LOCALIZATION OF DROSOPHILA EMBRYOS USING CONNECTED COMPONENTS IN SCALE SPACE

Zachary Bessinger, Guangming Xing, Qi Li

Western Kentucky University
Department of Mathematics and Computer Science
1906 College Heights Blvd, Bowling Green, KY

ABSTRACT

Localization of Drosophila embryos in images is a fundamental step in an automatic computational system for the exploration of gene-gene interaction on Drosophila. In this paper, we introduce a localization framework based on the analysis of connected components in the Gaussian scale space of an embryonic image. We propose three criteria for the selection of the optimal scale. The experiment results show the promise of the proposed methods.

***Index Terms*—** Localization, connected component, scale selection.

1. INTRODUCTION

Drosophila embryonic images provides detailed spatial and temporal information of gene expression, which become an important tools for micro-biologist to study gene-gene interaction [1]. High throughput drosophila embryonic images arises the need of automatic system of analyzing gene-gene interaction patterns. As well as many other micro-biology imaging circumstance, a Drosophila embryonic image contain a Region of Interest (ROI), i.e., the targeting embryo in imaging. Such embryonic images usually contain significant amount of variations: i) imaging conditions, such as the contrast, scale, orientation, and neighboring embryos, ii) gene expression patterns, and iii) developmental stages.

Localization of the ROI in Drosophila embryonic images is a fundamental step in an automatic computational system for the exploration of gene-gene interaction on Drosophila. Due to serious image variations in embryonic images, it has been observed that the straight application of existing techniques on edge detection and contour extraction failed to obtain desirable results [9, 8, 5]. Several approaches have been proposed to extract the ROI from embryonic images [4, 9, 8]. Peng et al. [9] proposed an approach that computes the standard deviation of the local windows of pixels to characterize pixels as foreground and background pixels, and applies 8-neighbor-connectivity region-growing method to localize the contour of the embryo. Pan et al. [8] used a variant of Marquardt-Levenberg algorithm to compute an optimal affine transformation to register localized embryos into an ellipsoidal region. Frise et al. [3] proposed a heuristic algorithm to separate the embryo of interest from multiple touching embryos, with the assumption that the center of the embryo of interest is the image center. Mace et al. [7] proposed an eign-embryo method to extract the contour of embryos, where particle swarm optimizer was used to reduce the computational cost of searching optimal eigen parameters. Puniyani et al. [10] proposed an edge detection based method that involves a set of heuristic constraints, including object size, convexity, shape features (e.g., ratio of the major over minor axis of an object), and the percentage of overlapping region. Li et al. [5] proposed a framework that consists of active contour and eigen-shape methods.

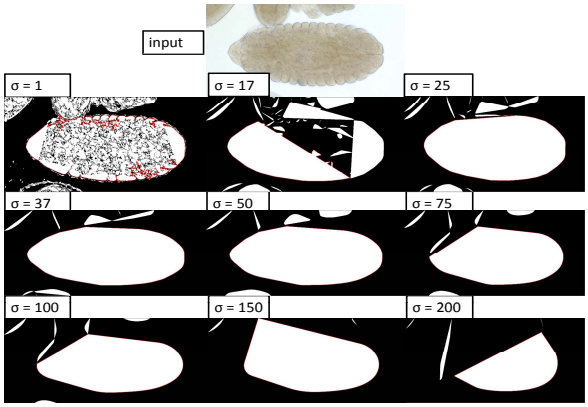


Fig. 1. Connected components of an embryonic image in a Gaussian scale space.

In this paper, we introduce a framework for the localization of the ROI of an embryonic image based on connected components in the Gaussian scale space. The scale space theory used in this paper is contributed by Lindeberg [6]. The rationale of the proposed framework is illustrated in Fig. 1 that shows connected components of an embryonic image in a subsampled scale space with scale σ ranging from 1 to 200.

