behind this is that Higgs boson decay energy distributions, while probabilistic, fall under a specific intensity range, which is what image histograms discretely quantify. We also utilized the image histograms as a way to gain intuition about the tonal distribution of our images.

These methods are explained with greater detail in the following sections.

4 Methods

4.1 Image Feature Extraction

4.1.1 Laplacian Operator

The Laplacian Operator is a robust method used primarily for edge-detection. The Laplacian Operator is given by the relation

$$\mathcal{L}(f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

where each partial is taken with respect to each axis [21]. In images, edges are typically areas with high variation of intensity between neighboring pixels. Thus, areas where the second derivative of a pixel array is equal to zero are indicators of areas where there is an edge. In order to minimize noise, we applied Gausssian smoothing before computing the Laplacian of the image [3].

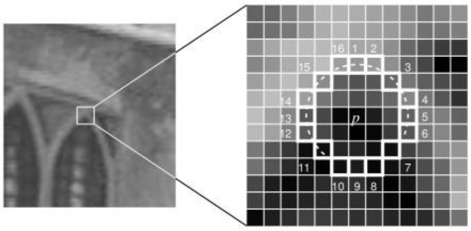


Figure 3: Visualization of FAST algorithm [3]

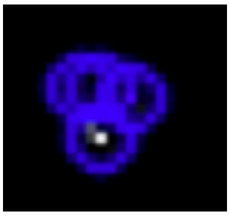


Figure 4: FAST keypoints of a signal event

4.1.2 FAST

The FAST algorithm for corner detection works by analyzing pixels and their surroundings. Given a pixel p with intensity I_p and a defined threshold value t, the pixel p is considered a corner if in the circle of 16 pixels p_n around it are darker or lighter than it, as is displayed in Figure 3. More specifically, a pixel p is considered a corner if $(\forall p_n . I_{p_n} \geq I_p + t)$ or $(\forall p_n . I_{p_n} \leq I_p - t)$ are found to be true.

The FAST detector outputs multiple keypoints for each colorflow energy image, with information that includes the coordinates and angle of each keypoint, the diameter of the keypoint neighborhood, the response by which the most strong keypoints were selected, and the octave from which the keypoints were extracted [22]. We used these keypoints as the features of our dataset of images. Figure 4 displays these keypoints for a sample Higgs boson particle decay event.

4.1.3 Image Histograms

An image histogram is a graphical representation of the distribution of tones in any given digital image. In our case, it is an array with pixel values (ranging from 0 to 255) in X-axis and corresponding number of pixels in the image on Y-axis. The typical histogram for our data contained a bin with a very high number of pixels of intensities 0, as it is the background intensity for the images. The number of non-zero bins was usually around 20, with very low intensities as well.

4.2 Classification Algorithms

We used a variety of classification algorithms dicussed in lecture, however, most of our work was with ensemble algorithms.

4.2.1 Adaptive Boosting

Adaptive Boosting (AdaBoost) is a meta-algorithm that combines multiple weak classifiers into a more accurate classifier. AdaBoost runs these weak classifiers multiple times, adapting each time so that subsequent classifiers are used to favor misclassified labels made by previous classifiers [8]. Algorithm 1 presents the algorithm with more detail and rigor. For our particular case, this weak classifier $\mathcal{L}$ was a Random Forest classifier and the number of iterations T was 50.

Algorithm 1 AdaBoost algorithm [15, 16]

```
Initialize the distribution as
    $D_1(i) = 1/M, i = 1, \ldots, M$
for $t = 1$ to $T$ do
    Get weak hypothesis $h_t : \mathcal{H} \mapsto -1, +1$ from training weak learner $\mathcal{L}$ using distribution $D_t$
    Compute the error rate
        $\varepsilon_t = \sum_{i=1}^{N} D_{t-1}(i) \mathbb{1}(h_t(x_i) \neq y_i)$
    Compute the weight $\alpha_t$ as
        $\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right)$
    Update the distribution, for $i = 1, \ldots, m$ as
        $D_t(i) = \frac{1}{Z_t} D_{t-1}(i) \exp(\alpha_t \mathbb{1}(y_i \neq h_t(x_i))$
    where
        $Z_t = \sum_{i=1}^{N} D_{t-1}(i) \exp(\alpha_t \mathbb{1}(y_i \neq h_t(x_i))$
end for
Construct and return the final classifier
    $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$
```

4.2.2 Random Forest

The Random Forest algorithm is a general ensemble learning classification method, relying on decision tree models [10]. The Random Forest classifier works by constructing multiple classification trees, where leaves represent either $+1$ or -1 labels and branches represent conjunctions of features. The classifier builds B trees during training. When an unseen sample requires classification, the input vector is passed through every single tree constructed, where every specific tree T_b outputs a prediction [4]. The classifier returns the label which most trees estimated. Given that an explicit algorithmic description of the method requires multiple pages of pseudo-code, we decided against presenting it in this paper. We invite interested readers who wish to get a thorough description of the classification method to refer to [7].

