Image Segmentation Assessment from the Perspective of a Higher Level Task

Mariela Atausinchi Fernandez*, Rubens M. Lopes† and Nina S. T. Hirata*
*Institute of Mathematics and Statistics, †Oceanographic Institute – University of São Paulo
São Paulo, Brazil
Email: mariela@ime.usp.br, rubens@usp.br, nina@ime.usp.br

***Abstract*—Image segmentation evaluation is usually performed by visual inspection, by comparing segmentation to a ground-truth, or by computing an objective function value for the segmented image. All these methods require user participation either for manual evaluation, or to define ground-truth, or to embed desired segmentation properties into the objective function. However, evaluating segmentation is a hard task if none of these three methods can be easily employed. Often, higher level tasks such as detecting or classifying objects can be performed much more easily than low level tasks such as delineating the contours of the objects. This fact can be advantageously used to evaluate algorithms for a low level task. We apply this approach to a case study on plankton classification. Segmentation methods are evaluated from the perspective of plankton classification accuracy. This approach not only helps choosing a good segmentation method but also helps detecting points where segmentation is failing. In addition, this more holistic form of segmentation evaluation better meets requirements of big data analysis.**

***Keywords*-holistic system evaluation; classification evaluation; plankton image classification; segmentation evaluation; plankton image segmentation;**

I. Introduction

Segmentation is one of the most studied problems in the fields of image analysis and computer vision. There are many algorithms for image segmentation and, although not equally numerous, studies concerned with segmentation evaluation are also receiving increasing attention [1], [2], [3], [4], [5], [6].

Most of the approaches for image segmentation evaluation can be divided into three major categories: subjective, supervised and non-supervised [1]. Subjective evaluation basically refers to visual inspection by a specialist in the application domain. Supervised evaluation refers to methods that objectively compare segmentation to a reference segmentation (usually known as ground-truth or gold standard). Unsupervised evaluation refers to objective metrics that are computed based solely on the segmented image.

One of the difficulties pointed in evaluating segmentation is the conflict between generality and objectivity [7]. For general purpose segmentation, ground-truth and the notion of segmentation accuracy may not be well defined, and thus even using objective metrics, evaluation is not completely objective. In fact, there are several situations in which more than one ground-truth may exist. Moreover, in many cases ground-truth is built subjectively. On the other hand, if evaluation is restricted to situations where ground-truth is well defined, generality is lost.

Generality and objectivity seek applicability to general segmentation problems. However, specificity in the sense of adequation to a specific application context is also an important characteristic. This is being achieved through the creation of a myriad of objective evaluation metrics. Additionally, we can argue that context information is in fact embedded both in supervised and non-supervised evaluation methods. In the supervised case, context information is embedded in the ground-truth. In the non-supervised case, the objective function can be designed to incorporate application specific segmentation properties.

How to deal, however, with problems in which ground truth information is not available or can not be easily obtained or when context information can not be easily embedded into an objective function? For instance, ground truth can not be easily built when contours of target objects can not be clearly delineated by visual inspection only (e.g., see plankton images in Fig. 1). In addition, manual delineation of contours would be particularly critical concerning workload when there are many types of targets and a very large number of images. A relatively large number of ground truth for every variation of the input must be available to achieve some statistical significance in the evaluation.

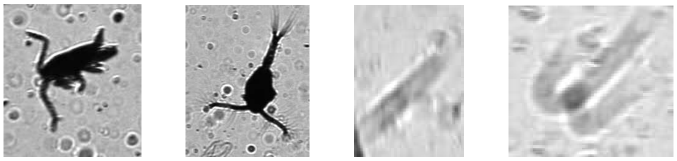


Fig. 1. From left to right: *Calanoida*, *Cyclopida*, Thick Fiber, and *Neoceratium*. The first two present good contrast and clear contour while the last two are blurred with no clear contour.

Given that segmentation is usually a step in a processing chain related to a specific application context, it is natural to expect that there should be connections between the segmentation task itself and the application goal. In fact, there are many ways to explore context in segmentation. For instance, interactive image segmentation methods exploit the fact that for human beings it is much easier to recognize an object/target than delineate its contours. Thus a common approach is to have users interactively placing rough marks on targets of interest directly on the image and then apply seed based segmentation algorithms as in [8]. Another example consists in adjusting parameters of a segmentation algorithm based on classification accuracy. In [9], from the observation that most classification errors are due to poor segmentation results, the authors propose to use classifier confidence as a measure for segmentation accuracy. When classifier confidence is low, segmentation is redone changing parameters of the segmentation algorithm and the classifier is applied to the new segmented image. The process is repeated iteratively, until a high classification confidence is achieved. Similarly, in [10], segments (superpixels) of satellite or aerial images that are generated by an unsupervised method are classified using some classification method. Classification accuracy is used to evaluate parameters of the segmentation method.

A method for evaluating solutions for low level tasks from the perspective of a higher level task, in a similar fashion as it is done to compare edge detection algorithm for object recognition in [11], may be an interesting approach to tackle some of the issues in image segmentation. A first issue, pointed above, is how to evaluate segmentation when ground-truth (or context information) can not be easily conveyed. A second issue is how relevant are the small differences among segmentations obtained with different segmentation algorithms from the point of view of the application.

