IMPROVING BACKGROUND ESTIMATION FOR FAINT ASTRONOMICAL OBJECT DETECTION

Paul Teeninga[1]*, Ugo Moschini*[1]*, Scott C. Trager*[2]*, and Michael H.F. Wilkinson*[1]

[1]Johann Bernoulli Institute, and [2]Kapteyn Astronomical Institute,
University of Groningen, P.O. Box 407, 9700 AK Groningen, The Netherlands

ABSTRACT

Estimation of the background is an essential step in automated extraction of faint, extended objects from large-scale, optical surveys in astronomy. In this paper we present an improvement on the background estimation method of a commonly used tool in this field: Source Extractor (SExtractor). We show that the original method suffers from bias caused by presence of extended sources, and present an alternative which greatly reduces this effect, leading to much better preservation of faint extended structures.

1. INTRODUCTION

Given the sheer size of modern astronomical surveys, automated detection of objects is an important processing task. A well-known example is the Sloan Digital Sky Survey [1] (SDSS) where the DR7 [2] catalogue contains 357 million unique objects. Manual extraction of such numbers of objects is not feasible. A commonly used tool is Source Extractor (SExtractor) [3], which uses a fixed threshold equal to 1.5 times the standard deviation of the background estimate with the purpose of image segmentation while avoiding false positives. An image background, caused by reflected light and photo-chemical reactions in earth's atmosphere, is estimated and subtracted before thresholding. SExtractor's estimate shows bias from objects as can be seen in Figure 1. Part (a) shows a pair of merging galaxies with a faint tail structure linking them, part (b) shows the contrast-stretched background estimate from SExtractor. Clearly, there is correlation with the objects. Subtracting this background, shown in part (c), with pixels below the background estimate set to zero, reduces their intensities, leading to failure in detecting faint outlying regions of galaxies, or tidal tails in galaxy mergers. Figure 1(d) shows the same result, but with the new, flat background estimate. Finally, part (e) and (f) show the different object detections from SExtractor and the latest version our own detection method[4]. Clearly, faint structures are detected better.

This work was funded by the Netherlands Organisation for Scientific Research (NWO) under project number 612.001.110.

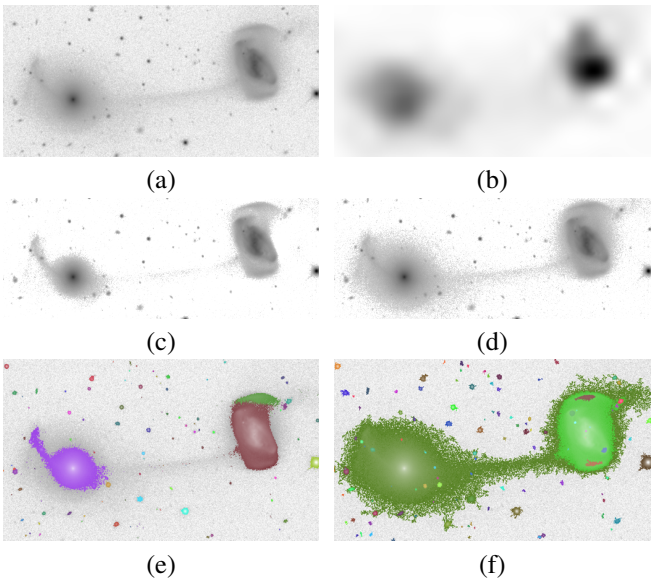

Fig. 1. Crop of SDSS file `fpC-002078-r1-0157.fit` showing merging pair of galaxies (a) original image; (b) background estimate of the image in by SExtractor, contrast stretched; (c) difference of (a) and (b), smoothed, logarithmic scale and showing values above $.75\hat{\sigma}_{\mathrm{bg}}$; (d) as (c) but with new background estimate subtracted. Part (e) shows source detections by SExtractor with default settings, while (f) shows the same results for our complete method from [4].

In this paper we present part of an ongoing project to provide robust methods for extraction of these faint structures, which are essential to the understanding of the evolution and morphology of galaxies. Initial results were presented in [5]. In this paper we focus on background estimation. We provide a heuristic which finds flat regions devoid of objects robustly, and determine that in the current data set, backgrounds turn out to be nearly flat. Unlike SExtractor's estimates that correlate with objects, our constant estimates derived from the detected empty areas provide much better results. The new method was validated on a data set of 254 monochrome images, a subset of the corrected images in SDSS DR7. Selection is based on the inclusion of merging and/or overlapping

galaxies which often include faint structures. Only images from the r-band are used which have the best quality [6]. The paper is organized as follows: first the method for estimating background is described, along with a discussion of parameter settings, and a critical analysis of the flat background model. This is followed by a comparison with the method in SExtractor, the conclusions and future work.

