# Classification and Indexing of Gene Expression Images

Karthik Jayaraman[α], Sethuraman Panchanathan[β], *Fellow SPIE*, Sudhir Kumar[γ]

[α] Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287,USA
[β] Department of Computer science, Arizona State University, Tempe, AZ 85287,USA
[γ]Department of Biology, Arizona State University, Tempe, AZ 85287,USA
Email: {s.kumar, panch}@asu.edu

**ABSTRACT**

In this paper, we present an approach for classification and indexing of embryonic gene expression pattern images using shape descriptors for retrieval of data in the biological domain. For this purpose, the image is first subjected to a registration process that involves edge fitting and size-standardization. It is followed by segmentation in order to delineate the expression pattern from the cellular background. The moment invariants for the segmented pattern are computed. Image dissimilarity between images is computed based on these moment invariants for each image pair. Area and Centroids of the segmented expression shapes are used to neutralize the invariant behavior of moment invariants during image retrieval. Details of the proposed approach along with analysis of a pilot dataset are presented in this paper.

Keywords: Gene expression image database, Shape features, Indexing, Content-based retrieval, Classification, Pattern recognition.

## 1. INTRODUCTION

Images are being generated at a rapid pace in many different areas of biological research. This requires concurrent development of efficient storage and retrieval systems in order to make use of these data effectively. In this paper we have focused on the area of biological research, which is concerned with elucidating interaction among genes in the early stages of the fruit fly development from a single cell to an adult. Biologists usually stain the developing embryo using gene specific probes in order to visualize the domain of gene expression. Any area of the embryo where the gene expression is turned on (or the gene product is localized) can be easily distinguished from areas where there is no expression. Figure 1a shows an image of the fruit fly embryo and Figure 1b shows an image in which the expression of gene is clearly visible when using a blue stain. Currently, scientists compare gene expression pattern images by visual inspection as there exists no computational framework to construct a visual-query to retrieve the best match image.

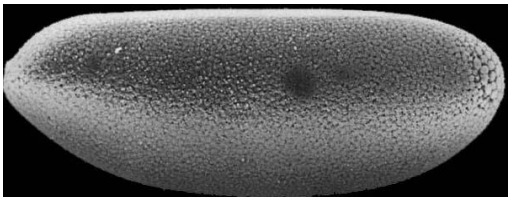

**Figure 1a:** Fruit fly embryo.

**Figure 1b:** Gene expression image.

Biologists require the retrieval of gene expression images to be based on shape similarity and amount and location of overlap. Color based indexing is not suitable as different patterns are stained using different colors. The conventional methodology of shape-based indexing employs global shape descriptors like Moment invariants [1], Fourier descriptors [2], etc. These approaches are typically size, rotation, and translation invariant, which lead to their failure in completely representing the amount and location of overlap. Therefore existing retrieval models are not sufficient for effective and efficient gene expression image retrieval. In this paper, we describe an indexing scheme based on the new set of shape descriptors, which are more suitable for the analysis of gene expression pattern embryo

Applications of Digital Image Processing XXIV, Andrew G. Tescher, Editor,
Proceedings of SPIE Vol. 4472 (2001) © 2001 SPIE · 0277-786X/01/$15.00

images. We begin with a detailed description of the image size standardization and segmentation procedure that has been mentioned in [11], and then discuss the classification and indexing scheme (section 2). The retrieval results from a prototype dataset are discussed in section 3 followed by the conclusions (section 4).

# 2. Classification and Indexing process

A typical Indexing and retrieval system consists of several modules namely data acquisition and standardization, segmentation, feature extraction, pattern recognition and query processing. Individual modules are discussed in detail in the following sections. The block diagram for the Indexing and retrieval system is shown in figure 2.