Such evaluation method would fall within the *system level evaluation* category listed in [1]. Although there are some attempts on employing such approach [11], [9], it seems it has not yet been established as a common evaluation method. A disadvantage of system level methods may be the fact that it only provides indirect evaluation. However, if the main interest of evaluation is to choose a good segmentation algorithm rather than to rank segmentation methods, we could take advantage of the fact that ground-truth at a higher level processing task is usually simpler to be provided. An immediate consequence is that there would be much more data for evaluating the algorithms. Furthermore, a system level evaluation stands aligned with current efforts for a holistic evaluation of systems, for instance those seen in scene understanding as in [12], [13]. As such, it may provide information about segmentation that is valuable from the application point of view.

In this work we present a case study on assessing segmentation methods from the perspective of a higher level task. Specifically, we consider the problem of plankton classification as the higher level task. Segmentation assessment is done based on classification accuracy. This is a particularly interesting case study since targets often present no clear contour and because there is a huge amount of data (making manual delineation of contours or visual inspection of segmentation results unfeasible). On the other hand, assigning class labels to the targets is a relatively much simpler task for a specialist. This enables the construction of a much larger dataset with ground-truth (for the higher level task).

These ideas are further elaborated and discussed in Section II. In Section III we describe our case study where the processing chain of interest is composed of three main components: target segmentation, feature extraction and classification. We list the evaluated segmentation algorithms, and examples of plankton segmentation with these algorithms. We also specify the training and classifier evaluation methods used in the experiments, and report classification results with discussions on what kind of information concerning segmentation can be extracted based on the classification results. In Section IV we present our concluding remarks.

II. Evaluation method

To further motivate the idea, we first take the problem of optical character recognition [14] as an example. In order to recognize, for instance, the contents in a scanned document page, a series of processing is usually performed. Typical processing includes binarization of the image, character segmentation, character recognition, and detection of page components such as figures, titles, paragraphs among others.

In this process, binarization is an extremely important step because it affects all subsequent steps. Due to such importance, many binarization algorithms have been developed and there are even competitions on binarization algorithms [15]. On the other hand, binarization is a relatively lower level task compared to the overall task of recognizing the document page content. Thus, choosing a binarization algorithm should take into account not only the performance of the algorithms in the specific binarization task itself, but also on how does it affect performance of higher level tasks. More specifically, although a "perfect" binarization is desirable, from the perspective of the final goal, a binarization only just close to perfect might be sufficient. For instance, how much slight differences in binarization of individual characters like the one shown in Fig. 2 would affect recognition rate?

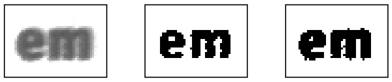


Fig. 2. Example of two slightly different binarization.

Supposing there are multiple binarization algorithms that lead to equivalent performance for a higher level task (e.g. in character recognition or even in word recognition), other aspects of the binarization algorithms such as processing speed, ease of implementation, availability, among others could be taken into consideration for selecting a particular one.

A. Procedure description

Let us consider a processing chain, composed of several components linked sequentially, where each component is responsible for a specific task. We assume that earlier components in the chain are responsible for semantically lower level tasks, while those at the end of the chain are responsible for higher level tasks. The output of one component feeds the input of the next component. Solutions for a specific component can be evaluated in terms of the performance of any of the subsequent components. A critical issue in this method is that the larger the distance between the two components, the larger the possible interference of other components in the evaluation.

To minimize possible interference of intermediary steps, the ideal is to evaluate solutions of a low level task in terms of the performance of the subsequent component in the chain. When this is not possible, the granularity of the components can be modified by fusing a subset of consecutive components into one single component. In this way, most processing chains can be reduced to a few components.

Under the above described assumptions, let then A represent the first task (the low level one) and B the second task (the higher level one) in the processing chain. Let us also suppose we are concerned in assessing different algorithms for task A, and that an input dataset D for the chain is available as well as ground truth for the output of the chain. Consider n solvers A_i for task A, a fixed solver B for task B, and a fixed evaluation metric for the output. The evaluation procedure is summarized below.

```
1) solve task A for input $D$ using each of the $n$ solvers, $A_i, i = 1, \ldots, n$, and denote the respective results as $A_i(D), i = 1, \ldots, n$
2) solve task B for each dataset $A_i(D), i = 1, \ldots, n$, using solver $B$, and denote the respective results as $B(A_i(D)), i = 1, \ldots, n$
3) evaluate performance measure for each resulting dataset $B(A_i(D)), i = 1, \ldots, n$
4) order each of the solvers of task A according to corresponding performance computed in the previous step and return
```

III. Case study on plankton classification

The discussed method is applied in a case study on plankton classification. The processing pipeline considered here consists of the following steps:

- target detection: individual target images are created by cropping targets from a large image;
- segmentation: contours of the target are delineated. This task may include pre-processing, a binary segmentation algorithm, and post processing;
- Feature extraction: several features used in [16] related to shape, size, color, Hu moments, among others, are extracted from the segmented targets;
- Classification: targets are classified using a previously trained classifier, having as input all or a selected subset of the features computed in the previous step.