2. BACKGROUND ESTIMATION

Images are acquired with a CCD, photons are converted to electrons which are counted per pixel. After subtracting the software bias from the corrected images, the pixel values are directly proportional to photo-electron counts[7]. Noise is mostly Poissonian due to photo-electron counts. The distribution is close to Gaussian; the sky (background) typically contributes 670 photo-electrons to the counts per pixel. In our method, background pixel values are assumed to be from a Gaussian distribution. The image is assumed to be the sum of a background image B, objects image O and Gaussian noise where the variance is equal to $g^{-1}(B+O)+R$, for per-image constants g, equivalent to $gain$ in the SDSS, and R, which is due to other noise sources; read noise, dark current and quantisation. A tile of the image will be called flat if the pixel values could have been drawn from a single Gaussian distribution, e.g. $B + O$ is close to constant in the tile. The background is approximated by the mean value of flat tiles and is subtracted from the image. To find flat tiles, we first split the image into tiles of the same size. The following statistical tests are applied to the tiles:

1. a normality test using the D'Agostino-Pearson K^2-statistic[8] which is based on the skewness and kurtosis.

2. t-tests of equal means for different parts of the tile.

The t-tests are used because the normality test does not take positions of pixels into account. Only using the normality test could lead to situations where tiles with a near-linear slope are accepted. The procedure is outlined in Algorithm 1.

Inverses of cumulative distribution functions (CDFs) are used to determine rejection boundaries in the tests. For example the K^2-statistic has approximately the χ^2 distribution with 2 degrees of freedom. The CDF inverse in this case simplifies to $-2\log(1-p)$ which gives a boundary of $-2\log(\alpha_1)$. A potential issue is that the statistics are not independent. There is a noticeable error in the actual rejection rate due to the χ^2 approximation, as seen in Figure 2, for $\alpha = 0.05$. However, the simulated rejection rate for a 16×16 flat tile is still close to 0.05 and the error decreases for larger tiles. The rejection rates are evenly divided between the K^2-test and the combined t-tests. Other ratios have not been tested. The whole procedure is outlined in Algorithm 2.

Algorithm 1 `IsFlat`(T, α)

In: $w \times w$ tile T. w is even. Rejection rate α.
Out: True if T is flat. False otherwise.

1: $\alpha_1 \leftarrow 1 - (1-\alpha)^{1/2}$
2: Perform the D'Agostino-Pearson K^2-test on the values of T with rejection rate α_1. Return false if rejected.
3: $(T_{1,1}, T_{1,2}, T_{2,1}, T_{2,2}) \leftarrow \frac{w}{2} \times \frac{w}{2}$ tiles partition of T.
4: $\alpha_2 \leftarrow 1 - (1-\alpha)^{1/4}$
5: Perform a t-test of equal means on the pairs $(T_{1,1} \cup T_{1,2}, T_{2,1} \cup T_{2,2})$ and $(T_{1,1} \cup T_{2,1}, T_{1,2} \cup T_{2,2})$ using rejection rate α_2. Return false if the null hypothesis of equal means (and variances), in any of the two tests, is rejected.
6: Return true.

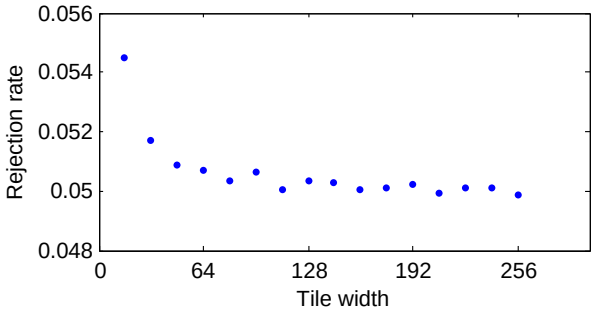

Fig. 2. `IsFlat`$(T, 0.05)$. Rejection rate for square flat tiles based on 1 million simulations for each size.

Algorithm 2 `BGMeanAndVariance`(I, α, w_0)

In: Image I with at least one flat $w_0 \times w_0$ tile. Rejection rate α. Minimum tile width w_0.
Out: Background mean and variance estimates.

1: $w \leftarrow w_0$
2: **while** the $2w \times 2w$ tiles partition of I contains a tile T where `IsFlat`(T, α) is true **do**
3: $\quad w \leftarrow 2w$
4: **end while**
5: Calculate and return the mean and variance of the pixels in the $w \times w$ flat tiles.

