Qualitative and Quantitative Evaluation of Two New Histogram Limiting Binarization Algorithms

Jan Brocher *jan.brocher@biovoxxel.de*
BioVoxxel
67112, Mutterstadt
Germany

Abstract

Image segmentation and thus feature extraction by binarization is a crucial aspect during image processing. The "most" critical criteria to improve further analysis on binary images is a least-biased comparison of different algorithms to identify the one performing best. Therefore, fast and easy-to-use evaluation methods are needed to compare different automatic intensity segmentation algorithms among each other. This is a difficult task due to variable image contents, different histogram shapes as well as specific user requirements regarding the extracted image features. Here, a new color-coding-based method is presented which facilitates semi-automatic qualitative as well as quantitative assessment of binarization methods relative to an intensity reference point. The proposed method represents a quick and reliable, quantitative measure for relative binarization quality assessment for individual images. Moreover, two new binarization algorithms based on statistical histogram values and initial histogram limitation are presented. This mode-limited mean (MoLiM) as well as the differential-limited mean (DiLiM) algorithms were implemented in ImageJ and compared to 22 existing global as well as local automatic binarization algorithms using the evaluation method described here. Results suggested that MoLiM quantitatively outperformed 11 and DiLiM 8 of the existing algorithms.

Keywords: Automatic Segmentation, Intensity Thresholds, Binarization Quality Assessment, Quantitative Segmentation Evaluation, ImageJ.

1. INTRODUCTION

Simple image segmentation by binarization methods have a long tradition and are broadly used in image processing. Generally, the binarization process is achieved by applying an intensity threshold dividing the image in two distinct regions, foreground and background. All objects of interest should thus be assigned to the foreground and the remaining parts of the image to the background. The aim of such methods in most cases is to extract features to be able to quantify object numbers, measure object sizes (e.g. area, length, area fractions, volumes surfaces), extracting text from documents or mask original images to specific regions retrieved from a binarized image. To achieve a high quality binarization with a reliable and good separation of foreground from background pixels depends first of all on the overall image quality. This is influenced by the lighting properties, objective quality, final resolution of the digital image, signal intensity, contrast and dynamic range as well as the objects' properties such as distance to neighboring objects, differences in intra-object intensities or the object-background intensity distribution. For most applications image pre-processing needs to be included to improve binarization. This might be done by using image filters to homogenize certain intensities while preserving object boundaries (e.g. applying a median convolution filter) or by background subtraction algorithms to eliminate unspecific signals. The latter is frequently necessary in fluorescent microscopic images from biological samples. Due to the availability of a plethora of processing methods the user always needs to be aware not to alter the primary image data too much to still preserve the objects properties before binarization to end up with a binary image which represents the objects of interest best. Manual segmentation in most image analysis software can be done by choosing a threshold value lying within the dynamic range of the image.

This theoretically leads to 255 different options in an 8-bit image and already 65536 possibilities in a 16-bit image. Thus, finding a reliable manual intensity threshold might be a tedious and user-biased procedure with the need to process each image individually. To be able to compare images of one series/experiment binarized by a manual threshold all images of that series would theoretically need to be segmented using the same upper and lower threshold values to prevent further increase in bias during image comparison. Those fixed thresholds will seldom lead to a satisfying overall outcome. The dependence on the image content and signal intensity rectifies the use of thresholds which take the pixel intensity distribution into account by basing the threshold choice on the images' histogram and its properties.

Therefore, many automatic thresholding methods of digital images have been proposed in the last decades [1], [2]. Automatic thresholds can be basically divided in global and local algorithms. Global algorithms normally use the complete image information for threshold determination and are applied to the complete image equally. Local algorithms, in contrast, determine various thresholds from different partitions of the image and/or contextual relations between those and allow varying threshold application in different image partitions.

The basic advantage of automatic thresholds is that they account for image variations and make a batch processing of many images easy without user bias.

Besides its long history and broad application, the topic of developing better new binarization algorithms as well as their qualitative and quantitative evaluation and comparison remains an active field of research and is extremely important to improve image analysis further.

Nevertheless, quality assessment of different binarization algorithms is not a simple task since certain evaluation criteria need to be specified in advance to allow quantitative evaluation of threshold performance. Thus, there is no real objective procedure to conduct such a comparison [3] since it furthermore depends on the nature of the scene or sample shown in the image and the needs of the user or the further analyses. There is also no perfect ground truth segmentation to which other methods could be compared [4].

Attempts of quality evaluation, mostly on grayscale images, have been made by using shape measures and region uniformity [1] or other criteria [5] as well as global image contrast and classification error calculations [6] or machine learning [7], [8]. The latter method for example is also based on the user input during the training session using example images and then performs an unsupervised evaluation based on the input. Other methods for completely unsupervised evaluation have a varying performance depending on the images and their content [7] and often need the previously created binary images.

The method presented here does not allow a completely unsupervised evaluation due to the necessity of a user-selected reference point based on pixel intensity. Besides a certain bias, this enables an individual quality assessment dependent on the image content as well as the user's needs for further analysis. The implementation in the Java based software ImageJ (and Fiji) allows a straight forward application to the original grayscale image. Several available binarization algorithms are applied at once to the image with the consecutive qualitative and quantitative evaluation and comparison.

The theory and algorithm of the qualitative and quantitative binarization evaluation method described here can be individually applied to all available automatic thresholding algorithms without any limitation to specific software, image contents or binarization performance.

2. MATERIAL AND METHODS

2.1 Qualitative and Quantitative Binarization Evaluation

The nature of any qualitative and quantitative evaluation procedure is to be able to separate true positive image areas from misclassified pixels. In the proposed method this separation is achieved by reference image pre-processing and color coding (2.1.1), definition of a relative quality reference intensity (2.1.2), color coding of the images to be evaluated and a consecutive (2.1.3) quantification of the pixel numbers in the individual image compartments classified as true positive, over-estimated (false positive), under-estimated (false negative) and background (true negative) (2.1.4).

2.1.1 Reference Image Pre-processing and Color-coding

The color coding is done by first applying a green look-up-table (LUT) to a copy of the original image (figure 1A) such that all shades of gray are transferred to different intensities of green (intensity 0 = black). Furthermore, the image is slightly contrast enhanced by using linear histogram normalization with an over-saturation of 1% of the total pixel number (figure 1C). The slight over-saturation enables a better qualitative visual evaluation, especially for images with a very low over-all intensity. An over-saturation, inverse proportional to the average intensity of the image was also considered for those adjustments but did not lead to good visualization in images with a low over-all mean intensity.

For color coding two 24-bit RGB images of the same size as the original image are created. One is filled with pure red pixels (every pixel (R/G/B) = 255/0/0), the other is equally filled with pure blue pixels (every pixel (R/G/B) = 0/0/255). Thereafter, the pre-processed copy of the original image is combined with those two color images separately by an addition of the colors (figure 1D and E). These images simulate the color coding range of the original image with the following partitions:

True positive signals are depicted in yellow and bright orange while over-estimated (false positive) areas/objects are displayed in red and dark orange (figure 1B left and D). Under-estimated (false negative) areas/objects are coded in cyan and bright blue and background (true negative) is shown in dark blue (figure 1B right and E).

2.1.2 Definition of the Relative Quality Reference Intensity

As a relative reference intensity in the given example in figure 1 the darkest intensity which is meant to still belong to the bright objects (and thus needs to be recognized as such) has to be carefully chosen by the user in the saturated copy of the original image (see 2.1.1). The color-coded images are then transferred into HSB color space and the hue channel serves for further processing.