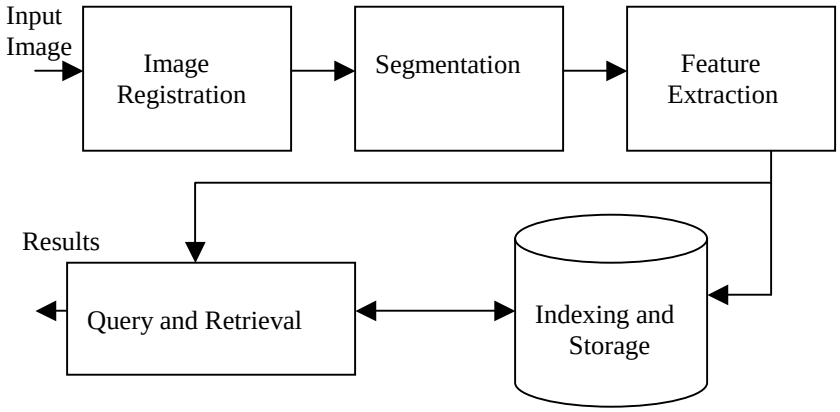


**Figure 2:** Block Diagram for the Indexing and retrieval system.

### *2.1. Data acquisition and standardization*

Gene expression pattern images are published in varying locations in differing formats by the biologists. We concentrated on the research articles available in the online scientific articles in order to construct a pilot data set. Images are often acquired under different orientations and are all from different developmental stages. Thus, there is a need to standardize the image prior to conducting any analysis. Image registration is a standard technique used for this purpose.

In general, the image registration problem can be stated as the determination of the suitable mapping (T) function in order to transform the input image to spatially correspond to reference image [6]. In a simple case, the image can be registered by combination of translation, rotation and scaling processes. In complex scenarios, the transformation can be non linear. Performing a non-linear registration operation can be computationally expensive. An optimal solution can be achieved by approximating mapping (T) function. Certain parameters have to be estimated to construct an approximate mapping (T) Function. These parameters can be determined by attempting to minimize the displacement between the images using certain feature points identifiable in both the images. These feature points are referred to as control points. The control points of the input image are mapped to the control points of the reference image by using mapping (T) function. All other points are mapped to lie between the control points.