Larger flat tiles are preferred to make detection of slopes due to objects easier. Some rows at the bottom of the image and columns at the right side of the image are ignored when the height and width are not divisible by w. Doubling w when searching for a tile size guarantees that the function runs in $\mathcal{O}(n \log n)$ time, with n the number of pixels. The noise variance is also returned because it is needed later for object segmentation. An important issue is the value of α. Assuming the background estimate does not have a negative bias, settings that give a lower average background estimate are better, as the only bias in the background estimate is a positive bias from objects. The only possible cause of a negative

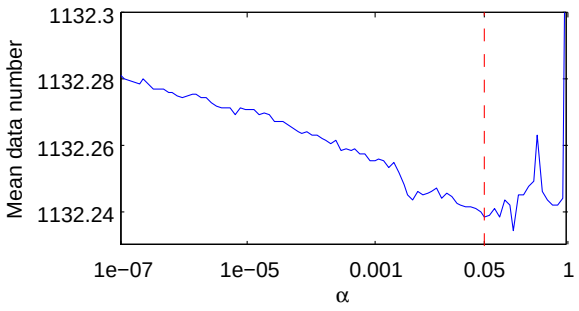

Fig. 3. Average background estimate for the data set as a function of α.

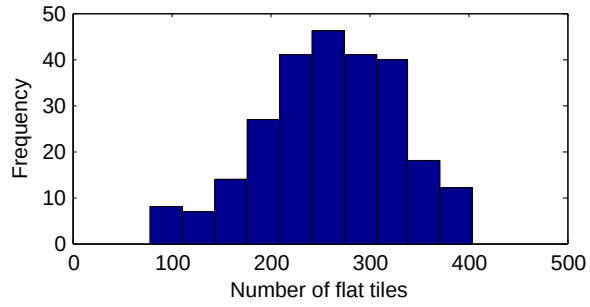

Fig. 4. Histogram of the number of flat 64×64 tiles.

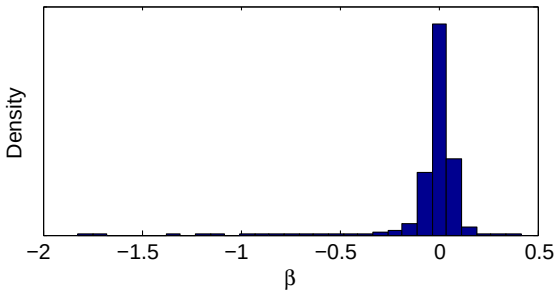

Fig. 5. Distribution of β, for all images combined. Tile size is 64×64.

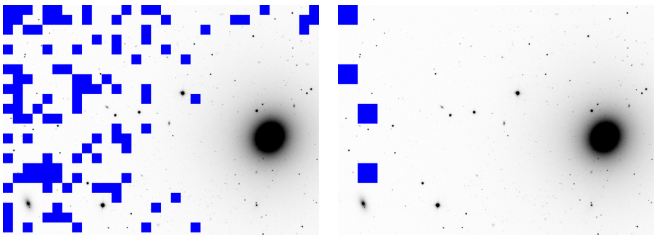

Fig. 6. Image with β outliers. Left: 64×64 flat tiles shown in blue; right: 128×128 flat tiles shown in blue. SDSS file: `fpC-003836-r4-0249.fit`

bias are object-like fluctuations in the background. However, such fluctuations would not be moving with stars and galaxies and would appear as artefacts. When α is too close to 1 the tile size decreases which results in more bias from objects. When α is too close to 0 more tiles are accepted which also results in more bias from objects. Figure 3 shows results for various settings of α. Considering the standard deviation of the noise at the background is approximately 5.4 for most images, there is not much difference between $\alpha = 10^{-7}$, $\alpha = 0.05$ and $\alpha = 0.5$: α is kept at 0.05. All images in the data set have 64×64 flat tiles. The minimum tile width w_0 is set to 64.