To put the users' pixel selection in perspective and avoid over-sensitivity a 3x3 pixel area around the selected pixel is evaluated in the hue channels of the two reference images (schematically depicted in figure 1B). The mean value of these 9 pixels is taken as reference limit (figure 1B, black line between 'positive' and 'over' or 'under' and 'background', respectively). See also table 1 for the assessment of different reference values. Using the basic RGB-to-HSB conversion in ImageJ, an 8-bit image is used to represent the different hue channel values with obvious accuracy limitations in comparison to calculated floating point pixel values. Furthermore, 6 color values are combined in 1 hue value to completely represent the full individual color channel range. This mapped the 255 different shades from yellow to red to the hue channel intensity values 0-42 and cyan to blue to the hue channel intensity values 127-170, respectively.

The quality cut-off values retrieved from the two color-coded reference images according to the users' original reference point selection (indicated by the two-headed arrow in figure 1B to refer to its flexibility) serve as cut-off values to divide the evaluation partitions as shown in figure1B.

A higher precision by calculating the individual hue channel floating point pixel values and using them as bases for the quantification was also evaluated but did not markedly improve the over-all evaluation performance (data not shown). In contrast, the time performance dropped massively which was further influenced by an increasing image size.

2.1.3 Color-coding of the Tested Image for Qualitative Evaluation

After performing the individual automatic thresholding methods the resulting binary images are color coded such that a red color (R/G/B = 255/0/0) is assigned to the foreground objects (figure 1F, G and H depicted in white) and a blue color (R/G/B = 0/0/255) is assigned to the background (figure 1F, G and H depicted in black). According to 2.1.1 the pre-processed copy of the original image is combined with those binary images resulting in the final color-coded images for qualitative visual evaluation (figure 1I, J and K). Those serve as a visual control for the final quantitative assessment.

2.1.4 Quantitative Assessment of Binarization Algorithms

According to the cut-off reference the pixel hue values ($H_{u,v}$) at position $I_{u,v}$ in the image are assigned to one of the four possible partitions:

$$\text{(Eq. 1)} \qquad I_{u,v} = \begin{cases} \text{over-estimated (false positive)} & \text{if} \quad 0 < H_{u,v} < \text{cut-off}_1 \\ \text{objects of interest (true positive)} & \text{if} \quad \text{cut-off}_1 <= H_{u,v} <= 42 \\ \text{under-estimated (false negative)} & \text{if} \quad 127 <= H_{u,v} < \text{cut-off}_2 \\ \text{background (true negative)} & \text{if} \quad \text{cut-off}_2 <= H_{u,v} <= 170 \end{cases}$$

with cut-off$_1$ = value between 0 and 42 (range from red to yellow) and cut-off$_2$ = value between 127 and 170 (range from cyan to blue)

The individual number of pixels assigned to the four partitions are used to calculate the percentage they cover in the complete image.
As a final quality parameter a measure was chosen equal to the one used by Shufelt [9]. This is:

(Eq.2) relative quality = 100% * true positive / (true positive + false positive + false negative)

The relative quality can be used to figure out the binarization method(s) which perform(s) best in respect to the reference point selection in comparison to other thresholds.

2.2 Mode-Limited Mean (MoLiM) and Differential-Limited Mean (DiLiM) Algorithms

The new binarization algorithms proposed within this manuscript first assumes some premises on the histogram of the image to be binarized.

(1) For the MoLiM algorithm the histogram needs to have a mode value (the intensity value which occurs most frequently in the image). For the second modified version of the algorithm (DiLiM) a median or intermediate mean intensity value (mean* see Eq. 3) might be calculated and used instead of the mode. The existence of a mode or a median value is true for all real images even in the artificial case of an image that contains the same intensity value in all pixels.

(2) It is made the assumption that the objects of interest cover the smaller part of the image (<50% of the total pixel number) while the background takes on >50% of all pixels. In the case of bright objects on a dark background this leads to a mode value smaller than the mean image intensity (mode<mean). In case the objects fill over 50% of the image area, the resulting binary image is inverted by default for final correct feature segmentation.

(3) The objects of interest are thought by default to be brighter than the background. For images which fulfill the opposite characteristics the mode value is most likely higher than the mean image intensity. For further processing, images which fulfill the latter characteristic will be simply inverted for the further processing to meet the criterion (2) mode < mean.

(4) Due to the assumptions (2) and (3) the mode will be most likely an intensity value which is present in the range of pixels to be finally assigned to the background in the binary image.

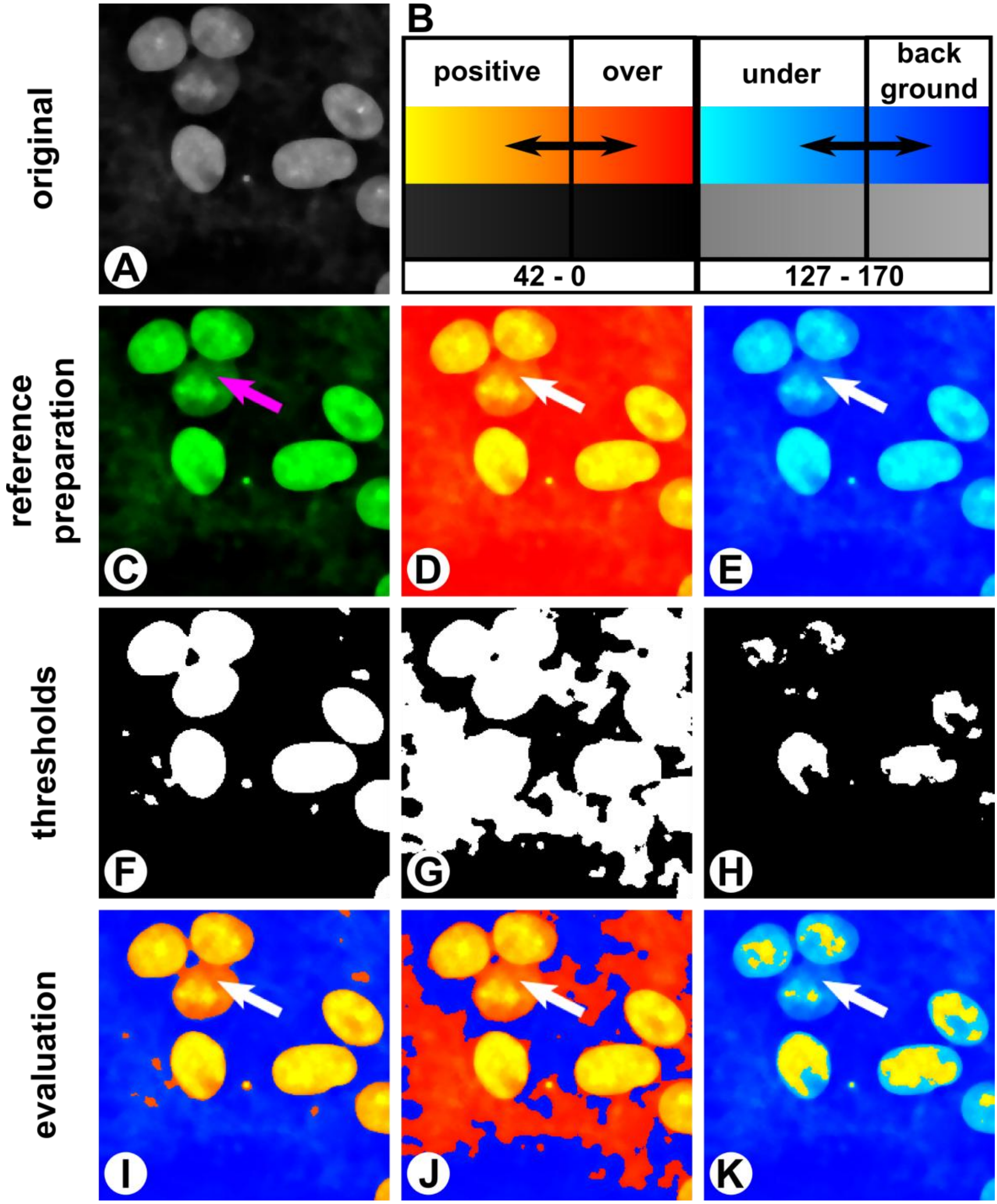


FIGURE 1: Qualitative and quantitative binarization evaluation procedure. (A) original image, (B) Image partitions scheme with assigned colors in RGB color space and the hue channel representation of HSB color space (gray scale below color representation) with the respective minimal and maximal pixel values assigned in the hue channel. (C) Pre-processed copy of original image, (D) reference image to determine cut-off value for separation of true positive pixels (yellow) from false positive (over-estimation, red) ones, (E) reference image to determine cut-off value for separation of true negative pixels (background, dark blue) from false negative (under-estimation, cyan) ones, (F-H) three different automatic binarization methods applied on the original image, (I-K) color-coded images for qualitative and quantitative evaluation of the different binarization methods shown in F-H. The arrow indicates the point selected for reference value determination. (I) shows an acceptable threshold, (J) an over-estimating and (K) and under-estimating threshold.

