Discriminative Segmentation of Microscopic Cellular Images

Li Cheng[1], Ning Ye[1], Weimiao Yu[2], and Andre Cheah[3]

[1] BioInformatics Institute, A*STAR, Singapore
[2] Institute of Molecular and Cell Biology, A*STAR, Singapore
[3] NUHS, Singapore

Abstract. Microscopic cellular images segmentation has become an important routine procedure in modern biological research, due to the rapid advancement of fluorescence probes and robotic microscopes in recent years. In this paper we advocate a discriminative learning approach for cellular image segmentation. In particular, three new features are proposed to capture the appearance, shape and context information, respectively. Experiments are conducted on three different cellular image datasets. Despite the significant disparity among these datasets, the proposed approach is demonstrated to perform reasonably well. As expected, for a particular dataset, some features turn out to be more suitable than others. Interestingly, we observe that a further gain can often be obtained on top of using the "good" features, by also retaining those features that perform poorly. This might be due to the complementary nature of these features, as well as the capacity of our approach to better integrate and exploit different sources of information.

1 Introduction

Cellular images segmentation is an indispensable step for modern biological research, and this is greatly facilitated by the recent development of fluorescence dyes and robotic microscopes. A number of unsupervised segmentation methods [1], such as thresholding, region-growing, watershed, level-set, and edge-based methods, have been developed to address this problem in scenarios where the foreground objects and the background regions have distinct color or textural properties. Many of these methods are dedicated to specific problems where domain knowledge is heavily exploited by tuning algorithmic parameters manually. This case-by-case approach could be very tedious. On the other hand, there are many images (such as in Figure 1) that turns to be rather difficult for unsupervised approaches [1], partly due to the variations of specimen types, staining techniques, and imaging hardware. This leads to a recent development in learning-based algorithms, including support vector machines (SVMs) [10], as well as conditional random fields and variants [4]. While they demonstrate that reasonable segmentation results can be produced for some difficult cases, these methods are often still dedicated to certain type of microscopic images for specific problems.

In this paper we propose a flexible learning framework for cellular image segmentation, and we intend to show that it is possible to develop a more generic segmentation framework that works effectively over a broader spectrum of microscopic images such as those displayed in Figure 1. This is achieved by carefully integrating information

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part I, LNCS 6891, pp. 637–644, 2011.

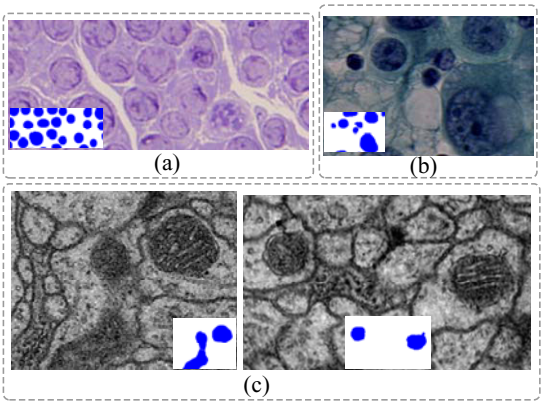

Fig. 1. Illustration of some difficult examples encountered in microscopic image segmentation. These are crop-out examples from three cellular image datasets: (a) hand, (b) serous, (c) ssTEM. Ground truth is shown in the inlet for each of the examples.

from both local and global aspects, as illustrated in Figure 2. In particular, three novel features are proposed: (1) An appearance feature that integrates both color and texture information; (2) A spoke feature that effectively encodes the shape of cellular foreground objects; (3) Meanwhile the detection score is also used to exploit the strength of object detection developed over the years in computer vision. Besides, a superpixel-based coding scheme is devised to incorporate higher-order scene context.

2 Our Approach

The flow chart of our approach is depicted in Figure 2. An image pixel is characterized by a set of features describing various local aspects in its neighborhood, such as shape, appearance, and context information. These pixel-based features are further pooled to form one vector for a superpixel or oversegment [7]. Finally, a global discriminative classifier is utilized to incorporate these superpixel-based shape, appearance, and context features to produce a segmentation prediction for the input image.

