MATHEMATICAL MORPHOLOGY APPLIED TO THE SEGMENTATION AND CLASSIFICATION OF GALAXIES IN MULTISPECTRAL IMAGES

E. Aptoula, S. Lefèvre and C. Collet

LSIIT UMR-7005 CNRS, Louis Pasteur University
Pôle API, Blvd Sébastien Brant, PO Box 10413, 67412 Illkirch, France
{aptoula, lefevre, collet}@lsiit.u-strasbg.fr

ABSTRACT

The automated segmentation and classification of galaxies still constitute open problems for astronomical imaging, mainly due to their fuzzy and versatile nature, as well as to the multitude of the available channels. In this paper, a mathematical morphology based approach is explored. First, a semi-automated method for multispectral galaxy segmentation, based on the marker controlled watershed transformation is proposed. Moreover, a novel and viewpoint independent morphological feature, based on the top-hat operator, is introduced for the distinction of spiral from elliptical galaxies. Illustrative application examples of the presented approach on actual images are also provided.

1. INTRODUCTION

The mathematical morphology (MM) theory constitutes a powerful image processing framework based on geometry and the theories of orders and sets. As a methodology, it is currently employed in various areas such as medical and geological image processing, automated industrial inspection, etc. However, despite its diversity of operators and application areas, the use of the morphological framework in the domain of astronomical imaging has been relatively limited.

Specifically, it was used for compression, particularly by means of skeletonization [5, 11]. Other important applications include the segmentation of multi-object astronomical images, realized in [9] with the help of a watershed transformation followed by region merging. Furthermore, another morphological tool, the pattern spectrum, has also found use in the astronomical context, by assisting the discrimination of stars from galaxies [6]. As far as noise reduction is concerned, opening and closing based filters have been explored in [12, 13, 14, 15]. Nevertheless, with the exception of [9] (2004), the aforementioned efforts date to the mid 90s and consequently do not take into account recent developments; besides, they are also all limited with monospectral images.

The work presented here, elaborates on a morphological approach to the problems of galaxy detection and subsequent classification, in a more general attempt to also underline the potential of MM within astronomical imaging. More precisely, a semi-automated morphology based method, designed with the end of detecting galaxies in single-object multispectral input is described. Moreover, as an important step toward obtaining a Hubble classification of galaxies, a viewpoint independent morphological feature based on the top-hat operator, is proposed for the distinction of spiral from elliptical galaxies. Test results of the overall process on actual galaxies are also provided.

The sequel of the paper is organized as follows: the next section details the problem context. Then, section 3 recalls briefly the definitions of certain morphological operators before elaborating on the proposed segmentation method, whereas section 4 describes the morphological feature employed for classification purposes. Finally section 5 is devoted to concluding remarks.

2. THE PROBLEM CONTEXT

The constantly increasing resolution of image acquisition systems provides the astronomical community with massive amounts of data containing numerous new celestial objects, of which galaxies are of special interest to the astronomers. Added the fact that multispectral images have also become widely available, the need for efficient and automated image processing methods has become greater than ever.

Most usually galaxies are extracted from the large sky maps together with their surroundings (figure 1). Yet, from an image processing point of view, the main interest of the astronomical community lies in their classification. To this end, a precise detection of the object's pixels along with pertinent classification features become indispensable. Considering additionally that far galaxies appear usually only a few tens of pixels wide, as well as the general nature of astronomical images characterized by faint objects on a noisy background, the overall task becomes at least challenging.

So far, mainly statistical methods have been employed for this purpose. However, besides being computation-intensive they also tend to concentrate on individual pixel values, while the multivalued structure of the input poses additional extension difficulties. That is why alternative processing options, such as the MM theory, have been considered. In particular, besides being computationally efficient, the main advantage of morphological operators lies in their inherent ability to exploit the spatial relationships of pixels. Moreover, they can also adapt to vectorial strategies in order to process multivariate data.

Figure 1: Galaxy HDF4-378 (left to right: 450, 814 and 1100nm)

Although the images of the same object acquired from different frequency bands provide a rich source of information, there are a couple of points that need to be taken into account. First, as in the case of HDF4-378 illustrated in figure 1 significant variations in the general noise level among

the available channels is quite possible. And secondly, there is no guarantee that all parts of a given galaxy will have a non-zero response in every wavelength; meaning that the object can be partially absent from one or more channels. The following chains of processing were directly influenced by these two properties.

