Classifying Fruit Fly Early Embryonic Developmental Stage Based on Embryo In situ Hybridization Images

Hua Zhong, Wei-Bang Chen, Chengcui Zhang
Department of Computer and Information Sciences,
The University of Alabama at Birmingham
Birmingham, AL 35294, USA
{hzhang, wbc0522, zhang}@cis.uab.edu

Abstract—In this paper, we present a supervised classification system for sorting Drosophila embryonic in situ hybridization (ISH) images according to their developmental stages. The proposed system first segments the embryo from an image and registers it for subsequent texture feature extraction. In order to extract the most distinguishing features for classifying developmental stages, we identify several areas of interest in an embryo with peculiar traits. Gabor filter is applied on these areas to extract texture features and Principal Component Analysis (PCA) is then performed on the extracted features to reduce dimensionality while retaining significant information. We adopt multi-class Support Vector Machine (SVM) as the classifier that learns model parameters from the training examples and classifies new examples with the trained model. We evaluate the system performance by comparing it to existing algorithms. The experimental results show that the proposed system achieves good performance in classifying Drosophila embryonic developmental stages and outperforms other state-of-the-art algorithms.

Keywords-Drosophila; in situ hybridization; Gabor filter; Support Vector Machine; Principal Component Analysis

I. INTRODUCTION

Gene expression directs embryo differentiation and development. To understand the embryogenesis, it is essential to know the roles of genes involved in the embryo differentiation and development [1, 2, 3]. In order to capture the spatial-temporal gene expression patterns, a series of time course in situ hybridization (ISH) experiments can faithfully record expression patterns across developmental stages [4, 5].

In developmental biology research, Drosophila melanogaster (the fruit fly) has been widely used as a model organism due to its short life cycle and large number of offspring. Developmental biologists procure the spatial-temporal gene expression patterns from Drosophila embryos in order to discover the fundamental developmental mechanisms of eukaryotic organisms. According to a staging system reported by Bownes (1975) and improved by Campos-Ortega and Hartenstein (1985), the Drosophila embryogenesis consists of 16 developmental stages whose durations vary from 10 minutes to 3 hours [6, 7].

Fig. 1 exhibits the in situ hybridization of CG7780 (DNaseII) in developmental stage 3 to stage 6. The dark blue staining indicates the expression of CG7780. It is obviously that CG7780 is silent in developmental stages 3 and 4, and starts expression in anterior and posterior endoderm in stage 5. In later stages, CG7780 continuously expresses in tissues differentiated from anterior and posterior endoderm.

The morphological changes of gene expression pattern of the same gene across developmental stages play an important role in understanding the interplay of genes in embryo differentiation and development [2]. These changes suggest the occurrence of specific events at a certain stage in the developmental cycle.

Currently, finding the correlation between gene expression and tissue differentiation from the in situ hybridization images heavily relies on human visual inspection. However, due to the fast growing of the Drosophila embryonic in situ hybridization images in the gene expression pattern database, manual inspection of these in situ hybridization images becomes increasingly infeasible and impractical.

In order to automate the visual inspection process, it is essential to determine the developmental stage of an embryo from in situ hybridization images. This paper presents an automatic supervised classification system for sorting Drosophila embryonic in situ hybridization images according to their developmental stages through image analysis.

As a matter of fact, the developmental stages in embryogenesis are typically defined on the basis of some morphological markers [8]. These markers provide dominant features to divide embryogenesis into a series of consecutive developmental stages, which motivates us to extract the

Figure 1. CG7780 gene expression patterns in Drosophila embryo at different developmental stages: (a) Stage 3, (b) Stage 4, (c) Stage 5, and (d) Stage 6
dominant features from embryonic *in situ* hybridization images.

