



Universiteit
Leiden
The Netherlands

Image Analysis and Platform Development for Automated Phenotyping in Cytomics

Yan, K.

Citation

Yan, K. (2013, November 27). *Image Analysis and Platform Development for Automated Phenotyping in Cytomics*. Retrieved from <https://hdl.handle.net/1887/22550>

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/22550>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/22550> holds various files of this Leiden University dissertation.

Author: Yan, Kuan

Title: Image analysis and platform development for automated phenotyping in cytomics

Issue Date: 2013-11-27

Chapter 2

Robust Image Segmentation for Cytomics

This chapter is based on the following publications

Yan, K., & Verbeek, J. F. (2012). Segmentation for High-throughput Image Analysis: Watershed Masked Clustering. *Proc. of ISoLA 2012*. LNCS 7610. Springer Berlin-Heidelberg 2012, pp. 25-41

Chapter Summary

Image segmentation is a crucial image analysis procedure in HT/HC screen study. It is often considered the first substantial step to convert image data into quantifiable elements. However, the selection of a proper segmentation algorithm is a nontrivial task. Often it requires an evaluation of both image quality and algorithm trait. To answer this question, this chapter illustrates the design of an innovative segmentation algorithm and its performance together with the assessment of several popular segmentation algorithms that have been used for HT/HC screen studies. The comparison provides an overview of the segmentation accuracy of each algorithm and a systematic assessment of the limitations of these algorithms. As a result, the comparison confirms that the watershed masked clustering (WMC), hysteresis threshold, and the level-set are good algorithms for image segmentation in HT/HC screen studies. The WMC algorithm shows the best performance with the selected HT/HC benchmark image sets.

2.1 Introduction

Image segmentation is defined as the procedure partitioning an image into multiple regions. It is often considered the most crucial step when converting image data into quantifiable elements [6][7][40][41]. Image segmentation algorithms are essentially built on two basic properties of image intensity: discontinuity and similarity. Algorithms relying on different discontinuity or similarity criteria can be further divided into the following types [42]: the edge based [43][44] and partial differential equation (PDE) [45] based algorithm are two major branches of the discontinuity-based algorithms, threshold- [46] and region-growing based algorithms are two major branches of the similarity-based algorithms.

These two branches of segmentation algorithms have proven their applicability in HT/HC screen studies. Threshold and region growing based algorithms are often employed in the image analysis of fluorescent and phase contrast microscopy due to its low computational cost and high convergence speed. However, region growing based algorithms [47] are vulnerable to local intensity variations when the size of object (foreground) is in the same range as the image noise (background).

The edge based and PDE based algorithms are used in magnetic resonance imaging (MRI) [48][49], fluorescent microscopy imaging [50][51], and phase contrast imaging. Algorithms in these two branches are considered local adaptive optimization procedures to evolve an initial curve towards the lowest potential of a cost function [52][53]. As a result, the design of a cost function often requires thorough understanding of image characteristics and cannot be easily adapted to exceptions.

Recent developments in imaging technology allow multi-channel visualization of cell structures using different microscope modalities, therefore it is important to estimate the image characteristics before selecting a proper segmentation algorithm. Our empirical study suggests that generic segmentation algorithms often prove less efficient when processing HT/HC screen images. Compared to standard imaging, quality of bio-imaging in HT/HC studies often suffers from technological complexities and experiment instabilities. In high-throughput imaging settings, these issues are the challenges to be resolved [54][55][56][57][58]. Some of the complexities may be corrected during the imaging by changing the mechanics of acquisition. Some may only be determinable after the imaging. Thus, it often relies on further digital image processing to compensate the image quality before performing segmentation. However, the processing does not necessarily guarantee an improvement in image quality.

As a result, a good selection of segmentation algorithm must be justified by heuristics; meaning to identify foreground without introducing background. Other than inventing a new principle of segmentation, here we propose a hybrid segmentation algorithm by merging existing segmentation algorithms to produce accurate masks. The **watershed masked clustering (WMC)** algorithm [9][10][29] improves segmentation efficiency by denying that all pixels may share similarity in intensity. Instead, it presumes that an image consists of a number of coarse regions in which all pixels may share the same intensity similarity. Unlike local-adaptive algorithms such as Bernsen's algorithm[59], the WMC iteratively trains only one threshold per coarse region using regional intensity values. Finally, the construction of each binary mask is tested based on morphological criteria, if necessary, a correction is imposed. By

considering both discontinuity and similarity of an image, WMC yields a significantly better performance in terms of both sensitivity and specificity. In addition, the WMC algorithm can preserve more morphological details such as cell protrusions.

In the following sections, existing HT/HC segmentation solutions together with the WMC algorithm will be introduced. Moreover, the robustness and applicability of all algorithms is estimated using ground truth image sets consisting of artificial objects and genuine HT/HC screen specimens.

2.2 Existing Segmentation Algorithms

In this section, the following algorithms are illustrated: Otsu's method [60], Bernsen's threshold [59], level-set method [45], hysteresis threshold [61], and watershed masked clustering [62]. Each algorithm is widely employed in biological image analysis. All algorithms are publicly available implementations as part of the the Fiji image analysis software [63][64]. The WMC algorithm is implemented using the ImageJ software [65].

2.2.1 Otsu's Method

In image processing, Otsu's method [60] is a threshold based segmentation algorithm that automatically performs image threshold optimization based on the histogram shape [66]. The algorithm assumes that the image contains two classes of pixels or a bi-modal histogram (e.g. foreground and background). It calculates the optimum threshold separating those two classes so that the combined spread (intra-class variance) is minimal [60]. Sometimes Otsu's method overlooks the intensity variation between individual objects (cf. Figure 2-1a). It is clear that Otsu's method can well detect objects of high intensity but cannot preserve detailed structure such as cell protrusions (cf. Figure 2-1b).

2.2.2 Bernsen's Threshold

Bernsen's Threshold or Local Method of Bernsen [59] is a local adaptive threshold-based segmentation algorithm. Instead of training a threshold for the whole image, the threshold is trained for each pixel using a certain similarity of the pixel neighborhood ω [34][60]. For example, a midrange based implementation calculates the threshold using the mean of the minimum $I_{low}(i, j)$ and maximum $I_{high}(i, j)$ gray value in a local window of a predefined kernel size ω in the image I [60]. In each neighborhood ω , the central pixel is labeled as either foreground or background when compared to the neighborhood threshold. However, if the contrast $C(i, j) = I_{high}(i, j) - I_{low}(i, j)$ is below a certain value, then that neighborhood is assumed to consist of only one class.

Compared to global thresholding solutions, Bernsen's threshold tries to adapt the threshold for each pixel. However, The Bernsen's threshold does not consider prior probabilities of two classes in the computation of the thresholds. It also does not adapt the window size when the sizes of objects vary. Therefore, Bernsen's threshold has limitations similar to Otsu's method. (cf. Figure 2-1c).

2.2.3 Level-set Segmentation

Level-set segmentation [45][67] is a segmentation algorithm based on partial differential equation (PDE). This segmentation algorithm is derived from the level-set method, which is

originally designed for tracking objects while later adapted to the application domain of image segmentation. The central idea of such adaption is to propagate a seed or a region contour inside an image with a propagation velocity that depends on similarity measurement, such as image gradient, until a boundary (discontinuity) is reached [68].