Appearance feature: Unary $\&$ **Binary Extensions of Color BoW model.** For a pixel in color images, its RGB and YUV color values are combined to form a 6D vector. For grayscale images, 1D intensity feature is used directly. As illustrated in the middle panel of Figure 3, a visual Bag-of-Words (BoW) model with K codewords is built, and these color vectors are thus mapped to the quantized space spanned by the codewords [11]. A novel appearance feature is proposed here by integrating BoW model with local neighboring information by means of unary/binary extensions: A unary extension partitions the local neighbors into disjoint parts. One scheme is to partition into concentric layers, as displayed in Figure 3(a). By pooling the codeword assignments of these features and normalizing to sum to 1, one partition is characterized by a histogram of length K codewords. A length $K \times S$ vector is thus produced by concatenating over S partitions ($S = 3$ in Figure 3(a)). Note that other partition schemes might also be possible.

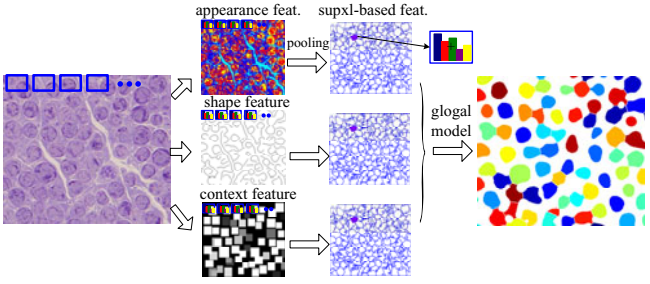

Fig. 2. An illustrative flowchart of the proposed approach. From a pixel of an input image, a set of features is extracted to capture various aspects in its neighborhood, including appearance, shape, and context information. A novel superpixel (over-segmentation) feature is devised to provide a more compact signature and to incorporate higher-order scene context. These superpixel-based features are fed to a global discriminative learning model to deliver a segmentation.

Meanwhile, a binary extension considers pairs of neighboring pixels, and similar to the concept of co-occurrence matrix, accumulates the counts into a 3D array indexed by (codewords, codewords, distance). Naively this leads to a vector of length $K \times K \times S'$, by accumulating the quantized distance of every feature pair with S' possible outcomes. Here we adopt hamming distance. In practice it is further summarized into a more compact vector representation: For a particular quantized distance, (a) a K-dim vector is extracted from the diagonal elements, and (b) a K-dim vector is obtained by summing over all the off-diagonal elements row-wise. For both cases the output vectors are normalized to sum to 1. As each case ends up giving a $K \times S'$ vector, a concatenation of both finally leads to a vector representation of length $2K \times S'$. Our final appearance feature is thus produced by concatenating both unary and binary extensions. In this paper, we fix $K = 100$, $S = 3$, and $S' = 3$.

Shape feature: Multi-scale Spoke Feature. as illustrated in Figure 4(a), for any location in an image, its spokes are equally sampled in angular space and each reach out until an edge is met. Determined by its local convexity (i.e. the orientation of its signed curvature), the spoke will contribute to one of the three bins: +, 0, and -, that encode the local shape as being convex, undecided, or concave, respectively. Therefore, the spoke feature essentially encodes the local shape information from the direct object boundaries surrounding this location, while being invariant to rigid transformation. As cellular objects often possess convex shapes, this feature ideally provide sufficient discrimination power to differentiate a location inside a cellular object from being outside. On the other hand, a single edge map usually does not faithfully retain object boundaries of the image. To address this issue, we use Canny edge detector together with its Gaussian smooth kernels at multiple scales. This gives rise to the multi-scale feature in Figure 4(c). In practice, for a pixel, 9 spokes are used and the number of scales is set to 5. The elements in the histogram vector of each scale are also normalized to sum to 1.

Context feature: Detection Score BoW model. Object detection is usually regarded as a separate problem from image segmentation, and thus dealt with by substantially

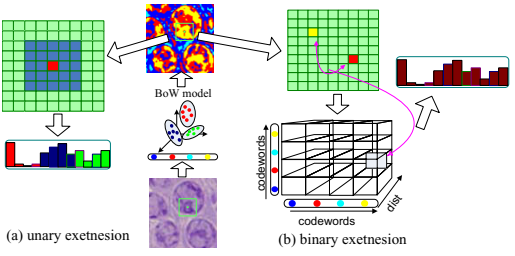

Fig. 3. Unary and binary extensions. (a) A unary feature extension partitions the window into concentric layers. By pooling the pixels' codeword assignments in the BoW model (length K) within each partition, a $K \times S$ length vector is produced as a concatenation of S histograms. Each histogram of length K comes from one partition. (b) A binary feature extension. Each pair of pixels in the window is used to accumulate the counts in a 3-dimensional array. Naively this leads to a vector of length $K \times K \times S'$, for a quantized distance of S' possible outcomes. In practice it is further summarized into a more compact vector of length $2K \times S'$. Then unary and binary extensions are concatenated together to form an appearance feature.