In this case study we start from the point where targets are already detected. Moreover, to fit the two task model described in the previous section, we consider segmentation as task A and feature extraction+classification as task B.

A. Segmentation methods

The segmentation algorithms considered consist of simple binarization and contour detection algorithms, favoring computational simplicity due to efficiency requirements in the application context. Let I denote an image, p a point in the image domain, $I(p)$ the intensity of I at p, $\overline{I}$ the mean intensity of image I and σ its standard deviation. Targets are assumed to be relatively darker than the background. The following methods are considered. The first three are taken from [16].

- **Fixed**: smoothing with a 5×5 Gaussian mask, fixed-intensity thresholding, and morphological closing. The adopted threshold value is $T = 170$ and the structuring element of the closing operator is a 5×5 elliptical kernel.
- **Dyn**: smoothing with a 5×5 Gaussian mask and dynamic intensity thresholding. The thresholding value is computed for each image as $T = \overline{I} - c\sigma$, using $c = 1.5$.
- **Waters**: smoothing with a 5×5 Gaussian mask and watershed from markers. Foreground and background markers are given by $M_f = \{p : I(p) < \overline{I} - 2\sigma\}$ and $M_b = \{p : I(p) > \overline{I} - \sigma\}$, respectively.
- **Yen**: histogram equalization, Yen's thresholding [17], and largest connected component selection.
- **Otsu**: histogram equalization, Otsu's thresholding [18], and largest connected component selection.
- **Isodata**: histogram equalization, Isodata thresholding [19], and largest connected component selection.

Figures 3 and 4 (the latter at the end of the paper) show a sample from each of the 16 classes considered in this study, together with segmentation results obtained using the above listed six methods. The 16 classes considered here do not necessarily correspond to distinct species of plankton. A same species may have been further divided based on some subjective or convenient criteria. We adopt the division used in [16].

Any visual inspection would require examination of a large number of segmentations. Considering that the number of classes may be very large (hundreds) and that each class may have hundreds or even thousands of images, the workload would be huge. More than that, if ground-truth were to be generated, a critical issue is not only the workload, but the need to repeat the work whenever there are significant changes in the characteristics of the images.

B. Classifier training and evaluation

Although manually classifying each target also represents a heavy workload, clearly it can be performed much faster than delineating the contours of the targets in the image. In order to evaluate different segmentation methods, for each experiment we fixed the feature extraction algorithms and the classifier model, and then a stratified 10-fold cross-validation (CV) of classification was performed for each of the segmented datasets. Feature extraction were performed using OpenCV [20]. Classifier training and cross-validation accuracy estimation were performed using WEKA [21], considering the 16 classes, each with 100 samples.

C. Results and discussions

In the first experiment, we used a set of 55 fixed features, extracted from the segmented images, and the SVM classifier (with $C = 13$, *RBF* kernel and $\gamma = 1$). As can be seen in