3. IS A CONSTANT A GOOD FIT?

An important question is whether the constant background gives a good model fit. Let μ_F be a statistic representing the mean of a flat tile, $\hat{\mu}_{\text{bg}}$ the background estimate (the hat indicates an estimate) and $\hat{\sigma}^2_{\text{bg}}$ the estimate of the noise variance at the background. If the background is flat, and the flat tiles are not biased by objects, $\mu_F \sim N(\mu_{\text{bg}}, w^{-1}\sigma_{\text{bg}})$, where w is the tile width, and let $\beta = (\hat{\mu}_{\text{bg}} - \mu_F)\hat{\sigma}^{-1}_{\text{bg}}$. β approximately $\sim N(0, w^{-1})$ with $\hat{\mu}_{\text{bg}}$ relatively constant. When $w = 64$, 95% of the absolute β values would be below 0.031 on average. In Figure 5 this is clearly not the case. 95% of the absolute β values are below 0.14. If changes in the background are the main cause (the background is not flat), a different fit closer to the local estimates could be better. The variations in the local estimates are still small, considering most detected objects contain pixel values above 3 ($\times$ standard deviation of the noise). Images with outliers are inspected to determine the cause. The background estimates in the image in Figure 6 at 64×64 tile size and 128×128 are 1137.1 and 1135.9 respectively, with $\sigma_{\text{bg}} \approx 5.4$, which shows the influence of the large object on the local estimates. The images in Figure 7, which also have been picked to inspect β outliers, have a similar situation. The main cause of the relatively large absolute values of β appears to be objects, not changes in the background. Experimentally, we verified that a non-constant background fit closer to the local estimates would increase the error at locations correlating with objects. For this data set, and these local background estimates, a constant is a good fit.

4. COMPARISON WITH SEXTRACTOR

SExtractor uses bi-cubic interpolation between local background estimates found by iteratively clipping pixel values above $3\times$ the sample standard deviation and recalculating the sample mean. SExtractor uses 64×64 tiles by default. Figure 8 shows that the constant estimate suffers less from object bias on average. The background estimate by SExtractor is 1.27 higher on average which corresponds approximately to $0.23\sigma_{\text{bg}}$, using $\sigma_{\text{bg}} \approx 5.4$. An image-sized tile in SExtractor also reduces bias, on average, but the higher noise standard

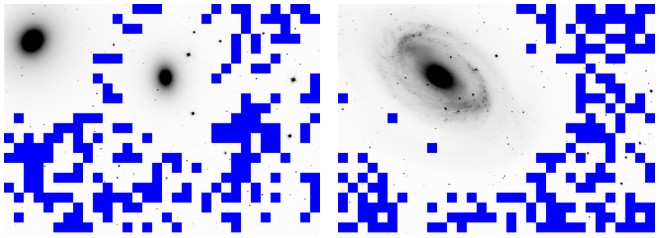

Fig. 7. Images with β outliers. 64×64 flat tiles shown in blue. Left: SDSS file `fpC-005313-r1-0067.fit`; right: SDSS file: `fpC-005116-r5-0148.fit`

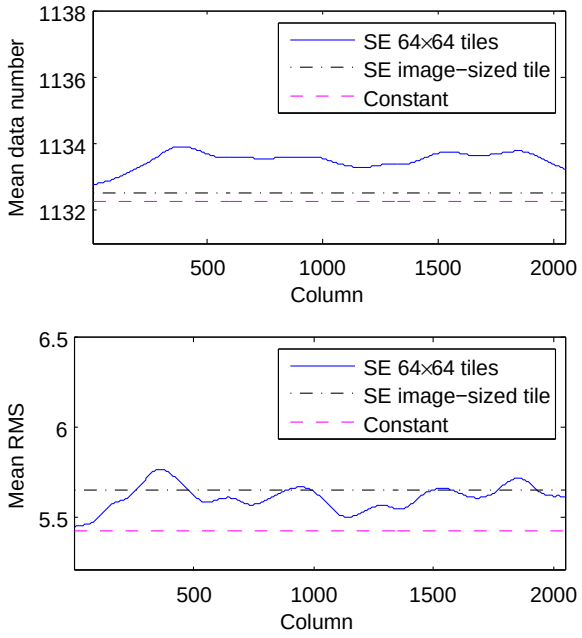

Fig. 8. Average estimate of the background (top) and noise standard deviation at the background (bottom) for all images, column-wise.

deviation, compared to the (other) constant, indicates a worse fit. The background estimate by SExtractor correlates with objects, see Figure 9, making it more difficult to detect (parts of) objects after subtraction. For example the connection between the merging galaxies in Figure 1(d) is more clear than in Figure 1(c). Another problem in the background estimation by SExtractor due to correlation with objects is distortion of object shapes, as seen in Figure 10, which appears to happen for every non-constant estimate. With the goal of having the least object bias and preserving object shapes, using the constant background estimation is clearly better.

