Published in final edited form as:
Proc Int Conf Image Proc. 2008 December 12; 1(1): 721–724. doi:10.1109/ICIP.2008.4711856.

PROCESS FLOW FOR CLASSIFICATION AND CLUSTERING OF FRUIT FLY GENE EXPRESSION PATTERNS

Andreas Heffel[1,4], **Peter F. Stadler**[1], **Sonja J. Prohaska**[2,3], **Gerhard Kauer**[4], and **Jens-Peer Kuska**[*,1]

[1]Interdisciplinary Centre for Bioinformatics, University of Leipzig, Härtelstraβe 16–18, 04107 Leipzig

[2]Department for Biomedical Informatics, School of Computing and Informatics, Arizona State Univeristy, 425 N. 5th Street, Phoenix, AZ 85004

[3]Center for Evolutionary Functional Genomics, The Biodesign Institute, PO Box 875001, Tempe, AZ 85287

[4]Faculty Technical Sciences, University of Applied Sciences O/O/W, Constantiaplatz 4, 26723 Emden

Abstract

The rapidly growing collection of fruit fly embryo images makes automated Image Segmentation and classification an indispensable requirement for a large-scale analysis of *in situ hybridization* (ISH) – *gene expression patterns* (GEP). We present here such an automated process flow for Segmenting, Classification, and Clustering large-scale sets of *Drosophila melanogaster* GEP that is capable of dealing with most of the complications implicated in the images.

Index Terms

Image segmentation; Image registration; Pattern recognition

1. INTRODUCTION

Microscopy and comparative analysis of images is an important tool to study the role of genes and their products in developmental processes of model organisms including the fruit fly *Drosophila melanogaster*. Techniques such as mRNA *in-situ* hybridization (ISH) have been developed to localize gene activity in space and to capture this information in form of spatio-temporal gene expression patterns. The boom in high throughput techniques has also seized modern developmental biology. Gene expression patterns for all (about 13,500) genes across a set of developmental stages are being generated systematically, leading to the accumulation of a wealth of high dimensional data with complex encoding. Bioinformatics currently helps with data management, storage, access, and integration (e.g. *BDGP* [1,2,3,4] and *FlyBase* [5]), but could also assist in data processing, data analysis and provide an observation tool for researchers in the field. Such an attempt has been made by *Fly Express*, which comprises 42,825 images for 2,871 genes as of January 17th, 2007. It is currently the largest database that computes and holds standardized images and offers a Basic Expression Search Tool, *BEST* [6], to retrieve genes with similar expression patters to a given query pattern. An alternative and more robust set of algorithms was proposed by Peng et al. [7,8]. Here we describe a

[*]Federal Ministry of Education and Research (BMBF), grant PTJ-BIO/31P4282 (MS CartPro)

processing pipeline for automatic segmentation, classification and clustering of ISH images. It is based on the representation of expression patterns by *Bessel eigenfunctions* which should allow a faster and easier pattern classification once the coefficients are calculated. We apply our approach to a subset of 681 genes with images for all 6 developmental stages and an average of 2.5 ISH images per gene and stage. In total, 10428 images were selected from *BDGP*[1].

2. PROCESSING PIPELINE

The goal of the image processing pipeline is to extract the Fourier coefficients for the gene expression pattern of every whole embryo. The processing pipeline is summarized in Figure 1.

First, a shading corrections is applied to the raw image to obtain a “clean image”. This is done by subtracting a Gaussian smoothed image from the raw image (filter width 64×64 pixel) and expanding the color range to the full 8 bit. Figure 2 shows an example of this process.

To find the transformation of the embryo shape onto a circle, we can only use the outline of the embryo because the color information contains the information of the gene expression and therefore differs from image to image. To extract the shape of the embryos from the image, we compute the magnitude of the gradient in the cleaned image. The gradient based segmentation produces a reliable and accurate separation on the border of the focal plane. Instead of a simple threshold, a Gaussian mixture model [9,10] in gradient space is used. The result of this fragmentation is a scalar image describing the probability for the membership to the two classes (background – embryo). This probability map is smoothed by a total variation filter. After filtering, a mask is obtained with an unpredictable count of holes in it, because some areas in the embryo have feature values similar to the background. We close the holes and obtain the binary Segmentation Result. Only the shape of a single isolated embryo and shapes of several embryos that touch each other will remain, see Figure 3 for an example.