In this study, we first normalize all the embryonic images. Then, we introduce Gabor filters in the proposed classification system for extracting the texture features that describe different developmental stages. In addition, according to our observations, not all the areas in an embryo can provide distinguishing texture features. Instead of using the texture features from the entire embryo image, we select several significant embryonic areas for extracting texture features. In addition, we apply principal component analysis (PCA) [9] on the extracted features, which reduces the number of dimensions of the feature space and eliminates some noises at the same time. Finally, a multi-class Support Vector Machine (SVM) is employed for classifying the image into the corresponding developmental stage. According to our experiments, the proposed classification system achieves high classification accuracy (93%) and outperforms several existing methods.

In the rest of this paper, we describe the proposed methodology in Section II. Section III presents the experimental results, and Section IV concludes this paper.

II. THE PROPOSED METHOD

The scheme of the proposed classification system comprises three main stages, including (1) image pre-processing, (2) Gabor texture feature extraction with principal component analysis, and (3) embryo image classification with multi-class support vector machine. We illustrate the high level architecture of the proposed *Drosophila* embryonic classification system in Fig. 2.

A. Image Pre-processing

1) Embryo Isolation and Image Registration: As shown in Fig. 1, the raw images of individual *Drosophila* embryos are mainly produced for the purpose of studying gene expression patterns during embryogenesis. Various tissues in embryos are colored due to the *in situ* hybridization, which may introduce noise in image analysis. In addition, these images are quite different in their shapes, orientations, and scales. Therefore, it is essential to normalize images for extracting comparable features for subsequent analysis. For this purpose, we apply the embryo isolation and image registration procedures reported in [4].

The goal of embryo isolation is to extract an individual embryo from an image. Once the embryo of interest is extracted from the image, we register and normalize the embryo image with dimensions of $600 \times 1200$. In addition, we normalize the embryo orientation in all registered images by orienting the anterior and dorsal of an embryo to the left and top, respectively. Finally, we normalize the image contrast by applying histogram equalization which conceptually uniform the histogram distribution, enhancing the texture in the image.

2) Significant Embryonic Areas: Various staging systems have been reported for studying *Drosophila* embryogenesis which describes the developmental process of a *Drosophila* embryo [10, 11, 12]. One of such staging systems reported by Bownes [12] (1975) and further improved by Campos-Ortega and Hartenstein (1985) has become the most commonly used staging system which divides the process of embryogenesis into 16 stages in three main phases, i.e., blastoderm, gastrulation, and organogenesis [6]. The blastoderm phase involves stages 1 to 5 where the egg starts cleavage division and gradually forming blastoderm cell. The gastrulation phase comprises stages 6 to 9 where the embryonic cells migrate and form three embryonic germ layers such as ectoderm, endoderm, and mesoderm. The last phase, the organogenesis phase, consists of stages 10 to 16 where the three embryonic germ layers gradually differentiate into internal organs of the organism.

In this study, we mainly address classifying *Drosophila* embryonic *in situ* hybridization images in the early
developmental stages, i.e., the stages prior to stage 6. In particular, we only adopt developmental stage 3 to stage 6. This is because in the blastoderm phase, stage 1 and stage 2 represent the process from fertilization to cleavage division of the egg. Since the embryogenesis events from these two stages are hardly visible and are not of much interest for further study, we do not include Drosophila embryonic in situ hybridization images from these two stages.

In order to automatically determine the developmental stage by means of image analysis, we observe that Drosophila embryonic in situ hybridization images obtained from different developmental stages give different visual representations of their patterns in embryo development. These different patterns can be described by texture features which can be used as criteria for staging embryo development. However, our observations on the embryonic images across different developmental stages show that not all the areas in an embryo at a given stage possess the distinguishing texture features that differentiate them from images in other stages.

In addition, two consecutive developmental stages often have highly similar texture features with only a few differences since they are adjacent in time. For example, existing algorithms usually have poor performance in distinguishing embryos from stages 4 and 5 since they have highly similar visual features. Thus, this motivates us to find those significant embryonic areas with distinguishing features which can be used to characterize embryos at different stages.

After visually examining embryo images carefully, we identify four significant embryonic areas as shown in Fig. 3. In our observations, we find that one major visual difference between the stages 4 and 5 is located at the rim of the embryos, which is actually the cytoplasmic rim at the periphery of the egg as shown in Fig. 4 (see the second and the third images in the first row).