3. SEGMENTATION

In this section the steps taken during the proposed segmentation method are described. We begin by recalling the definitions of some basic greyscale morphological operators. For the theory of MM the reader should refer to [16, 17].

3.1 Preliminary definitions

Let $f : E \to T$ be a greyscale image, with E the discrete coordinate grid while T represents the set of possible grey values. In the present case, a multispectral image of n channels is considered as $\mathbf{f} : \mathbb{Z}^2 \to \mathbb{Z}^n$ with $f_i : \mathbb{Z}^2 \to \mathbb{Z}$ denoting its i-th channel.

In short, MM studies the transformations of an image, when it interacts through operators with a matching pattern B, called *structuring element* (SE). Throughout the sections that follow only *flat* SEs are employed, in other words B is considered as a subset of E. Consequently, the definitions of the two basic morphological operators, *dilation* and *erosion*, become respectively:

$$\delta_B(f)(x,y) = \bigvee_{(r,s)\in B} f(x-r,y-s), \ (x,y) \in E \quad (1)$$

$$\varepsilon_B(f)(x,y) = \bigwedge_{(r,s)\in B} f(x+r,y+s), \ (x,y) \in E \quad (2)$$

A multitude of operators is then derived from dilation and erosion; such as *opening* and *closing*, used extensively with smoothing ends and defined respectively as:

$$\gamma_B(f) = \delta_B(\varepsilon_B(f)) \quad (3)$$

$$\phi_B(f) = \varepsilon_B(\delta_B(f)) \quad (4)$$

Opening and closing form in their turn the basis of more advanced morphological tools (e.g. granulometry).

3.2 Preprocessing

The aim of the segmentation process is to extract the galaxy, in other words to obtain a monospectral image with two labels (object and background). Given the multivalued input, several morphological processing strategies can be considered [7, 10]; for instance the marginal approach, which consists in processing independently each channel, the vectorial approach that processes all channels simultaneously usually by means of vector orderings, or even reducing the multivariate data into a single channel by means of fusing.

In the present case, it was chosen to employ the marginal approach. Hence, the operators are applied independently to each channel of the multispectral input. Despite the additional computational burden, the advantage of having the entire set of morphological tools available for use with no need for special adaptation, is in fact invaluable. Moreover, the inter-channel information is not wasted either, as merging operations of different types take place before the final segmentation step.

For an initial noise reduction the OCCO (open-close close-open) filter is employed. Although it affects the object borders, the level of detail preservation is nevertheless at a satisfactory level.

$$\mathrm{OCCO}_B(f) = \left\lfloor \frac{1}{2}\gamma_B(\phi_B(f)) + \frac{1}{2}\phi_B(\gamma_B(f)) \right\rfloor \quad (5)$$

In this section as in the following all SEs are to be considered as 5x5 disk approximations, unless otherwise specified.

3.3 Watershed transformation

The segmentation process, or more likely "detection" because of the presence of a single galaxy, is realized in our case by means of the watershed transformation [4]. The principle of this powerful and unsupervised segmentation tool, based on geodesic operators, becomes most clear in terms of a flooding simulation of the topographic relief of the input, which must be processed in such a way that the regions of interest correspond to dark areas. Supposing that water comes through the regional minima, the formed basins would fill up. And whenever two distinct basins are about to merge, a dam is built to prevent it. After the entire relief is flooded, the resulting dam lines provide the sought borders.

In practice however, the high number of image minima usually yields an oversegmentation. In order to overcome this eventual problem, the marker controlled version of the transformation was chosen. Hence, as the flooding is realized only from the selected "markers" (each marker being a set of connected pixels), the result is bound to have as many regions as initial markers. Besides being also unsupervised, the use of markers introduces the additional advantage of integrating a priori knowledge into the process.

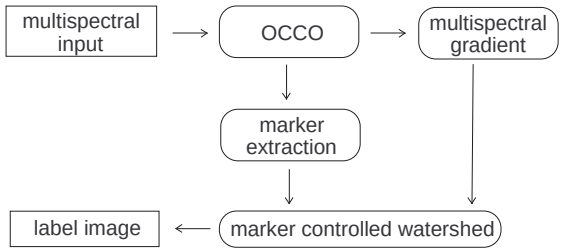

Figure 2: Outline of the segmentation method

Therefore the core of the problem consists in extracting the two markers, denoting the galaxy and its exterior, from the smoothed input. The main challenge however, lies in the choice of the internal marker, representing the galaxy; which has to be not only large enough to allow the flood water to overcome the internal complexity of the galaxy, so as not to result in a subset of the sought object, but also small enough to be restrained inside it. While figure 2 illustrates the principle of the entire method, the outline of the proposed marker extraction scheme is given in figure 3.