(5) If assumptions (2) - (4) are true, all values lying left of the mode (lower intensities) will finally be assigned to the background. Due to this fact, they might negatively influence the determination of some thresholds by shifting them towards the background intensities and away from the objects intensities. An elimination up to and including the mode intensity value initially limits the image content towards the intensities of the objects of interest and might thus further improve optimal threshold determination (see figure 2).

For the mode-limited mean (MoLiM) algorithm, a new mean value is calculated after initial intensity limitation from the restricted area which is then taken as the final threshold value.

In the case of the differential-limited mean (DiLiM) algorithm the initial limitation is done depending on the histogram context.

$$\text{(Eq. 3)} \quad T_{init} = \begin{cases} 1 - 255 \rightarrow \text{mean}^* - 255 & \text{if} \quad \text{mode} = 0 \;\&\; \text{median} = 0 \\ \text{median} - 255 & \text{if} \quad \text{mode} = 0 \;\&\; \text{median} > 0 \\ \text{median} - 255 & \text{if} \quad \text{mode} > 0 \;\&\; |\text{mode-median}| < |\text{median-mean}| \\ \text{mode} - 255 & \text{if} \quad \text{mode} > 0 \;\&\; |\text{mode-median}| > |\text{median-mean}| \end{cases}$$

mean* = intermediate mean of the histogram range 1-255.

After this limitation to T_{init} - 255 the new mean value is calculated as mentioned above and taken as the final threshold value before binarization. A comparison between a mean threshold and the MoLiM is schematically outlined in figure 2.

2.3 Test Images

The battery of test images for the evaluation of the mode-limited and differential-limited mean algorithm performance and the threshold evaluation method described above were chosen by including images with different histograms. In total 18 different images were evaluated. An excerpt of those test images is depicted in figure 3. The red arrows indicate the features of interest. The corresponding histograms are depicted next to or below the respective images.

2.4 Software

The method proposed here, makes use of the open source image processing and analysis software ImageJ [10] or Fiji [11] and its macro scripting language but could easily be transferred into a Java-plugin or using any other programming language. The decision to use the ImageJ software was also to make the evaluation method available for a big community of users without previous knowledge on the algorithms. The additional ease of using it inside ImageJ or Fiji is the presence of 16 global and 6 local automatic thresholding algorithms (at the time of manuscript preparation). Information and source code for the presented evaluation method can be downloaded and used in ImageJ or Fiji [12]–[14]. The two new binarization algorithms described here were implemented as an ImageJ/Fiji Java-plugin and will be publically available with publication of this manuscript as a Fiji update from the BioVoxxel update inside the Fiji software [14].

For programming and evaluation of the algorithm Fiji with the ImageJ version 1.48f-h under Java 1.6.0_24 [64-bit] was used. This version applied the AutoThreshold v1.15 from Gabriel Landini (19th February 2013) and the AutoLocal Threshold v1.4 from Gabriel Landini (2nd November 2011) [15], [16]. Three more local thresholds were available after manuscript preparation (updated on 18th November 2013) which were not included in this evaluation. Binarization was performed using the default settings for the local thresholding algorithms which were: radius = 15 pixels, parameter 1 = 0, and parameter 2 = 0.

2.5 Statistics

To compare the average performance quality measure of MoLiM and DiLiM an unpaired two-tailed student's t-test was performed due to variance homogeneity (determined using an F-test). The confidence interval was chosen at 95% (alpha = 0.05).

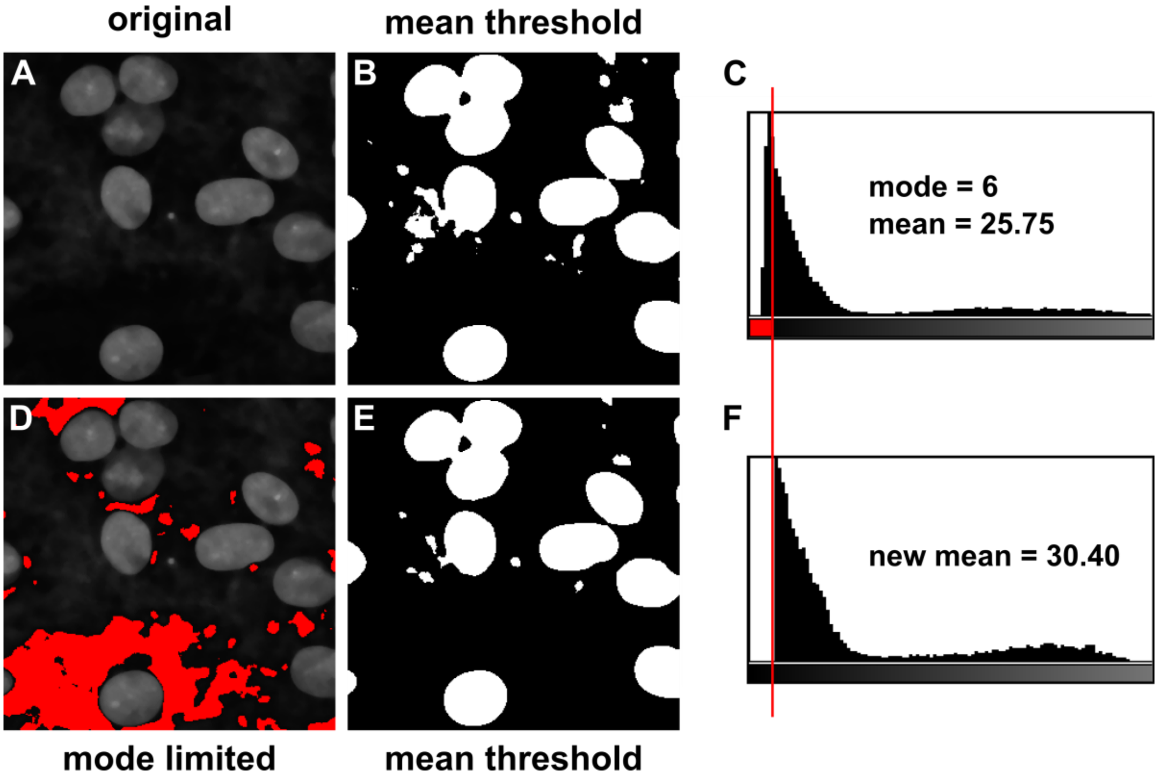


FIGURE 2: Outline of the mode limited mean algorithm. If the mean histogram value is taken as threshold on the original image (A) areas outside the wanted regions of interest are binarized (B). Initial limitation to the value above the mode value (7 in the example above) (C) excludes all pixels with a lower intensity value (D shown in red). This restricts the remaining image area and the histogram leading to an increase of the new histogram intensity mean (mode limited mean) and to a more specific binarization (E and F).