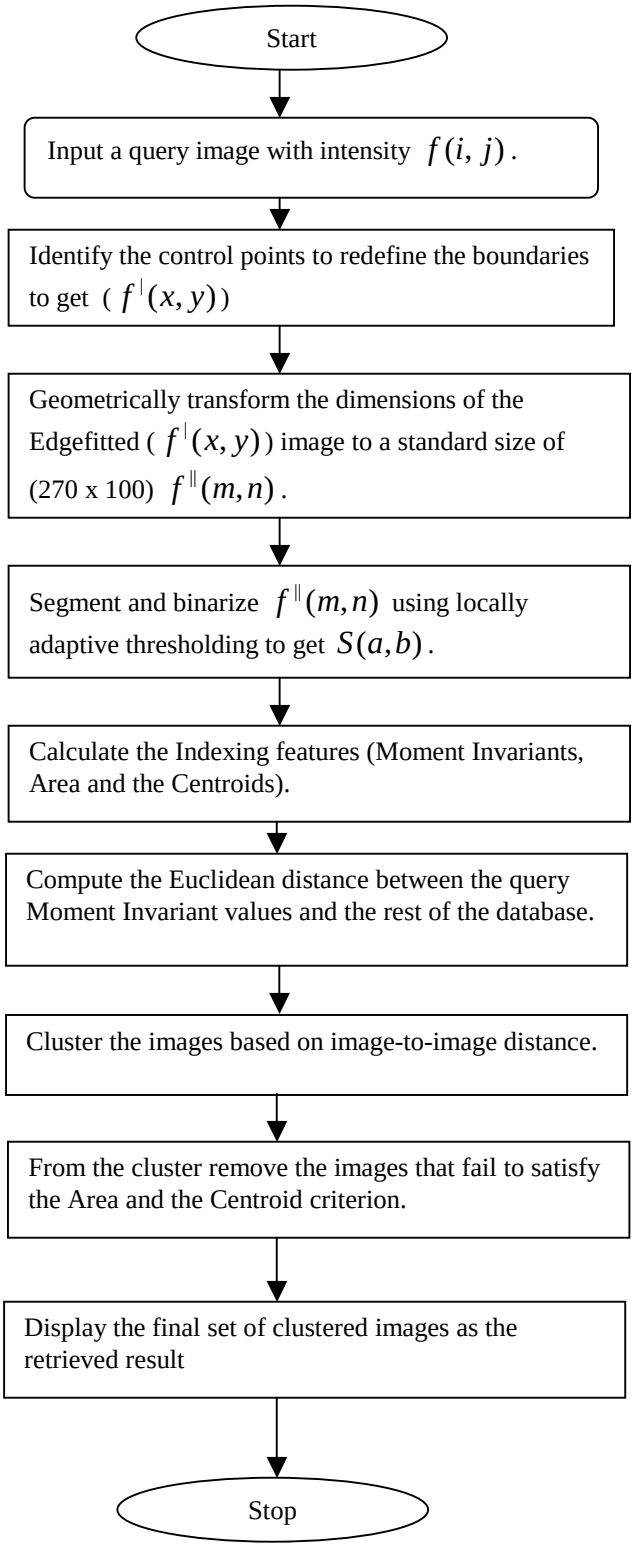


**Figure 3:** Flow chart Diagram for the Proposed indexing and retrieval algorithm.

Let $f(i, j)$ and $g(x, y)$ be the input and reference images, respectively. The set of control points C = ($i$, $j$, $x$, $y$). The point ($i$, $j$) in $f$ corresponds to ($x$, $y$) in $g$. The mapping function $T$ geometrically registers $f$ with $g$ if

$$T(i, j) = (x, y) \text{ for every point in set C.} \quad (1)$$

$$f^{\parallel}(x, y) = f(T(x, y)) \quad (2)$$

The image $f^{\parallel}$ defined in equation (2) is called the registered image of $f$ with respect to $g$. The points $T(x, y)$ can be non-integers. Therefore interpolation is used to estimate the value of the $f$ using the neighboring points. This process is referred to as resampling. Therefore, the image registration process includes identification of control points, estimation of mapping functions, and resampling of the images.

The existing image registration approaches [7,8] in the biology domain are specific to certain type of images namely CRT, MRI, etc. Kumar et al. have presented the outlines of first image registration algorithms specifically designed for embryonic gene expression images in [11]. Here, a new image registration algorithm called "Edge-fitting" is used. This algorithm assumes that the images were acquired under uniform illumination conditions. The Edge-fitting process is divided into three steps: (1) determine the control points from the boundaries of the embryo, (2) map these control points to form the new image boundaries, and (3) resize the Edge fitted image to a standard size (we choose 270 (Width) x 100 (height)). Figures 4a and 4b illustrate the Edge fitting concept.

## 2.1.1. Edge Fitting

### 2.1.1.1. Determine control Points

The control points are determined using the following algorithm:

1. Take 5 by 5 pixel samples of the image from the four corners of the image.
2. Calculate the mean ($\mu$) and the standard deviation ($\sigma$) for this sample of 100 pixels.
3. For each row, traverse from the left boundary of the image and find the first point ($x$, $y$), where the following equation is satisfied

   $$f(x, y) - \mu \geq \xi\sigma \,, \quad (3)$$

   where $2 < \xi < 4$. In general, a value of 3 was found to produce good results. The ($x$, $y$) point is noted for all the rows of the image. The smallest value of the $y$ is chosen as the "Left" control point.

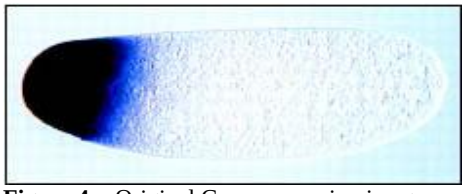

**Figure 4a:** Original Gene expression image.

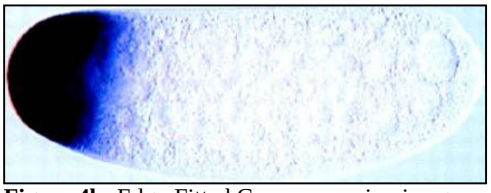

**Figure 4b:** Edge-Fitted Gene expression image.

4. Similarly, for each row traverse from right to left boundary of the original image and find the first point ($x$, $y$) that satisfies equation (1). This point is termed as the "Right" control point.
5. For each column, traverse from top to bottom and find the first point ($x$, $y$) that satisfies equation (1). This point is termed as the "Top" control point.
6. Similarly for each column traverse from bottom to top and find the first point ($x$, $y$) that satisfies equation (1). This point is termed as the "Bottom" control point.

The four control points are taken as the image boundaries of the gene expression image to derive $f^{|}(x, y)$. Image $f^{|}(x, y)$ is geometrically transformed to a standard image size of 270x100 to obtain $f^{||}(x, y)$. This size was chosen based on the average size of the images available in the scientific journals. This helps in defining the position of a gene expression pattern with respect to the edges of the embryo. The other advantage is that images of the same size are easier to handle. Geometric transformation is the process that alters the size of the image to the required dimensions. It consists of two basic operations (1) spatial transformations, which defines the "rearrangement" of pixels on the image plane, and (2) gray level interpolation which deals with the assignment of gray levels to pixels in the spatially transformed image.

