Automatic Annotation Techniques for Gene Expression Images of the Fruit Fly Embryo

Madhusudhana Gargesha[1,3*], Jian Yang[1,2], Bernard Van Emden[1,2],
Sethuraman Panchanathan[1,4], Sudhir Kumar[1,2*]

[1]Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, AZ 85287-5301
[2]School of Life Sciences, Arizona State University, Tempe, AZ 85287-4501
[3]Department of Electrical Engineering, Ira A. Fulton School of Engineering, Arizona State University, Tempe AZ 85287-5706
[4]Department of Computer Science and Engineering, Ira A. Fulton School of Engineering, Arizona State University, Tempe AZ 85287-8809

ABSTRACT

We present an application of image analysis techniques to automatically annotate biological images depicting gene expression patterns in developing embryos of fruit fly (*Drosophila melanogaster*), a model organism to study gene interaction. The aim is to determine the view (lateral versus dorsal/ventral [non-lateral]), orientation (anterior-left or anterior-right), and the developmental stage of the embryo. We employed contour curvature analysis, symmetry of the gene expression patterns, and shape differences at the anterior and posterior ends of the embryo, among others, for these purposes. An analysis of a pilot database of 3500 images indicates that view was correctly identified in 62%, orientation in 85%, and developmental stage in 73% of the images. We observed that correct inferences had better separation in feature space than incorrect inferences. This means that, although these methods do not exhibit very high classification accuracy, they could be employed to identify images which need manual intervention, thereby reducing the target set for biologists. The novelty in this work is in the integration of well-established image analysis with the biological knowledge for annotating the embryos. Our examinations show that features that provide discrimination ability among different views, different orientations, and different developmental stages are often restricted to certain regions of the embryo, which agrees with the longstanding knowledge in the developmental biological community.

Keywords: view, orientation, stage, curvature analysis, gene expression pattern, symmetry, texture analysis

1. INTRODUCTION

Currently, biologists spend significant resources (in terms of time) in manually annotating images derived from large-scale experimental techniques (for e.g., the Berkeley *Drosophila* Genome Project (BDGP)[1]). This annotation includes data such as view, orientation, and stage of development of the embryos, among other biological features. The primary objective of our work is to develop image analysis techniques to automate some of these annotation tasks in order to reduce, or eliminate, time and resource requirements. We begin with a description of what is meant by "view", "orientation", and "stage" of development for the images in which the fruit fly gene expression patterns are captured. View refers to the positioning of the embryo during the image acquisition process. It could take two states: lateral and non-lateral (dorsal/ventral) (Figure 1). Orientation refers to the positioning of the anterior and posterior ends of the embryo in the image. If the anterior end of the embryo is at the left hand side of the image, we call the orientation anterior-left. Otherwise, we call it anterior-right. Stages in *Drosophila melanogaster* development denote the time after fertilization at which certain specific events occur in the developmental cycle[2]. Before any useful features could be extracted, we need to apply some preprocessing operations to standardize these images. We specifically perform

*Correspondence should be addressed to M.G. (mgargesha@asu.edu) or S.K. (s.kumar@asu.edu). The National Institutes of Health (S.K.) and Center for Evolutionary Functional Genomics, The Biodesign Institute, ASU, provided support for this research.

background removal, edge fitting and resizing operations on the image to accomplish this task following the procedures outlined by Kumar *et al.*[3].

Our technique for view determination can be classified into the same category of image processing techniques as pose/view determination techniques that have been applied to other types of images, including pose detection from human face images[4,5], pose estimation for textured surfaces using vanishing points[6], and automatic determination of view angles for tomography images[7]. However, unlike the natural world images, the presence of gene expression patterns in biological images of the same view, orientation, and stages poses special challenges and makes many of the traditional solutions unsuitable. In our approach, we examine the usefulness of gene expression staining as an augmentative technique. Our primary technique for automatic lateral/non-lateral view determination employs shape analysis of the contours of the embryo (see Figure 1). We have specifically examined the amount of discriminatory information present in various vertical and horizontal slices along the anterior-posterior axis, and verified our results with the known biological knowledge. The augmentative technique for view determination employs difference in gene expression areas for the upper and lower halves of the embryo.