First, given the i-th channel f_i, an estimation of the object's whereabouts is obtained. Thanks to the distinctive brightness of the center of the galaxy (i.e. bulge), the center of gravity provides an easy way of marking the object. Considering that the background pixels are not only in majority but also relatively close in value after the smoothing step,

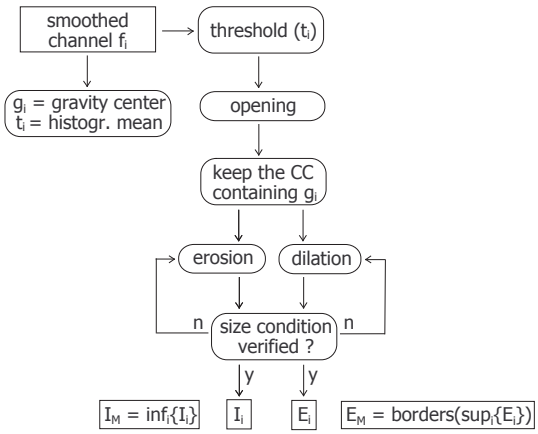

Figure 3: Outline of the marker extraction method, I_M and E_M denoting respectively the internal and external markers

the threshold employed during the binarization that follows is chosen as the arithmetic mean of the channel's histogram.

An opening is then applied to the binary result with the end of eliminating any eventual excessive parts. After which the connected component marked by the gravity center is extracted (CC_i). At this point CC_i is expected to represent a "rough" approximation of the sought galaxy. In order to evaluate the level of approximation, we exploit the general elliptical shape of galaxies and calculate the convexity of CC_i, assuming that high concavity corresponds to inadequate approximations of the actual galaxy.

The convex hull CH is computed by intersecting the half-planes (in 16 directions) containing the extracted component. As defined in [18]:

$$\mathrm{CH}(X) = \bigcap_{\theta} \left[\phi_{\pi_\theta^+}(X) \cap \phi_{\pi_\theta^-}(X) \right] \tag{6}$$

with π_θ^+ the half plane of direction θ and π_θ^- its complement. Hence, the convexity c is simply the surface (i.e. number of white pixels) ratio of the extracted component to its convex hull:

$$c = \frac{\mathrm{surface}(CC_i)}{\mathrm{surface}(\mathrm{CH}(CC_i))} \tag{7}$$

with $0 < c \leq 1$.

Given c as an evaluation measure of CC_i, we employ a number of successive dilations (δ_b^k) and erosions (ε_b^k) in order to compensate for an eventually inadequate threshold. More precisely, in the case of the internal marker, CC_i is eroded, whereas for the external marker it is dilated until the expressions (8) and (9) are respectively verified:

$$\frac{\mathrm{surface}(\varepsilon_b^k(CC_i))}{\mathrm{surface}(CC_i)} \leq c \tag{8}$$

$$\frac{\mathrm{surface}(\delta_b^k(CC_i))}{\mathrm{surface}(CC_i)} \geq (2-c) \tag{9}$$

Finally, the marginal binary results of dilation(s) are merged by means of a supremum, and then its borders are extracted as external markers. The supremum ensures that the maximum available area from all channels is taken, and so the galaxy is most probably entirely inside the obtained marker. On the other hand, the infimum of the results of erosion(s) is used. Of course this last step renders necessary the presence of at least some part of the galaxy in all channels; however taking the intersection of the eroded areas from each channel not only decreases the possibility of having an excessive internal marker but also relaxes the effect of an inadequate threshold due to high noise level. Consequently, for highly concave threshold results, representing inaccurate approximations of the sought borders, several erosions and dilations take place, in order to increase the distance between the markers thus compensating for the threshold error. Likewise, for high values of c, CC_i is hardly modified as it is assumed to already represent relatively well the sought borders. The obtained markers are then employed as minima for the supremum based multispectral gradient:

$$\nabla_B^{sup}(\mathbf{f}) = \vee [g_B(f_1), \ldots, g_B(f_n)] \tag{10}$$

with g_B being the standard morphological gradient:

$$g_B(f) = \delta_B(f) - \varepsilon_B(f) \tag{11}$$

which yields a result topographically suitable for a watershed transformation. And the flooding that follows provides the sought borders.