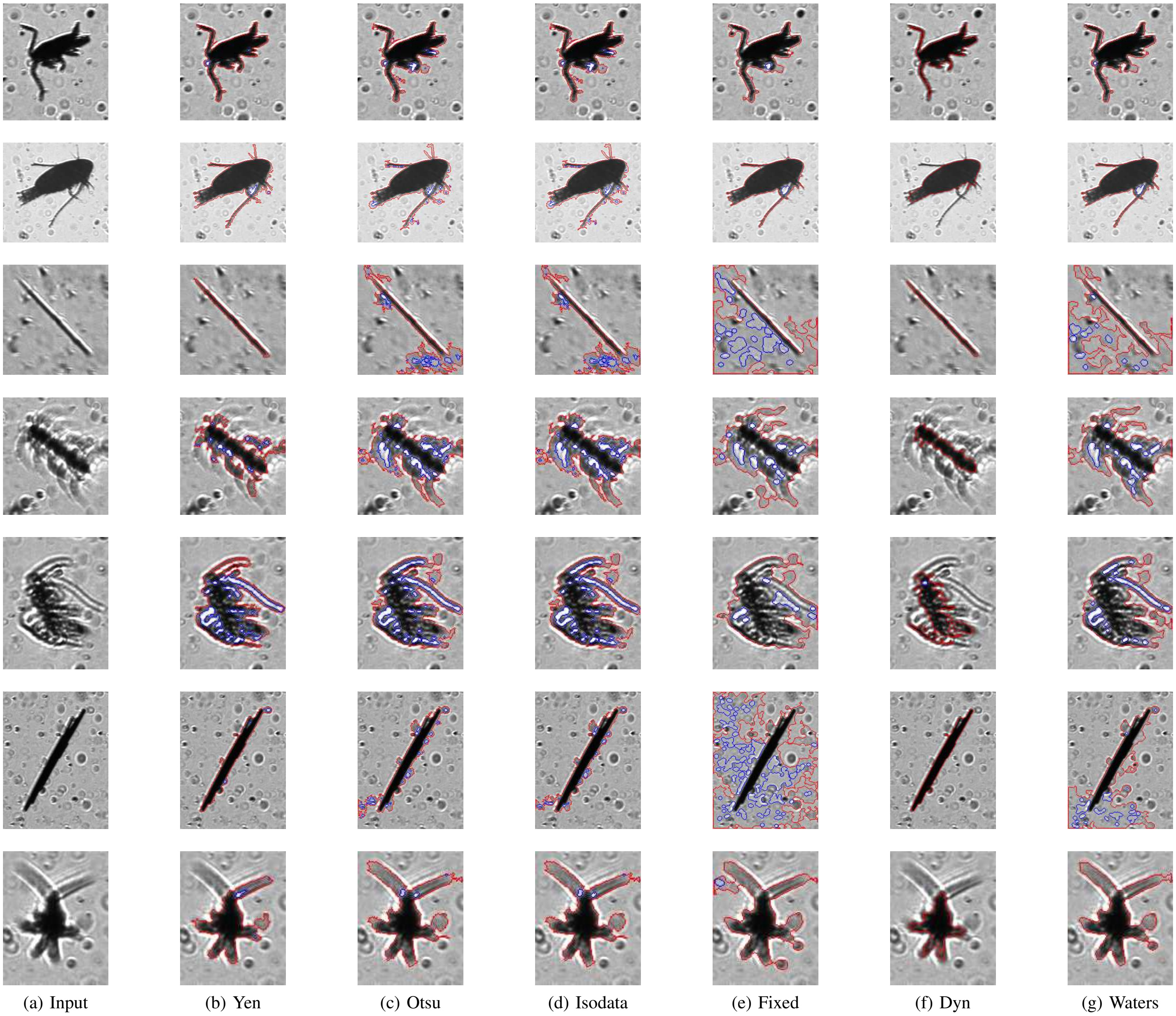

(a) Input (b) Yen (c) Otsu (d) Isodata (e) Fixed (f) Dyn (g) Waters

Fig. 3. Samples of 7 classes of plankton and respective segmentations using the six methods. From top to bottom, *Copepod Calanoid (Acartia)*, *Copepod* jumping, Fine fibers, *Chaetoceros* out of focus, *Copepod* dead, Thick Fibers, and *Nauplius* out of focus.

Table I, top accuracies (even considering variations among folders) were achieved when using **Waters** or **Yen** as the segmentation method. This is consistent with the visual evaluation of segmentation results (see Figs. 3 and 4).

In order to reinforce this finding of the best segmentation methods, we did a second experiment that consisted on varying the fixed part B of the chain by replacing the classifier model. The results using Random Forest (RF, with Seed = 4, numTrees = 40) and K-nearest neighbor (KNN, with k=10) models are shown also in Table I. Classification results have the same behavior of the one obtained with SVM, that is, best mean accuracies are obtained again using **Waters** and **Yen** as the segmentation methods, for both classifier models. This is a reassuring result in the sense that all three models indicate that **Waters** and **Yen** are the two best segmentation methods.

Table II details the classification accuracy per class, with respect to the SVM model. It shows the true positive rates with respect to each of the classes. Note that when comparing **Waters** with **Yen**, there is a large difference in the classification rate of class *Chaet. out of focus*. Actually, we decided to consider **Yen** as an alternative segmentation method after observing in a preliminary study that **Waters** resulted in poor classification rate for that class. After manually trying several segmentation methods, **Yen** was chosen by visual inspection as one of the algorithms that most improved segmentation of samples in that class. In fact, the classification result reinforces that perception.

Without information on classification accuracy per class, it

TABLE I
10-FOLD CV ACCURACY VARYING CLASSIFIER MODELS

Methods	**Cross-validation accuracy** (% out of 1600)		
	SVM	RF	KNN
Fixed	86.81±2.90 (1389)	84.13±2.47 (1346)	79.75±2.55 (1276)
Dyn	80.13±1.83 (1282)	75.06±1.10 (1201)	71.81±3.61 (1149)
Waters	**89.81**±2.36 (1437)	**86.56**±2.81 (1385)	**84.00**±2.17 (1344)
Yen	**90.31**±2.79 (1445)	**86.38**±2.69 (1382)	**84.63**±3.13 (1354)
Otsu	85.63±2.34 (1370)	79.56±3.06 (1273)	77.63±3.27 (1242)
Isodata	86.19±1.71 (1379)	81.13±3.69 (1298)	78.69±3.81 (1259)