### 2.1.1.2.Spatial transformations

Let the original image of height X and width Y be transformed spatially to some new dimensions say A and B. The transformations can be expressed as

$$A = R(x, y) \tag{4}$$

$$B = S(x, y) \tag{5}$$

Where $R(x, y)$ and $S(x, y)$ represent the spatial transformations that result in the new image.

We can solve for $R(x, y)$ and $S(x, y)$ using the following equations:

$$R(x, y) = C_1 x + C_2 y + C_3 xy + C_4 \tag{6}$$

$$S(x, y) = C_5 x + C_6 y + C_7 xy + C_8 \tag{7}$$

The unknowns $C_1$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$, $C_7$ and $C_8$ define the transformation of the pixels coordinates.
We use four tie points to solve for the eight unknowns as shown below (figure 5)

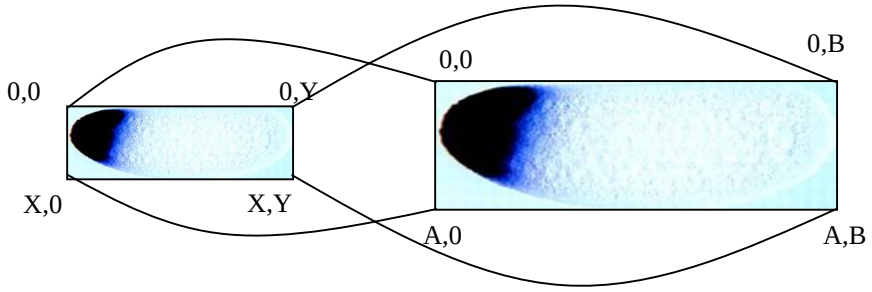


**Figure 5:** Tie points for spatial transformation.

### 2.1.1.3.Gray level interpolation

The method used for gray level interpolation is known as bilinear interpolation [3]. It is a first order interpolation. The pixel coordinates of the transformed image are mapped back into the original image. The points that are mapped back usually get placed in between pixel positions in the original image. Then the bilinear transformation determines the pixel color value for the transformed image pixel based on four neighbor pixels.

## *2.2. Segmentation*

Image Segmentation subdivides the image into its corresponding foreground and background objects. The level at which this subdivision is done depends on the problem being solved. In our case, we are interested in separating the regions of expression (referred to as expression objects), which are the stained regions of gene activity from the embryo background. Segmented expression objects represent the pattern of gene expression in the embryo. These stained regions have the information regarding the location, intensity and the activity of the gene.

### 2.2.1. Segmentation by Locally Adaptive Automatic Thresholding

Thresholding is one of the most basic methods of extracting the foreground and background objects. Thresholding is fast and is effective in situations where objects are clearly distinct from their backgrounds. Hence, their pixel values are different and form different groups. Thus we can select a threshold to extract the objects from the background. Conventional histogram based or global methods are deficient in detecting small targets of possibly low contrast [4]. The digitization of such images in image space might have a poor target /background contrast and unwanted background clutters making extraction difficult [4]. In order to solve the above problems we employ automatic object extraction by local adaptive image thresholding scheme [4]. The result of the automatic segmentation (in binary format) for image in figure 1b is shown in figure 6.

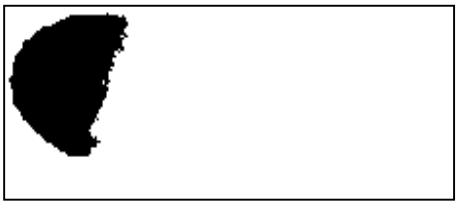

**Figure 6:** Segmented and binarised Gene expression image.

### 2.2.2. Extraction by User Guided Region Growing

Region growing is a procedure that groups pixels or sub regions into larger regions. The process starts with a set of seed points from which we grow regions by appending to each seed point those neighboring pixels that have similar properties [2]. This process is called as pixel aggregation. The seed points are usually obtained by taking random pixel clusters in the image. This results in multiple sub-regions that have similar pixel properties. These sub-regions are combined or collapsed to form single regions, which results in an increased computational complexity. We propose a user guided region growing scheme, which eliminates the need to collapse, or combine regions. In this scheme, the user can pick the seeds contained in the Region of interest (ROI). The user selects two sets of seed points representing the foreground and background properties.