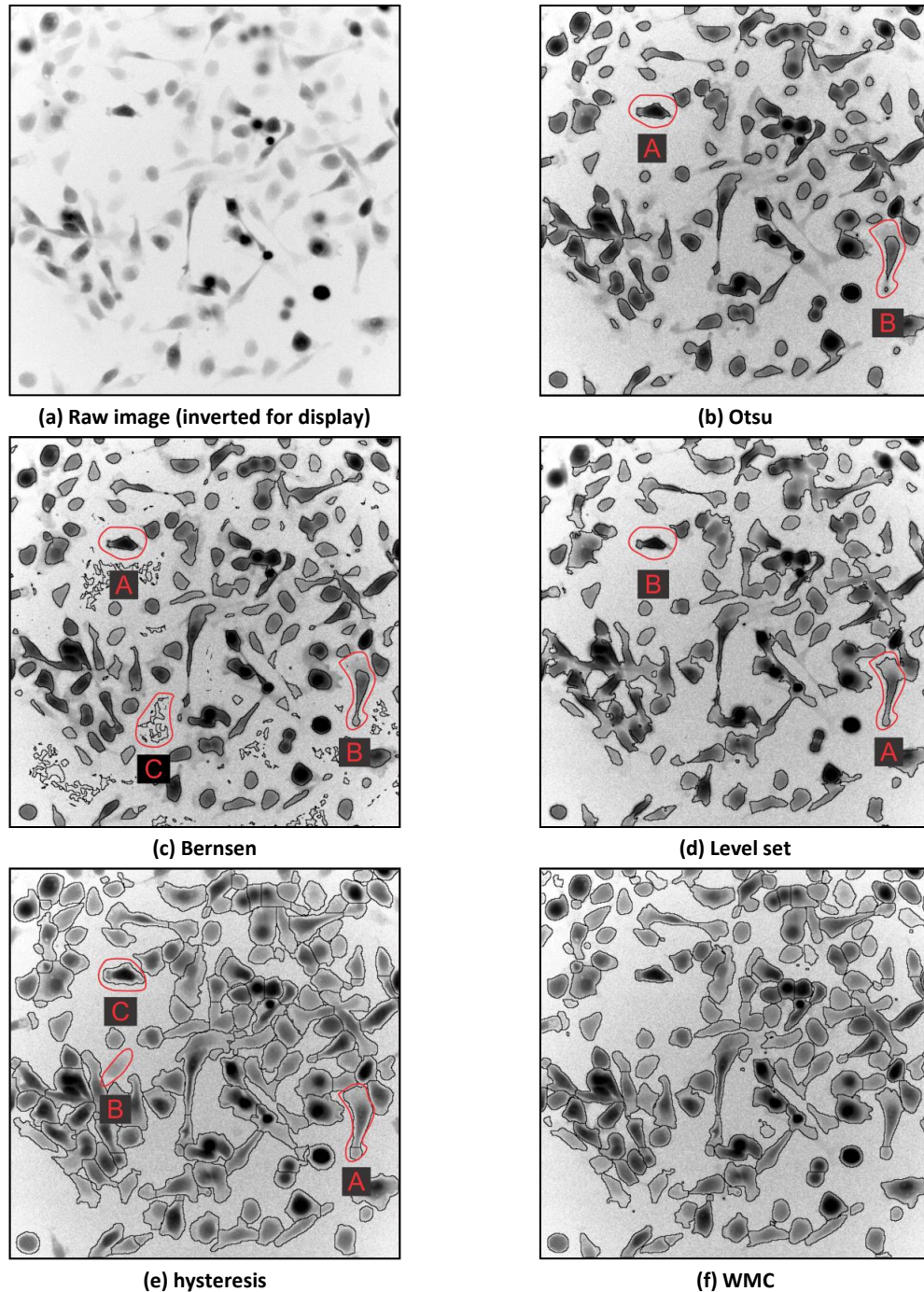


Figure 2-1 (a) Epifluorescence microscopy images from live tumor cells. These are used for research in migratory behaviors of cancer cells. (b) – (f) Segmentation of sample images using different segmentation algorithms [A]: correctly segmented, [B]: under-segmented, [C]: over-segmented

Compared to threshold-based algorithms, the level-set considers both similarity and discontinuity of an image. As a result, it results in better recognition of object details and more robust to intensity variation (cf. Figure 2-1d). Still, the robustness of the level-set method is depending on the direction estimation and stop criterion.

2.2.4 Hysteresis Thresholding

The hysteresis thresholding [44][61] is a bi-thresholding procedure typical employed in two class segmentation problem [69]. In hysteresis thresholding, image is first segmented by an upper threshold known as “high-edge” in order to obtain only the confidence object pixels. These pixels are guaranteed to be the true foreground pixel with a higher false negative ratio. Then a second threshold is introduced as a lower threshold known as “low-edge”, which obtains probable object pixels but with a higher false positive ratio. From the “low-edge” and “high-edge” pixels, the segmentation is achieved by connecting “high-edge” pixels with “low-edge” pixels using predefined kernel; i.e. linear or quadric.

The robustness of the hysteresis method is depending on the choice of both the low-edge threshold and the high-edge threshold. Existing studies also demonstrate the possible to perform an adaptive selection of both thresholds [43]. Compared to Otsu, Bernsen, and level set, the hysteresis thresholding method yields a significantly higher masking accuracy (cf. Figure 2-1e). However, the choice of both thresholds requires constant tuning due to variation between images which is not allowed in HT/HC image study. Moreover, due to the fixed thresholds, hysteresis method tends to undertrain the masking (cf. Figure 2-1e) and object containing only weak edges is still undetectable using hysteresis thresholding (cf. Figure 2-1e).

2.2.5 Watershed Masked Clustering Algorithm

The watershed masked clustering (WMC) algorithm is designed to be a robust and dedicated solution to the application domain of HT/HC studies [10][29][9][15][70] using fluorescence microscopy. The WMC algorithm consists of three sequential steps (cf. Figure 2-2b). At each step, the segmentation result is recursively refined based on image heuristics. (1) In the first step, the WMC algorithm starts by employing a region selection mechanism to find several coarse regions considered as a rough mask that requires further optimization. (2) In the second step, a more precise masking is obtained from each coarse region using machine learning: e.g. fuzzy C-means clustering. (3) In the final step, the refined masks are reassessed based on multiple phenotypical criteria and corrected by merging, if necessary. Following this workflow (cf. Figure 2-2), the WMC algorithm converts a multimodal optimization problem into a collection of local optimization problems.

Compared to other segmentation approaches, the WMC algorithm is very robust (cf. Figure 2-1f) to regional variation of intensity in images (cf. Figure 2-1a). Moreover, unlike the hysteresis algorithm, WMC does not introduce a higher rate of false foreground when increasing its sensitivity. This trait is particular important for images with a low signal-to-background ratio. The WMC algorithm will be discussed in detail in the next section.

2.2.5.1 Region Selection

The WMC algorithm starts by defining coarse regions (cf. schema 1) using a region selection mechanism. In the current implementation, this is accomplished using the maxima-seeded watershed algorithm [30], in which the growing of the watershed region is initialized from a pixel with the highest intensity compared to its neighboring pixels: this particular pixel is referred to as the local maximum. In order to define a valid local maximum, the intensity of such a pixel must exceed the lowest pixel intensities by a threshold value h , where h is an estimated level of noise tolerance in terms of pixel intensity, h is commonly referred to as the

h -maximum (cf. Figure 2-3c). Higher values of h provide a less sensitive watershed separation and *vice versa* (cf. Figure 2-3a & b). In practice, a higher value of h often leads to incomplete separation of the objects in the image. Moreover, objects that occur in clusters are often not sufficiently separated (cf. Figure 2-3b). We can derive the range for the value of h , since the h -maximum is considered a reference relative to the intensity value of the pixels. Let I_M be the maximum intensity in the dynamic range of the sensor, and I_{max} the maximum intensity in the region under study, the h -maximum is typically in $[1, (I_M - I_{max})]$. In Figure 2-3, the results of the maxima-seeded watershed for different values of h are depicted. From empirical observations in the images typical to HT/HC experiments ($I_M=255$), a value $h=10$ provides satisfactory watershed regions.

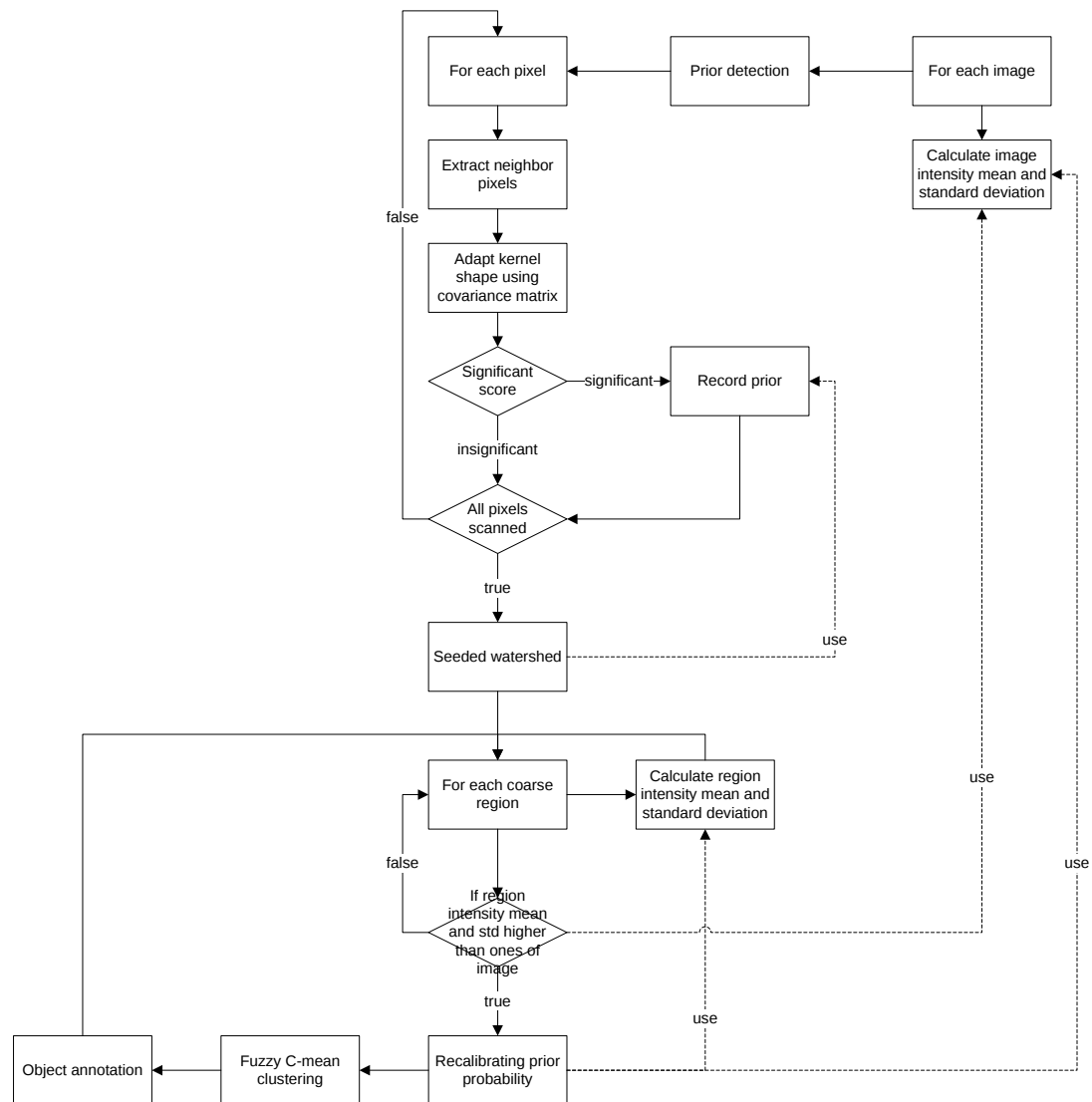


Figure 2-2 is the illustration of the three main steps of the Watershed Masked Clustering Algorithm. In the schema, the subsections are indicated in which that particular step is discussed in detail. As part of the automation process, at completion of the loop there is always a quality check Q to prevent wrongly processed images to be part of the analysis.