Next, we observe that the anterior-end is usually narrower compared to the posterior-end of the embryo, a feature which is more typical of lateral than non-lateral views (Figure 1). We have based our primary technique for orientation annotation determination on differences in spatial extent of the embryo at the anterior and posterior ends. An augmentative technique, based on multi-scale curvature analysis[8] of the embryonic contours at the anterior and posterior ends of the embryo, aids the annotation process.

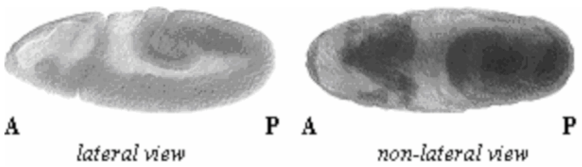

Figure 1. Lateral and non-lateral views of *Drosophila melanogaster* embryos. TheAnterior (A) and posterior (P) ends have also been labeled for each of these views.

Our stage determination technique is similar to texture-based analysis and classification techniques that have been applied, for instance, to meteorological images[9], satellite SAR imagery[10], classification of carotid plaque images[11] and digitized mammograms[12]. All of these techniques derive feature vectors from block-based analysis of the images, as is obvious in any texture analysis technique. However, unlike these images, biological images have distinct regions that are expected to provide most of the discriminatory information. The feature vector for stage determination was based on textural features (extracted using Gabor filters[13]) of image sub-blocks, because image texture at sub-block level changes as embryonic development progresses (Figure 2). In order to enhance local features in the image that are necessary for discriminating between the stage groups, we perform a preprocessing on the images by employing an image enhancement technique called CLAHE[14]. Finally, a discriminant classifier[15] was employed for annotation.

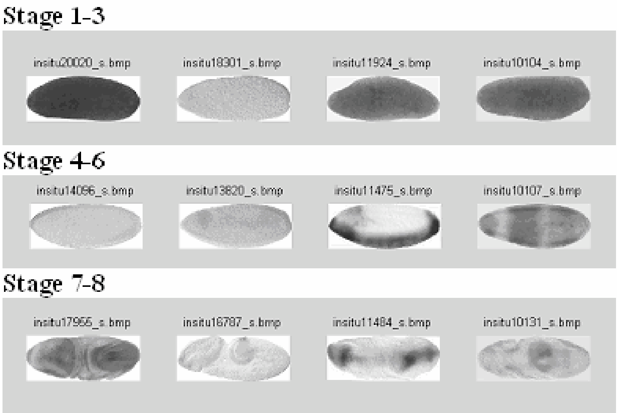

Figure 2. Drosophila images across different stages of development.

Large-scale validation experiments were conducted on about 3500 *in situ* hybridization images of the *Drosophila melanogaster* from the BDGP project[1]. These images were standardized to a predetermined spatial size and the manual annotations of view, orientation and stage done by experts were used as the ground truth. We also determined a measure of the degree of separation for the results obtained from the various systems and observed that the correctly annotated results had a higher degree of separation than the wrongly annotated ones. Our results demonstrate that the image analysis techniques can be useful in significantly reducing annotation time and they agree with known biological understanding regarding the amount of discriminatory information contained in various embryonic sub-regions.

The rest of the paper is organized as follows. We discuss automatic techniques for view, orientation and stage determination, in sections 2, 3 and 4, respectively. We present experimental results in section 5 and discuss them in section 6. Finally, section 7 concludes the paper.

2. LATERAL/ NON-LATERAL VIEW DETERMINATION

We have based our automatic techniques for lateral/non-lateral view determination on the curvature of the outer contour. To determine the axis of symmetry (see Figure 3), we find the centroid and orientation of the embryo using image analysis tools[16] and determine the line that passes through the centroid with the orientation computed in the previous step. We then perform a multi-scale curvature analysis[8] on the upper and lower contours of the embryo. A feature vector based on differences in curvature values of corresponding points in the upper and lower contours is used to train a discriminant analysis-based classifier[15].