TABLE II
DETAILED TP RATE PER CLASS USING SVM

Classes	**Percentage of Instances Correctly Classified per Class** (% relative to 100 instances per class)					
	Fixed	**Dyn**	**Wat.**	**Yen**	**Otsu**	**Isod.**
Chaetoceros (Chaet.)	83.00	73.00	**87.00**	80.00	77.00	76.00
Chaet. out of focus	67.00	66.00	67.00	81.00	**83.00**	80.00
Copepod calanoida	87.00	82.00	**90.00**	85.00	79.00	84.00
Copepod cyclopoida	95.00	94.00	95.00	**96.00**	84.00	85.00
Copepod out of focus	91.00	86.00	**95.00**	94.00	86.00	86.00
Copepod jumping	92.00	85.00	91.00	**94.00**	89.00	88.00
Copepod dead	84.00	73.00	83.00	**87.00**	79.00	80.00
Copepod (no antenna)	92.00	87.00	**93.00**	92.00	87.00	87.00
Coscinodiscus T.	97.00	92.00	**99.00**	96.00	97.00	97.00
Fine Fibers	88.00	91.00	**92.00**	**92.00**	92.00	93.00
Thick fibers	88.00	70.00	83.00	**85.00**	76.00	78.00
Nauplius out of focus	**92.00**	88.00	90.00	89.00	87.00	86.00
Neoceratium (Neoc.)	88.00	83.00	**95.00**	92.00	90.00	91.00
Neoc. out of focus	80.00	73.00	87.00	**88.00**	78.00	82.00
Odontella sinensis	83.00	66.00	91.00	**95.00**	90.00	88.00
Pyrocystis	90.00	73.00	**99.00**	**99.00**	96.00	98.00
Average (Avg.)	**86.81**	**80.13**	**89.81**	**90.31**	**85.63**	**86.19**

would probably be much harder to reach the same diagnosis. Classification results, separated by class, can also provide another valuable information. Let us suppose that, rather than classifying each target plankton, we are interested in detecting plankton of a specific class. Let us also suppose that such detection needs to be performed on the fly, with no time for training a specific classifier to recognize only plankton of that class. In such a case, we can decide to use the segmentation method that most favors the identification of samples of that class, without concerning how well it performs with regard to other classes. It would only require changing the segmentation component in the recognition system. For instance, suppose we were interested in detecting occurrences of *Nauplius out of focus*. From the above results, using the **Fixed** segmentation method could be an adequate choice.

A third experiment was performed changing again the fixed part B. This time, the classifier model was fixed to SVM with the same configuration used before, and distinct feature subsets were used. To define these subsets, first a feature selection method available in WEKA were applied on 60% of the segmented images for each segmentation method, generating a total of 6 feature subsets. Then, three subsets were defined from these 6 subsets as follows. The first subset, F_1, was built taking those features that were in at least 5 of these 6 subsets. The second and third subsets, F_2 and F_3, were built in a similar way, considering those features that were in at least 3 and 2 groups, respectively. The point that is noteworthy is the fact that once again **Waters** and **Yen** were the best performing segmentation methods as can be seen in Table III.

TABLE III
10-FOLD CV ACCURACY USING SVM WITH SELECTED SUBSETS OF FEATURES

Methods	**Feature subset** (number of features)			
	F_1 (14)	F_2 (29)	F_3 (34)	**All** (55)
Fixed	81.94% ±2.58	85.32% ±2.71	85.69% ±2.72	86.81% ±2.90
Dyn	73.75% ±1.51	78.31% ±1.51	79.06% ±1.19	80.13% ±1.83
Waters	**85.88%** ±2.31	**88.94%** ±1.94	**88.81%** ±1.95	**89.81%** ±2.36
Yen	**85.56%** ±3.13	**89.25%** ±2.86	**89.63%** ±1.84	**90.31%** ±2.79
Otsu	78.81% ±4.84	84.44% ±2.67	84.44% ±3.09	85.63% ±2.34
Isodata	78.88% ±4.30	83.81% ±2.90	85.06% ±2.76	86.19% ±1.71

In the fourth experiment we assess the relevance of the pre-processing steps in the segmentation methods. Results, using SVM classifier with the same configuration used before, are shown in Table IV. Four of the tested segmentation methods, except **Dyn** and **Yen**, resulted in a slight better classification accuracy when segmentation were applied without the pre-processing step (histogram equalization or smoothing). However, this experiment alone is not sufficient to conclude that pre-processing is or is not relevant. It only indicates that pre-processing may be unnecessary for four of the methods while it may be important for two. For a solid conclusion, additional assessment is necessary and, if possible, using a much larger amount of data.

TABLE IV
10-FOLD CV ACCURACY USING SVM ON IMAGES SEGMENTED WITH/WITHOUT PRE-PROCESSING

Methods	**Percentage of Instances Correctly Classified**	
	With Pre-Processing	**Without Pre-Processing**
Fixed	86.81% ±2.90	**87.00**% ±2.60
Dyn	**80.13**% ±1.83	79.56% ±2.11
Waters	89.81% ±2.36	**90.25**% ±1.75
Yen	**90.31**% ±2.79	87.25% ±2.43
Otsu	85.63% ±2.34	**87.88**% ±1.99
Isodata	86.19% ±1.71	**88.13**% ±2.32

In the fifth experiment we explored the idea of adjusting parameter values of segmentation algorithms based on classification accuracy, in a similar way as in [9], [10]. In particular, we tried different thresholding values for the fixed threshold (**Fixed**) method, and the results are shown in Table V. As can be seen, $T = 180$ yields better classification accuracy than the originally chosen $T = 170$.