different techniques. Nevertheless, detection outputs possesses important information about the locations and sizes of the foreground objects that can be utilized to help segmentation. In addition, as generated through top-down schemes, the detection scores carries context information over to pixel level. Here we adopt a dedicated mixture model-based object detector [6]. The bounding box detections are overlayed onto a two dimensional space with each assigning its detection score. This is treated as a separate channel of the input image. Then the context feature is produced through a BoW model, similar to that of the color BoW model as previously shown in the middle panel of Figure 3. Here the number of bins in the BoW model is set to 4.

Superpixel-based Feature. An image is usually represented as a two-dimensional lattice where each node corresponds to a pixel. However a pixel by itself contains limited information. Alternatively, an image can be expressed as a general planar graph, and each node is now a superpixel or oversegmentation [7]containing a set of nearby pixels, usually obtained using an unsupervised segmentation. As depicted in Figure 2, for pixels within a superpixel, their features are pooled to form a higher-order feature vector that describes the entire superpixel. Similarly, the output vectors are each normalized to sum to 1. While providing a more compact feature representation, this superpixel-based feature is also able to capture higher-order scene context.

Global Model and Postprocessing. The discriminative learning method we have utilized is a Structured Support Vector Machine (SSVM) [12], where the optimal assignment problem is solved by the graph-cuts algorithm [2] that incorporates both node and edge energies to ensure local and global compatibilities. The proposed features are concatenated as a long feature vector used for the node energy; While our edge energy adopts a simple Ising model [2]. Our postprocessing step utilizes distance transform and generalized Voronoi diagram [13], to remove tiny segments and separate those lightly touched objects.

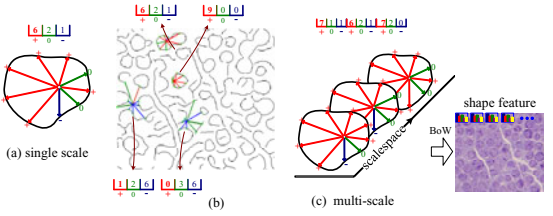

Fig. 4. (a) An illustration of the proposed *spoke feature*, (b) its usages in an edge map of the image in Figure 2 on four image locations, and (c) a *multi-scale spoke feature*. Spokes are equally sampled in angular space, and are further mapped into a histogram vector of three bins marked by +, 0, and -, which denote locally convex, undecided, and concave, respectively. To alleviate the issues introduced by edge detection, edge maps are extracted at multiple scales, and the associated histograms are concatenated to form a multi-scale spoke feature. A common procedure in edge (e.g. Canny) detectors is to convolve raw image with Gaussian kernel of certain width, which is regarded as selecting a scale-space [9]. Multi-scale here refers to applying kernel of multiple widths, leading to multiple edge maps. For a fixed image location, a multi-scale spoke feature is obtained by concatenating the spoke feature vectors obtained over scales.

3 Experiments

Image Datasets. Three image datasets are used during the experiments, where for each dataset, half of the images are used for training and the rest images are retained for testing purpose. The *hand dataset* contains images of hand nerve endings harvested from fresh frozen adult human cadavers. They are preserved in gluteradehyde and refridgerated, then embedded in liquid wax to form a block to facilitate microtome sectioning of the specimen perpendicular to the longitudinal axis of the axons. They are then stained with methylene blue and photographed using a light microscope to facilitate the process of histomorphometry. They are further partitioned into 24 smaller images of similar sizes, with ground-truth provided for nerve endings. To study the drosophila first instar larva ventral nerve cord (VNC), the *ssTEM dataset* is generated, which has 30 images[3][1] taken from a serial section Transmission Electron Microscopy (ssTEM), with image resolution 4x4 nm/pixel. Ground-truth annotations are provided for mitochondria. Finally, the *serous dataset* [8] contains 10 microscopic images[2] from serous cytology. Ground-truth annotations are provided for cell nuclei.