Given coarse regions, it is guaranteed that:

1. In each watershed region, the intensity landscape is always unimodal [30][71].
2. Seeded watershed implements a restriction on the possible starting point of path searching. An empty region usually does not contain valid seed, thus no watershed region will be formed in an empty region.

Schema 1 Watershed Masked Clustering Algorithm

Perform maxima-seeded watershed segmentation

Reverse watershed line into coarse region

for each coarse region r **do**

 False-check on coarse region

if region is valid **then**

 Perform weighted fuzzy C-means clustering in intensity space I of r

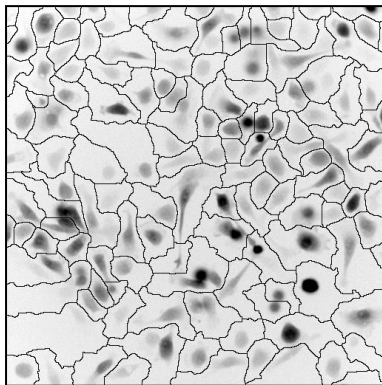
 Obtain labeling

 Create regional mask

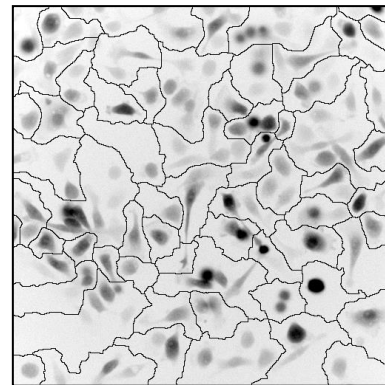
end if

end for

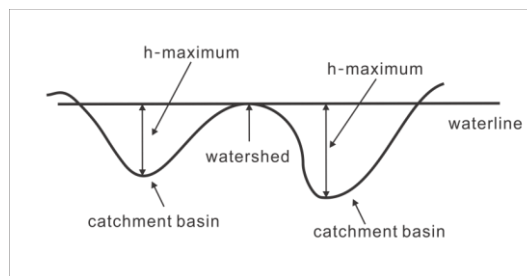
Combine regional mask into final object label



(a)



(b)



(c)

Figure 2-3: The (a) shows the definition of h -maximum. The (b) illustrates the watershed cutting lines at $h=20$. The (c) illustrates the watershed cutting lines at $h=50$.

Regarding variation in the extreme case, the result of the region selection may still contain invalid regions. Therefore, similar to Bernsen's threshold, the WMC algorithm implements a false-check mechanism for invalid regions that survive the region selection process. The false-check mechanism is constructed using intensity information including both standard deviation and mean of regional intensity values (cf. Equation 2-1). A valid candidate (coarse) region

should fulfill the first criterion, meaning the region should have large intensity diversity that suggests the presence of both sufficient foreground and background pixels. In the case that a region fails the first criterion, the second criterion will further distinguish the situation and make a final decision: i.e.

$$\begin{cases} \sigma(I_r) \geq \min_{std} \\ \mu(I_r) \geq \min_{mean} \end{cases} \quad \text{Equation 2-1}$$

, where I_r is the intensity values of one coarse region. Within these coarse regions, a threshold is trained based on local intensities. The threshold training procedure will be illustrated in the next section.

2.2.5.2 Threshold Training

With the coarse region, binary masking can be further refined from each coarse region. In order to perform such refinement, an approach is required that is capable of establishing a local adaptive threshold while being computational finite. Such can be accomplished by a weighted fuzzy C-means clustering algorithm (WFCM) [72][73]. This clustering is applied sequentially and an optimal threshold is calculated within each of the regions. Consequently, each region has its own threshold value taking into account local conditions, i.e. the local variation in image intensity.

In addition, the WFCM method has a set of weighting factors ω that allows the introduction of prior probability of the membership of the pixels in the clusters. The definition of such a weighting factor is similar to the reverse version of the prior probability in the Bayesian theorem. A smaller weighting factor is assigned to the cluster having, potentially, a larger standard deviation and vice versa. The sum of all weighting factors is always one. The weighting factor ω can be directly derived from the data [71][74], however, with a known type of image data, e.g. HT/HC images, commonly a preset value is used. The implementation of the WFCM method is described in the pseudo-code as:

Schema 2 Weighted Fuzzy C-means Clustering algorithm

```

Given weighting factor  $\omega$ 
Initial membership matrix  $u$  at step  $k=0$ 
for each  $k$  step do
    Calculate the center vector  $c_j$  for each cluster  $j$  given Equation 2-2
    Update the membership matrix  $u$  to  $(k+1)$  given Equation 2-3
    Creating regional mask
end for
Combing regional mask into final mask

```

The WFCM method is formalized as:

$$u_{ij} = \left(\sum_c^{k=1} \left(\frac{\omega_j \cdot \|x_i - c_j\|}{\omega_k \cdot \|x_i - c_k\|} \right)^{\frac{2}{m-1}} \right)^{-1}, \quad \text{Equation 2-2}$$

where u_{ij} denotes the membership matrix, c_j is the j^{th} cluster, x_i is the data vector i and ω_j is the weighting factor for cluster j . Empirically, it was established that for cell imaging a value of $\omega = 0.2$ for the foreground and a value of $\omega = 0.8$ for the background is sufficient. This should be interpreted as: (1) there is an 80% chance a certain pixel belongs to the foreground and (2) there is a 20% chance that a certain pixel belongs to the background. By increasing the weighting factor for the foreground, less intense structures such as protrusions of a cell or objects with a low overall intensity will be preserved. In this manner, the weighting factor works similarly to the parameter for the degree of sensitivity in the fuzzy c-means clustering (FCM) algorithm [74]. Along with Equation 2-2, the clusters are formalized as following:

$$c_j = \frac{\sum_{i=1}^N (u_{ij}^m \cdot x_i)}{\sum_{i=1}^N (u_{ij}^m)}, \quad \text{Equation 2-3}$$

where u_{ij} denotes is the membership matrix at step k and m is the, so called, fuzzy coefficient that expresses the complexity of the model, by default $m=2$. In our algorithm, we strive at a quick convergence of the WFCM and therefore the initial seeds for c are defined as follows:

$$c_{Foregroundseed} = \bar{I} + (2^{nb} - 1) \cdot \frac{I_{max} - \bar{I}}{\sigma(I)} \quad \text{Equation 2-4}$$

$$c_{Backgroundseed} = \bar{I} - (2^{nb} - 1) \cdot \frac{\bar{I} - I_{min}}{\sigma(I)} \quad \text{Equation 2-5}$$

where I_{min} , I_{max} denote the minimum/maximum intensity in the image I , $\bar{I}$ denotes the mean of the intensities in image I , $\sigma(I)$ denotes the standard deviation in the intensities of the image I and nb denotes the dynamic range of the intensity expressed in number of bits. In the standard case of unsigned 8-bit images $nb=8$.

The flexibility of weighted fuzzy C-means clustering is a more robust solution when addressing the complexity in the HT/HC images. The application of this step results in a binary object in each of the regions of step 1 (cf. §2.2.5.1), if correct, shape features can be derived. However, the watershed method might have introduced some irregularities in the establishment of the coarse regions, which requires an additional evaluation; this evaluation is elaborated in the next section.

2.2.5.3 Object Optimization

At onset of our algorithm, the watershed segmentation is applied resulting in regions that are individually processed. Depending on the variation in the data, the watershed algorithm is known to result in an overcut of the segmentation; overcut is commonly referred to as the situation in which the watershed segmentation produces more regions than actually present in the image [75]. This overcut might affect the individual objects, as a result of which the objects need be split or merged (cf. Figure 2-4). Therefore, the last step in our algorithm is to compensate for the possible overcut caused by the watershed segmentation. We refer to this process as an object optimization as we evaluate the results obtained in the object segmentation. In this procedure, only the objects that share a border with a watershed line are evaluated, as these objects are the candidates for overcut. The procedure is summarized as follows:

Schema 3 Object Optimization

```
Given watershed line in step 2
for each pixel  $l_i$  in watershed line do
  for each pixel  $l_n$  in the 4-connected neighbor of  $l_i$  do
    if pixel  $l_n$  is overlapping with the binary mask of an object then
      Object is sharing the watershed line pixel
    end if
  end for
  if pixel  $l_i$  is shared by more than two objects then
    Calculate all criteria
    if all criteria are true then
      Discard the pixel  $l_i$ 
    end if
  end if
end for
Combing regional mask into final mask
```

The solution for the object optimization is a merging mechanism that uses multiple criteria; currently, two criteria are implemented but depending on the type of data; more can be added. The two criteria are:

1. Evaluation of the strength of watershed line; the objects are merged based on a local difference in maximum and average intensity in the object.
2. Evaluation of the orientation of the objects; the object are merges based on assessment of the difference in orientation of their principal axes.