In stage 6, the ventral furrow begins to invaginate, which forms two distinct textures in the ventral area. This area can serve as a distinguishing feature for stage 6 embryos when compared to the corresponding embryo areas in other stages as shown in Fig. 4, though the dorsal areas in stage 5 and 6 are highly similar.

Therefore, one major innovation in our proposed system is that we locate and combine several significant embryonic areas in Drosophila embryonic images and construct feature vectors with sufficient distinguishing power for classifying embryos in different developmental stages. These features represent and reflect the visual difference in raw images. In this study, we extract four significant embryonic areas from images and refer to them based on their location in the embryo, namely dorsal area, ventral area, posterior area, and anterior area. Each of these areas possesses at least one distinguishing texture from one stage.

In order to extract these areas, we first apply contrast-limited adaptive histogram equalization (CLAHE) on the raw image to enhance the contrast between embryo and background. Next, we convolute a 3-by-3 range filter on the enhanced image to emphasize the contour of the embryo. Subsequently, a segmentation mask is created for the embryonic area through threshold-based segmentation, with the threshold value set to the averaged image intensity (between 0 and 1) plus 0.005. According to our experiments, this threshold value achieves the best result in separating the embryo from the background. Based on the extracted embryo segment, a minimum bounding ellipse is calculated based on which a registered ellipse mask is created and stored together with the normalized image (e.g., the second image in Fig. 5).

Ideally, the significant embryonic areas should be located in the same place in the registered images. However, the shape of the embryo is not exactly the same. In this study, we propose to determine the significant embryonic areas adaptively with the following approach.

![Figure 4. Four significant embryonic areas](image_url)
To retrieve the dorsal area, we start from the top middle pixel in the minimum bounding ellipse of an embryo segment, and search downward in the ellipse mask image until we find the first pixel in the actual embryonic segment. That pixel is then used as a reference point to extract a dorsal area of size 128×128 (from the registered image) in which that reference point is the top middle pixel of the sub-area (Fig. 3).

The posterior and the anterior areas are extracted in a similar way, starting from two pixels with fixed locations \((x:1, y:1112)\) for posterior area and \((x:1, y:90)\) for anterior area in the registered/normalized image, where the upper-left corner of the image is the origin in the image coordinate system; \(x\) and \(y\) indicate the row index and column index, respectively. Based on these two pixels, two reference points can be retrieved by searching down the ellipse mask until a true embryonic pixel is reached. The two retrieved reference points are then used as the top-right corner and the top-left corner to retrieve the posterior and the anterior areas, respectively. All retrieved significant areas have the same size of 128×128.

As we mentioned above, one major distinct texture feature in stage 6 images is the ventral furrow located in the ventral area of an embryo. However, the ventral area cannot be simply retrieved by the above method. The main reason is that images from stage 6 have ventral furrows with different sizes and shapes due to the different time points at which the images are obtained. Through visual inspections, we locate a relatively stable sub-area in the ventral whose top-left corner is located around the location \((330, 540)\) in the registered embryo image. Our experiments demonstrate that the extracted ventral area provides a reasonably good representation of the ventral furrows in stage 6. Again, all the four extracted significant embryonic areas are of the size 128×128. The basic process of extracting significant embryonic areas is illustrated in Fig. 5.

### B. Feature Extraction

As aforementioned, since an embryo gradually differentiates into various kinds of tissues and organs, producing unique textures, embryonic images from the same developmental stage usually share similar texture patterns (e.g., the morphological formation of the yolk in the middle of the embryo in stage 3 and the cytoplasmic rim in stage 4 as shown in Fig. 4). Therefore, these unique patterns can be used as distinguishing features for staging embryo development.

However, these patterns often vary in their shapes, orientations, and sizes, and cannot be easily extracted through some conventional image feature extraction methods such as histogram distribution. In order to extract these unique patterns as texture features, we apply Gabor filters in the proposed feature extraction and classification system. The details of Gabor filters are briefly described as follows for the sake of completeness.