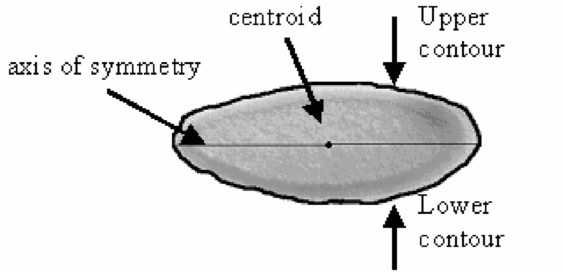

Figure 3. Illustration of upper and lower outer contour of the Drosophila (lateral view), obtained by determining the central axis of symmetry

A secondary method for lateral/non-lateral view annotation employs the gene expression pattern contained in the image. The image is divided into a predetermined number of vertical strips. The gene expression binary pattern for the whole image is automatically derived by employing the CLAHE technique[14], followed by global thresholding[17]. The difference in the amount of expression on either side of the horizontal axis of symmetry was determined for each of the vertical strips (Figure 4). The number of vertical strips for which this quantity exceeds an empirically determined threshold is used as the basis for supervised classification using a discriminant analysis-based classifier[15]. When we combine the two techniques suitably (by simply taking the logical OR of the decisions of the classifiers based on the two features), we observed that the error rate was lower than that for each of the individual techniques (see experimental results in section 5).

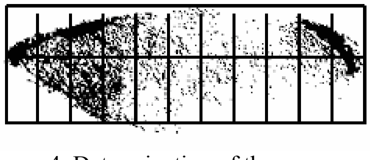

Figure 4. Determination of the gene expression symmetry.

If the classifier employs distance measure $\mathbf{D^i_L}$ from every sample **i** (in the test dataset) to the training set of laterals, and distance measure $\mathbf{D^i_{NL}}$ from every sample **i** to the training set of non-laterals, then we can define the following measure of degree of separation for every sample **i** (which is directly indicative of the degree of separation in the results):

$$D^i_{L-NL} = \left| D^i_L - D^i_{NL} \right| \tag{1}$$

3. ANTERIOR/POSTERIOR ORIENTATION DETERMINATION

The anterior-end of the fruit fly embryo is usually narrower than the posterior-end. Therefore, the primary technique for orientation determination uses pixel ratio comparison (number of pixels lying within the embryo boundaries versus the number outside it) at the two ends of the embryo image. Considering the standardized image (Figure 5), the anterior end of the embryo would have a smaller ratio of area of foreground pixels (i.e., pixels belonging to the embryo) to total area of a rectangular box enclosing part of the embryo, as compared to the posterior side. A difference in this ratio for the anterior and posterior ends is used as the parameter for an empirically-determined threshold-based classification scheme. This technique is augmented by a secondary technique that employs a multi-scale curvature-based descriptor[8] derived from the embryonic contour. Parts of the embryonic contour lying within rectangular boxes shown in Figure 5 are used to compare curvature at the anterior and posterior ends, and empirical thresholds are used for the difference in curvature thus determined, to perform classification (the classifier for A/P orientation determination needs to assign images to two classes – "anterior-left" and "anterior -right").

We conjectured that a combination of the two techniques would result in better A/P orientation determination. The pixel ratio-based technique was used as a primary technique, with the curvature-based technique as an auxiliary technique. The threshold value from the primary technique was used to decide whether or not to employ the auxiliary technique. When the decisions of the two techniques agreed, the classification was trivial. When the primary and auxiliary techniques differed in their decisions, the images were classified as uncertain, which would require manual intervention.

However, we observed from experimental results that the uncertain cases usually constituted a small ratio of the total number of images. Hence, this classification mechanism automates the decision process to a large extent.

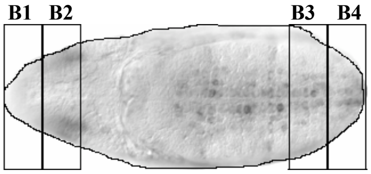

Figure 5. Boxes for computing pixel ratios for the purpose of A/P orientation detection. (The outer contour of the embryo has been traced in black)

4. DETERMINATION OF STAGE OF DEVELOPMENT

For distinguishing between different stages of development, we need to extract features from internal organs of the embryo (unlike view determination where we extracted a descriptor for external contour shape). We observed that a distinguishing feature across the various developmental stages is image textural properties at a sub-block level. We examined a subset of images drawn from three stage groups (1-3, 4-6 and 7-8) and observed that textural features at the sub-block level could be used for discrimination. However, features obtained from some regions of biological images provide most of the discriminatory information in some stages of development. For example, we know that morphological changes at the anterior and posteriors end of the embryo occur during stages 4-6, e.g., formation and shifting of pole cells at posterior end, prominent displacement of cell membranes at the anterior and posterior ends.