For criterion 1, we implemented an intensity-based merging algorithm so as to estimate the necessity of merging the objects through the evaluation of the strength of the watershed lines. In this function all watershed lines are evaluated. This criterion can be generalized with the evaluation function K :

$$K(l_i) \rightarrow \min\left(\frac{\delta_1}{\tau_1}, \frac{\delta_2}{\tau_2}\right) > T_k \quad \text{Equation 2-6}$$

where the l_i denotes the i^{th} watershed line, δ_1 denotes the difference between the average intensity under the watershed line and maximum intensity of object on one side of the watershed and similarly, δ_2 represents the object on the other side of the watershed line; where τ_1 and τ_2 denote the difference between the maximum and minimum intensity value within one object on either side of the watershed line l_i . A valid watershed line should fulfill the condition given in Equation 2-6. If $K(l_i)$ exceeds a threshold T_k then the objects on either side of the line are merged to one and the watershed is neglected. In Figure 2-4a, the intensity-based merging criterion is illustrated.

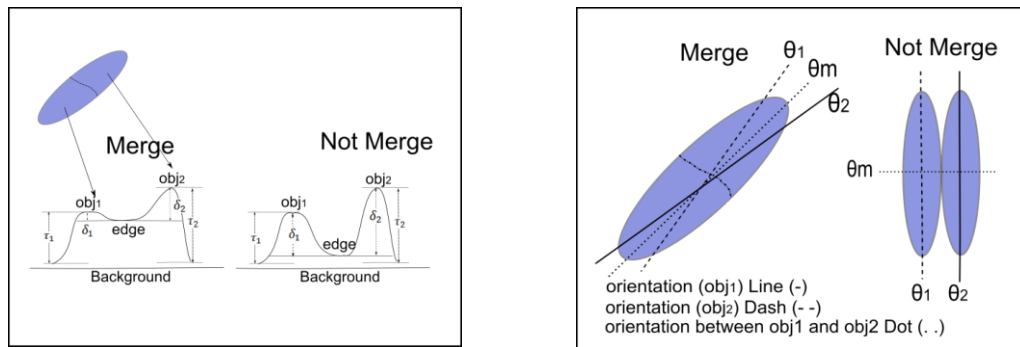
For criterion 2, we implemented an orientation-based merging algorithm [76][77][78], which provides a unique possibility to split/merge large structure complexes or elongated objects (e.g. protrusions). At watershed line l_i we consider the principal axis of the objects on either

side of the line. A two component Boolean function is designed so that when true, i.e. both components are true, the objects will be merged and the watershed line will be neglected. This function P is written as:

$$P(l_i) \rightarrow \begin{cases} |\theta_1 - \theta_2| < T_p \\ |\theta_1 - \theta_m| + |\theta_2 - \theta_m| < T_p \end{cases} \quad \text{Equation 2-7}$$

where θ_1 denotes the angle between the horizontal image axis (x-axis) and the principle axis of object 1, similarly θ_2 is defined for the object on the other side of watershed line l_i . The θ_m is the angle between the horizontal image axis and the line crossing the centers of mass of the two objects (cf. Figure 2-4b). The components in $P(l_i)$ are separately evaluated; so, if the principle axis of each individual object spans a minimum angle T_p while the line crossing the centers of mass of the two objects lies within the angular wedge T_p of the two principle axes, only then these two objects will be merged. In Figure 2-4b, the orientation-based merging is illustrated by two cases.

In studies where the objects are cells orientation merging is used less frequent whereas in the studies on analysis of protein expression in endocytosis or cell signaling it is often applied.

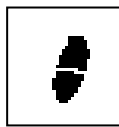


(a) Two typical cases of intensity based merging; (left) a merge is realized and (right) a merge is not realized using $K(l_i)$ (cf. Equation 2-6).

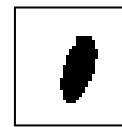
(b) Two typical examples of orientation based merging; (left) a merge is realized and (right) a merge is not realized using $P(l_i)$ (cf. Equation 2-7).



(c) Sample object



(d) Overcut from step 1



(e) Merged from step 3

Figure 2-4: Illustration of the merging of objects based on a combination of criteria; in (c,d,e) a specific case for one object (cell) is illustrated.

Once the object optimization is applied, one can be certain that all objects are correctly extracted and these can be subject to a characterization of the shape. For the specific case of the time-lapse images in HT/HC, both the shape of the binary object and the intensity profile can be measured. The intensity profile of an object is derived by applying the (final) binary mask to the original image. In addition to standard features, higher order features can be used [78]. In the case of HT/HC, the features are used to discriminate between the experimental conditions that are applied [10][4]. Examples of the application of this step of the WMC

algorithm are worked out in the next section (cf. §2.3) where performance of WMC is compared to other segmentation algorithms.

2.3 Performance Evaluation

This section will address the performance assessment of the WMC segmentation algorithm together with several popular HT/HC-proven algorithms. In order to get a good impression of their robustness and accuracy, four image sets (cf. Figure 2-5) including image sets with artificially generated objects and images from genuine HT/HC experiments are produced. The ground truth image sets employed are the following:

1. Artificial image set with artificial objects (cf. Figure 2-5a)
2. HT29 phalloidin image set (cf. Figure 2-5b)
3. MTLn3 GFP image set (cf. Figure 2-5c)
4. MA image set (cf. Figure 2-5d)

The artificial image contains a number of randomly generated ellipsoid objects. The HT/HC image set is based on imaging of living migrating tumor cells or subcellular structures that ectopically express fluorescent protein [18][80]. The performance estimation for each algorithm is derived from the comparison between the binary mask obtained by the algorithm and the corresponding ground-truth binary mask for each image (cf. Figure 2-5e-h).

In the generation of the test images, the ground-truth masks for the artificial test images are explicitly constructed. The usage of such artificial image provides an image test set with an unbiased ground-truth and controllable noise, emulating a common case in fluorescence microscopy. The employment of genuine HT/HC image sets illustrates the empirical performance of the segmentation algorithm under different experiment designs or image modalities. The construction of a ground truth in these image sets are accomplished via manual delineation.

The image quality of each test set, in terms of segmentation complexity, is measured by image coefficient of variance (*image-CV*). The *image-CV* [81] is an alternative definition of the signal-to-background ratio in biomedical image processing. It is measured as the ratio between average foreground intensity and standard deviation of background intensity (cf. Equation 2-8).

$$CV = \frac{\mu_f}{\sigma_b} \quad \text{Equation 2-8}$$

A higher *image-CV* suggests that the distribution of foreground intensity value is far from the distribution of background intensity value. Therefore, it measures the significance of the difference between foreground intensity and background intensity.

The image set with artificial objects (cf. Figure 2-5a) is an image set which consists of images containing artificially generated objects resembling the basic phenotype of the cells. With artificially generated objects, here we eliminate the observation bias frequently occurring in manual ground truth construction. Moreover, we use image set containing artificially generated objects to introduce hypothetical noise to test algorithm robustness. The *image-CV* (cf. Figure 2-6a) shows that the image quality of this image set is similar to HT/HC screen studies.

The HT29 phalloidin image set (cf. Figure 2-5b) [82] is a public available ground truth image sets from a HT/HC screen study. Here we choose the image channel containing phalloidin (cytoplasm staining) channel as the test images due to its complexity in term of segmentation. However, the image quality of HT29 is still higher than both artificial image set and MTLn3 GFP image set (cf. Figure 2-6a), representing the difference between foreground and background is higher in HT29 images.

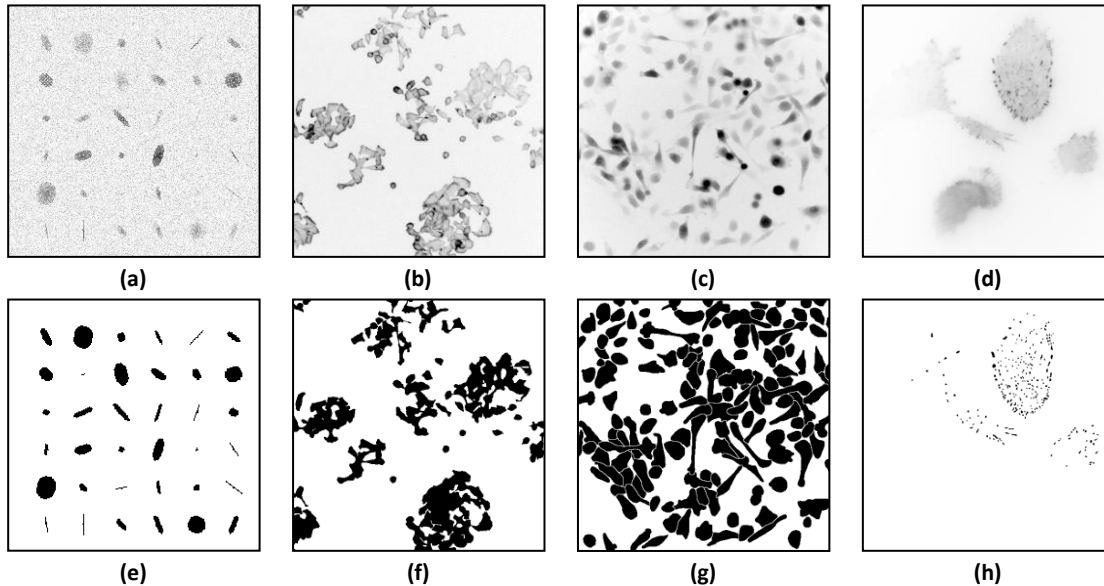


Figure 2-5 (a) artificial image set and ground truth mask (e); (b) HT29 phalloidin channel and ground truth mask (f); (c) MTLn3 GFP channel and ground truth mask (g); (d) MA image and ground truth mask (h)

The MTLn3 image set (cf. Figure 2-5c) is a ground truth image set from a HT/HC screen study using an aggressive cancer cell line. In terms of migration speed, the rat breast carcinoma MTLn3 cell-line is considered one of the most aggressive in vitro assays. The MTLn3 cells migrate as individual cells with an average velocity of 40 $\mu\text{m/hr}$ [9][83] whereas the HT29 cell-line only migrates with an average velocity of 4 $\mu\text{m/hr}$. GFP signal distribution of MTLn3 cells is however more complex since the GFP protein expression is variable from cell to cell. The *image-CV* (cf. Figure 2-6a) shows that image quality of the MTLn3 GFP image set will not be easy to segment.