	Laplace	FAST	Histogram
AUROC	0.542	0.553	0.660

Table 1: AUROC scores for image processing using AdaBoost classifier

4.3 Training and Testing Methods

Since we had large datasets at our disposal we used hold-out cross validation. Namely, we split our dataset into two sets, a training set which composed of 75% of the data, and a testing set which comprised of the remaining 25% of the data. We would train each model on the training set and would then evaluate the hypothesis returned from that model on the test set. Classifiers will be evaluated by their receiver operating characteristics, or ROC curves. We will also be quantifying the performance of all binary classifiers tested by calculating the area under the ROC curve, or AUROC. We use this method because the AUROC represents the probability that a signal example will be classified correctly, which is exactly what we wish to optimize towards.

5 Experiments and Results

5.1 Image Processing

While testing with both datasets, the results presented are those run on the finer granularity data in $\mathbb{R}^{100\times100}$. This is because the colorflow images in the lower granularity dataset had too low of a resolution for many of these methods to provide any meaningful outputs.

We ran an AdaBoost classifier for each of the image processing feature extraction methods detailed previously. Results for these are presented in Table 1. As is evident, all image descriptors tested fall short of current state of the art methods. We hypothesize that the reason why these image descriptors performed so poorly was that the images that we are working with are not typical photographic images. Very few pixels of these images have non-zero value. Thus, descriptors such as the Laplace Operator provide very limited information about the image since there are many sparse energy jets captured in the colorflow images themselves. The image histogram, while outperforming the other two extraction methods, was not a particularly meaningful feature since, unlike a typical image with a very rich corresponding histogram, our images were so sparse that the histograms themselves were heavily skewed to 0.

5.2 Vectorized Image Classification

5.2.1 Classifier Selection

Our first task was to find the classification method best suited for the vectorized images as image features. For this we utilized the smaller granularity images since, given that the number of features for each sample is a number of orders of magnitudes smaller than the higher resolution images, classifiers ran much faster, allowing us to perform more tests.

Figure 5 presents the ROC curves for our tests using multiple classifiers – namely, Linear Discriminant Anal-

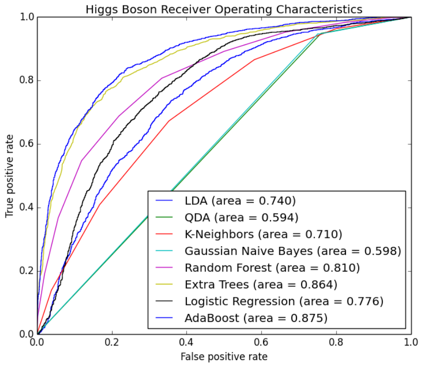


Figure 5: ROC curves for various classification methods for 10,000 samples of lower resolution images

ysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (K-Neighbors), Gaussian Naive Bayes, Random Forest, Extra Trees, Logistic Regression and AdaBoost with Random Forest classifiers as its base estimators [5]. Evidently, ensemble learning methods such as Extra Trees and Adaboost and classification tree methods such as Random Forest outperformed other predictive models. With an AUROC of 0.875, the AdaBoost classifier not only surpassed all other classifiers tested, but was also a far better model that current state of the art classifiers running FDA on jet pull data.

Given the exceptional results achieved with vectorized images in comparison to those attained through various image descriptors and image processing techniques, we decided to keep exploring vectorized images as our featureset.

5.2.2 Granularity Analysis

The next step in our experimentation was to compare image granularity and its effect on classification performance. Figure 6 displays the ROC curves for an AdaBoost classifier with Random Forest estimators, presenting the curve for FDA using jet pull data for comparison purposes. It's interesting that the lower resolution images outperformed the finer granularity images. This indicates that the higher resolution did not offset the loss of information that came as a result of larger images. Given its poorer performance and much greater training computation time, we proceeded to keep exploring methods using only the coarse dataset.

5.2.3 Feature Selection

The AdaBoost classifier, we found, was completely overfitting to the training data, to the extent that it was perfectly classifying all training examples. Thus, we set out to select only the most important features, allowing us to enhance our classifier's generalization. The importance of a figure can be calculated given the depth of the feature in a classification tree. Features found at the top of a decision tree have a much higher contribution to the final result than features at the bottom, and thus have a larger effect on the final prediction decision [2]. We would take the average depth of each feature in each decision tree in every Random Forest base estimator and use that calculated estimate as the relative importance of that feature [5]. Figure 7 demonstrates the relationship between feature size and the predictive quality of our classification methods. Our intuition that the most important features are in the center of the image are confirmed in Figure 8.