3. RESULTS

3.1 Influence of Reference Point On Evaluation Output

First, the threshold quantification algorithm was checked for the influence of the reference point position and thus its intensity. Five different reference points were chosen in the same image as depicted in figure 4. The intensity differences in the selected point were as follows (TABLE 1):

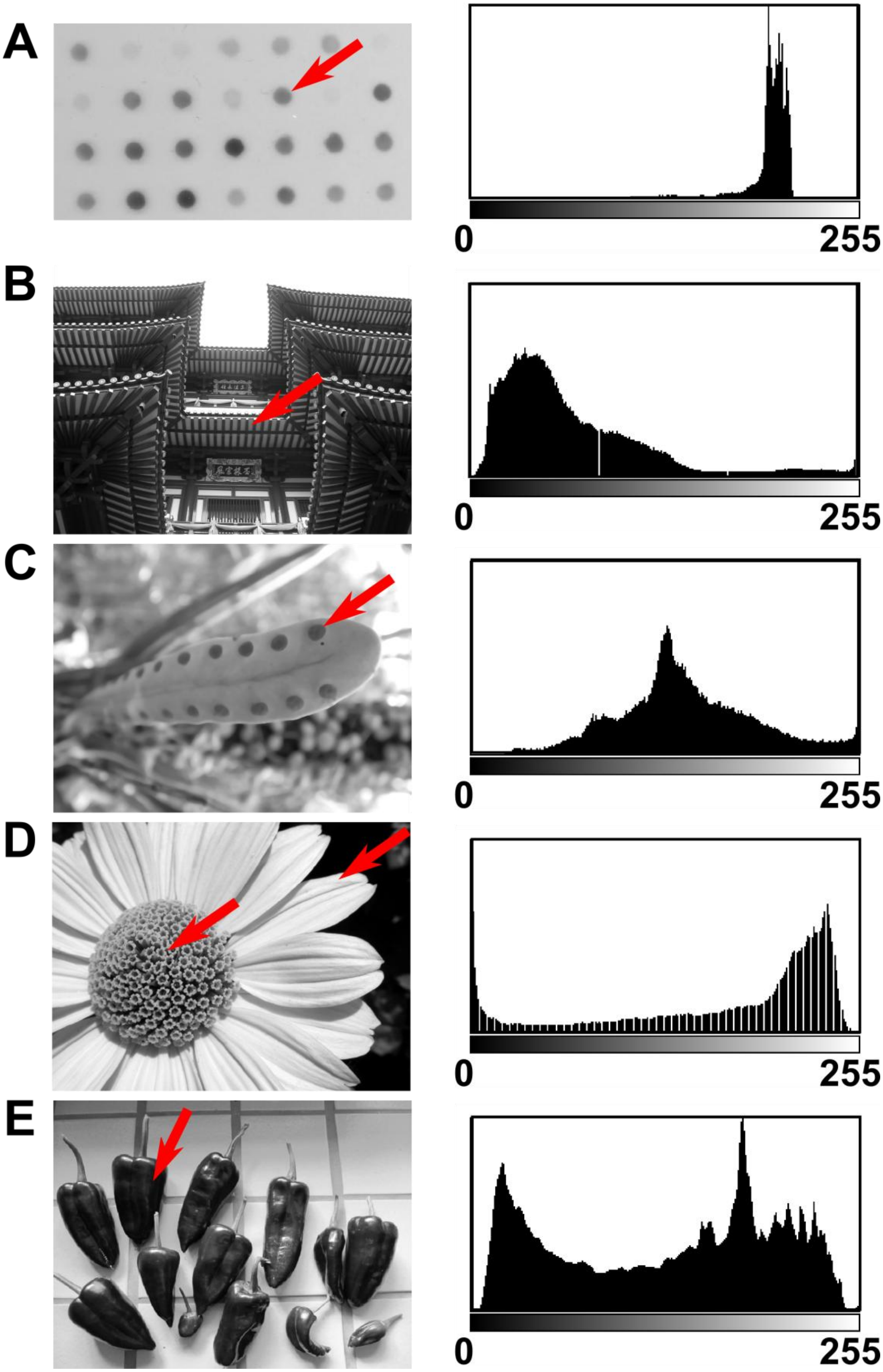
reference position	mean of reference pont	mean of 3x3 reference area
1	30	30.222
2	45	45.333
3	63	63.667
4	90	89.667
5	26	26.444

TABLE 1: Mean values of the reference point selection and the average intensity of the 3x3 area around the reference point.

3.2 Robust High Performance of MoLiM and DiLiM Algorithms

The application of either the MoLiM or the DiLiM on 18 different test images revealed a good performance in feature extraction of both algorithms. Nine of those individual performance analyses are shown in figure 5. In 2 out of the 18 cases MoLiM performed better than all other methods (on images': Chillis and Muscle). The performance of MoLiM and DiLiM was above the average performance of all methods for the individual images with the exception of one image in the case of MoLiM and 3 images for DiLiM. Both methods also were performing superior than the over-all cumulative performance quality for the complete set (all methods tested on all test images) of 62.82%. None of the local thresholds had an average performance better than the over-all performance quality. The global thresholds "Minimum", "Percentile" and "Shanbhag" also performed worse than this limit. The best thresholds in the current test set besides MoLiM was "RenyiEntropy" and "Li". MoLiM had a tendency to give the best binarization results under the chosen conditions while this was not significant for all comparisons as shown in figure 6. The latter indicates that both, MoLiM and DiLiM, performed significantly better than all the local thresholds as well as the global thresholds "Percentile" and "Shanbhag". MoLiM was further superior to "Intermodes", "MinError" and "Minimum".

Thus, in the context of specific feature extraction, the MoLiM and DiLiM algorithm can be considered to be a robust binarization method over a wide range of different digital images and histogram shapes.