**Figure 7a:** Original image.


**Figure 7b:** Segmented region using user picked seeds.

This helps in obtaining better seed (i.e. with different pixel properties) points and thus eliminates the need to combine the sub-regions. The other advantage is that the user has the control over the regions that are extracted. These regions can be used as the basis for retrieval. We have employed a seeded region growing algorithm described in [10]. The seed values chosen by the user are employed for growing the regions. We note that other user guided segmentation technique proposed by Perry *et al.* [9] employ the user input to merge the regions after region growing and is not suitable for our purpose. The image in figure 7a has two different genes expressed namely *sg* and *yo*. The extraction of gene expression of a single gene is achieved through user guided region growing (figure 7b).

## *2.3. Pattern Descriptors*

Feature extraction facilitates compact representation of an image in terms of its contents. These features facilitate pattern-based recognition. As mentioned earlier the retrieval of gene expression images must be based on shape similarity and amount and location of overlap. Moments Invariants are the most commonly used shape descriptors. They are computationally inexpensive and are invariant to changes in scale, translation and rotation of the shapes. Area and Centroids of the shapes are used to neutralize the invariant behavior of Moment invariants. The proposed indexing model uses Moment invariants, Area and Centroid features extracted from the segmented images.

### 2.3.1. Moment Invariants

Moment invariants can be obtained from second and third order moments of any shape [1,3]. There are seven invariant parameters that can be used to calculate the distances between two different shapes.

Moments of order (p+q) are defined as

$$M_{pq} = \sum_i \sum_j x_p y_q f(x, y) \tag{8}$$

Where x and y are the coordinates of the image and f(x,y) is the image intensity value. p,q = 0,1,2…
The central moments are expressed as

$$\mu_{pq} = \sum_i \sum_j (x - x_m)(y - y_m) \tag{9}$$

Where p, q = 0,1,2 and

$$x_m = m_{10} / m_{00} \tag{10}$$

$$y_m = m_{01} / m_{00} \tag{11}$$

The seven sets of moments [1,2] can be calculated from $\mu_{pq}$ by using the following equations. The calculated moment invariant values ($\phi_1 \ldots \phi_7$) are stored in a database.

$$\phi_1 = \mu_{20} + \mu_{02} \tag{12}$$

$$\phi_2 = (\mu_{20} - \mu_{02})^2 + 4(\mu_{11})^2 \tag{13}$$

$$\phi_3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \tag{14}$$

$$\phi_4 = (\mu_{30} - \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \tag{15}$$

$$\phi_5 = (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \tag{16}$$

$$\phi_6 = (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \tag{17}$$

$$\phi_7 = (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + (3\mu_{12} - \mu_{30})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \quad (18)$$

### 2.3.2. Area of the shape

All the images have a standard size after image registration. Therefore, expression pattern area can be used to compute the amount of overlap of the gene expression shapes. The area of the expression pattern is obtained from the $m_{00}$. Therefore, there is no additional computation required for area calculation. Area measure helps us to refine the results by eliminating similar shapes with different scaled sizes.

### 2.3.3. Centroid of the shape

$x_m$ and $y_m$ are the Centroids of the shape. The Centroids help in eliminating similar shaped expression patterns that are not present in the same region as the query expression pattern (Figure 8a and 8b).

**Figure 8a:** Expression pattern on the left corner.

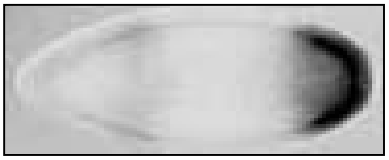

**Figure 8b:** Expression pattern on the right corner.

## *2.4. Retrieval system*

An Image-image Distance (ID) was calculated between the query image(i) and database images(j) using equation (19). An initial set of images was retrieved based on low ID scores. Images from the initial set that satisfied the equation (20) and (21) were retained. Area (i) and Area (j) denote the areas of the gene expression shape. ( $X_i$ , $Y_i$ ) and ( $X_j$ , $Y_j$ ) denote the Centroid points of the query and database images. The images retained in the initial set are displayed as retrieved results.