Further changes occur in stages 7-8 which are mainly restricted to the middle regions of the embryo, e.g., formation of amnioserosa and amnioproctodeal invagination[18].

We performed CLAHE[14] as a preprocessing step, followed by a feature extraction step where Gabor filters[13] at different scales and orientations were convolved with the image and the resulting data was extracted from different regions at a sub-block level (Figure 6). Finally, a discriminant analysis-based classifier[15] was employed with suitable training to determine the stage of development of embryo images.

In a generic sense, let $st_k, k = 1,2,...n$ denote n stage groups. If the classifier employs distance measure $D^i_{st_k}$ from every sample **i** (in the test dataset) to the training set corresponding to stage group st_k . The degree of separation for each sample **i** in the test set from each of the stage groups is given by:

$$d - sep^i_{st_k} = \frac{1}{n-1} \sum_{\forall j \neq k} \left| D^i_{st_j} - D^i_{st_k} \right| \tag{2}$$

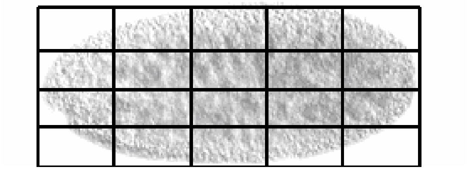

Figure 6. A rectangular grid overlaid on the CLAHE image, to compute Gabor features from each sub-block.

5. EXPERIMENTAL PROCEDURES AND RESULTS

5.1 Results from lateral/non-lateral view determination

Experiments were conducted on a database of 3526 images that had lateral/non-lateral annotations provided by a set of developmental biology experts at the ASU School of Life Sciences. The dataset consisted of 77% laterals and 23% non-laterals. We manually picked 100 images for training – 50 laterals and 50 non-laterals, and used the remaining 3426 images for testing. We extracted vertical slices along the anterior -posterior axis of the embryo consisting of the medial 10%, 20%, 30%, 40%, 50% and 100% of embryo contour features, for the curvature-based primary technique. We extracted gene expression from the entire embryo for the secondary technique. We then combined the decisions of the classifiers for the two techniques, to obtain overall classification rates (Figure 7a-c). Figure 7d shows the measure discussed in equation (1) averaged over all test dataset samples.

5.2 Results from anterior/posterior orientation determination

Experiments were conducted on a database of 3665 images for which the manually annotated orientations were available. Figure 8 shows the performance of the two individual techniques on this dataset. When we combined the two techniques as described previously (section 3), the overall performance obtained was better than any of the individual techniques. However, we ended up with 802 images in the uncertain set, which meant that these images needed manual examination to properly determine their anterior/posterior orientation.