TABLE V
10-FOLD CV ACCURACY USING SVM ON IMAGES SEGMENTED WITH **Fixed** METHOD FOR DIFFERENT THRESHOLD VALUES (T)

Threshold (T)	**170**	**175**	**180**	**185**	**190**
Correctly Classified Inst. out of 1600	**86.81**% ±2.90 1389	86.93% ±2.38 1390	**88.19**% ±3.71 1411	87.81% ±2.90 1405	87.44% ±2.58 1399

Note, however, that adjusting the threshold value by hand is not an easy task due to great variations among images. In Figure 5 we show some examples where $T = 180$ works better than $T = 170$ while in Fig. 6 we show some examples where the inverse happens.

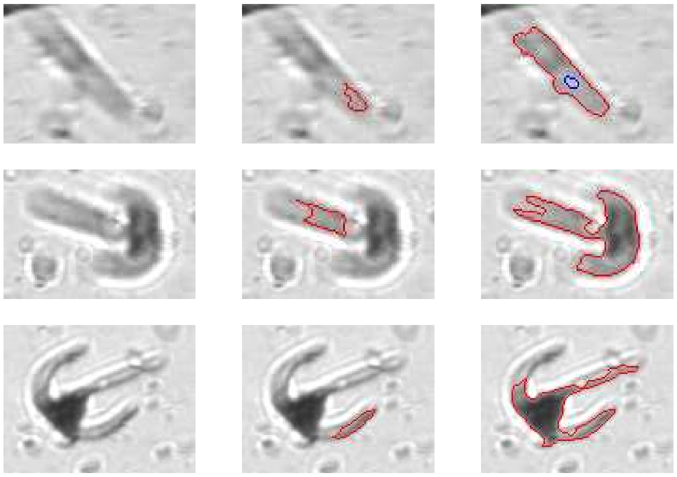

(a) Input (b) T = 170 (c) T = 180

Fig. 5. Samples of 3 classes of plankton, segmented using the **Fixed** method, with different thresholding values: from top to bottom, Thick Fibers, *Neoceratium* out of focus, and *Neoceratium*. Better results are obtained with $T = 180$.

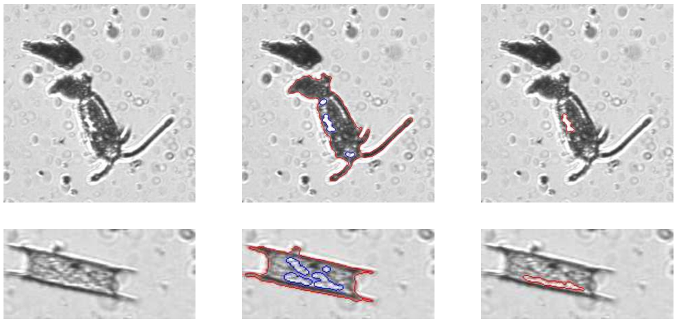

(a) Input (b) T = 170 (c) T = 180

Fig. 6. Samples of 2 classes of plankton, segmented using the **Fixed** method, with different thresholding values: *Copepod* dead (top row), and *Odontella sinensis* (bottom row). Better results are obtained with $T = 170$.

In the last experiment we try to assess how contour precision affects results. For that, based on visual inspection, we choose 314 well segmented images (157 *calanoidas* and 157 *cyclopoidas*). Then, we systematically dilated and contracted the boundary of the objects, applying morphological dilation and erosion using structuring elements (SE) of increasing size. The classification accuracies with respect to SEs of different size are shown in Table VI. As can be seen, the classification accuracies relative to eroded boundaries decreases faster than those for dilated ones as we increase the SE size. This can be explained by the fact that erosions remove important thin features such as antennas and legs, while dilation better preserves the general shape for the species considered in this experiment. This experiment shows that contour precision, as far as the shape is preserved, does not affect accuracy significantly.

TABLE VI
10-FOLD CV ACCURACY USING SVM ON ERODED/DILATED BOUNDARIES WITH STRUCTURING ELEMENTS (SE) OF DIFFERENT SIZE

Erosion SE radius			**Original**	**Dilation SE radius**		
10	**5**	**3**		**3**	**5**	**10**
87.90% ±4.73	91.06% ±5.37	90,44% ±7.02	92.99% ±6.35	92.69% ±4.73	92.37% ±4.97	90.48% ±5.64

IV. Concluding remarks

In our case study, by analyzing classification accuracy, we find out that **Waters** and **Yen** are the two segmentation methods that consistently presented best results. In order to reinforce this finding, we have varied the fixed part of the chain (feature extraction+classification) with respect to the number of features and also with respect to the classifier model. A reassuring fact is that the finding is, in general, in agreement with visual perception we have from the segmented images.

Altogether, we believe that the case study on plankton classification presented in this work supports the applicability of a system level evaluation method for assessing image segmentation methods. The evaluation method not only helps choosing a good segmentation method, but points where segmentation may be failing, favoring a more holistic evaluation of the process.