The MA image set (cf. Figure 2-5d) is a ground truth set consists of subcellular structure known as the matrix adhesion (MA). Due to the imaging settings, the image contains some bleed-through signal from other channels. Thus, it presents a more complex segmentation problem (cf. Figure 2-6a).

The four image sets are believed to be a reasonable representation of the quality of HT/HC images. The Figure 2-6a shows that HT29 phalloidin has the best image quality for segmentation (*image-CV*=27.85). It confirms the quality of the HT29 image set over the MTLn3 GFP image. The artificial image set has a CV similar to MTLn3, suggesting the artificial image set does emulate the real-world scenario of HT/HC screening. Among all ground truth image

set, the MTLn3 GFP image set has the worst image quality for segmentation. The MA image set shows a slightly higher ($image-CV=8.16$) than MTLn3 GFP.

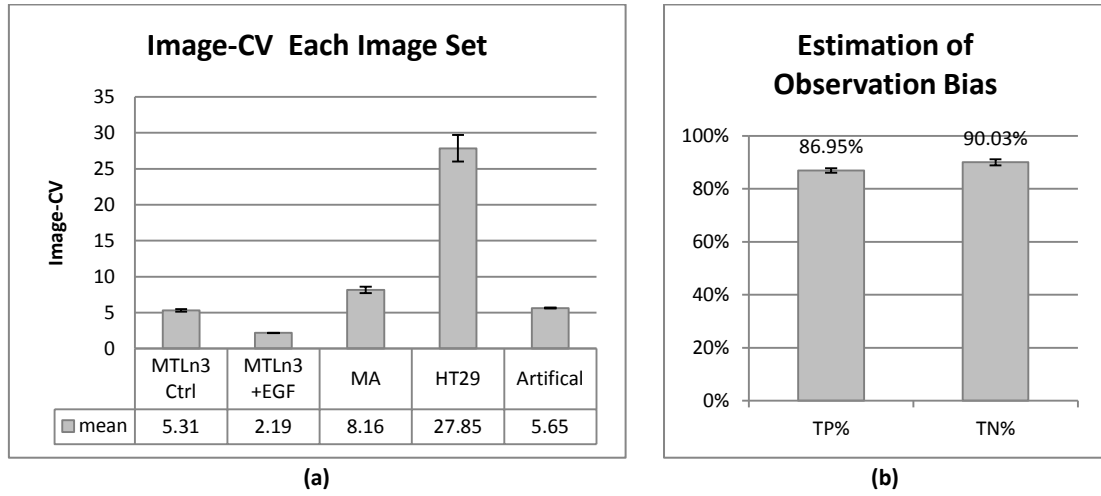


Figure 2-6 (a) *image-CV* estimation for each image set, (b) estimation of observation bias over MTLn3 GFP image set

Additionally, manual delineations are repeated with a random shuffling and rotation of the images to cover observation bias. From the estimation of the observation bias with MTLn3, it shows that on average the observers have a F1-score of 89.87% (cf. Figure 2-6b). From the observation bias, it is clear that ground truth mask is not 100% reproducible, which sustains the necessity of the artificial image set. The observation bias of HT29 is not available. With the ground truth image set, we will further introduce the assessment metrics and methodology in the next section.

2.3.1 Assessment Metrics and Methodology

In this section, the pixel-level mismatch is calculated for all segmentation algorithms. The rationale behind this test is to simulate the typical data processing workflow for HT/HC, therefore the parameters used for each of the algorithms are optimized only once and henceforth applied to the whole image set in the experiment. For none of the algorithms in the experiment an individual tuning is applied. The parameters for all algorithms (cf. § 2.2) were obtained from the high-content screening literature [14][84][17][12] and existing software [63][64][82].

Before introducing the error estimation methodology, a clarification of the assessment metrics is presented. Segmentation algorithms are often considered simplified versions of two-class classifiers that are trained in intensity space [60][59][46]. Therefore, similar to the error estimation for the classifier, the error test normally covers both type-I error (False Positive) and type-II error (False Negative). The performance of a segmentation algorithm can be assessed using the number of correct and incorrect segmented pixels [74]. This definition only covers the type-I error (FP) which may lead to an overtraining of the algorithm [85]. For a balanced conclusion, we take into account both type-I error (FP) and type-II error (FN). Furthermore, instead of just using the two errors types, we introduce the F1-score [86] which is derived from the types-I/II errors, as the major criterion of segmentation performance.

The two types of errors for different algorithms are defined in terms of the True Positive and True Negative. True positive (TP) is defined as the ratio of pixel overlap between the ground-truth mask and the segmented mask by each algorithm. This ratio is expressed as:

$$TP = \frac{M \cap M'}{M}, \quad \text{Equation 2-9}$$

where M' is the set of pixels belonging to the foreground of binary mask provided by the algorithm and M is the set of pixels belonging to the foreground of the ground-truth mask. In similar fashion, the true negative (TN) is calculated as:

$$TN = \frac{\bar{M} \cap \bar{M}'}{\bar{M}}. \quad \text{Equation 2-10}$$

In this way, TP represents the percentage of correctly segmented foreground pixels whereas TN represents the percentage of correctly segmented background pixels. From the values of TP and TN, the false positives (FP) are derived, i.e. $FP = 1 - TP$ (percentage of incorrectly segmented foreground pixels), and likewise the false negatives (FN) are derived, i.e. $FN = 1 - TN$ (incorrectly segmented background pixels). From these values, the sensitivity and the specificity [87] are calculated by:

$$sensitivity = \left(\frac{TP}{TP + FN} \right) \quad \text{Equation 2-11}$$

$$specificity = \left(\frac{TN}{FP + TN} \right) \quad \text{Equation 2-12}$$

Given the results, the specificity and the sensitivity for all of algorithms of a particular set of test images can be computed. In addition, from the specificity and sensitivity, the F1-score is derived by:

$$F1 = 2 \cdot \frac{specificity \cdot sensitivity}{specificity + sensitivity}. \quad \text{Equation 2-13}$$

A good segmentation algorithm should yield the highest F1-score but this only occurs when both specificity and sensitivity are approaching 100%. Our choice for the F1-score aims to enforce a balanced performance in terms of preventing either oversegmentation or undersegmentation.

In the next few sections, the performance assessment of the following algorithms will be exemplified. These algorithms include Bernsen local threshold algorithm, Otsu threshold algorithm, Level-set algorithm, hysteresis threshold algorithm, together with the WMC algorithm. All of the algorithms have claimed the intrinsic capacity of performing well under noisy conditions typical to HT/HC imaging [78][14][84][17][12]. For these algorithms, open-source plug-ins are available in Fiji [63][65][64] and CellProfiler [82] and these implementations were used without modifications and the tuning of the parameters for each algorithm is accomplished by cell biology specialists. These algorithms will be tested for all four ground truth image sets.

2.3.2 Artificial Objects and Test Images

In order to understand and verify the behavior and performance of segmentation algorithms without introducing observation bias, ground-truth images with objects resembling the shapes, which are normally found in high-content imaging, are constructed (cf. Figure 2-7). Each image consists of a number of ellipsoid objects and each object has a unique intensity profile. The intensity landscape is generated through an exponential decay function that is initiated at the centre of each object. The minimum and maximum value of an intensity profile of an object are generated using a uniform distributed random generator and scaled in the range of (20, 255). In this way, it resembles a random intensity variation in HT/HC screen. In addition, the orientation of each of the objects is varied by applying a rotation to each of the object. The rotation is in the range of $[-30^\circ, 30^\circ]$ using the center of mass as the pivot; the rotation angle is selected from a uniform random generator that is scaled to the rotation range. The random rotation of objects aims to test whether the segmentation algorithm is sensitive to direction.

The original binary image with all the objects is kept as the absolute ground-truth mask for the segmentation so that error estimation can be applied over a range of test images. In this test, a total amount of 30 images is generated. To simulate image noise typical to HT/HC and fluorescence microscopy, Poisson noise is generated by giving each pixel a random intensity (Poisson) [35][88] oscillation. In the noise-added images, the intensity oscillation is ranged from 0 to 255 while the size is approximately 3 to 5 pixels. In the original image, the object intensity will be above zero and therefore a true global threshold can always be found.