Performance Evaluation. Performance of a cellular segmentation method is often quantitatively assessed by two types of metrics: those of pixel-based and those of object-based. We follow the PASCAL VOC evaluation criteria [5]: For pixel-level, we directly adopt the criteria of its image segmentation task [5]. The metric for object-level evaluation is an adaptation of the criteria used in its object detection task [5]. For *Pixel-based*

[1] The dataset is downloaded from
`http://www.ini.uzh.ch/~acardona/data/tifs.tar.bz2`
[2] The dataset is downloaded from
`http://users.info.unicaen.fr/~lezoray/databases/SerousDatabase.zip`

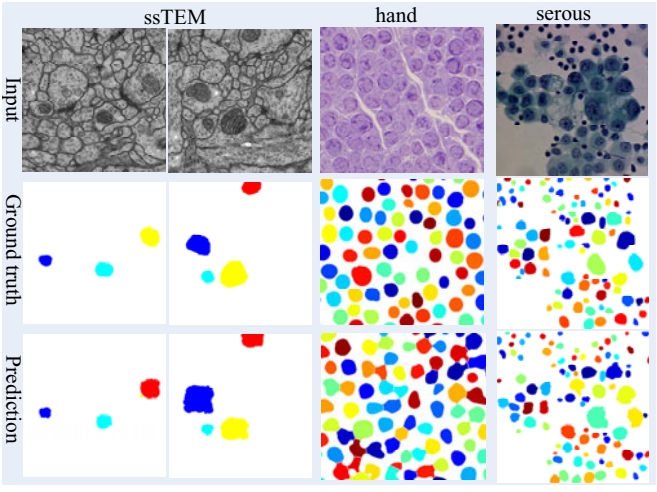

Fig. 5. Sample experimental results on the three datasets

Evaluation , a common accuracy measures the average percentage of pixels being correctly classified for both foreground and background classes. This metric however can be misleading when class distribution is unbalanced, e.g. when the dataset contains fewer foreground object pixels and a larger percentage of background pixels (the ssTEM dataset). To rectify this issue, the PASCAL image segmentation task advocates to compute the accuracy as (Eq.(4) of [5]) $\frac{\mathrm{TP}}{\mathrm{TP+FN+FP}}$[3]. For *object-level Evaluation* , object-based image segmentation can be considered as a special object detection task, where in addition to location and scale, it also demands the detailed shape of a foreground object. Following the object detection task of PASCAL [5], we use an intersection/union ratio to determine a correct object-level match: Given a pair of objects consisting of a prediction area O_{p} and a ground-truth O_{gt}, there exists a match if the overlap ratio, $\frac{\mathrm{area}(O_{\mathrm{p}} \cap O_{\mathrm{gt}})}{\mathrm{area}(O_{\mathrm{p}} \cup O_{\mathrm{gt}})}$ exceeds a threshold t, where $\cap$ denotes the intersection, and $\cup$ the union. t is set to 0.5 as in [5]. Similarly we define object-level accuracy as $\frac{\mathrm{TP}}{\mathrm{TP+FN+FP}}$.

Experiments. Throughout the experiments, the unsupervised method of [7] is used to partition an image into superpixels, and the C value of the linear SSVM [12] is fixed to 100.

As expected, a multi-scale shape feature leads to improved performance, when comparing to its single-scale counterpart. This is demonstrated in hand dataset, where the pixel- (and object-) level accuracy is around 68% vs. 51% (and 76% vs. 73%), when a multi-scale shape feature is compared to a single-scale one. We also observe that *no single* feature excels in all datasets. For example, the appearance feature dominates the performance for the serous dataset, while it works less well in the hand dataset, and leads to the worst results for the ssTEM dataset. Meanwhile, it is mostly preferable to

[3] TP, TN, FP, and FN refer to True Positive, True Negative, False Positive and False Negative, respectively.

Table 1. Comparisons of pixel- & object- level accuracies. Here 'three features' refers to the full version of our proposed method; 'appearance', 'shape', and 'context' are variants where only one type of feature is used; Meanwhile, 'dedicated unsupervised seg.' refers to the segmentation method in [13]. See text for more detail.