$$ID = (\phi_{1i} - \phi_{1j})^2 + (\phi_{2i} - \phi_{2j})^2 \quad (19)$$

$$(Area(i) - Area(j)) / Area(i) < 0.1 \quad (20)$$

$$\sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2} < 50 \quad (21)$$

# 3. Results

Experiments were carried out to prove the effectiveness of the proposed classification and indexing approach. The database consists of 97 gene expression images containing simple and complex shapes. The image shown in figure 9 was used as query image for a simple shape. The results obtained using Moment invariant indexing is shown in figure10. The retrieved results for our method are shown in figure11. It is obvious that our method retrieved images that matched

in shape and overlap. Image shown in figure 12 was used as a query image for a complex or occluded shape. The results obtained using Moment invariant indexing and our method are shown in figures 13 and 14 respectively. The results show that in case of complex shapes the indexing method retrieves shapes that are rotated but similar. This occurs, as the centroids of those images are similar to that of the query image. This is still significant to a biologist as the amount of overlap is significant.

**Figure 9:** Simple shaped query image.

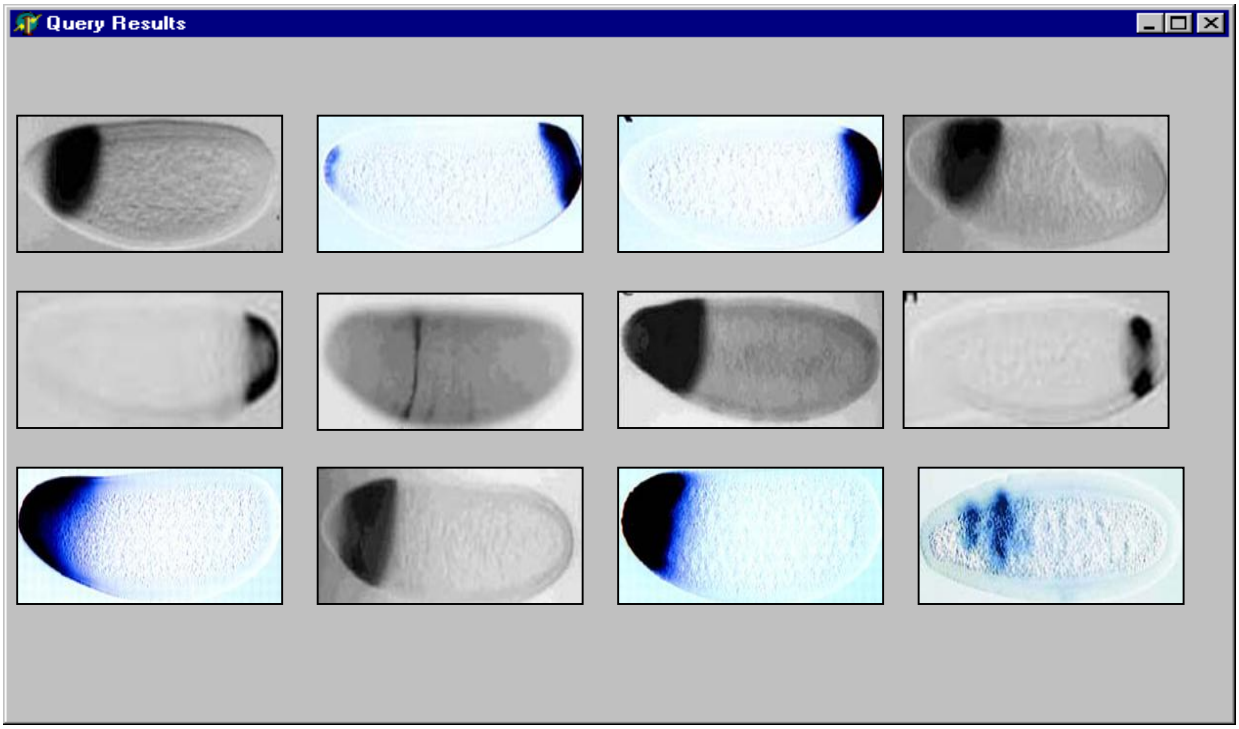


**Figure 10:** Retrieval results based on Moment invariant indexing for figure 9.

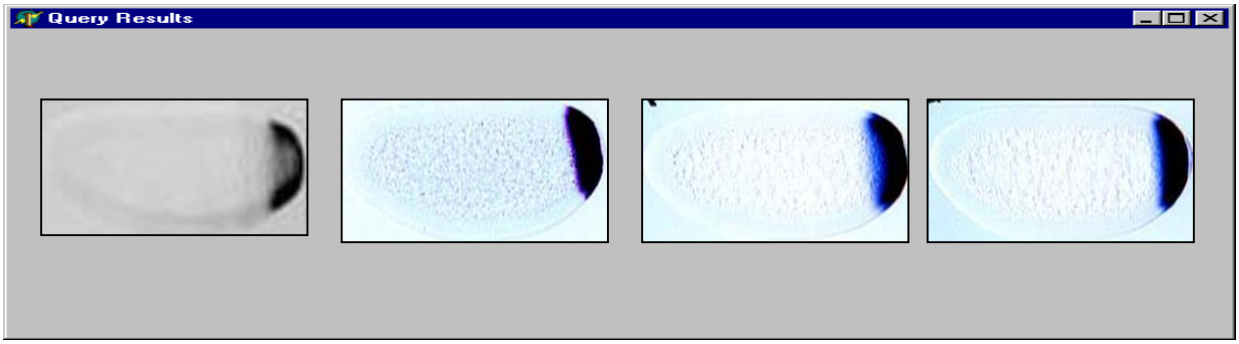