One potential drawback of the approach is the fact that it is an indirect evaluation of solutions. In processings where obtaining accurate segmentation is not a pursuit, a direct measure of the performance of algorithms are not really needed. Besides that, by repeating the evaluation process, changing the fixed part, a more robust evaluation can be achieved. This could be easily implemented using black-box solutions for the fixed part. A second potential drawback is computational cost. However, modern computing facilities such as cloud and distributed computing can be used to mitigate the computationally intensive part.

On the other hand, advantages and new evaluation possibilities enabled by this type of approach surpass these drawbacks. Some of the advantages and possibilities, are highlighted here. With respect to segmentation, if there is interest in extracting precise measures of a target, it may be helpful to first identify its class and then apply a segmentation method customized to that class. In this sense, then, we first need to find a segmentation algorithm that favors correct classification of the targets. Moreover, as already pointed, it may be possible to identify where segmentation should be improved by analyzing classification errors.

The evaluation approach studied in this work is particularly useful in cases where ground-truth for a higher level task can be obtained in a much easier way than for the low level task. Since human beings, the agents that usually provide the ground-truth, deal much better with semantically higher level data, it is likely that much more training and validation data will be available for the higher level tasks than the amount that would be possible for the low level task. Amount of training and validation data is important for statistically sound evaluation.

Moreover, ground-truth is a means to embed context into evaluation. This way of connecting evaluation to context is, as already mentioned, more easy for semantically higher level tasks. A direct consequence of this fact is that this system level evaluation method is better suited than other methods to big data analysis. In big data analysis, systems not only need to process data efficiently but they also have to be able to quickly adapt to variations in observed data. For instance, as image acquisition devices are frequently improving, a particular image segmentation algorithm that had best performance at a given moment may no longer be the best one after a device is improved. In such situations, being able to quickly replace the segmentation component is an important issue in many applications. In order to do such replacement, a quick evaluation of segmentation algorithms is also needed. The time required to prepare ground truth for the segmentation task is not feasible but the time required to manually classify a set of images may be acceptable.

Although our case study is concerned with evaluation of image segmentation methods, the same principle applies to any low-level/high-level pair of processing tasks. We believe that the existence of computational resources and technologies such as cloud and distributed computing, and the emergence of several software tools that can be used as black-boxes for solving specific tasks will enable practical system level evaluation of algorithms and methods in general. The case study presented in this work is a contribution toward advances in this type of evaluation.

Acknowledgment

M. A. Fernandez thanks CAPES for the financial support, R. M. Lopes thanks CNPq (grant 311936/2013-0) and N. S. T. Hirata thanks CNPq (grants 484572/2013-0 and 307357/2012-0). The authors thank Dr. Paulo Miranda for helpful suggestions and Leandro T. De La Cruz for the help with implementations.

References

[1] H. Zhang, J. E. Fritts, and S. A. Goldman, “Image segmentation evaluation: A survey of unsupervised methods,” *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260 – 280, 2008.

[2] 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, 2006.

[3] A. Martin, H. Laanaya, and A. Arnold-Bos, “Evaluation for uncertain image classification and segmentation,” *Pattern Recognition*, vol. 39, no. 11, pp. 1987–1995, 2006.

[4] J. S. Cardoso and L. Corte-Real, “Toward a generic evaluation of image segmentation,” *IEEE Transactions on Image Processing*, vol. 14, pp. 1773–1782, nov 2005.

[5] 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. 929–944, 2007.

[6] Y. J. Zhang, “A survey on evaluation methods for image segmentation,” *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.

[7] F. Ge, S. Wang, and T. Liu, “Image-segmentation evaluation from the perspective of salient object extraction,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, June 2006, pp. 1146–1153.

[8] A. Falcão, J. Udupa, and F. Miyazawa, “An ultra-fast user-steered image segmentation paradigm: live wire on the fly,” *IEEE Transactions on Medical Imaging*, vol. 19, no. 1, pp. 55–62, Jan 2000.

[9] Y. Ding, G. Vachtsevanos, A. Yezzi, Y. Zhang, and Y. Wardi, “A recursive segmentation and classification scheme for improving segmentation accuracy and detection rate in real-time machine vision applications,” in *14th International Conference on Digital Signal Processing*, vol. 2, 2002, pp. 1009–1013 vol.2.

[10] T. Kavzoglu and M. Yildiz, “Parameter-Based Performance Analysis of Object-Based Image Analysis Using Aerial and Quikbird-2 Images,” in *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. II-7, Sep. 2014, pp. 31–37.

[11] M. C. Shin, D. B. Goldgof, and K. W. Bowyer, “Comparison of edge detector performance through use in an object recognition task,” *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 160 – 178, 2001.

[12] G. Heitz, S. Gould, A. Saxena, and D. Koller, “Cascaded classification models: Combining models for holistic scene understanding,” in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. Curran Associates, Inc., 2009, pp. 641–648.

[13] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh, “Analyzing Semantic Segmentation Using Hybrid Human-Machine CRFs,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3143–3150.

[14] R. Kasturi, L. O’Gorman, and V. Govindaraju, “Document image analysis: A primer,” *Sadhana*, vol. 27, no. 1, pp. 3–22, 2002.