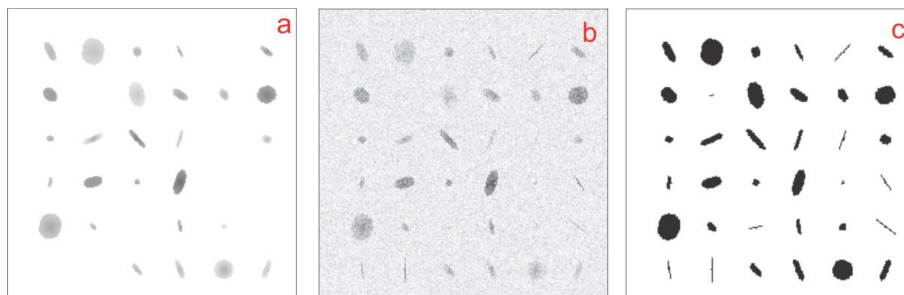


Figure 2-7: Artificial image set being used for assessing the efficiency of the segmentation. (a) Original Image in 256x256 (b) Noise-added Test Image (c) Ground-truth Mask

2.3.3 HT29 Phalloidin Images

The HT29 phalloidin image set is known as the “Human HT29 Colon Cancer” dataset [84] (cf. Figure 2-8); a publically available image dataset including ground-truth image set provided by the Broad Institute of MIT. This set contains 12 images of human HT29 colon cancer cells (cf. Figure 2-9a & c). Each image consists of three channels including phalloidin (channel 1), pH3 (channel 2), and nucleus (channel 3 Hoechst). Hoechst labels DNA present in the nucleus. Phalloidin labels actin present in the cytoplasm. The pH3 stain indicates cells that are in division and will not be used in this test. For the benchmark study, we focus on image segmentation performance with the phalloidin channel (cf. Figure 2-8c) because it contains several complexities including intensity variation, uneven illumination and phenotypical variation in object size (cf. Figure 2-8a & c).

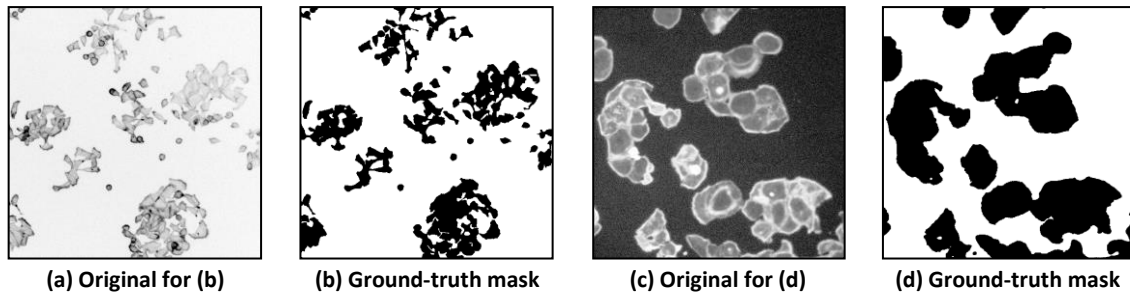


Figure 2-8: images (512x512) from the HT29 phalloidin channel, which (a) represents phenotype of control cells, which (c) represents phenotype of treated cells. There is an observable size variation between (a) and (c).

2.3.4 MTLn3 GFP Images

The MTLn3 GFP image set is a time-lapse image sequence, i.e. a dynamic process, of the breast carcinoma MTLn3 line used to understand tumor cell migration in the context of breast cancer metastasis (cf. Figure 2-9). The set is provided by the Leiden Academic Center for Drug Research (LACDR). It consists of 96 time-lapse image sequences, each of 75 frames in 5 minute sampling intervals. Each sequence portrays an *in vitro* cell migration pattern typical in HT/HC experiments. The GFP (Green Fluorescent Protein), was ectopically expressed to label the cell body (cytoplasm + nucleus), enabling fast fluorescent imaging [80]. For the performance tests, we will only use the first 14 images of the sequence to reduce the size of the image set to reasonable for proportions for this test. In addition, for this image set also a ground-truth image is required. The MTLn3 ground-truth images were obtained by manual segmentation performed by biologists through tracing on a digitizer tablet (WACOM, Cintiq LCD-tablet) (cf. Figure 2-9c). In contrast to the artificial image set, manual segmentation may contain bias between and within observers. To that end, the manual segmentation is replicated a few times. Moreover, to improve visibility of objects in each image, the observer will draw the cells using images to which an intensity equalization is applied [89] (cf. Figure 2-9b).

The MTLn3 set is a good representation of a high throughput screen with *in vitro* live cell migration (cf. Figure 2-14) for it contains intensity variation, uneven illumination, phenotypical variation and object overlapping.

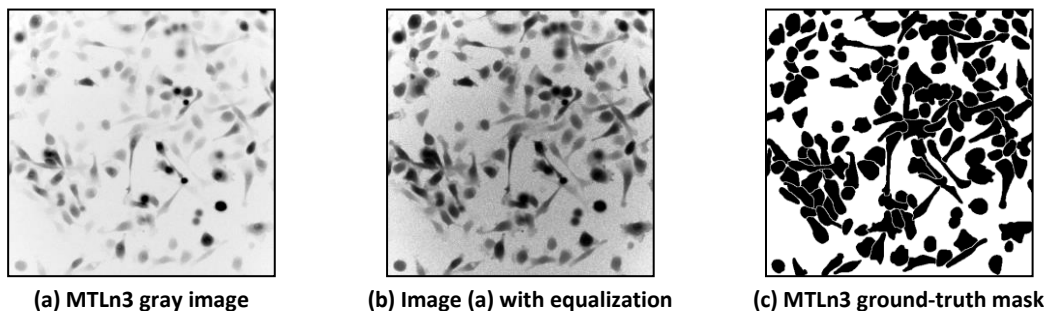


Figure 2-9: Typical image (512x512) from the MTLn3 GFP set

2.3.5 MA Images

The MA image set is a time-lapse image sequence of matrix adhesion (MA) dynamics captured using total internal reflection fluorescence (TIRF) microscopy [25]. This set is provided by the Leiden Academic Center of Drug Research as a study of correlation between matrix adhesion regulation and cell migration. The MA image set consists of subcellular macromolecular complexes visualized using TIRF microscopy (cf. Figure 2-10). The magnification of MA TIRF is

higher than other sets but the object size is relatively small (approx. 5~65 pixels). Due to the smaller size, each object contains less intensity information for the segmentation algorithm. In terms of ground truth construction, observers reported difficulty in distinguishing between background and foreground. The procedure for the ground truth set is the same as the MTLn3 image set.

The MA image set is included as a case study for algorithms' extensibility to both relatively small objects (MA) and relatively large objects (cell body) (cf. Figure 2-11 a & b). Here we define relatively small objects as objects occupying less than 0.01% of whole image and the relatively large objects as objects occupying more than 1% of image. For a 512x512 image, it means a small object will be approximately 10~50 pixels while a large object will be approximately 5000~10000 pixels.

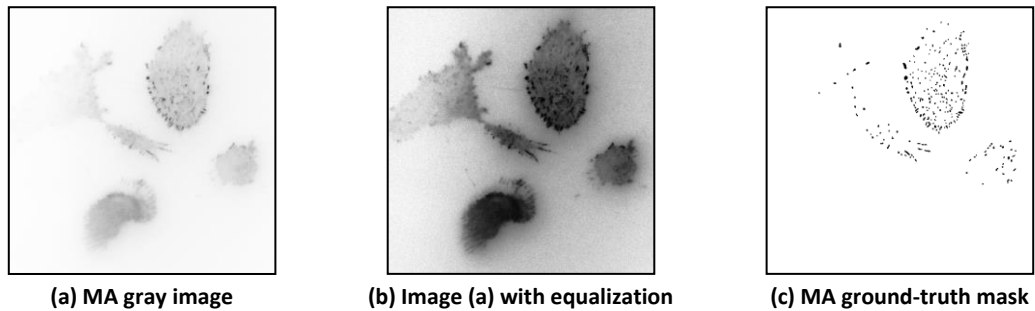


Figure 2-10: Typical image (512x512) from the MA TIRF set

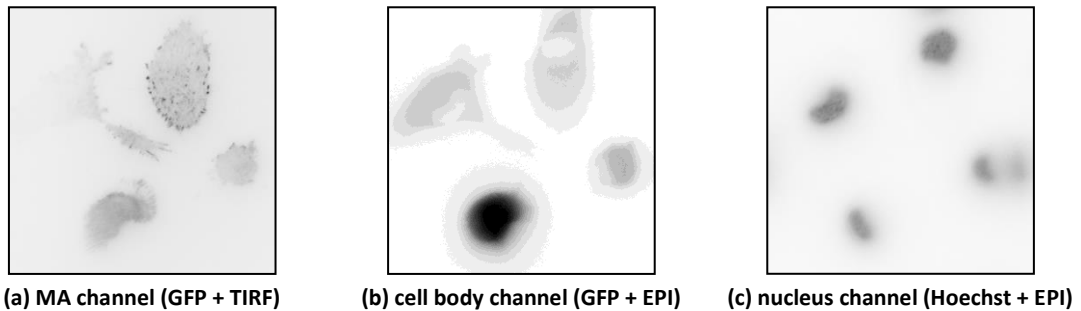


Figure 2-11 complete design of the imaging acquisition for MA image set

2.3.6 Result of Benchmark Study

Artificial Objects and Test Images

All algorithms are applied over the same 30 test images (cf. Figure 2-12). In the Table 2-1, the results for the true-positives are listed. In the Table 2-2, the results for the true-negatives are listed. The F1-scores are listed in the Table 2-3. The object merging accuracy in WMC is also tested using the same image set. An overcut object is defined as a group of objects obtained by segmentation algorithm share the same object in ground truth mask. A total amount of 238 overcut objects are detected in this image set. Using object optimization, the WMC recovers 202 out of 238 overcut objects, i.e. approximately 85%.

From Table 2-3 we can conclude that the WMC yields the highest performance while Otsu, Bernsen and Hysteresis produce acceptable results. The level-set shows the lowest performance. From the image (cf. Figure 2-12), it is noticeable that none of the segmentation algorithms has successfully captured objects smaller than noise. The WMC can well preserve shape of object regardless of noise and intensity variation, but it has a tendency of

undertraining the threshold in some cases. Both Bernsen and Otsu can also preserve objects with higher intensity homogeneity but cannot adapt to intensity variation. The Hysteresis algorithm can well preserve objects with higher intensity homogeneity but cannot preserve objects with an extreme morphology, i.e. an elongated object or a small spherical object. The level-set algorithm fails to adapt its propagation due to the intensity variation.