5. CONCLUSIONS AND FUTURE WORK

The results show that object bias is clearly reduced in our method, in comparison to SExtractor, although a slight positive bias is still expected. We feel this is better than a negative bias, which could lead to many false positive detections. There are of course many other methods to which our approach should be compared before we can conclude that this is the best way forward, including recent techniques based on sparse representations and wavelets [9]. However, the improvement on the SExtractor scheme in this application is evident, and we are currently using it within an improved statistical attribute filtering framework for faint object detection [4]. The new version shows far better detection of faint structures than SExtractor, as was shown in Figure 1(e) and (f). Full results are in [4].

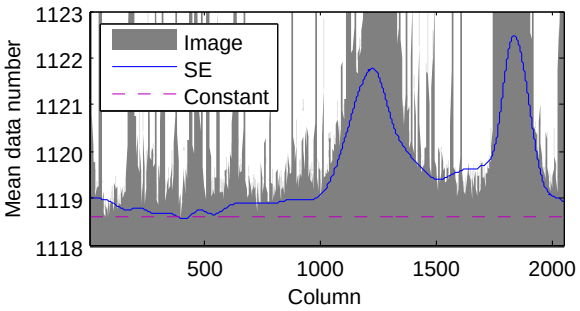

Fig. 9. Background estimate by SExtractor compared to the constant estimate, averaged column-wise, for Figure 1(a).

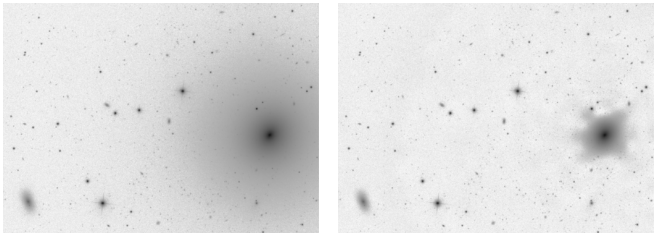

Fig. 10. Left: Constant background estimate, subtracted from the image. Right: SExtractor background estimate, subtracted from the image. Logarithmic scale. SDSS file `fpC-003836-r4-0249.fit`.

Future work would include different shapes (instead of squares) of areas representing the background, which could give better local estimates. One possible way is to start with tiny square flat tiles and iteratively merge similar sized areas if they pass a flatness test. We are also considering extensions to non-flat background estimates for data sets which require this. One option is using weighted (by range) k-nearest neighbours. This works even if only a few local background estimates are available and, depending on k, is less affected by outliers than bi-linear or bi-cubic interpolation.

6. REFERENCES

[1] C. Stoughton, R. H. Lupton, M. Bernardi, M. R. Blanton, S. Burles, F. J. Castander, A. J. Connolly, D. J. Eisenstein, J. A. Frieman, G. S. Hennessy, et al., "Sloan digital sky survey: early data release," *The Astronomical Journal*, vol. 123, no. 1, pp. 485, 2002.

[2] K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. A. Prieto, D. An, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, et al., "The seventh data release of the sloan digital sky survey," *The Astrophysical Journal Supplement Series*, vol. 182, no. 2, pp. 543, 2009.

[3] E. Bertin and S. Arnouts, "Sextractor: Software for source extraction.," *Astronomy and Astrophysics Supplement Series*, vol. 117, pp. 393–404, 1996.

[4] P. Teeninga, U. Moschini, S. C. Trager, and M. H. F. Wilkinson, "Improved detection of faint extended astronomical objects through statistical attribute filtering," in *Proc. ISMM 2015*, 2015, Submitted.

[5] P. Teeninga, U. Moschini, S. C. Trager, and M. H. F. Wilkinson, "Bi-variate statistical attribute filtering: A tool for robust detection of faint objects," in *11th International Conference "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-11-2013)*, 2013, pp. 746–749.

[6] J. E. Gunn, M. Carr, C. Rockosi, M. Sekiguchi, K. Berry, B. Elms, E. De Haas, Ž. Ivezić, G. Knapp, R. Lupton, et al., "The Sloan digital sky survey photometric camera," *The Astronomical Journal*, vol. 116, no. 6, pp. 3040, 1998.

[7] sdss.org, "Photometric flux calibration," Published online: http://www.sdss2.org/dr7/algorithms/fluxcal.html.

[8] R. B. D'Agostino, A. Belanger, and R. B. D'Agostino Jr, "A suggestion for using powerful and informative tests of normality," *The American Statistician*, vol. 44, no. 4, pp. 316–321, 1990.

[9] E. Martnez-Gonzlez, J. E. Gallegos, F. Argeso, L. Cayn, and J. L. Sanz, "The performance of spherical wavelets to detect non-gaussianity in the cosmic microwave background sky," *Monthly Notices of the Royal Astronomical Society*, vol. 336, no. 1, pp. 22–32, 2002.