1) **Gabor Filters**: Gabor filters, a well-known and commonly used texture feature extraction method, are very effective in extracting the characteristics of spatial frequencies and orientations in textures [13]. Given the rich texture information present in ISH images and their implications for staging embryo development, we believe that texture features extracted via Gabor filters can be a good representation of the visual features in raw ISH images.

The Gabor function is obtained by modulating a sinusoid with a Gaussian envelope. We demonstrate a two-dimensional Gabor filters over the image domain \((x, y)\) as the function \(g(x, y)\) and the Fourier transform \(F(u, v)\) in Equations (1) and (2), respectively [13].

\[
g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[ -\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right) \right] \exp\left[ 2\pi j W_x x + 2\pi j W_y y \right] \tag{1}
\]

\[
F(u, v) = \exp\left[ -\frac{1}{2} \left( \frac{(u-W)^2}{\sigma_u^2} \right) + \frac{v^2}{\sigma_v^2} \right], \quad \begin{align*}
\sigma_u &= \frac{1}{2} \pi \sigma_x \\
\sigma_v &= \frac{1}{2} \pi \sigma_y
\end{align*} \tag{2}
\]

We can obtain a certain Gabor filter with a given scale and orientation by simply dilating and rotating the generic Gabor function \(g(x, y)\). The complete Gabor filter family in
all scales and orientations can be obtained through the following generating function:

\[
g_{mn}(x, y) = a^{-m}g(x', y'), \quad a > 1, m, n = \text{int} \tag{3}
\]

where \( l \) is the total number of orientations, \( n \) is the number of orientations. The scale factor \( a^m \) in Equation (3) is meant to ensure that the energy is independent of \( m \) (the \# of scales). This collection of functions forms a non-orthogonal basis of functions for the multi-resolution decomposition [13].

2) Feature Representation: Let \( I(x, y) \) be a given image and \( g_{mn}(x, y) \) be the filter obtained from the basic Gabor function. The filter convoluted image \( J \) is defined in Equation (4).

\[
J_{mn}(x, y) = \int I(x, y)g_{mn}^*(x-x_1, y-y_1)dx_1dy_1 \tag{4}
\]

where \( * \) indicates the complex conjugate. We represent the textures features as the mean \( \mu_{mn} \) and the standard deviation \( \sigma_{mn} \) of the magnitude of the transform coefficients as formalized in Equations (5) and (6).

\[
\mu_{mn} = \iint |J_{mn}(x, y)| \, dx \, dy \tag{5}
\]

\[
\sigma_{mn} = \sqrt{\iint (|J_{mn}(x, y)| - \mu_{mn})^2 \, dx \, dy} \tag{6}
\]

In this study, we generate a series of Gabor filters with 12 scales and 16 orientations as the texture features, producing a feature vector \( \pi \) (shown in Equation (7)) of length 384 for each significant embryonic area.

\[
\pi = [\mu_1, \sigma_1, \mu_2, \sigma_2, \ldots, \mu_{192}, \sigma_{192}] \tag{7}
\]

3) Principal Component Analysis: As described in the last section (II.B.2), Gabor filters perform analysis in different scales and in various orientations, producing feature vectors with high dimensionality. The analysis on high dimensional feature space usually suffers from excessive noise and is not efficient. Therefore, we introduce principal component analysis to make the feature space more compact while retaining the most distinguishing features. Through the use of principal component analysis, we can also reduce data dimensions and eliminate some of the noise.

In this study, for each image, we apply Gabor filters with 12 scales and 16 orientations on the three selected significant embryonic areas, obtaining a feature vector of \( 384 \times N \) dimensions (\( N \) is the number of significant areas used). Principal component analysis is then applied on this high dimensional data, which projects the data to a lower dimensional space from the most informative viewpoints. We select the top 100 ranked features produced from principal component analysis as the most significant features for the subsequent developmental stage classification. By using principal component analysis, we alleviate the problem of the curse of dimensionality, reduce the noise from the high dimensional data, and keep the most informative features.