In case of several touching embryos, additional knowledge about the expected shape must be included to obtain the mask of a single embryo (see Figure 4). Gradient vector field snakes are then used to isolate the embryo [11,12]. To make an automatic placement of the initial contour possible, the images are rigidly registered onto an ellipse prior to the active contour segmentation.

The next step in the image processing pipeline is to register the extracted embryo outline onto an ellipse. This is done in two steps: first, we apply a rigid registration [13] and second, a nonlinear registration onto the ellipse [14,15]. This minimizes the distortions produced by the nonlinear registration. The transformations onto an ellipse, computed from the registration of the embryo outlines, are applied to the masked embryo images (Figure 5 left) to obtain the expression pattern mapped onto an ellipse (Figure 5 right). Finally, the ellipse is stretched to circular shape.

After the geometric transformation the expression pattern can be extracted. Color segmentation as in refs. [16,17] was used to handle the wide variation in the staining of the image and to detect the pattern. To find the Gaussians and their weights, an Expectation-Maximization (EM) method was used [10]. This method requires pre-setting of the expected number of Gaussians. We assume more Gaussians than we need for representing the pattern, i.e. *oversegmentation*, and drop Gaussians that belong to the background to separate staining patters from possible artifacts. We initialize the EM algorithm with 7 expected Gaussians and keep 4 of them to represent the gene expression pattern. The result is one scalar image per Gaussian describing the probability for the membership to the regarded Gaussian. These four probability maps are combined by taking the maximum for each individual pixel.

The patterns $\mathcal{P}(r, \varphi)$ are described by a set of Fourier coefficients:

$$\mathcal{P}(r,\varphi)=\sum_{j=1}^{\infty}\sum_{k=0}^{\infty}a_{j,k}\psi_{j,k}(r,\varphi) \tag{1}$$

As basis, the eigenfunctions of the Laplace operator on a circle of radius ℓ,

$$\psi_{j,k}(r,\varphi)=N_{j,k}e^{ik\varphi}J_k\left(\frac{r j_{k,j}}{\ell}\right), \tag{2}$$

are used, where ℓ is the radius of the circle, $J_k(z)$ are the k-th Bessel function, $j_{k,j}$ is the j zero of the k-th Bessel function and $N_{j,k}$ is a normalization factor so that

$$\int_0^{\ell}\int_0^{2\pi}\psi^*_{j\prime,k\prime}(r,\varphi)\psi_{j,k}(r,\varphi)r\mathrm{d}\varphi\mathrm{d}r=\delta_{j\prime,j}\delta_{k\prime,k}.$$

The $\psi_{j,k}$ form a complete orthonormal system. Two examples of the basis functions are shown in the top row of Figure 6. We found that every pattern can be adequately expressed by a set of 420 eigenfunctions ($k \in [0,\cdots, 20]$, $j \in [1,\cdots, 20]$). The lower row of Figure 6 shows an example image (left) and its reconstruction by the Fourier series (right).

3. VERIFICATION

To check how well the patterns are represented by the Fourier coefficients, a hierarchical clustering on a set of 7 images is shown in Figure 7. For the clustering the Euclidean norm in the truncated 420 dimensional space of Fourier coefficients $\|a_{j,k}\|$ is used. Notice that this representation groups similar patterns together. Larger data sets of images show the same agreement with the visual expectation (data not shown). In the application, the orientation of the embryo w.r.t. the two axis of symmetry of the ellipse is important. At present, the correct orientation is not generated automatically. Unless annotated in the initial image, all four orientations are included in the clustering and in the end an expert control [1] is inevitable.

4. CONCLUSIONS AND FUTURE WORK

We have presented an automatic image processing pipeline for the classification of gene expression patterns of fruit fly images. The image processing steps use the outline of the fly embryos to obtain a transformation to a circular shape and apply this transformation to the gene expression pattern. The transformed intensity pattern is expressed in terms of Bessel functions and the metric space of the expansion coefficients is used to identify similarities in the patterns.

The metric space of the expansion coefficients will be used in image retrieval applications and in relation to genomic meta data of the images.

Acknowledgments

We thank Sudhir Kumar, Bernard van Emden and the *Fly Express* team for the presentation of the problem and initial help with the data set. Research conducted by SJP was in part supported by NIH and the *Fly Express* project.