[15] I. Pratikakis, B. Gatos, and K. Ntirogiannis, “ICDAR Document Image Binarization Contest (DIBCO 2013),” in *12th International Conference on Document Analysis and Recognition (ICDAR)*, Aug 2013, pp. 1471–1476.

[16] D. J. Matuszewski, “Computer vision for continuous plankton monitoring,” Master’s thesis, University of São Paulo, April 2014.

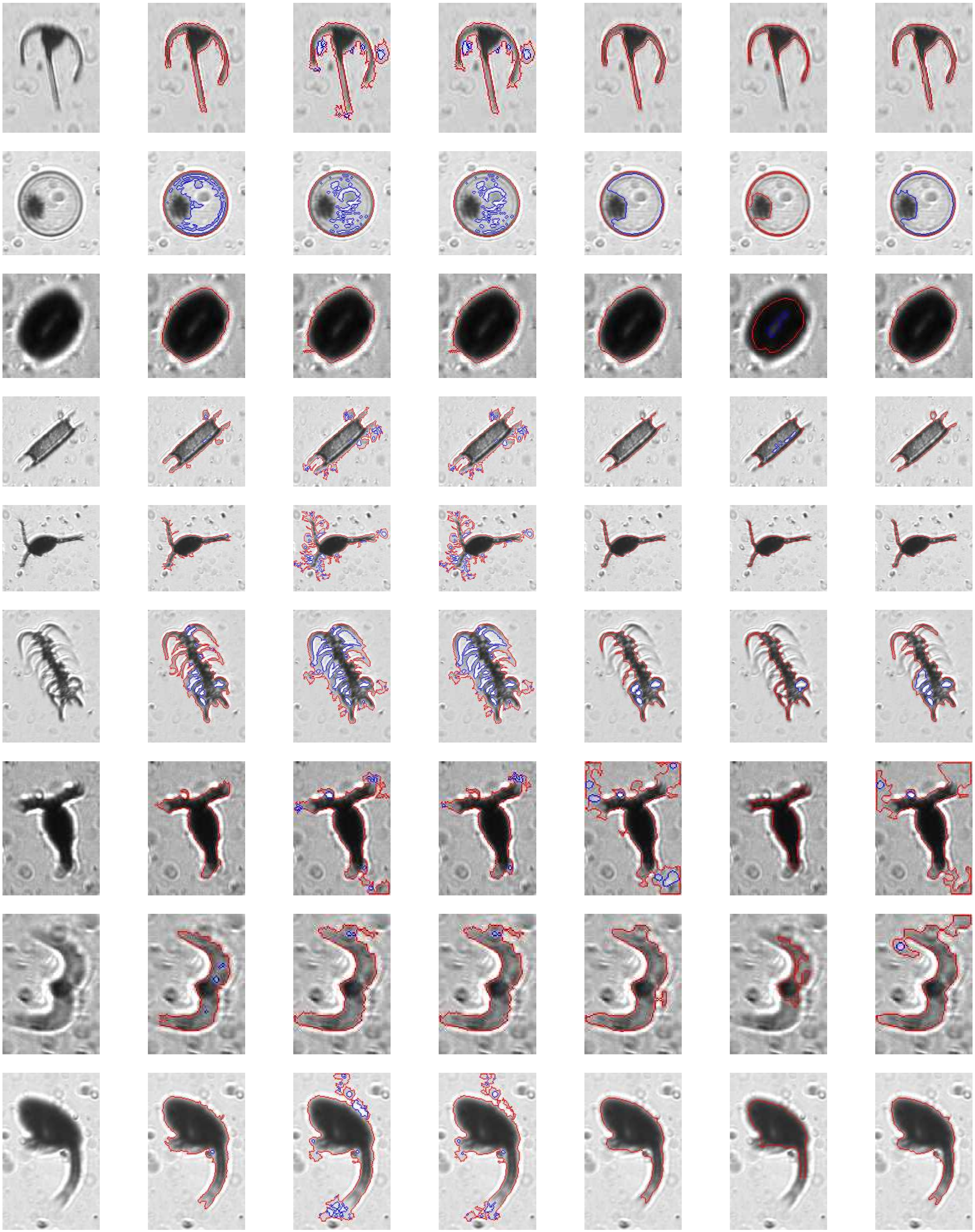
[17] J.-C. Yen, F.-J. Chang, and S. Chang, “A new criterion for automatic multilevel thresholding,” *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 370–378, Mar 1995.

[18] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan 1979.

[19] T. Ridler and S. Calvard, “Picture thresholding using an iterative selection method,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, no. 8, pp. 630–632, Aug 1978.

[20] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.



(a) Input (b) Yen (c) Otsu (d) Isodata (e) Fixed (f) Dyn (g) Waters

Fig. 4. Samples of 9 classes of plankton and respective segmentations using the six methods. From top to bottom, *Neoceratium*, Pyrocystis, Coscinodiscus, *Odontella sinesis*, *Copepod Cyclopoida*, *Chaetoceros*, *Copepod* (Oithona) out of focus, *Neoceratium* out of focus, and *Copepoda* without antenna.