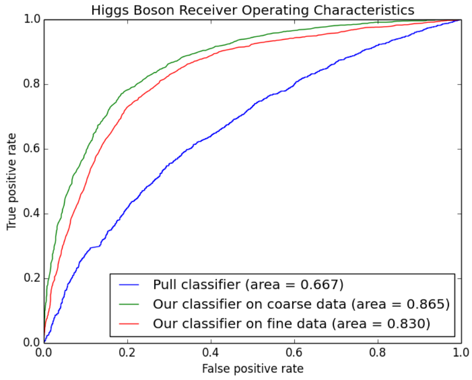


Figure 6: ROC curves for AdaBoost classifier with Random Forest estimators for lower resolution images (10,000 samples) and higher resolution images (30,000 samples) and FDA classifier with pull data

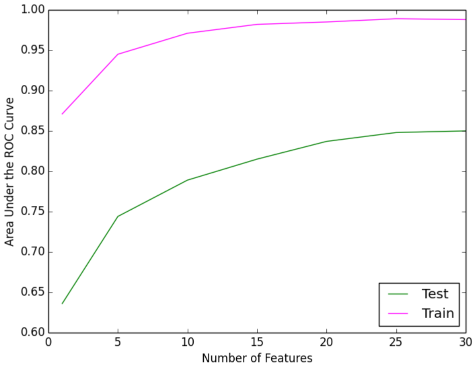
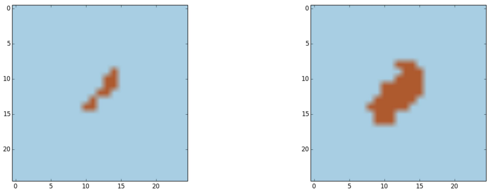


Figure 7: AUROC scores for top features

5.2.4 Training Size Analysis

An evaluation of the dataset size would give us valuable information as to whether more training data was required. Interestingly, Adaboost was able to get a very high AUROC score even with a small training size. Already for training sizes of around 1000 we achieved an AUROC of 0.862 and after 2000 examples the marginal gain of adding more samples was zero. Thus, there is no need for more training samples. Our intuition here is that since our images are generated from the same event and



(a) Ten features

(b) Forty features

Figure 8: Most important features

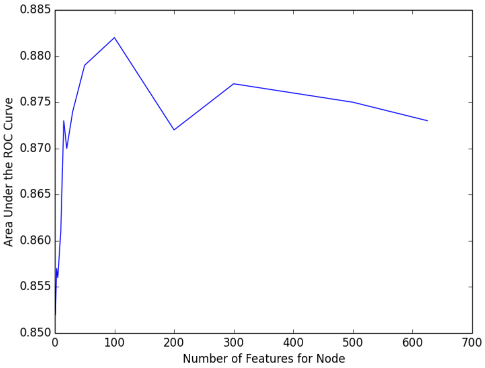


Figure 9: Varying number of features for split at nodes

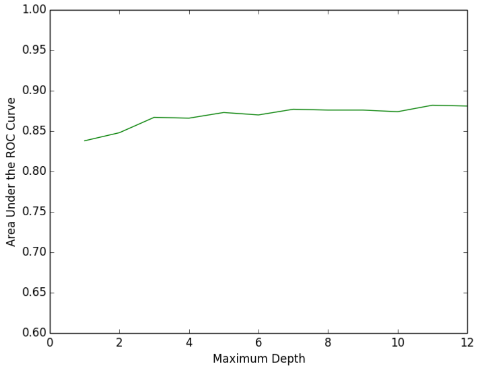


Figure 10: Varying classification tree maximum depths

there are only few important features, training and testing examples are very similar. Thus few training examples is enough to learn.

5.2.5 Base Estimator Hyperparameter Optimization

While Random Forest classifiers have multiple parameters associated to them that can be adjusted to maximize performance, the adjustable parameter which Random Forest classifiers is most sensitive to is the number of variables selected at each node used to calculate the split [4]. Increasing it amplifies the correlation between any two trees in the forest, but also increases the strength of every individual tree [4]. Thus, there is an optimal value that balances both effects. Figure 9 shows us that the optimal value for this parameter is 100. We ran these tests on a third, development set, to make sure that our final numbers still reflected a generalized error.

We also tested various maximum depth constraints for the decision trees in the Random Forest classifiers. When no maximum is provided, then the nodes are expanded until all leaves are pure [5]. As is evident in Figure 10, while after a depth of 7 the parameter began to have only a small effect on classification power, larger depths increase power of classification. The most optimal maximum depth found was 11.