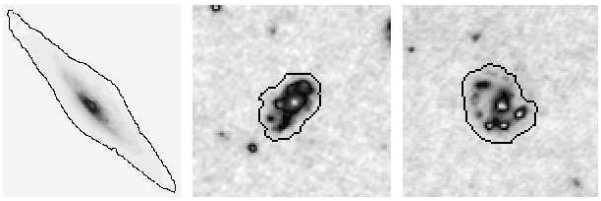

Figure 4: Detection results for NGC 2683, HDF4-378, HDF4-550 (left to right)

Judging from the results illustrated in figure 4, although the quality of the borders can be considered satisfactory the important remark concerns the robustness and adaptability of the method to different noise levels and galaxy types; which is underlined by the absence of any direct user intervention, with the exception of the structuring element form and size choice. With no optimization in place, for an image of 313×313 pixels with 2 channels the segmentation process takes just over 6 seconds with a Java based implementation on an Intel 2.1GHz system equipped with 512MB of memory. Nevertheless, considerable optimizations can be realized and will be the subject of future work.

4. CLASSIFICATION

Having obtained the pixels belonging to the galaxy, in this section the focus is on their exploitation with the aim of classifying the galaxy by morphological methods, according to the only overall scheme still in widespread use, the one proposed by E. Hubble in the early 1900s (figure 5). With regard to Hubble's classification scheme, galaxies are grouped roughly as *elliptical* (E0 to E7), *spiral* with (SBa, SBb, SBc) or without bars (Sa, Sb, Sc) and *irregular* (Irr). Spectral features and brightness profiles have long been employed for the classification of galaxies in combination with several statistical and data analysis methods (e.g. matched filters, wavelets,

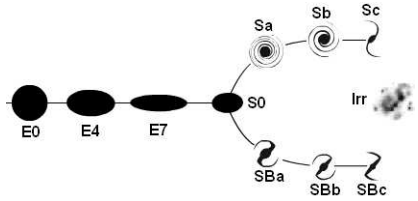

Figure 5: The Hubble tuning fork

shapelets, neural networks, etc). Extensive details on the properties of each galaxy group and on the methods of classification can be found in [1, 2, 3, 8, 19].

Yet, spectral features concentrate on individual spectral signatures, which becomes a drawback given the fact that the distinction among the galaxy classes depicted in figure 5 is due mainly to shape based properties (spirals, bars, etc). That is why the morphological framework, thanks to its ability to exploit the spatial relationships of pixels, constitutes a complementary approach to spectral methods. Nevertheless, as far as shape-based features are concerned, unless the viewpoint of the galaxies is taken into account, it can present a serious obstacle. For instance, if the only distinctive feature is ellipticity, an E0 elliptical galaxy can easily be misclassified as E7 when viewed from edge-on. A handicap which renders unsuitable several powerful morphological tools that otherwise would have contributed considerably to the resolution of the problem (e.g. skeletonization).

In the light of these facts, as a first step toward a complete Hubble classification based on MM methods, the problem of the distinction of elliptical from spiral galaxies was examined. From an astronomical point of view only spiral galaxies, either with bars or not, have active star formation sites whereas old and cold stars are dominant in elliptical galaxies. Thus, the number of these sites can be considered as a pertinent feature. Furthermore, a distinctive characteristic of star formation sites and young stars is their relatively high brightness level in the infrared domain, as a result they correspond to peaks in the topographic relief of the input. Moreover, these high brightness sites continue to be visible independently from the actual view angle of the galaxy. An important morphological operator capable of extracting such peaks is *top-hat* (TH); which is defined as the difference between the input and its opening:

$$\mathrm{TH}_B(f) = f - \gamma_B(f) \tag{12}$$

Specifically it extracts the contrasted components with respect to the background. For the same reason given in section 3.2, the peak extraction method is applied in a marginal fashion (figure 6). More precisely, after having segmented the multispectral input only the pixels belonging to the detected galaxy are kept. Next, in order to decrease the perturbation caused by noise, the image is smoothed with a *median* filter.