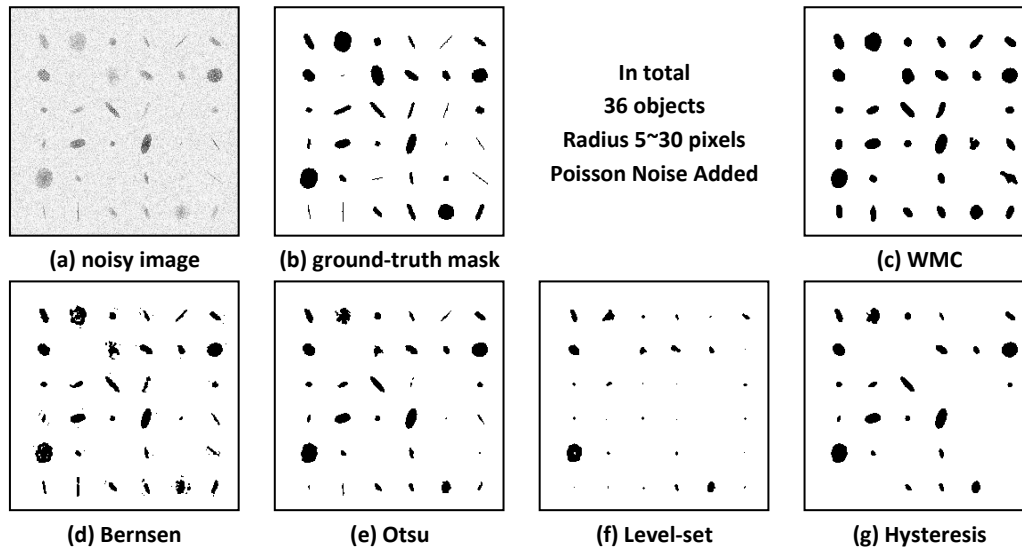


Figure 2-12: (a) noise-added test image, (b) ground-truth masks for the object, (c) to (g) are binary images obtained by corresponding segmentation algorithms.

Table 2-1 True positive rate

True Positive	WMC	Bernsen	Otsu	Level-set	Hysteresis
Avg	96.67%	91.83%	98.48%	60.57%	86.39%
Std	3.81%	6.23%	1.68%	28.14%	2.74%

Table 2-2: True negative rate

True Negative	WMC	Bernsen	Otsu	Level-set	Hysteresis
Avg	95.34%	74.73%	84.00%	99.01%	78.10%
Std	5.45%	28.36%	18.46%	1.36%	12.64%

Table 2-3: Specificity and sensitivity of segmentation efficiency using artificial images

Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	96.62%	90.14%	98.22%	71.52%	85.16%
Specificity	95.40%	78.42%	86.02%	98.38%	79.78%
F1 Score	95.99%	84.60%	91.83%	74.98%	82.95%

HT29 Phalloidin Images

From Table 2-4, we can conclude that nearly all algorithms can perform well with HT29 image set (cf. Figure 2-13d). It is unclear which algorithm is better since WMC, Otsu, and Hysteresis are all producing good results. In general, WMC (cf. Figure 2-13d), Hysteresis (cf. Figure 2-13h) and level-set method (cf. Figure 2-13g) are all performing well with the raw images. The Otsu segmentation result (cf. Figure 2-13f) shows that a global threshold algorithm is less robust to intensity variation but its performance can be improved by intensity equalization. The halo

structures in Bernsen (cf. Figure 2-13e) is believed to be associated with a localization of staining at the cell border region, which is smaller than the kernel size chosen in Bernsen.

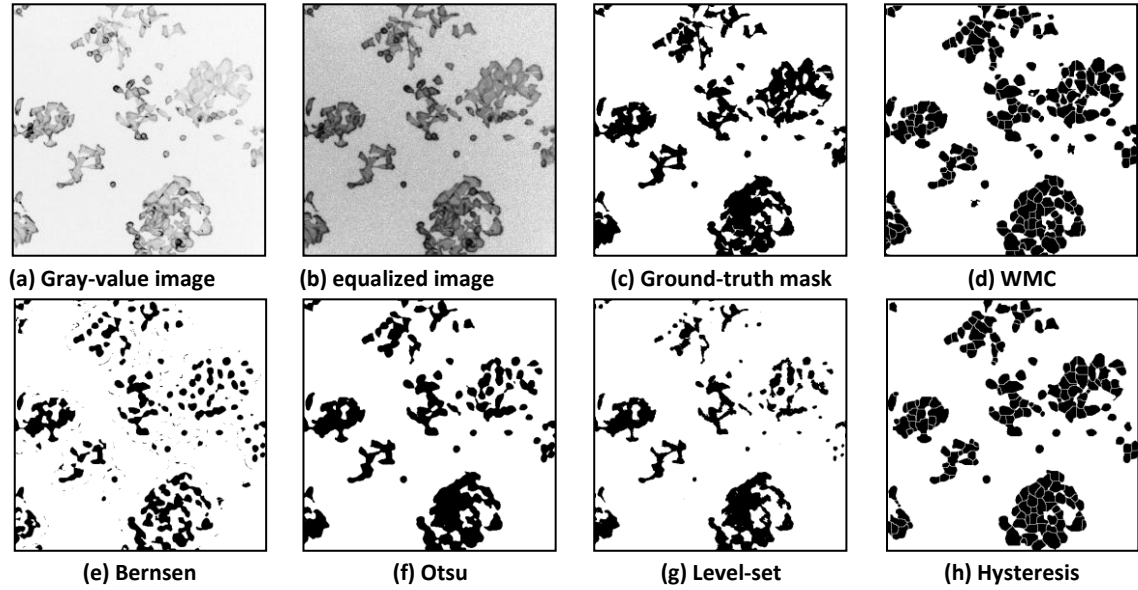


Figure 2-13: (a) Original HT29 image acquired with a 10x lens; image size is 512x512 pixels, 8 bit. b) the image (a) with intensity equalization c) Ground-truth masks and (d-h) masks obtained by the segmentation algorithms.

Table 2-4: Specificity and sensitivity of the segmentation algorithms in the HT29 set

No Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	92.49%	95.49%	99.13%	98.24%	97.51%
Specificity	91.71%	63.71%	79.48%	84.57%	83.29%
F1 Score	92.10%	76.43%	88.23%	90.89%	89.84%
Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	85.63%	86.16%	97.06%	80.27%	84.70%
Specificity	88.21%	66.43%	93.81%	79.04%	82.59%
F1 Score	86.90%	75.02%	95.41%	79.65%	83.63%

MTLn3 GFP Images

From Table 2-5, we conclude that MTLn3 is a difficult image set for image segmentation. It is immediately clear that the overall performance is much lower compared to the experiment with the HT29 set (cf. Table 2-4). Most methods are able to extract the brighter region around the nucleus (cf. Figure 2-9) while the fuzzier cytoplasm region is not detected. With MTLn3 set, the WMC algorithm still shows the highest performance. The Hysteresis thresholding algorithm is able to portray a good and stable performance. Compared to the previous experiment (cf. Table 2-4), the performance of each algorithm is decreased. The WMC algorithm, however, performs quite stable under these different circumstances.

Although WMC and Hysteresis are able to preserve the fine details with comparable F1-score, the Hysteresis demonstrated an unbalanced performance by yielding a low sensitivity but high specificity, meaning it misses many foreground pixels. The WMC, on the other hand, shows a

more balanced performance in both sensitivity and specificity. All algorithms show improvements after intensity equalization.

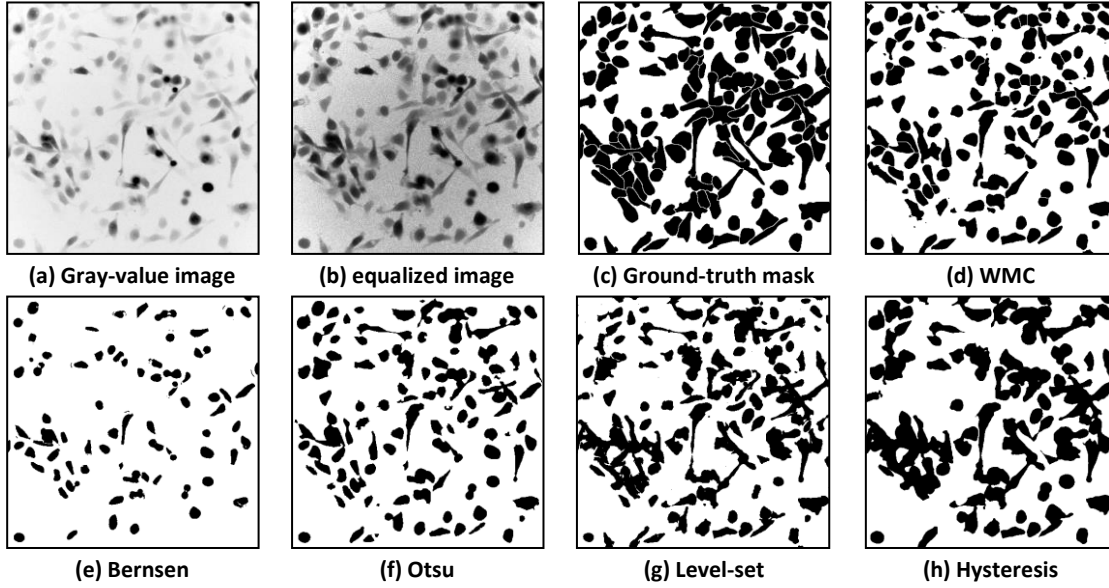


Figure 2-14: (a) Original MTLn3 image acquired with a 20x lens (NA 1.4), image size 512x512 pixels, 8-bit. (b) the image (a) with intensity equalization. (c) ground-truth masks and (d-h) masks obtained by the segmentation algorithms.