**Figure 11:** Retrieval results for our set of descriptors for figure 9.

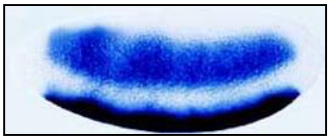

**Figure 12:** Complex shaped query image.

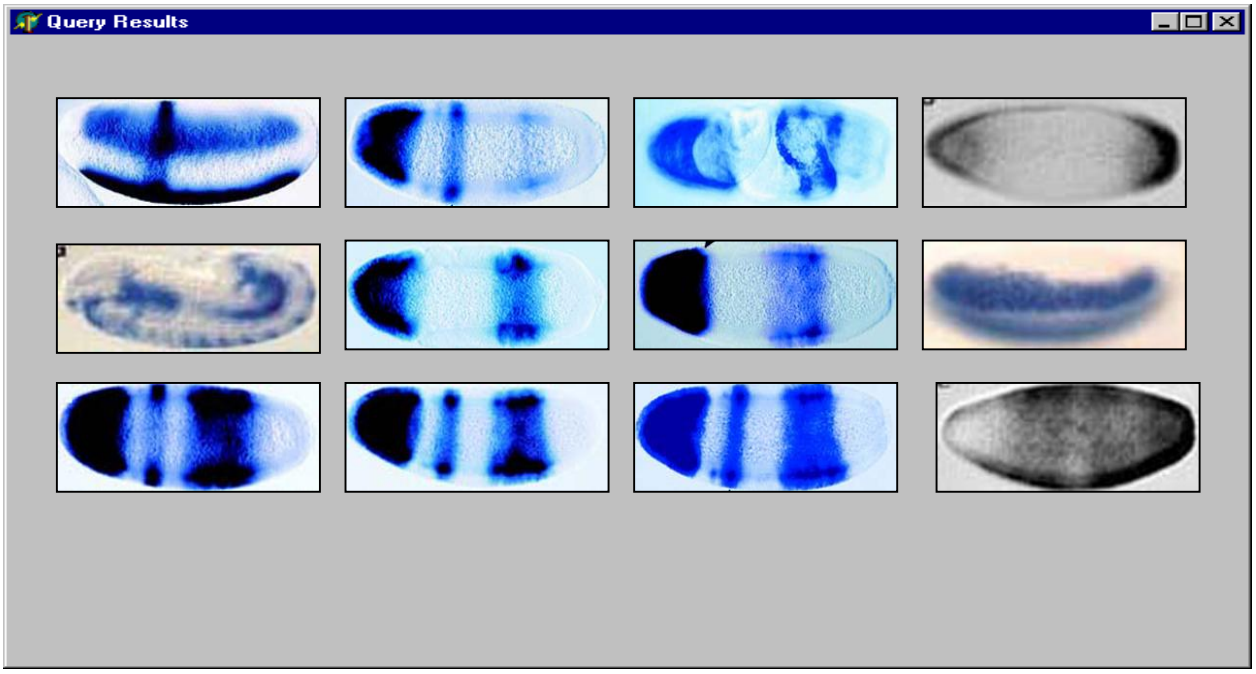


**Figure 13:** Retrieval results based on Moment invariant indexing for figure 12.

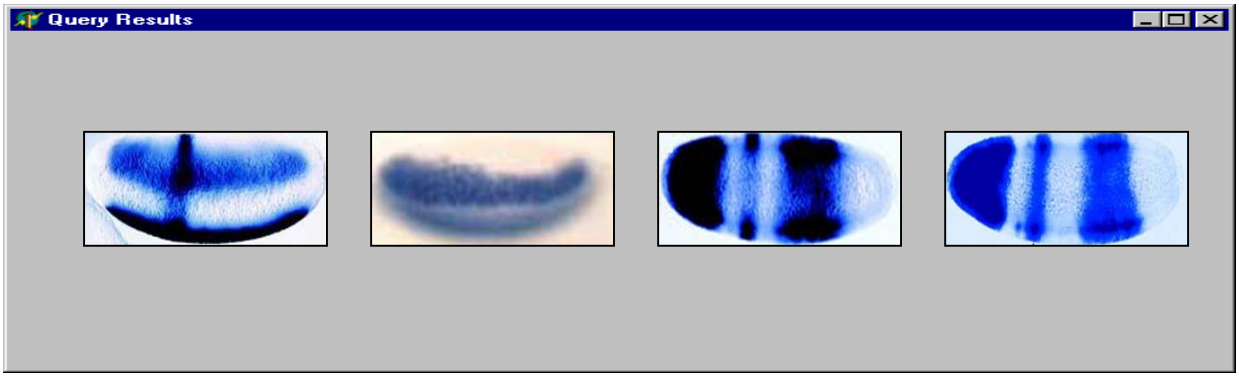


**Figure 14:** Retrieval results for our set of descriptors for figure 12.

# 4.Conclusions

We have presented a classification and indexing approach based on multiple descriptors for the retrieval of gene expression images that have been transformed to a standard size of 270x100. A user guided region growing algorithm that gives the user greater control over segmentation process and enables segmentation of single gene expression pattern in a multiple gene expression image has been presented. The indexing algorithm effectively utilizes a set of shape descriptors (Area and Centroid) enabling biologically meaningful retrieval. Results show that the proposed indexing approach has a superior retrieval performance in contrast to the moment invariant indexing method.

# References


1. M.K. Hu, *Visual pattern recognition by Moment Invariants*, IRE, 179-187, 1962.
2. R.Gonzalez and R.Woods, *Digital Image processing*, Addison -Wesley, Reading MA, 1993.
3. K.R. Castleman, *Digital Image Processing*, Prentice-Hall Inc., New Jersey, 1979.