The median belongs to a more general class called "rank filters". Given a rank filter of rank k along with a SE, in order to calculate its output, the pixels contained within the SE, centered on the considered pixel, are first sorted in ascending order. Then the output of the filter is simply the pixel at the k-th position. Specifically, if $\rho_{B,k}$ represents the rank filter of rank $k \in \{1,2,\ldots,n\}$ with a SE B and $n = card(B)$, then the median becomes $\rho_{B,(n+1)/2}$, with n being odd.

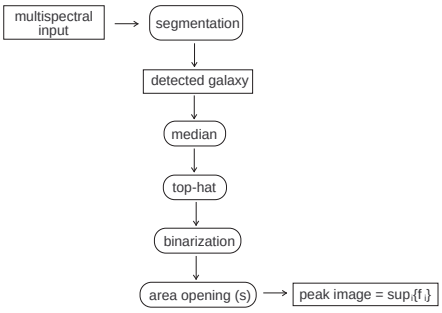

Figure 6: Outline of the peak extraction method

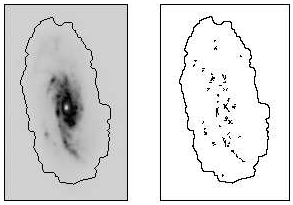

Figure 7: NGC 3672: (left) the detected galaxy and (right) the detected peaks (s = 2)

Subsequently the top-hat operator is applied, after which the result is binarized by considering as dark the zero valued pixels. The area opening γ^a that follows, eliminates all components having a number of pixels inferior or equal to a user defined limit s:

$$\gamma_s^a(f) = \bigcup_{B \in \mathcal{B}} \gamma_B(f) \tag{13}$$

with $\mathcal{B}$ being the set of all SEs that are connected and with an area equal to s. As this area argument aims to remove components due to noise, it should be chosen according to the noise level as well as size of the galaxies in question, which is already available since the detection process. At this point, the marginal results are combined by means of a supremum, in order not to miss the peaks detected only in certain channels; which on the other hand also increases the number of false positives.

Given the result illustrated in figure 7, several actual measures of star formation activity can be considered; in our case we chose to employ the number of connected components. The classification based on this *single* feature that followed, as well as the preceding detection process were tested on a total of 64 images[1] of mostly nearby bright galaxies, purely elliptical or spiral and in their majority with relatively negligible noise level; of which 45 were observed on 2 channels, 16 on 3 and 3 on exactly 6 channels, recorded with charge coupled devices in various wavelengths between optical and near infrared. More precisely, their sizes varied from 101×101 to 561×561 pixels, with 8 bits per channel. The classification was realized with the help of the k-means algorithm (random seeds, 2 classes) and the area parameter was set to $s = 2$.

Of course the number of subjects is relatively very small but sufficient for the purpose of illustration. The resulting accuracy was 79% (table 1), computed as the fraction formed

[1]The images were kindly provided by the Strasbourg Astronomical Observatory (http://astro.u-strasbg.fr/)

Actual \ Predicted	Spiral	Elliptical	Total
Spiral	43	10	53
Elliptical	3	8	11
Total	46	18	64

Table 1: Classification based on peak count

by the number of succesful classifications divided by the total number of subjects. The classification errors were mainly due to the varying noise level of the input and to normalization issues. Specifically, as the top-hat operator removes all objects larger than the SE in use, a normalization step is necessary in order to take into account the varying galaxy sizes.

5. CONCLUSION AND PERSPECTIVES

In this work, the preliminary results of an application of the morphological framework on the problems of segmentation and classification of galaxies in multispectral images was presented. Besides pointing out the potential of the MM theory as far as astronomical imaging is concerned, it also deals with issues of multivalued morphological processing. The marginal approach was preferred throughout this work; however it can be disadvantageous if the channels are significantly correlated. In which case decorrelating transformations or vectorial morphological processing strategies should be considered.

The proposed segmentation method, free of an eventual oversegmentation, depends uniquely on the chosen markers hence allowing the insertion of a priori knowledge into the procedure. Automation and execution speed were prioritized, hence user intervention is limited with the choice of the SE form and size. The method is based on the fact that galaxies are generally elliptical objects and thus highly convex, which is used as an auto-evaluation scheme during marker extraction; nevertheless not all galaxies verify this assumption (e.g. irregular). Along with this last problem which remains to be investigated, a more elaborate study with a larger image set, concentrating particularly on precision and robustness is also necessary.