A
0
255
B
0
255
C
0
255
D
0
255
E
0
255

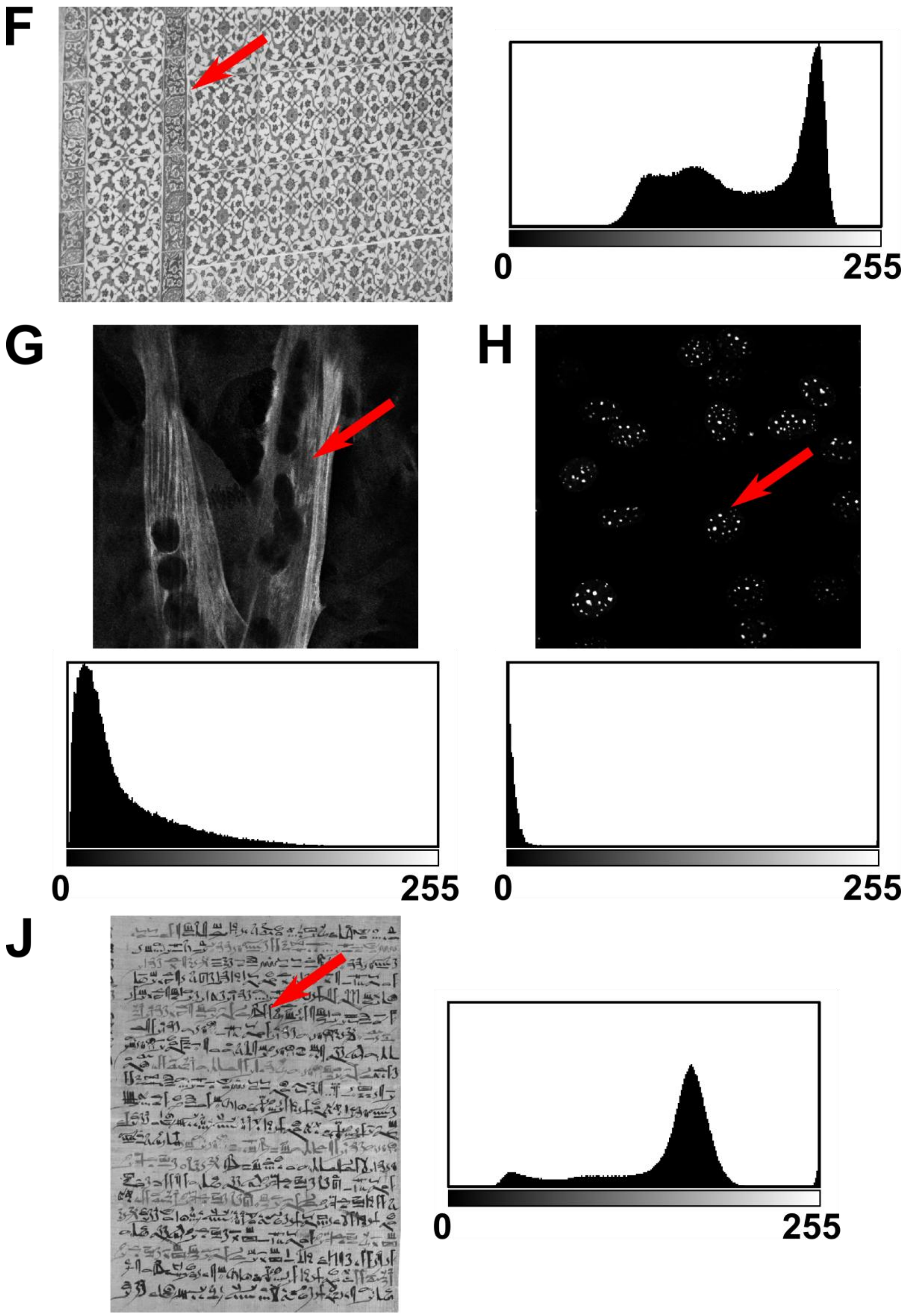

F
0
255
G
H
0
255
0
255
J
0
255

FIGURE 3: Excerpt of test images to investigate the performance of the "mode limited mean" and "differential limited mean" algorithms. Images were chosen to reflect different histogram shape distributions to account for a variety of images where binarization might be applied, such as landscapes, natural scenes (B-E), patterns (F), biological images (A, G and H), and text documents (J, [17]). The corresponding histograms are depicted on the right hand side or below the respective images.

As expected, the reference point selection showed an influence on the evaluation and thus also on the suggestion of the respective threshold to be considered the best. The thresholds performing with a similar qualitative appearance and quantitative value (Li and MaxEntropy in figure 3B and F) obviously are those with reference mean intensity values lying close together.

This indicates, that the choice of the intensity reference cut-off value needs to be done very thoroughly to retrieve reliable results. Additionally, the quantitative results need to be seen in the context of the qualitative visual output in the color-coded images as well as the analytical context in which the images will be analyzed. This also permits the influence of the user which thus can figure out thresholds fulfilling his/her needs apart from pure intensity separation quality.

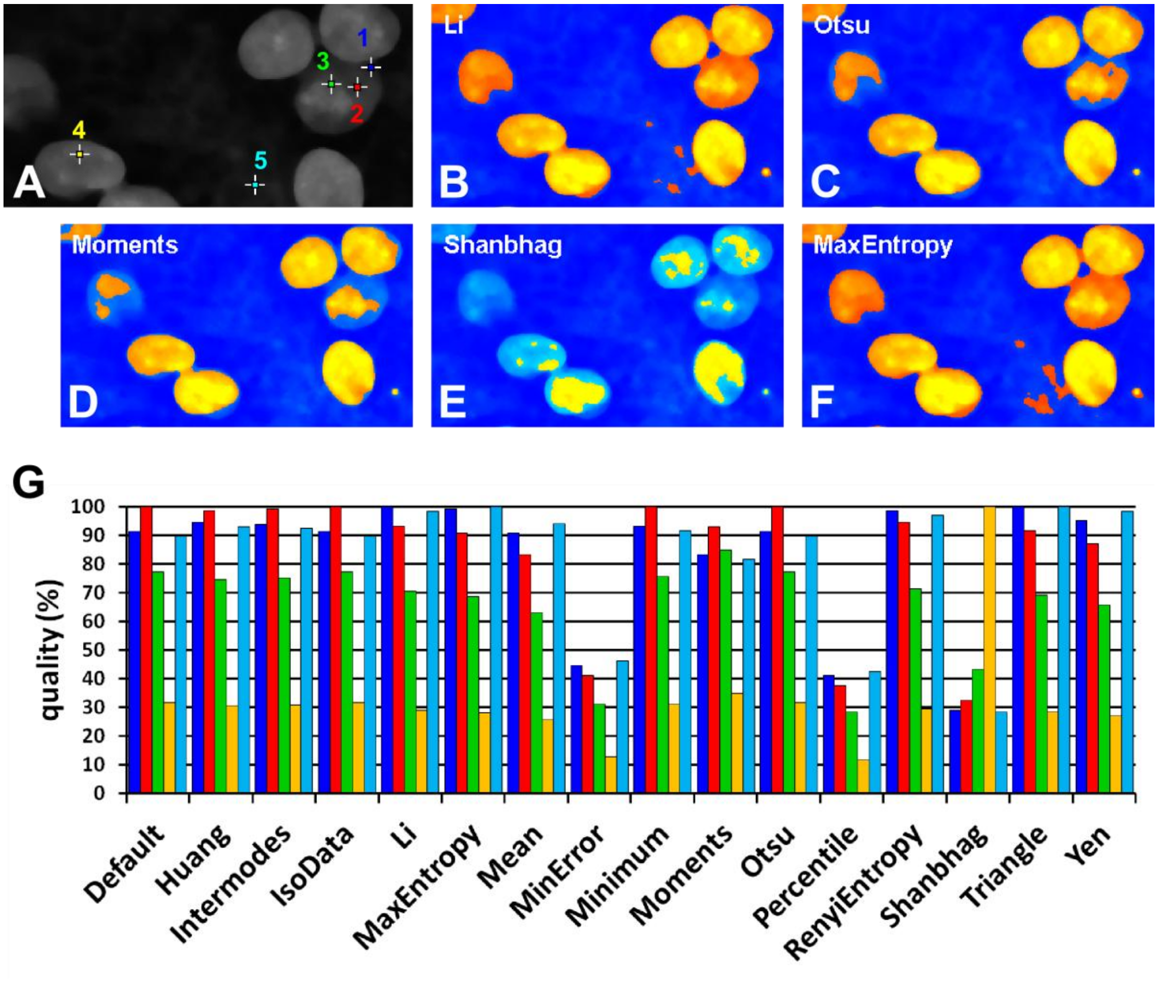


FIGURE 4: Influence of reference point position on evaluation results. (A) indication of reference point position. (B-F) Thresholds quantified as the best thresholds relative to the reference point selection: (B) reference point 1, threshold: Li, (B) reference point 2, threshold: Otsu, (B) reference point 3, threshold: Moments, (B) reference point 4, threshold: Shanbhag, (B) reference point 5, threshold: MaxEntropy. (G) Graph showing the relative quality measure in percentage in relation to the reference point selections (indicated by the same colors).

4. DISCUSSION

4.1 Semi-automatic Quantitative and Qualitative Binarization Evaluation

Objective evaluation of binarization techniques is a difficult task, since specific reference criteria need to be set fixed beforehand or need to be calculated from the original image content itself in a way which enables realistic quality assessment. User-specific requirements on the extracted objects further make this more difficult. Besides the advantage of unsupervised binarization analysis, methods with either fixed quality reference parameters or those applying machine learning algorithms on representative images might suffer from a decrease in reliability depending on several factors. Those include an increasing variability in image content or intensities as well as contrast differences in a series of images to be evaluated together or difficult automatic shape or edge determination of objects.

The method presented here does not apply complete unsupervised analysis but allows an easy and quick evaluation of several binarization methods for untrained users. Obviously, a certain user bias is unavoidable and the reliability of the quantitative output relates to the carefulness during reference intensity cut-off value selection. It should also be mentioned that the method could in theory be implemented as unsupervised batch processing procedure with a fixed reference intensity value given as delineation criteria. Of importance is that comparison of all binarization algorithms present in the respective implementation is done in relation to the very same, previously chosen reference value and thus represents an unbiased evaluation of different binarization results from the same original image. In general, it is less biased and superior to manual threshold definition as well as to simple visual comparison of the binary results of several automatic segmentations with the original image counterparts.