C. Embryo Image Classification

In the proposed classification system, we apply a multi-class Support Vector Machine (SVM) classifier as the supervised machine learning mechanism in order to differentiate the developmental stages of Drosophila embryos on the basis of the extracted texture features.

Support Vector Machines (SVMs), originally designed for binary classification and regression, are a set of supervised machine learning methods which classify data by transforming training examples into a higher dimensional space and finding a linear separating hyper-plane with the maximal margin in the space for partitioning dataset into corresponding groups. In addition, a SVM classifier can be used to group linear and non-linear dataset by using various kinds of kernel functions. For instance, linear kernel is commonly used in classifying linear dataset while the radial basis function (RBF) kernel and polynomial kernel are widely used in grouping non-linear dataset. Moreover, as aforementioned, SVMs were originally designed for binary classification. However, in recent years, SVM classifiers have been extended and applied to multi-class classification problems.

In this study, we adopt a multi-class SVM as the classifier in the proposed system for determining the developmental stage of Drosophila embryos from Drosophila embryonic in situ hybridization images. The reason we choose multi-class SVM in the proposed system is that the use of multi-class SVM enables the system to classify all embryonic images with a single classifier, and thus, avoids conflicting class labels which is usually the case when multiple binary classifiers are used to predict class labels in a multi-class dataset.

In practice, we use OSUSVM, a commonly used SVM toolbox for the MATLAB environment, as the multi-class SVM classifier in the proposed system. The OSUSVM toolbox adopts LIBSVM package [14, 15] as its core, and provides three different kernels, including linear kernel, RBF kernel, and polynomial kernel. In this study, we test the proposed classification system with both linear and non-linear kernels since we have no prior knowledge of the underlying data distribution. Specifically, we apply linear
kernel and polynomial kernel multi-class SVMs in classifying developmental stages.

As a supervised machine learning algorithm, SVM requires training images and their texture features from each developmental stage with their ground-truth class labeled. To construct a SVM classifier, the proposed system learns the essential parameters which can be used for creating the optimal separating hyper-planes from the training examples. The trained SVM classifier is then able to partition the testing examples with the separating hyper-planes aforementioned where each partition corresponds to a developmental stage. In this way, a testing Drosophila embryonic in situ image receives its class label from the partition where it resides.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

The 1475 Drosophila embryonic images used in this study were procured from the Berkeley Drosophila Genome Project (BDGP) [1]. Four different early developmental stages, including stage 3 (487 images), 4 (485 images), 5 (257 images), and 6 (246 images), are included in our experimental dataset. We do not include any embryonic image from stages 1 or 2 due to no significant tissue differentiation occurs at these stages.

We evaluate the proposed system with the repeated random sub-sampling validation which arbitrarily splits dataset into training and testing datasets (with a training-testing ratio of 2:1 in this study). In order to reduce bias, we perform 60 iterations of the random sub-sampling validation and obtain the overall performance by averaging all the classification accuracy values.

A. Combination of Significant Embryonic Areas

As aforementioned, we identify four significant embryonic areas in embryo with peculiar traits for differentiating developmental stage. In order to evaluate the effectiveness of these significant embryonic areas, we test the proposed system with 16 different combinations, including the whole embryo and the combinations of the four significant embryonic areas extracted from the dorsal, ventral, posterior, and anterior (as shown in Fig. 3).

Table I presents the classification accuracies for the whole embryo and the two best significant embryonic area combinations. The combination A includes the anterior, dorsal, and ventral areas, while the combination B comprises the dorsal, ventral, and posterior areas.

From

TABLE , we can see that the classifiers that use features extracted from significant areas outperform the classifier that use global features (‘Whole embryo’ in Table I). We can also observe that the SVM classifier with linear kernel significantly outperforms the SVM classifier with non-linear polynomial kernel. This indicates that the Drosophila embryonic in situ hybridization images might be linearly distributed due to the temporal relation of these images. In addition, the experimental results show that the significant embryonic area combination B produces the highest classification accuracy among all combinations.