We can observe that some scales in the middle (say, σ=25 or 37) achieve a good balance between the completeness of the ROI (the central embryo) and separability of the ROI from non-ROIs (i.e., neighboring partial embryos). However, Fig. 1 also shows the complexity of image structures of embryonic images in scale space. Note that there is no obvious superset/subset relationship among connected components in different scales.

We will propose three criteria to select a good scale for the embryo localization: i) minimization of the number of connected components, ii) maximization of the area of the largest connected component, and iii) minimization of the displacement of the contour of the largest connected components to a pre-defined shape model (e.g., ellipse). The proposed criteria are application oriented and thus different from the general principle of automatic scale selection proposed by Lindeberg [6], such as: "A powerful approach to perform local and adaptive scale selection is by detecting local extrema over scales of normalized differential entities." [6]. In experiment, we give a comparison of three proposed criteria. We also compare the proposed framework with previous works, and obtain a higher localization accuracy.

2. CRITERIA FOR SCALE SELECTION

Denote $I_\sigma = I * G_\sigma(x, y)$ the convolution of an image I and the 2d Gaussian filter with scale σ. Denote E_{I_σ} a binary image of I_σ. Assume that E_{I_σ} consists of a number of connected components, that is

$$E_{I_\sigma} = \bigcup_i C_{i,\sigma}, \qquad C_{i,\sigma} \bigcap C_{j,\sigma} = \emptyset, i \neq j,$$

where $\{C_{i,\sigma}\}$ is a set of connected components (in terms of 8-connectivity) in scale σ, and $\emptyset$ indicates an empty set. Furthermore, we denote $C_{i_\sigma,\sigma}$ the *largest connected component* in the binary image E_{I_σ}, i.e.,

$$i_\sigma = \text{argmax}_i |C_{i,\sigma}|,$$

where $|C|$ represents the area of a connected component C.

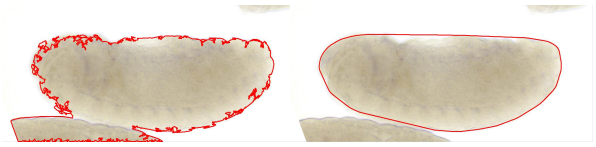
In the following, we propose three criteria for the selection of the optimal scale to localize the ROI in an embryonic image.

Criterion 1: Minimization of the number of connected components

As illustrated in Fig. 1, an embryonic image tends to be over-segmented if scales are relatively small. As the scale increases, the number of connected components decreases to a stable number. Based on this observation, we propose a criterion that aims to minimize the number of connected components, i.e.,

$$\sigma^* = \text{argmin}_\sigma |\{C_{i,\sigma}\}|, \tag{1}$$

where $|\{C_{i,\sigma}\}|$ indicates the number of connected components. Moreover, if there are multiple scales leading to the minimum number of connected components, the smallest scale will be selected.



(a) (b)

Fig. 2. Largest connected components: a) without, b) with smoothness constraint.

Criterion 2: Maximization of the largest connected components

A different strategy to select the optimal scale is to focus on analyzing largest connected components across scales. Our second criterion is to maximize the area of the largest connected component, i.e.,

$$\sigma^* = \text{argmax}_\sigma |C_{i_\sigma,\sigma}| \tag{2}$$

Furthermore, we add a smoothness constraint on the boundary of the largest connected component $C_{i_\sigma,\sigma}$. Denote $c = \{p_0, p_1, \ldots, p_n, p_0\}$ the boundary of a connected component $C_{i_\sigma,\sigma}$, and $a_i = \cos^{-1}\langle \frac{p_{i+1}-p_i}{\|p_{i+1}-p_i\|}, \frac{p_{i-1}-p_i}{\|p_{i-1}-p_i\|} \rangle$ the cross angle of point p_i. A connected component $C_{i_\sigma,\sigma}$ is said to satisfy the *smoothness constraint* if $\max a_i \geq 135^\circ$.