The morphological feature that was subsequently presented, constitutes a first step toward a complete Hubble classification of galaxies by means of morphological methods. The extracted number of peaks represents the number of star formation sites in galaxies, and it is this firm physical property that asserts its pertinence as a classification feature for distinguishing elliptical from spiral galaxies. Furthermore, its relative independence from the view angle constitutes an important advantage; future work will concentrate particularly on exploiting the spatial organization of the obtained pixels. However, the size of the SE in use continues to be the cause of a normalization issue which can be resolved by choosing it proportionally to the size of the detected galaxy.

In conclusion, despite the discouraging fuzziness of the astronomical data, this work shows once more that the morphological framework can be used in this application domain as an invaluable complementary approach to spectral and statistical methods.

REFERENCES

[1] R. G. Abraham, F. Valdes, H. K. C. Yee, and S. van den Bergh. The morphologies of distant galaxies. 1: an automated classification system. *Astrophysical Journal*, 432:75–90, September 1994.

[2] A. Adams and A. Woolley. Hubble classification of galaxies using neural networks. *Vistas in Astronomy*, 38:273–280, 1994.

[3] M. A. Bershady, A. Jangren, and C. J. Conselice. Structural and photometric classification of galaxies. I. Calibration based on a nearby galaxy sample. *Astronomical Journal*, 119:2645–2663, June 2000.

[4] S. Beucher and F. Meyer. The morphological approach to segmentation: the watershed transformation. In E. R. Dougherty, editor, *Mathematical Morphology in Image Processing*, pages 433–482. Dekker, New York, 1993.

[5] A. Bijaoui, Y. Bobichon, and L. Huang. Digital image compression in astronomy: morphology or wavelets. *Vistas in astronomy*, 40(4):587–594, 1996.

[6] A. J. Candeas, U. B. Neto, and E. C. Filho. A mathematical morphology approach to the characterization of astronomical objects. *Journal of the Brazilian Computer Society*, 3(3), April 1997.

[7] M. Comer and E. Delp. Morphological operations for color image processing. *Journal of Eletronic Imaging*, 8(3):279–289, July 1999.

[8] A. J. Connolly, A. S. Szalay, M. A. Bershady, A. L. Kinney, and D. Calzetti. Spectral classification of galaxies: an orthogonal approach. *Astronomical Journal*, 110:1071–1082, September 1995.

[9] M. Frucci and G. Longo. Watershed transform and the segmentation of astronomical images. In *Proceedings of Astronomical Data Analysis III*, Naples, Italy, April 2004.

[10] J. Goutsias, H. J. A. M. Heijmans, and K. Sivakumar. Morphological operators for image sequences. *Computer Vision and Image Understanding*, 62(3):326–346, November 1995.

[11] L. Huang and A. Bijaoui. Astronomical image data compression by morphological skeleton transformation. *Experimental Astronomy*, 1(5):311–327, 1991.

[12] S. M. Lea and L. A. Kellar. An algorithm to smooth and find objects in astronomical images. *The Astronomical Journal*, 97(4):1238–1276, April 1989.

[13] M. Lybanon, S. M. Lea, and S. M. Himes. Segmentation of diverse image types using opening and closing. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, volume I, pages 347–351, Jerusalem, Israel, 1994.

[14] M. Maccarone. Fuzzy mathematical morphology: concepts and applications. *Vistas in astronomy*, 40(4):469–477, 1996.

[15] M. Maccarone, M. Tripiciano, and V. Di Gesu. Fuzzy mathematical morphology to analyse astronomical images. In *Proceedings of the IAPR-IEEE 11th International Conference on Pattern Recognition (ICPR'92)*, volume 3, pages 248–251, The Hague, The Netherlands, September 1992.

[16] J. Serra. *Image Analysis and Mathematical Morphology Vol I*. Academic Press, London, 1982.

[17] J. Serra. *Image Analysis and Mathematical Morphology Vol II: Theoretical Advances*. Academic Press, London, 1988.

[18] P. Soille. Grey scale convex hulls: definition, implementation, and application. In *Proceedings of the International Symposium on Mathematical Morphology and its Applications to Image and Signal Processing*, pages 83–90, Amsterdam, The Netherlands, June 1998.

[19] S. van den Bergh. *Galaxy Morphology and Classification*. Cambridge University Press, New York, May 1998.