Dataset	Brief Description	Pixel Acc.	Obj. Acc.
ssTEM	appearance	45.06%	10.01%
	shape	51.14%	15.37%
	context	72.03%	56.25%
	three features	75.07%	64.71%
hand	appearance	78.18%	82.07%
	shape: only single-scale	51.16%	73.33%
	shape	68.02%	75.69%
	context	74.46%	83.06%
	three features	79.05%	85.10%
	dedicated unsupervised seg.	56.99%	50.85%
serous	appearance	83.91%	81.43%
	shape	71.61%	64.07%
	context	65.15%	47.60%
	three features	85.11%	83.98%
	dedicated unsupervised seg.	62.92%	38.28%

consider all the complementary features. Since our discriminative learning approach is able to perform an implicit feature selection, through learning it usually allocates higher weights to the "good" features. Interestingly, even when some features fail or perform less well when being used alone, a further gain can usually be obtained by retaining those features: Empirically this phenomenon is observed for all three datasets. Consider the ssTEM dataset for example, by employing the best (context) feature, a pixel- (object-) level performance of about 72% (56%) is attained; which is much better than considering the other two features where the corresponding results are merely no more than 51% (15%). However, when considering all three features jointly, the performance is further improved to 75% (65%). We think this might be attributed to the complementary nature of these features.

As a comparison method, a state-of-the-art *dedicated unsupervised Segmenter* [13] has been implemented. This method contains a few steps including noise removal, Gaussian smoothing, and thresholding based on the color histogram. After converted to binary images, distance transform is applied, and object centers are detected by seed finding. Then the generalized Voronoi diagram is applied to separate the touching objects. Unfortunately we can not produce a reasonable result for ssTEM dataset using the method of [13], despite significant effort in tuning the internal parameters. We speculate this is because the mitochondria objects are not sufficiently distinct in any of the color channels. Notice for each of the datasets, the internal parameters are *manually* adjusted to attain best performance. Nevertheless, as demonstrated in Table 1, this method performs considerably inferior to those from our approach.

We also compare our results to those of [8], which describes a supervised method combining pixel classifier and watershed, and has its results on the serous dataset: This

method reports an accuracy of 93.67% when a K-means RGB is used, and 96.47% if a Bayes RGB is deployed instead. Note [8] uses the traditional pixel-level accuracy, as the average percentage of pixels being correctly classified for both foreground and background classes. Meanwhile our approach on serous dataset achieves a better result of 98.12% under this evaluation criteria (85.11% under our evaluation criteria). Note that for serous dataset the overall class distributions are 7% for the nuclei pixels and 93% for the background pixels. As explained previously, these scores therefore tend to be saturated and more biased toward the background class.

4 Outlook

For future work we plan to conduct extensive evaluation of our approach on different microscopic datasets, and extend to work with 3D.

References

1. Bengtsson, E., Wahlby, C., Lindblad, J.: Bobust cell image segmentation methods. Pattern Recognition and Image Analysis 14(2), 157–167 (2004)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE TPAMI 23(11), 1222–1239 (2001)
3. Cardona, A., Saalfeld, S., Preibisch, S., Schmid, B., Cheng, A., Pulokas, J., Tomancak, P., Hartenstein, V.: An integrated micro- and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. PLoS Biol. 8 (2010)
4. Chen, S., Gordon, G., Murphy, R.: Graphical models for structured classification, with an application to interpreting images of protein subcellular location patterns. J. Mach. Learn. Res. 9, 651–682 (2008)
5. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) challenge. Int. J. of Computer Vision 88(2), 303–338 (2010)
6. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Int. Conf. Computer Vision and Pattern Recognition (2008)
7. Levinshtein, A., Stere, A., Kutulakos, K., Fleet, D., Dickinson, S., Siddiqi, K.: Turbopixels: Fast superpixels using geometric flows. IEEE TPAMI 31, 2290–2297 (2009)
8. Lezoray, O., Cardot, H.: Cooperation of color pixel classification schemes and color watershed: a study for microscopical images. IEEE TIP 11(7), 783–789 (2002)
9. Lindeberg, T.: Edge detection and ridge detection with automatic scale selection. IJCV 30, 117–154 (1998)
10. Marcuzzo, M., Quelhas, P., Campilho, A., Mendonca, A.M., Campilho, A.: Automated arabidopsis plant root cell segmentation based on svm classification and region merging. Comput. Biol. Med. 39, 785–793 (2009)
11. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering object categories in image collection. In: ICCV (2005)
12. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. 6, 1453–1484 (2005)
13. Yu, W., Lee, H., Hariharan, S., Bu, W., Ahmed, S.: Evolving generalized voronoi diagrams of active contours for accurate cellular image segmentation. Cytometry 77, 379–386 (2010)