4. L wen-Nung, Automatic Target Segmentation by Locally Adaptive Image Thresholding, IEEE transactions on Image processing, Vol. 4,No 7, 1995.
5. L.F. Costa and R.M. Cesar Jr., *Shape Analysis and Classification theory and practice*, CRC Press, Florida, 2001.
6. J.Owczarczyk. W.J.Welsh and S.Searby, *Performance Analysis of Image registration techniques*, Image Processing and its Applications, 1989, Third International Conference on, 1989,Page(s): 10 –13.
7. M. Moshfeghi, *Multimodality Image registration Techniques in Medicine*, Engineering in Medicine and Biology Society, 1989. Images of the Twenty-First Century, Proceedings of the Annual International Conference of the IEEE Engineering in, 1989,Page(s): 2007 -2008 vol.6.
8. G.P. Penney, J Weese,J A little, P Desmedt, D.L.G.Hill and D. Hawkes, *A comparison of similarity measures for use in 2D-3D medical Image registration*, IEEE transactions on Medical Imaging, Vol. 17,No. 4 ,August 1998.
9. S.T.Perry and P.H.lewis, A Novel image viewer providing fast object delineation for content based retrieval and navigation, http://www.mmrg.ecs.soton.ac.uk /publications/archive/perry1998/html/
10. R. Adams and L. Bischof, *Seeded Region Growing*, IEEE transactions on PAMI, Vol. 16,No. 6, June 1994.
11. S.Kumar, S. J. Newfeld, A. Marti-Subirana, K. Jayaraman, and S. Panchanathan, *Analysis of embryonic gene expression patterns in silico*. (Submitted), 2001.