Table 2-5: Specificity and sensitivity of segmentation algorithms in MTLn3 image set

No Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	73.02%	21.50%	17.25%	27.86%	66.86%
Specificity	90.75%	99.62%	99.67%	99.13%	94.21%
F1 Score	80.92%	35.36%	29.42%	43.50%	78.21%
Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	85.70%	36.03%	39.17%	30.37%	79.94%
Specificity	82.54%	98.35%	98.80%	99.20%	87.25%
F1 Score	84.09%	52.75%	56.09%	46.50%	83.43%

MA Images

The performance assessment (cf. Table 2-6) shows that WMC demonstrates a stable and robust performance similar to the previous image sets (cf. Table 2-4 and Table 2-5). This holds also for the hysteresis thresholding and level-set algorithm. Both Bernsen and Otsu algorithm show a lower performance. The smaller object size poses difficulty for all algorithms since there is less intensity information available for the threshold training or contour propagation. This has been foreseen in the test with artificial objects (cf. Table 2-3).

Table 2-6 Specificity and sensitivity of segmentation algorithms in MA TIRF image set

No Intensity Equalization					
Performance	WMC	Bernsen	Otsu	Level-set	Hysteresis
Sensitivity	84.00%	67.16%	62.20%	74.85%	80.70%
Specificity	97.21%	99.09%	99.74%	99.84%	98.91%
F1 Score	90.12%	80.06%	76.62%	85.56%	88.88%

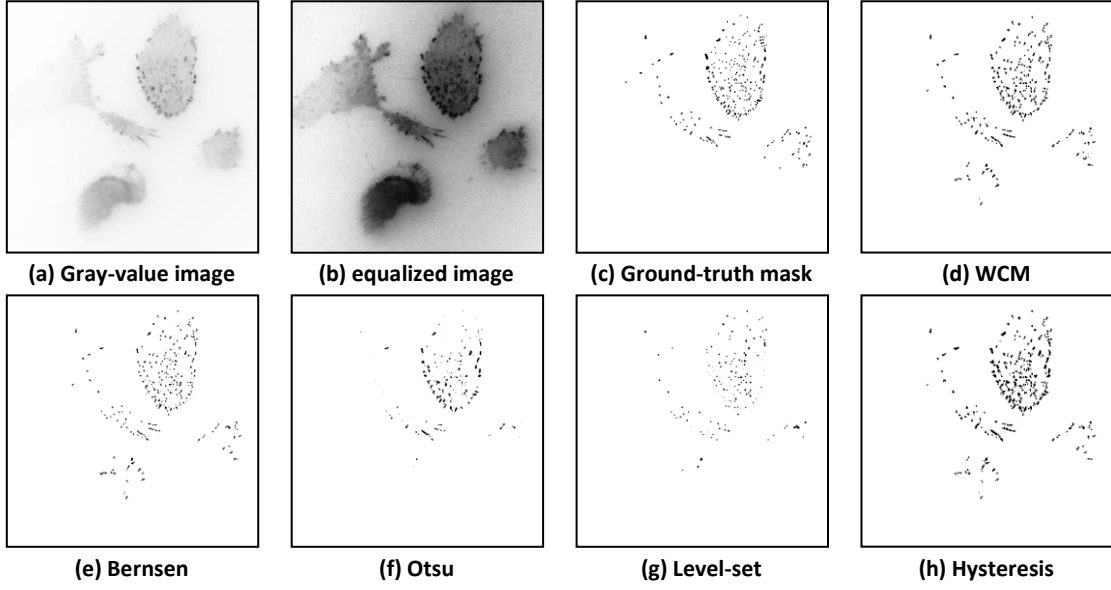


Figure 2-15 (a) Original MTLn3 image acquired with a 20x lens (NA 1.4), image size 512x512 pixels, 8-bit. (b) the image (a) with intensity equalization. (c) ground-truth masks and (d-h) masks obtained by the segmentation algorithms.

2.4 Computation Complexity

To further verify the high-throughput nature of the WMC algorithm, its computational complexity is studied and compared to the same selection of segmentation algorithms in previous sections (§2.3). The computational complexity of the WMC algorithm is derived from its two major components, namely watershed segmentation and fuzzy C-means clustering. Both algorithms are known to be NP-complete problems. In an NP-complete problem, the computational cost grows in polynomial order, but the total amount of computational cost is nondeterministic. We will therefore focus on the computational complexity in fashion of a single-iteration. In the watershed segmentation (flooding based), worst case, the searching of all descending paths is in the order of $O(n^2)$ [71].

In iteration of fuzzy C-means (FCM) clustering, the computations are divided into three steps and the total complexity is a joint sum of all individual steps.

Step 1: update membership matrix

$$u_{ij} = \left(\sum_c^{k=1} \left(\frac{\omega_j \cdot \|x_i - c_j\|}{\omega_k \cdot \|x_i - c_k\|} \right)^{\frac{2}{m-1}} \right)^{-1} \in O\left(n^{\frac{4}{m-1}}\right) \quad \text{Equation 2-14}$$

Step 2: calculate new seeds

$$c_j = \frac{\sum_{i=1}^N (u_{ij}^m \cdot x_i)}{\sum_{i=1}^N (u_{ij}^m)} \in O(n) \quad \text{Equation 2-15}$$

Step 3: validation

$$\|u^{k+1} - u^k\| \in O(n^2) \quad \text{Equation 2-16}$$

The total complexity of FCM clustering in a single iteration is $O\left(n + n^{\frac{4}{m-1}} + n^2\right)$. Together with the computational complexity of watershed segmentation, the total complexity of WMC can be written into $O\left(n + n^{\frac{4}{m-1}} + n^2\right)$. With fuzzy factor $m=2$, the computational complexity of the WMC is in the order of n^4 . The practical computational performance of each algorithm is given in the Table 2-7. The performance is based on a test with 14 images of the MTLn3 set.

Table 2-7: Computational performance in seconds; executed on a 2.4 GHz P4 (single-thread) with 4GB RAM. All implementations of the segmentation algorithms are in Java.

average time (sec)	WMC	Bernsen	Otsu	Level-set	Hysteresis
mean	4.46	754.3	1.14	2.34	0.06
std	0.22	7.58	0.02	0.17	0

From this analysis, we can deduce that the computational load of the algorithm will be in reasonable bounds and suitable for the type of application it was initially designed. Further implementation of concurrent computation in WMC has significantly reduced the computation time (cf. Table 2-8) by a linear factor equals to the number of threads.

Table 2-8: Computational cost in seconds; executed on a 2.7 GHz i7 (8-thread) with 8GB RAM. All implementations of the segmentation algorithms are in Java.

average time (sec)	WMC (multi-thread)	WMC (single-thread)
mean	0.42	3.37
std	0.08	0.12

2.5 Conclusion and Discussion

This chapter illustrated a number of image segmentation algorithms, which are applied in the image analysis procedures for high-content screens. In conclusion, from Table 2-3, Table 2-4, Table 2-5, and Table 2-6, it can be established that the WMC algorithm outperforms the other algorithms regardless image modality and image quality issues. Moreover, WMC has demonstrated a stable and robust performance for each image sets compare to other algorithms. This is in particular important for HT/HC screen studies since it is impossible to predict the output of treatments. Therefore, WMC algorithm is considered a good solution for HT/HC screen studies.

The major advantage of the WMC algorithm is that it can deal with variations in staining intensity typical for bio-imaging and specific to high-throughput in vitro experiments (cf. Figure 2-1). The local intensity variations in the image limit application of Otsu segmentation; it requires a global optimum for the threshold, which may not be possible. Along the same line, the level-set method is not suitable as it presumes a consistent intensity for the objects in the image. The regional approach in WMC followed by a local clustering transforms the segmentation to a local problem so that threshold levels can be found efficiently. For segmentation in cytomics edge based methods are noise susceptible, therefore intensity variations necessitate region based approaches. This is confirmed from our findings comparing Hysteresis segmentation to WMC, especially with more artificial noise or staining variations in the image (cf. Figure 2-1).

The WMC consists of three independent steps and if we consider these individually further improvements can be formulated. In step 1, the watershed algorithm, the initialization of the watershed algorithm is currently based on local maxima; other schemas must be investigated to render a better initialization. Now, a priori knowledge is not used whereas this might facilitate a better estimate for the initialization. In step 2, fuzzy weighted C-means clustering is used, however, other clustering approaches can be probed; similarly to step 1, a priori knowledge on the intensity distribution might be supportive in finding a better clustering approach. Regarding step 3, we implemented only a few of the situations of oversegmentation. This particular step of the algorithm can be adapted to experimental conditions, i.e. a priori knowledge can be tuned with respect to the experiment so as to overcome certain imperfections of earlier steps. In future research this will be elaborated, however, the global idea of the WMC algorithm will stand its case.

The WMC has been successfully applied to other experiments in the domain of bio-imaging, e.g. detection of small vessels [18], cell membrane [70] and cytoskeleton formation [90]. With further generalization, the algorithm can be engaged in a broader scale of imagery. The future research on the tuning of the subsequent steps of the WMC algorithm will contribute to this generalization.