The current literature contains several reports on different evaluation methods regarding image segmentation in general, while many of those deal with multi-class segmentation algorithms such as "Mean Shift", other statistical region merging methods which are mostly applied to color images [8], [18]–[21]. Those analyses are unequally more difficult compared to the problem discussed here. This originates from the fact that the classification into different regions which in the best case resemble separated real objects is a difficult task itself while the individual perception on those might additionally be highly variable. Thus, a decision on the correct or incorrect pixel classification needs to include many different criteria such as intra- as well as inter-region variances (e.g. by using entropy calculations) [19]. Of note, while the present report rather deals with the separation of information into 2 classes only, including techniques of regional combination as pre-processing step in grayscale images might in turn improve the object separation during binarization. Ge and coauthors present evaluations according to the extraction of the most salient object in the image [22] which is not necessarily possible by separation of intensities above a certain threshold from those below as mostly aimed for during binarization. Therefore, such a method needs to consider different image properties regarding reliable segmentation events. Most publications evaluating binarization algorithms do this in the limited context of document binarization and thus text extraction [3], [23]–[25] which in comparison to the above mentioned methods as well as the use of natural or biological images might show a rather low complexity. One of those reports compared binarization results of synthetically created images with their corresponding original counterparts [3] which makes the evaluation during a test setup less error prone but is not transferable to real life examples. This applicability to real images was an important prerequisite during the development of the proposed method. The presented evaluation is independent of the image context and only limited to the intensity distribution in grayscale images.

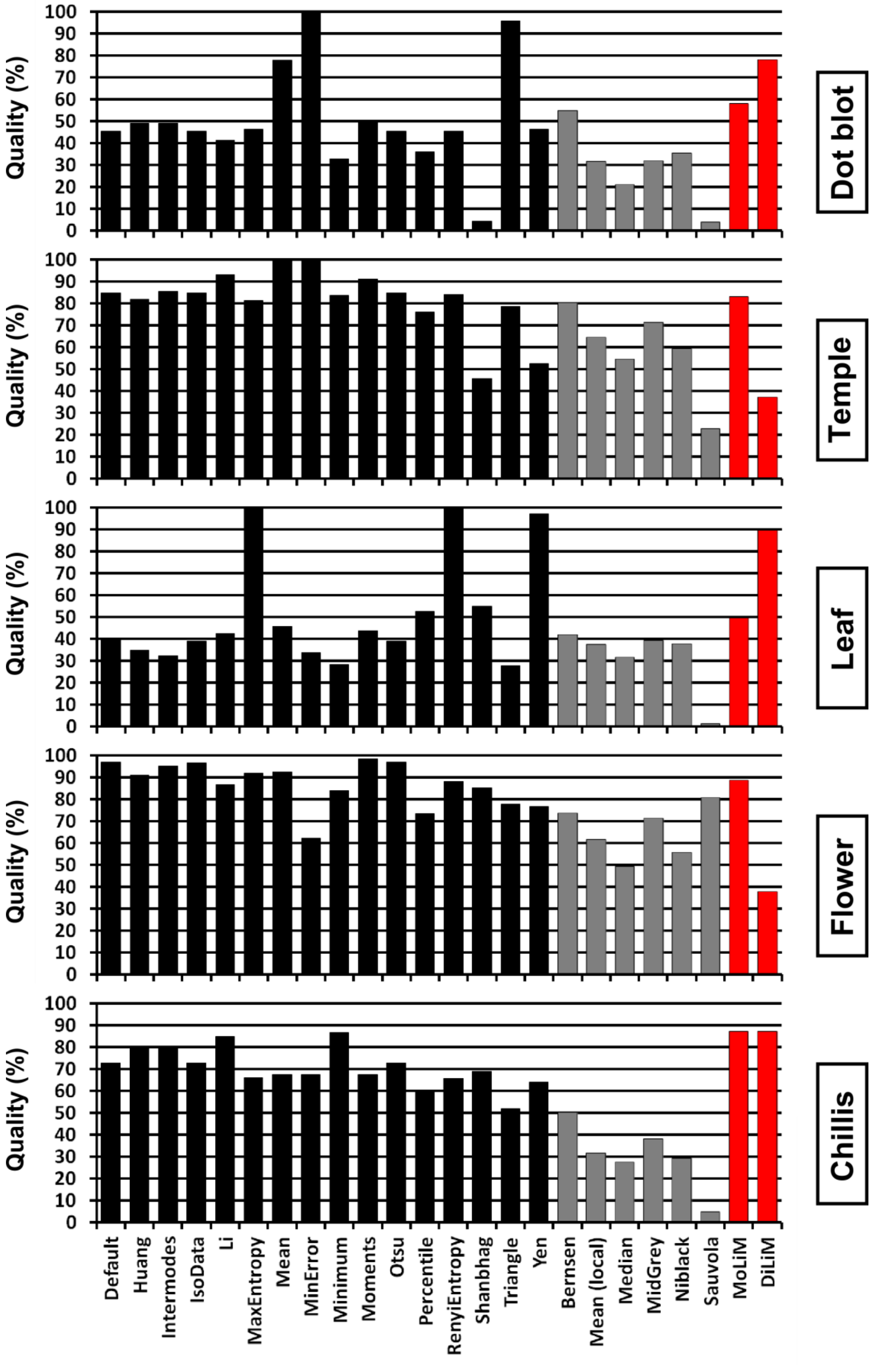
Furthermore, available methods often rely on the comparison of the binarization result with one or more ground-truth images mostly retrieved by previous manual segmentation [22], [26]. This makes the procedure time intensive and needs one or more experts to create more or less reliable ground truth samples. Moreover, those examples not necessarily present the best basis for comparison in a context with high inter-image variability. Different unsupervised methods compare regions to e.g. edge maps determined by edge finding algorithms [21] and thus strongly rely on the performance quality of those techniques.

The proposed method takes the original image information as the best and unaltered ground truth which perfectly accounts for the different intra- as well as inter-image variability in addition. Hence, there is no time consuming search or biased creation of ground truth segmentations necessary which as mentioned naturally neglect a certain amount of variability during comparison. Nevertheless, the reference intensity cut-off needs to be chosen in the described method individually which, without a doubt, includes a certain amount of user bias as well. Besides this, a basically infinite amount of known binarization algorithms can be compared straight forward based on the same reference point and as such the proposed method represents a quick, robust and reliable, quantitative measure for relative binarization quality assessment for individual images.

The described comparison of binarization methods also might highlight certain limitations of a pool of automatic segmentation algorithms regarding a specific image when reaching relative qualities markedly below 100% as well as non-satisfying visual output. The latter might directly indicate that a complete or partial feature extraction in the current image under the chosen conditions might not be achievable without further image pre-processing (e.g. filtering) or image improvement and thus saves time by avoiding a trial and error adjustment e.g. using manual thresholds.

Therefore, a thorough selection of the reference value is crucial for the performance of the quantitative evaluation. The reference point has to be positioned as close as possible to the intensity value which should still be recognized (darkest value to be accepted for bright objects and brightest value to be accepted for dark objects) and extracted.

In summary, the described algorithm is based on a qualitative output for visual analysis of color-coded images and a numerical quantification of the binarization performance. In combination, both output formats provide an easy evaluation method to test the quality of individual image segmentations and reliably compare them to each other under set conditions. This is especially useful to determine the best segmentation algorithm when several of them perform visually similar. Then, the quantitative evaluation might point out slight superior binarization performances not visually determinable by the user in the context of the reference point selection.