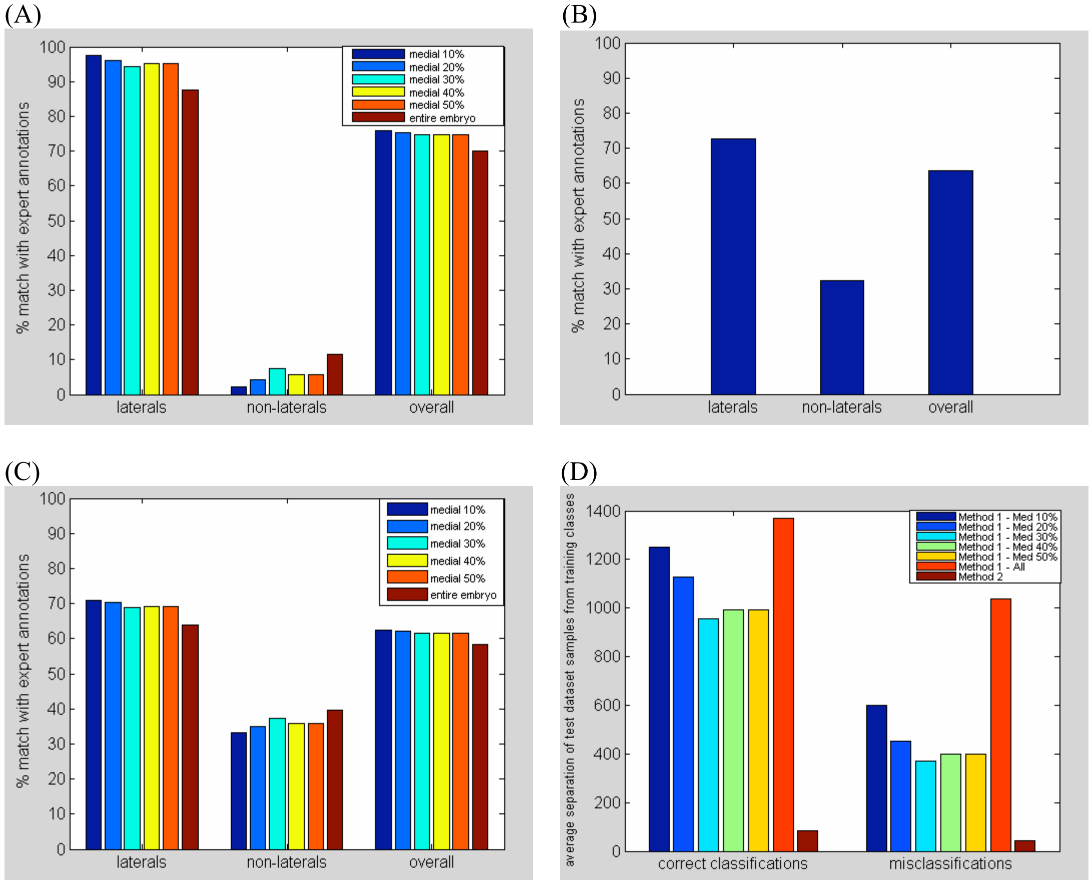

Figure 7. Correct Classification rates of (a) curvature analysis method for view determination with features obtained from various medial slices, (b) gene expression symmetry method for view determination (with features obtained from the entire image), (c) combined method for view determination (a logical AND combination of the classifier decisions for the two techniques) . (d) Classifier performance in terms of degree of separation of samples from lateral and non-laterals (equation 1).

5.3 Results from stage determination

Experiments were conducted on 3608 standardized images for which stage annotations were available; there were 460 images from stages 1-3, 1974 images from stages 4-6, and 1174 images from stages 7-8. The training set for each stage group consisted of 10% of the total number of images for that group; the remaining images were used for testing. We divided the standardized embryo into sub-blocks as shown in Figure 6. We chose a 32 x 32 sub-block size and ignored the outer-most ring of sub-blocks as they were not important for classification. From the remaining sub-blocks, we extracted features from 0-25%, 25-50%, 50-75% and 75-100% on either side of the embryo vertical midline. We then averaged the classification performance across 25 random choices of test and training data sets. Figure 8b shows the classification performance. Figure8c shows the measure discussed in equation (2) above averaged over all samples in the test dataset.