TABLE II shows the confusion matrix which presents the classification results produced from the significant embryonic area combination B. In

TABLE II, each row corresponds to a predicted class and each column corresponds to an actual class. The number in the $i$th row and the $j$th column is the percentage of images in class $j$ that is classified as class $i$. From

TABLE II, we can observe that all the values along the main diagonal are greater than 80%. In addition,

TABLE II shows that the classification of stage 5 and stage 6 images results in lower accuracies than that of the stage 3 and stage 4 since some images in stage 5 and stage 6 visually resemble each other.

The above results indicate that more than 80% of Drosophila embryonic in situ hybridization images can be correctly classified with the use of the texture features extracted from dorsal, ventral, and posterior areas. Therefore, in the rest of the experiments, we compare the performance based on the texture features extracted from the significant embryonic area combination B, i.e., the dorsal, ventral, and posterior areas.

B. Performance Comparisons

Table III compares the proposed system with the Regularized Uncorrelated Linear Discrimination Analysis (RULDA) which is an existing data reduction approach for Drosophila embryonic developmental stage classification reported in [8, 16]. In this experiment, we replace PCA in our proposed system with RULDA for dimension reduction.

We can observe from Table III that the propose system adopting SVM classifier with the linear kernel and PCA achieves the best classification accuracy of 93% with a standard deviation of 0.93%, while RULDA-based classifier can only achieve 65% (standard deviation: 2.49%) of classification accuracy. In addition, when using the polynomial kernel, the proposed method achieves 89% (std:

| TABLE I. COMPARISON OF SIGNIFICANT EMBRYONIC AREA COMBINATIONS |
|-------------------|-------------------|-------------------|
|                   | Whole embryo | Combination A | Combination B |
| Linear            | 85.42%       | 86.32%       | 93.27%       |
| Polynomial        | 81.58%       | 85.18%       | 89.10%       |

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Ground-truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>99%</td>
</tr>
<tr>
<td>4</td>
<td>1%</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
</tr>
</tbody>
</table>
0.98%) accuracy while the the classification accuracy of RULDA is 81% (std: 1.78%). The proposed system outperforms RULDA-based classification system in both linear and polynomial kernel. Moreover, from the experimental results, we can observe that RULDA performs better in classifying stage 3 images than it does for images from stages 4-6. Aside from that, in the framework proposed in [8], the authors consider the stages 4, 5 and 6 as a single class, while we further refine the stage classification in this study.

**IV. CONCLUSIONS**

In this paper, we propose a supervised classification system to determine the developmental stage via image analysis in an automatic manner. The proposed system comprises three main steps: (1) image pre-processing, (2) Gabor texture feature extraction with principal component analysis, and (3) embryo images classification with multi-class support vector machine. More specifically, in the image pre-processing step, we normalize all embryo images, and register those normalized images in our database. In the second step, we apply Gabor filters to extract texture features from selected significant embryonic areas, and then, perform principal component analysis to reduce the data dimension and possible noise. In the third step, a multi-class SVM classifier learns the essential parameters from training examples, and then, predicts the class label as the developmental stage of unknown testing examples.

We test the proposed system on a dataset containing 1475 images from the early developmental stages, i.e., stage 3 to stage 6. The experimental results show that the proposed system successfully classifies *Drosophila* embryonic *in situ* hybridization images and outperforms the RULDA-based method reported in [8].

In our future work, we will expand our dataset and include more *Drosophila* embryonic *in situ* hybridization images from other developmental stages. Once we can identify the developmental stages for *Drosophila* embryonic images, we will further recognize and locate various kinds of tissues in the embryo according to the stage information. As a long term goal, we plan to find the relationship between tissue development and the gene expression patterns which will help us to better understand how gene regulates embryo development.

**ACKNOWLEDGMENT**

The work of Chengcui Zhang was supported in part by the UAB ADVANCE program and NSF DBI-0649894.

**REFERENCES**