Fig. 2 shows a comparison of the Criterion 2 without and with the smoothness constraint. Without the smoothness constraint, the jaggy boundary of the largest connected component extracted by Criterion 2 shows the "over-detailed" connected component, which in turn indicates that Criterion 2 selects a relatively small scale as the optimal scale. With smoothness constraint, the largest connected component extracted by Criterion 2 is consistent with the real boundary.

Criterion 3: Minimization of shape inconsistency

Ellipse is a shape model that has been used to model the boundary of Drosophila embryos [8]. Our third criterion aims to maximize the fitness of the boundary of the largest connected component to the ellipse model, equivalently, to minimize the their inconsistency. Specifically, we will use a quadratic equation to model the shape of an embryo, for the convenience of numerical computation, as follows:

$$S(p) = ax^2 + by^2 + cxy + dx + ey + f = 0,$$

where $p = (x, y)$ is a point, and $a, b, \ldots, f$ are 6 parameters. Given a point set P (the boundary of a connected component), $a, b, \ldots, f$ can be estimated by linear fitting via eigen-decomposition [2].

The inconsistency (or offset) between P and its estimated shape S, denoted as $d(P, S)$, measures how well the point set P is consistent with a shape of ellipse. Under the assumption of a single ROI (targeting embryo), our focus is on the

measurement of the inconsistency between the largest connected component and its estimated shape, i.e., $d(C_{i_\sigma,\sigma}, S)$. (For simplicity, $C_{i_\sigma,\sigma}$ denotes the boundary of a connected component.) Formally, our third criterion aims to minimize above-mentioned inconsistency as follows:

$$\sigma^* = \text{argmin}_\sigma d(C_{i_\sigma,\sigma}, S) \quad (3)$$

Fundamentally, the inconsistency between two different point sets P and S is decided by the statistics of point-to-set distances, e.g., $\{d(p, S)\}_{p\in P}$. There can be different interpretation of these statistics towards the inconsistency measurement of two point sets, including mean, median, worst case, etc. In this study, the inconsistency between the largest connected component and its estimated shape is defined by the worst case of point-to-set statistics, i.e.,

$$d(C_{i_\sigma,\sigma}, S) = \max_p \{d(p, S)\}_{p\in C_{i_\sigma,\sigma}}.$$

Recall that the ellipse is a rough model of drosophila embryos, and therefore the general statistics such as mean or median may not be the most appropriate choice. The rationale of the worst case based distance lies in the observation of scale-space behaviors: With the increment of the scale σ, neighboring objects tend to "diffuse/merge" into the ROI. This brings an "expanded" connected component that contains not only the ROI but also non-ROI objects.

3. EXPERIMENTS AND CONCLUSIONS

We will test the proposed framework on BDGP (Berkley Drosophila Genome Project) [1]. BDGP images were captured using high throughput in situ protocol [1] for the determination of gene expression patterns of Drosophila embryos during different developmental stages. Each image is a high-resolution spatial representation of an embryo that might be neighbored by some other embryos. BDGP images are available at a public webpage [1].

We first give an experimental analysis of three proposed criteria along with the localization framework on the BDGP embryonic images. We then present a set of successful results achieved by the best criterion, and an estimation of the localization accuracy. To have a quantified evaluation of the proposed method, we define a successful localization if the overlapping region between the algorithmically extracted ROI and manually extracted ROI (i.e., the groundtruth) is less than 95%. We will also present an analysis of a failure case to illustrate the limitation of the proposed framework.

The scale σ in our experiment ranges 1 to 200 in a sampled space, which is, [1, 13, 25, 38, 50, 75, 100, 150, 200]. Sampled scale space is used to reduce the computational cost.