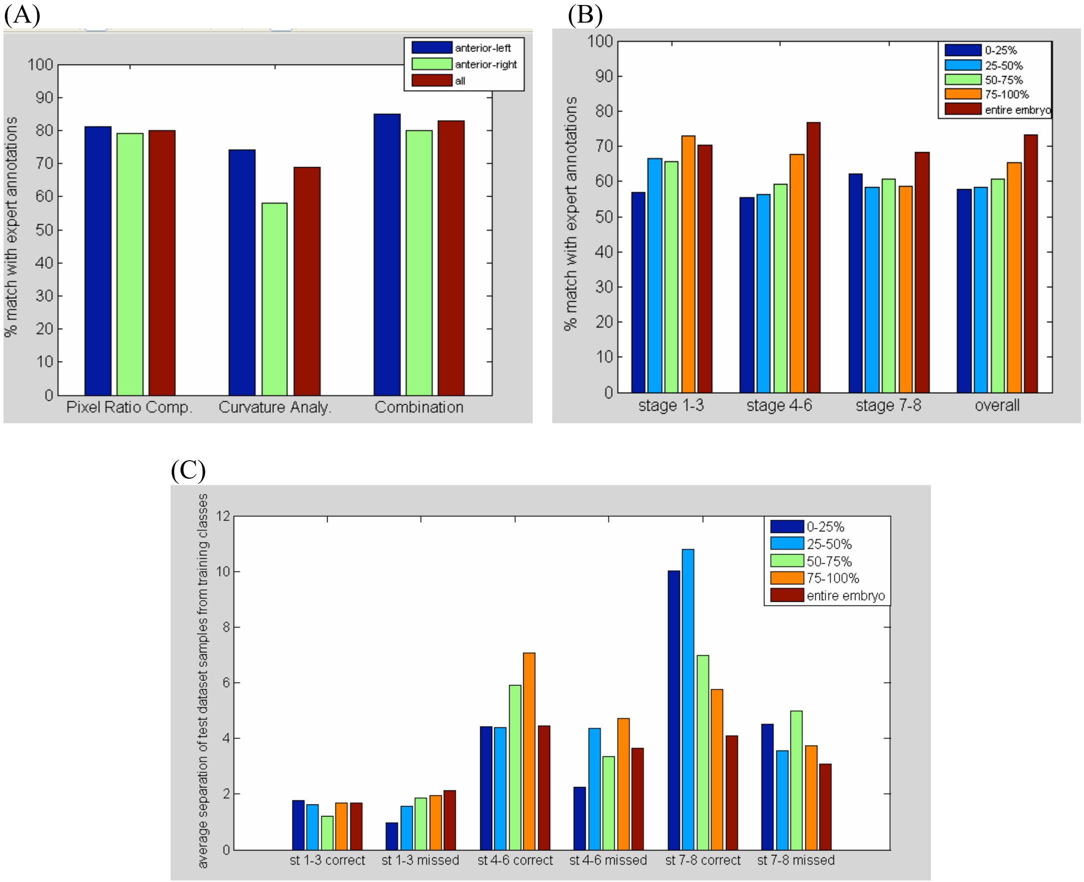

Figure 8. (a) Performance of the orientation determination technique. (b) Correct Classification Rates of stage determination technique for various sub-regions within the embryo image. (c) Classifier performance for stage determination in terms of degree of separation of samples from various stage groups (equation 2).

6. DISCUSSION OF RESULTS

6.1 View Determination

From Figures 7 (a-c), we observe that our methods perform very well on lateral views, but a large percentage of non-lateral views are wrongly annotated. This is because most non-lateral views in our current dataset have morphological distortions in their shape, and have little or no gene expression. The curvature analysis based method is heavily biased towards laterals. The gene expression symmetry method performs better with non-lateral views. A combination of the two methods leads to lower overall performance compared to the curvature analysis method, but reduces the bias towards laterals. Also, we observe that extracting data from narrower medial slices leads to an increase in classification performance of lateral views. We note the measure of separation plotted in Figure 7d is very different for correct and missed classifications (it is higher for the former case as expected). From this Figure, it is evident that this measure could be used with a suitable threshold value to decide which images to inspect visually for deriving the annotations.

6.2 Orientation Determination

We observe that the pixel ratio comparison method performs better than the curvature analysis method for both anterior-left and anterior-right images (Figure 8a). However, the combination of the two methods performs better than any individual method because it increases the correct classification rate of both anterior-left and anterior-right.

6.3 Stage Determination

We observe that the stage determination performance depends on the medial 0-25% (central part) and the 75-100% (anterior and posterior ends) features to a large extent (Figure 8b, c). In other words, these features are a large contributor to the overall performance. Specifically, the separation of stages 1-3 and 4-6 from each other and from stage group 7-8 is largely dependent on features from 75-100% on either side of the vertical central line, which corresponds to the anterior and posterior ends. This result tallies with biological domain knowledge as discussed in section 4. Also, the separation of stages 7-8 from the rest of the stage groups is dependent on features from 0-25% on either side of the vertical central line, which corresponds to the middle of the embryo. This, again, is consistent with the knowledge of *Drosophila* development. We note that, with the exception of stage group 1-3, the measure of separation plotted in Figure 8c is much higher for correct classifications than for misclassifications. It is obvious that the measure could be used with a suitable threshold value to decide which images to inspect visually for deriving the annotations.