6 Conclusion

We set out to construct a binary classifier for the identification of Higgs boson decay events from colorflow jet images as modeled by the ATLAS experiment. We were able to construct a classifier which was able to achieve an AUROC of 0.882. This is substantially better than state of the art systems utilizing jet pull information. The work detailed in this paper provides a glimpse of the incredible potential that decision tree and ensemble methods have for this classification problem. It is interesting that our experiments related to computer vision were not able to achieve higher results, but that it was simply vectorizing the colorflow images as is that generated the best AUROC. As described above, there are a few reasons for our results. On the one hand it is because our training and testing images were very similar, which made classification inherently easier. Second, there is clear distinction between the Higgs and non-Higgs energy decay rate and these differences can be characterized by few features. To see this, simply observe Figures 1 and Figures 8 to see that she shape of the higgs boson energry and the shape of the most important features are very similar. Nonetheless, it is important to stress that this is paper may represent a breakthrough in the Higgs boson identification process as we were able to vastly increase the probability of correctly identifying the presence of such a particle. At the same time, there is still a lot of work that lies ahead.

7 Future Work

First, the most important next step is to test our findings on new data that comes from different events. While the FDA baseline and our results use the same data, it is still unclear whether our findings are useful when it comes to distinguishing Higgs boson decay from in different contexts.

Furthermore, it may be interesting see what happens when more data or different data from the same events are added. The coloflow images we used are all preprocessed, and while this is not a computationally heavy process, it would be interesting to see whether our algorithms are able to detect rotations and translation. In addition, adding more features apart from simply the energy pattern may be interesting and we can imagine that this may lead to even higher accuracy in identification.

Finally, there are more sophisticated algorithms that we would wish to test. The first thing that comes to mind is a Convolutional Neural Network, these have been seen to lead to improved image recognition algorithms [12]. However, there are a number of issues here that arise due to the high pixelation in the images.

We very much look forward to running these experiments in the upcoming months and hope to publish our results as soon as possible.

References

[1] G. Aad, E. Abat, J. Abdallah, A. Abdelalim, A. Abdesselam, O. Abdinov, B. Abi, M. Abolins, H. Abramowicz, E. Acerbi, et al. The atlas exper-

iment at the cern large hadron collider. *Journal of Instrumentation*, 3(08):S08003, 2008.

[2] S. Bernard, L. Heutte, and S. Adam. Influence of hyperparameters on random forest accuracy. In *Multiple Classifier Systems*, pages 171–180. Springer, 2009.

[3] G. Bradski. *Dr. Dobb's Journal of Software Tools*.

[4] L. Breiman and A. Cutler. Random forests-classification description. *Department of Statistics, Berkeley*, 2007.

[5] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013.

[6] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman. Jet-images: computer vision inspired techniques for jet tagging. *Journal of High Energy Physics*, 2015(2):1–16, 2015.

[7] M. Denil, D. Matheson, and N. de Freitas. Consistency of online random forests. *arXiv preprint arXiv:1302.4853*, 2013.

[8] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156, 1996.

[9] J. Gallicchio and M. D. Schwartz. Seeing in color: jet superstructure. *Physical review letters*, 105(2):022001, 2010.

[10] T. K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

[11] A. Hook, M. Jankowiak, and J. G. Wacker. Jet dipolarity: top tagging with color flow. *Journal of High Energy Physics*, 2012(4):1–15, 2012.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] J. Lucio et al. Proceedings of the ii mexican school of particles and fields. Technical report, Teaneck, NJ; World Scientific Pub. Co., 1987.

[14] L. Mackey and A. Schwartzman. Physics event reconstruction at the large hadron collider. *Stanford Data Science Workshop*, 2015.

[15] J. Matas and J. Sochman. Adaboost. *Center for Machine Perception, Czech Technical University, Prague*, 2001.

[16] R. E. Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

[17] B. Scholkopft and K.-R. Mullert. Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*, 1:1, 1999.

[18] J. Shelton. Tasi lectures on jet substructure. *arXiv preprint arXiv:1302.0260*, 2013.

[19] S. M. Smith and J. M. Brady. Susana new approach to low level image processing. *International journal of computer vision*, 23(1):45–78, 1997.

[20] M. SStrassler. The known particles - if the higgs field were zero. October 2011.

[21] L. J. van Vliet, I. T. Young, and G. L. Beckers. A nonlinear laplace operator as edge detector in noisy images. *Computer Vision, Graphics, and Image Processing*, 45(2):167–195, 1989.

[22] D. G. Viswanathan. Features from accelerated segment test (fast), 2009.