[1] http://www.fruitfly.org/insituimages/insitu_images/

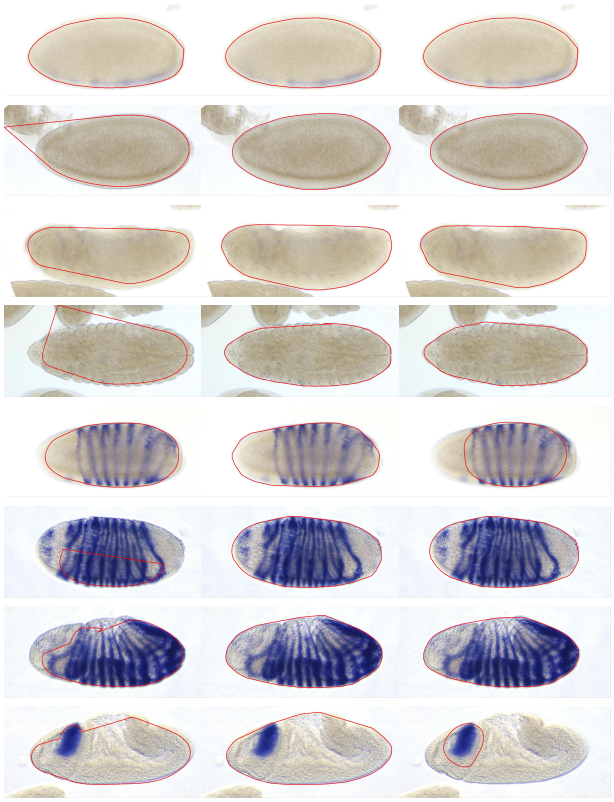


Fig. 3. Comparison of three criteria. Overall, Criterion 2 achieves better results than the other two.

3.1. Experimental analysis of three criteria

Fig. 3 presents the results of four embryonic images achieved by the three criteria. From Fig. 3, we obtain two main observations. First, 1st-4th images contain no gene expressive (blue) regions. Criterion 2 and 3 are comparable with each other. 5th-8th images contain gene expressive (blue) regions. Criterion 3 (ellipse oriented) localizes a sub-region (an expressive region) instead of the entire ROI. Second, Criterion 2 outperforms Criterion 3 that in turn outperforms Criterion 1. The superiority of Criterion 2 over Criterion 3 implies that the ellipse model seems not a very effective shape model in the context of localization of ROIs in drosophila embryonic images. It is also a natural expectation that Criterion 2, without any shape constraint has better potential to be applicable to the localization of ROIs of other type (micro-scopic) images.

3.2. Experimental results

We tested the proposed method with Criterion 2 on a dataset of 1000 BDGP images, and obtained the localization accuracy of 91% in terms of the successful localization defined above. As a comparison, we tested connected components (without scale space) on the dataset, and obtained the accu-

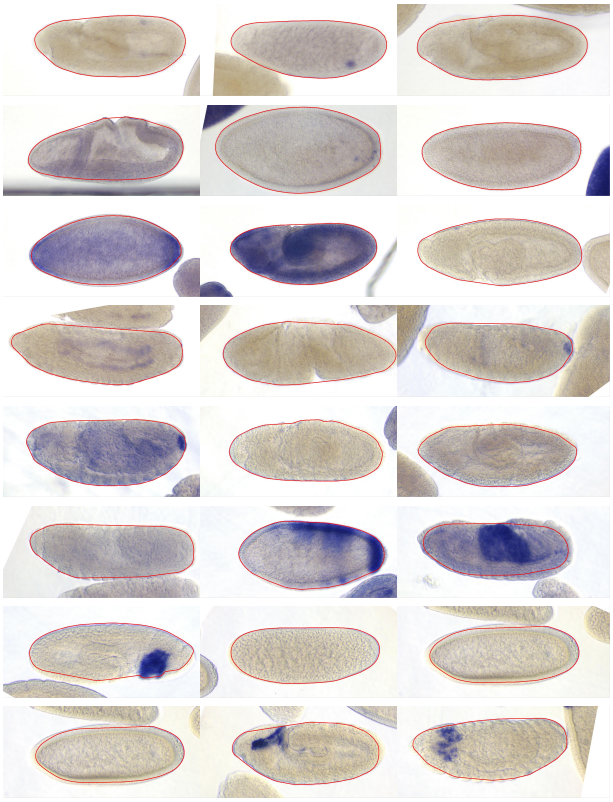


Fig. 4. Samples of successful localization.