7. CONCLUSIONS AND FUTURE WORK

Experimental results have proved that the techniques for view, orientation and stage determination are efficient in automating the annotation process for stage 1-8 gene expression images of the fruit fly (*Drosophila melanogaster*) to a large extent. We find that for each annotation task (view, orientation and stage) a combination of two techniques performs better than any of the individual techniques. However, there exist some drawbacks in our framework. The performance of our framework is greatly affected by preprocessing stages (e.g., the accuracy of image segmentation for extracting gene expression binaries, the parameters that control the CLAHE preprocessing step) and also the choice of empirical thresholds used in orientation determination. Therefore, more robust methods for preprocessing need to be explored. Also, the performance of the both view and stage determination technique is dependent on the staining present in the embryos; results suggest that it aids view determination but that it is detrimental to stage determination. This points to the need for developing staining-independent techniques for stage determination.

REFERENCES

1. P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S.E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S.E. Celniker, and G.M. Rubin, *Systematic determination of patterns of gene expression during Drosophila embryogenesis*, Genome Biol. 3(12):research0088.0081-0088.0014, 2002.

2. The Interactive Fly: Stages of Development and Mitotic Domains.
(URL: http://sdb.bio.purdue.edu/fly/aimain/2stages.htm>)

3. S. Kumar, K. Jayaraman, S. Panchanathan, R. Gurunathan, A. Marti-Subirana, and S.J. Newfeld, *BEST: A novel computational approach for comparing gene expression patterns from early stages of Drosophila melanogaster development*. Genetics 2002, 162(4):2037-2047, 2002.

4. H. Saito, A. Watanabe, and S. Ozawa, *Face pose estimating system based on eigenspace analysis*, Proceedings of the International Conference on Image Processing, 1:638-642, 1999.

5. A. Tsukamoto, C.W. Lee, and S. Tsuji, *Pose estimation of human face using synthesized model images*, Proceedings of the International Conference on Image Processing, 3: 93-97, 1994.

6. E. Ribeiro and E.R. Hancock, *Improved pose estimation for texture planes using multiple vanishing points*, Proceedings of the International Conference on Image Processing, 1:148-152, 1999.

7. S. Basu and Y. Bresler, *Tomography with unknown view angles*, IEEE International Conference on Acoustics, Speech, and Signal Processing, 4: 2845-2848, 1997.

8. L.F. Costa and R. M. Cesar, *Shape Analysis and Classification: Theory and Practice*, CRC Press, Cleveland, Ohio, 484-500, 2000.

9. L. Alparone, G. Benelli, and A. Vagniluca, *Texture-based analysis techniques for the classification of radar images*, IEEE Proceedings of Radar and Signal Processing, 137(4): 276-282, 1990.

10. F. Dell'Acqua, and P. Gamba, *Texture-based characterization of urban environments on satellite SAR images*, IEEE Transactions on Geoscience and Remote Sensing, 41(1): 153-159, 2003

11. C.I. Christodoulou, C.S. Pattichis, M. Pantziaris, and A. Nicolaides, *Texture-based classification of atherosclerotic carotid plaques*, IEEE Transactions on Medical Imaging, 22(7): 902-912, 2003.

12. S. Baeg, and N. Kehtarnavaz, *Texture based classification of mass abnormalities in mammograms*, Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems, 163-168, 2000.

13. B.S. Manjunath and W. Y. Ma, *Texture features for browsing and retrieval of image data*, IEEE T-PAMI special issue on Digital Libraries, Nov. 1996.

14. S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, and K.E. Muller, *Contrast-limited adaptive histogram equalization: speed and effectiveness*, Proceedings of the First Conference on Visualization in Biomedical Computing, 337-345, 1990.

15. R. Baldock and J. Graham, *Image Processing and Analysis: A Practical Approach*, Oxford University Press Inc., New York, 114-121, 2000.

16. MATLAB Image Processing Toolbox. (URL: http://www.mathworks.com/products/image/)

17. N. Otsu, *A threshold selection method from gray-level histograms*, IEEE Transactions on Systems, Man, and Cybernetics, 9: 62-66, 1979.

18. V. Hartenstein, Atlas of *Drosophila* Development, Cold Spring Harbor Laboratory Press, 1993.