Quality (%)
0
10
20
30
40
50
60
70
80
90
100
Dot blot
Temple
Leaf
Flower
Chillis
Default
Huang
Intermodes
IsoData
Li
MaxEntropy
Mean
MinError
Minimum
Moments
Otsu
Percentile
RenyiEntropy
Shanbhag
Triangle
Yen
Bernsen
Mean (local)
Median
MidGrey
Niblack
Sauvola
MoLiM
DiLiM

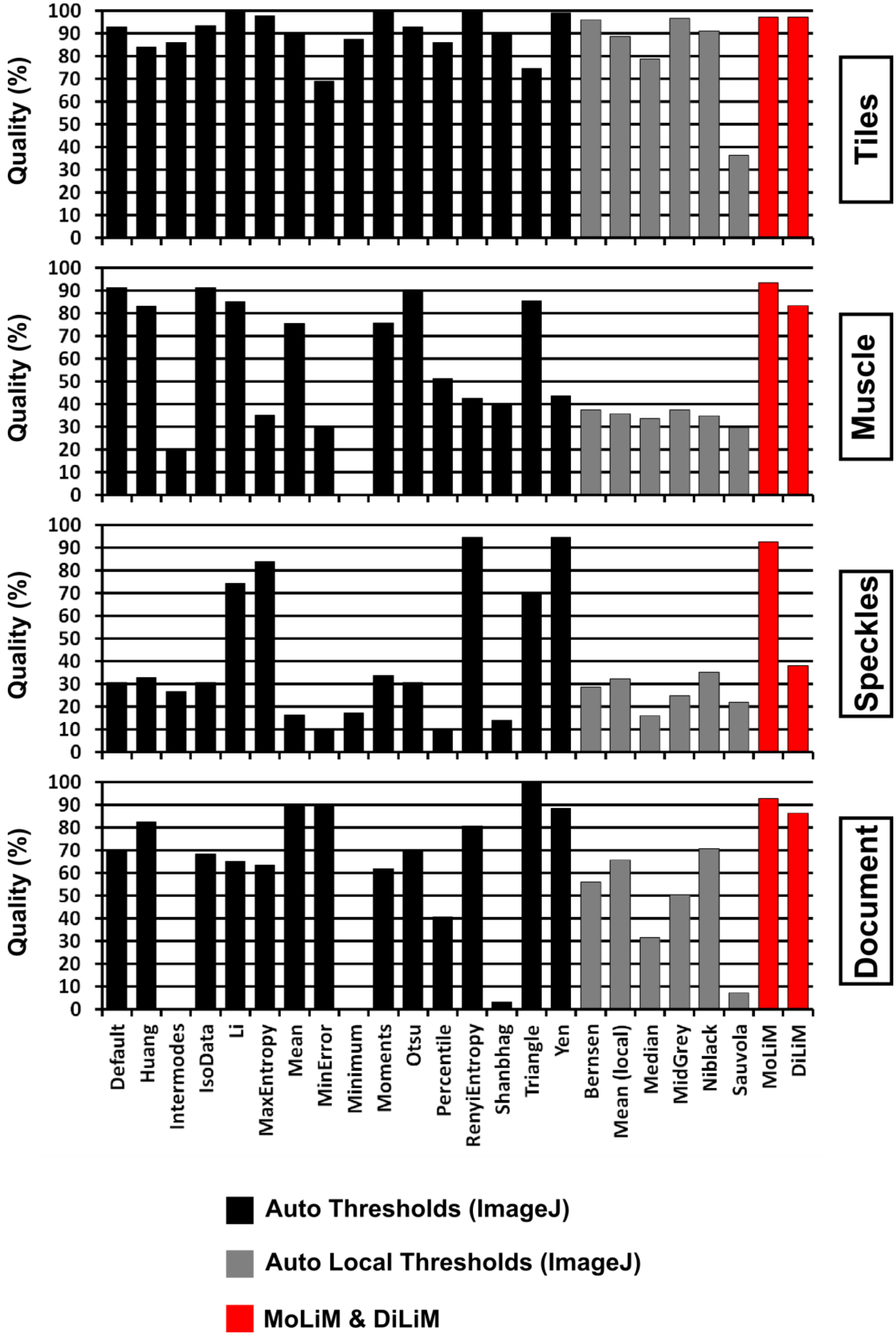

Quality (%)
0
10
20
30
40
50
60
70
80
90
100
Tiles
Muscle
Speckles
Document
Default
Huang
Intermodes
IsoData
Li
MaxEntropy
Mean
MinError
Minimum
Moments
Otsu
Percentile
RenyiEntropy
Shanbhag
Triangle
Yen
Bernsen
Mean (local)
Median
MidGrey
Niblack
Sauvola
MoLiM
DiLiM
Auto Thresholds (ImageJ)
Auto Local Thresholds (ImageJ)
MoLiM & DiLiM

FIGURE 5: Exemplary performance evaluations of nine test images using the "Threshold Check" proposed above. Quality output is shown in percent (%) relative to the chosen intensity reference cut-off value. Black columns represent ImageJ AutoThreshold methods, gray columns the Auto Local Thresholds implemented in ImageJ and red columns indicate the newly proposed methods MoLiM and DiLiM. The graphs correspond to the test images depicted in figure 3.

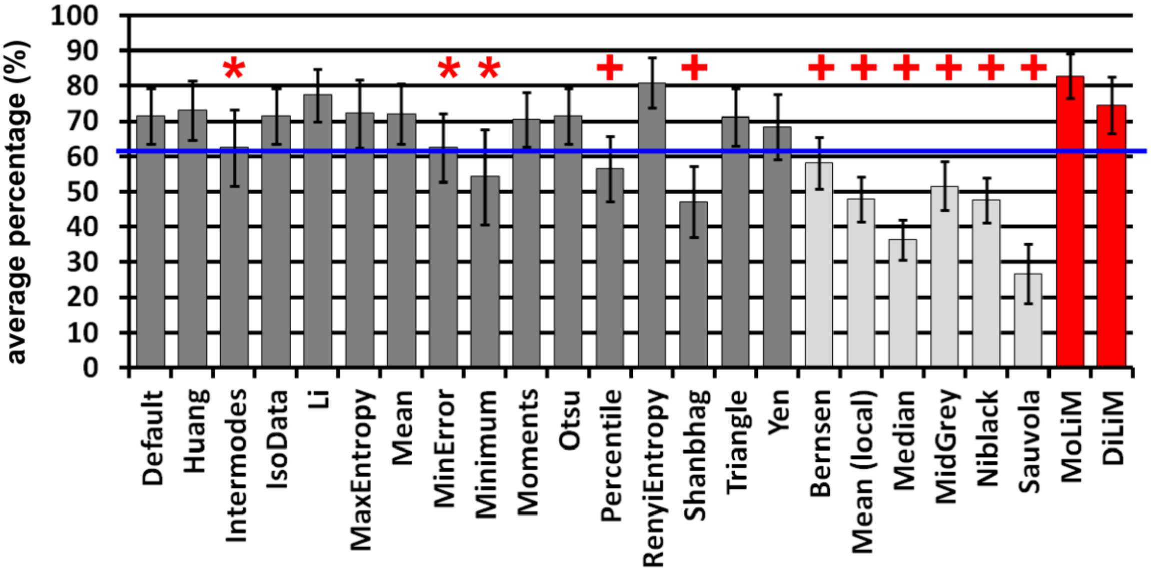


FIGURE 6: Average performance evaluation of MoLiM and DiLiM algorithms (red) in comparison with 16 established global (dark gray) and 6 local (light gray) automatic binarization methods. Depicted is the individual average relative quality measure in percent using n=18 test images with different histogram appearances. Error bars indicate the standard error of the mean (S.E.M.). The blue line indicates the average performance calculated from all auto threshold algorithms over all tested images together. Asterisks indicate significant better performance of the MoLiM, while crosses indicates a significantly better performance of both, MoLiM and DiLiM, in comparison to the indicated methods. Significance was accepted for p-values <0.05.

4.2 MoLiM and DiLiM Performance

The newly proposed binarization algorithms MoLiM and DiLiM use an initial limitation of the histogram and the respective image area according to the histograms mode or median value and hence achieve a good separation of potential background from foreground (region of interest) pixels.