Fig. 5. Samples of failedlocalization.

racy of 65%. We also test an active contour based localization approach [5] using the same dataset, and obtain accuracy 86%, which demonstrate the improvement of the proposed framework over existing work.

Fig. 4 and Fig. 5 present successful and failed localization of the proposed method, respectively. In Fig. 5, the first one is caused by the occurrence of various gene-expressive subregions. The second one is caused by a touching neighboring embryo. The third one is caused by the complexity of non-touching neighboring embryos.

Fig. 6 shows an embryonic image that fails the proposed criterion in our current implementation of scale space. We check structures of connected components in scale space. Recall that scale σ ranging from 1 to 200 non-continuously. We analyze that the underlying reason for the failure may be caused by insufficient quantification of the scale space. More essentially, it is caused by the complex image structure that is in turn caused by a number of gene expression regions.

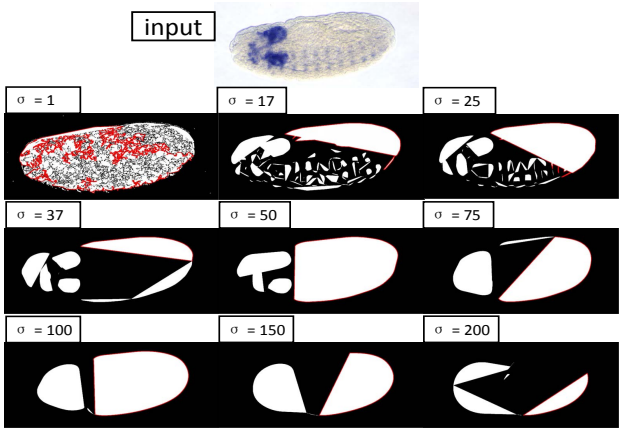


Fig. 6. Analysis of a failure case.

4. REFERENCES

[1] P. T. et al. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 3(12):1–14, 2002.

[2] A. Fitzgibbon, M. Pilu, and R. Fisher. Direct least square fitting of ellipses. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5):476–480.

[3] E. Frise, A. Hammonds, and S. Celniker. Systematic image-driven analysis of the spatial drosophila embryonic expression landscape. *Molecular Systems Biology*, 6:345, 2010.

[4] M. Gargesha, J. Yang, B. V. Emden, S. Panchanathan, and S. Kumar. Automatic annotation techniques for gene expression images of the fruit fly embryo. In *Visual Communications and Image Processing 2005. Edited by Li, Shipeng; Pereira, Fernando; Shum, Heung-Yeung; Tescher, Andrew G. Proceedings of the SPIE. Vol. 5960*, pages 576–583, 2005.

[5] Q. Li and C. Kambhamettu. Contour extraction of drosophila embryos. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 8(6):1509–1521, 2011.

[6] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[7] D. Mace, N. Varnado, W. Zhang, E. Frise, and U. Ohler. Extraction and comparison of gene expression patterns from 2d rna in situ hybridization images. *Bioinformatics*, 15(26(6)):761–9, 2010.

[8] J. Pan, A. Balan, E. Xing, A. Traina, and C. Faloutsos. Automatic mining of fruit fly embryo images. In *KDD*, pages 693–698, 2006.

[9] H. Peng and E. W. Myers. Comparing *n situ m*rna expression patterns of *drosophila* embryos. In *RECOMB*, pages 157–166, 2004.

[10] K. Puniyani, C. Faloutsos, and E. P. Xing. Spex2: Automated concise extraction of spatial gene expression patterns from fly embryo ish images. *Bioinformatics*, 26(12):i47–i56, 2010.