The method tested here included a low precision hue value determination assigning one hue value to six consecutive colors in RGB color space. Nevertheless, the application of a high precision hue value determination in the reference and test images had only little influence on the relative quality assessment of the individual thresholds (data not shown). Besides that, they resulted in a considerable drop of the algorithms' time performance in the current test setup. This performance decreased strongly with increasing image size. In this case, processing time strongly out-performed precision. Furthermore, the slight reduction in precision might improve the robustness regarding the intensity reference point selection, since individual pixel differences will have a reduced impact on the evaluation.

The advantage of the proposed algorithm is that it does not depend on the histogram shape such as the existance of two mode values (peaks) [27], a fixed percentage of foreground pixels [28] or the necessity to iteratively converge two threshold values [29]. Thus, besides being computationally efficient, the algorithm never fails to achieve a binarization.

MoLiM significantly out-performed 11 and DiLiM 8 out of 22 tested binarization algorithms under the chosen test conditions and in respect to the relative reference point. The latter was chosen as objective as possible to indicate a cut-off value which leads to good separation of the objects of interest in the individual images from the supposed background.

5. CONCLUSION

The semi-automatic segmentation performance evaluation presented here can be applied to all types of binarization algorithms as long as the original image is available. Furthermore, a set of automatic binarization methods can be included directly before the analysis as implemented in an ImageJ macro script prepared to produce the results described in this manuscript. This makes the evaluation process extremely time efficient. It is easy to use and simply implies some knowledge about the objects of interest and the requirements for further analyses.

The MoLiM and DiLiM binarization algorithms evaluated with the described quality assessment provided very robust and computationally low-cost segmentation results as alternatives to many other existing thresholding methods.

6. FUTURE DIRECTIONS

Besides the fact of the often stated importance of such segmentation evaluation techniques there is only a limited number available so far which still holds potential for future related research.The presented method will be tested for application in a batch processing setup to see if, besides comparison of several binarization algorithms using one image, a reliable and robust comparison can also be achieved when comparing a set of similar images. Therefore, a fixed reference value as ground truth criteria needs to be chosen. The studies will include trials on reference intensity choices according to intra-image intensity distribution and inter-image intensity fluctuations.

7. REFERENCES

[1] P. . Sahoo, S. Soltani, and a. K. . Wong, “A survey of thresholding techniques,” *Comput. Vision, Graph. Image Process.*, vol. 41, no. 2, pp. 233–260, Feb. 1988.

[2] C. A. Glasbey, “An Analysis of Histogram-Based Thresholding Algorithms,” *CVGIP: Graphical Models and Image Processing*, vol. 55, no. 6. pp. 532–537, 1993.

[3] P. Stathis and N. Papamarkos, “An Evaluation Technique for Binarization Algorithms,” vol. 14, no. 18, pp. 3011–3030, 2008.

[4] R. Unnikrishnan, C. Pantofaru, and M. Hebert, “A Measure for Objective Evaluation of Image Segmentation Algorithms,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - Workshops*, vol. 3, pp. 34–34.

[5] J. S. Weszka and A. Rosenfeld, “Threshold Evaluation Techniques,” *IEEE Trans. Syst. Man. Cybern.*, vol. 8, no. 8, pp. 622–629, Aug. 1978.

[6] A. Najjar and E. Zagrouba, “An Unsupervised Evaluation Measure of Image Segmentation : Application to Flower Image,” pp. 448–457, 2012.

[7] H. Zhang, S. Cholleti, S. A. Goldman, and J. E. Fritts, “Meta-Evaluation of Image Segmentation Using Machine Learning,” in *EEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR’06)*, vol. 1, pp. 1138–1145.

[8] M. Borsotti, P. Campadelli, and R. Schettini, “Quantitative evaluation of color image segmentation results,” *Pattern Recognit. Lett.*, vol. 19, no. 8, pp. 741–747, 1998.

[9] J. A. Shufelt, “Performance evaluation and analysis of monocular building extraction from aerial imagery,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 4, pp. 311–326, Apr. 1999.

[10] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, “NIH Image to ImageJ: 25 years of image analysis,” *Nat. Methods*, vol. 9, no. 7, pp. 671–675, Jun. 2012.

[11] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, “Fiji: an open-source platform for biological-image analysis.,” *Nat. Methods*, vol. 9, no. 7, pp. 676–82, Jul. 2012.

[12] J. Brocher, “BioVoxxel Toolbox,” 2014. [Online]. Available: http://fiji.sc/BioVoxxel_Toolbox. [Accessed: 05-Mar-2014].

[13] J. Brocher, “Theshold Check (BioVoxxel Toolbox),” 2014. [Online]. Available: http://fiji.sc/BioVoxxel_Toolbox#Threshold_Check. [Accessed: 05-Mar-2014].

[14] J. Brocher, “BioVoxxel Fiji update site,” 2014. [Online]. Available: http://sites.imagej.net/BioVoxxel/. [Accessed: 05-Mar-2014].

[15] G. Landini, “Auto Thresholds,” 2013. [Online]. Available: http://fiji.sc/wiki/index.php/Auto_Threshold. [Accessed: 29-Nov-2013].

[16] G. Landini, “Auto Local Thresholds,” 2013. [Online]. Available: http://fiji.sc/Auto_Local_Threshold. [Accessed: 29-Nov-2013].

[17] “Edwin Smith Papyrus.” [Online]. Available: http://upload.wikimedia.org/wikipedia/commons/b/b4/Edwin_Smith_Papyrus_v2.jpg. [Accessed: 24-Jan-2014].

[18] H. Zhang, J. E. Fritts, and S. a. Goldman, “Image segmentation evaluation: A survey of unsupervised methods,” *Comput. Vis. Image Underst.*, vol. 110, no. 2, pp. 260–280, May 2008.

[19] H. Zhang, J. E. Fritts, and S. a. Goldman, “An entropy-based objective evaluation method for image segmentation,” no. 1, pp. 38–49, Dec. 2003.

[20] R. Unnikrishnan, C. Pantofaru, and M. Hebert, “Toward objective evaluation of image segmentation algorithms.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 929–44, Jun. 2007.

[21] J. Freixenet, X. Mu, D. Raba, J. Mart, and X. Cuf, “Yet Another Survey on Image Segmentation : Region and Boundary Information Integration,” *ECCV*, pp. 408–422, 2002.

[22] F. Ge, S. Wang, and T. Liu, “Image-Segmentation Evaluation From the Perspective of Salient Object Extraction,” *2006 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Vol. 1*, vol. 1, pp. 1146–1153, 2006.

[23] O. D. Trier and T. Taxt, “Evaluation of binarization methods for document images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 3, pp. 312–315, Mar. 1995.

[24] A. Kefali, T. Sari, and M. Sellami, “Evaluation of several binarization techniques for old Arabic documents images,” in *The First International Symposium on Modeling and Implementing Complex Systems MISC*, 2010, no. 1, pp. 88–99.

[25] K. Ntirogiannis, B. Gatos, and I. Pratikakis, “An Objective Evaluation Methodology for Document Image Binarization Techniques,” in *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, 2008, pp. 217–224.

[26] J. K. Udupa, V. R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn, “A framework for evaluating image segmentation algorithms,” *Comput. Med. Imaging Graph.*, vol. 30, no. 2, pp. 75–87, Mar. 2006.

[27] J. M. S. Prewitt and M. L. Mendelsohn, “THE ANALYSIS OF CELL IMAGES*,” *Ann. N. Y. Acad. Sci.*, vol. 128, no. 3, pp. 1035–1053, Dec. 2006.

[28] W. Doyle, “Operations Useful for Similarity-Invariant Pattern Recognition,” *J. ACM*, vol. 9, no. 2, pp. 259–267, Apr. 1962.

[29] J. Kittler and J. Illingworth, “Minimum error thresholding,” *Pattern Recognit.*, vol. 19, no. 1, pp. 41–47, Jan. 1986.