REFERENCES

1. Bdgp: Berkeley drosophila genome project. http://www.fruitfly.org/

2. Tomancak, Pavel; Beaton, Amy; Weiszmann, Richard; Kwan, Elaine; Shu, ShengQiang; Lewis, Suzanna E.; Richards, Stephen; Ashburner, Michael; Hartenstein, Volker; Celniker, Susan E.; Rubin, Gerald M. Systematic determination of patterns of gene expression during drosophila embryogenesis. Genome Biology 2002;vol. 3(no 12)
3. Tomancak, Pavel; Berman, Benjamin P.; Beaton, Amy; Weiszmann, Richard; Kwan, Elaine; Hartenstein, Volker; Celniker, Susan E.; Rubin, Gerald M. Global analysis of patterns of gene expression during drosophila embryogenesis. Genome Biology 2007;vol. 8(no 7)
4. Montalta-He, Haiqiong; Reichert, Heinrich. Impressive expressions: developing a systematic database of gene-expression patterns in drosophila embryogenesis. Genome Biology 2003;vol. 4(no 2):205–205. [PubMed: 12620112]
5. Flybase: A database of drosophila genes and genomes. http://flybase.bio.indiana.edu/
6. Kumar, Sudhir; Jayaraman, Karthik; Panchanathan, Sethuraman; Gurunathan, Rajalakshmi; Marti-Subirana, Ana; Newfeld, Stuart J. Best: a novel computational approach for comparing gene expression patterns from early stages of drosophila melanogaster development. Genetics 2002;vol. 162(no 4): 2037–2047. [PubMed: 12524369]
7. Peng, Hanchuan; Myers, Eugene W. Comparing in situ mRNA expression patterns of drosophila embryos. Proceedings of the eighth annual international conference on Resaerch in computational molecular biology; 2004. p. 157-166.
8. Peng, Hanchuan; Long, Fuhui; Zhou, Jie; Leung, Garmay; Eisen, Michael B.; Myers, Eugene W. Automatic image analysis for gene expression patterns of fly embryos. BMC Cell Biology 2007;vol. 8
9. Pernkopf, Franz; Bouchaffra, Djamel. Genetic-based EM algorithm for learning Gaussian mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 2005;vol. 27(no 8):1344–1348. [PubMed: 16119273]
10. Moon, Todd K. The expectation-maximization algorithm. IEEE Signal Processing Magazine 1996:47–60.
11. Xu, Chenyang; Prince, JL. Snakes, shapes, and gradient vector flow. IEEE Transactions on Image Processing 1998;vol. 7(no 3):359–369. [PubMed: 18276256]
12. Kass, Michael; Witkin, Andrew; Terzopoulos, Demetri. Snakes: Active contour models. International Journal of Computer Vision 2004 November;vol. 1(no 4):321–331.
13. Gottesfeld Brown, Lisa. A survey of image registration techniques. ACM Computing Surveys 1992;vol. 24(no 4):325–376.
14. Braumann, Ulf-Dietrich; Kuska, Jens-Peer. Influence of the boundary conditions on the result of non-linear image registration. Proceedings of the IEEE International Conference on Image Processing; 2005 Sept.. p. I-1129-I-1132.
15. Braumann, Ulf-Dietrich; Kuska, Jens-Peer. A new equation for nonlinear image registration with control over the vortex structure in the displacement field. Proceedings of the IEEE International Conference on Image Processing; 2006 Oct.. p. 329-332.IEEE Signal Processing Society
16. Braumann, Ulf-Dietrich; Einenkel, Jens; Horn, Lars-Christian; Kuska, Jens-Peer; Löffler, Markus; Scherf, Nico; Wentzensen, Nicolas. Registration of Histologic Colour Images of Different Staining. In: Handels, Heinz; Ehrhardt, Jan; Horsch, Alexander; Meinzer, Hans-Peter; Tolxdorff, Thomas, editors. Bildverarbeitung für die Medizin 2006 - Algorithmen, Systeme, Anwendungen; Springer-Verlag; 2006 Mar. p. 231-235.Informatik aktuell
17. Kuska, Jens-Peer; Braumann, Ulf-Dietrich; Scherf, Nico; Löffler, Markus; Einenkel, Jens; Höckel, Michael; Horn, Lars-Christian; Wentzensen, Nicolas; von Knebel Doeberitz, Magnus. Image registration of differently stained histological sections. Proceedings of the IEEE International Conference on Image Processing; 2006 Oct.. p. 333-336.IEEE Signal Processing Society

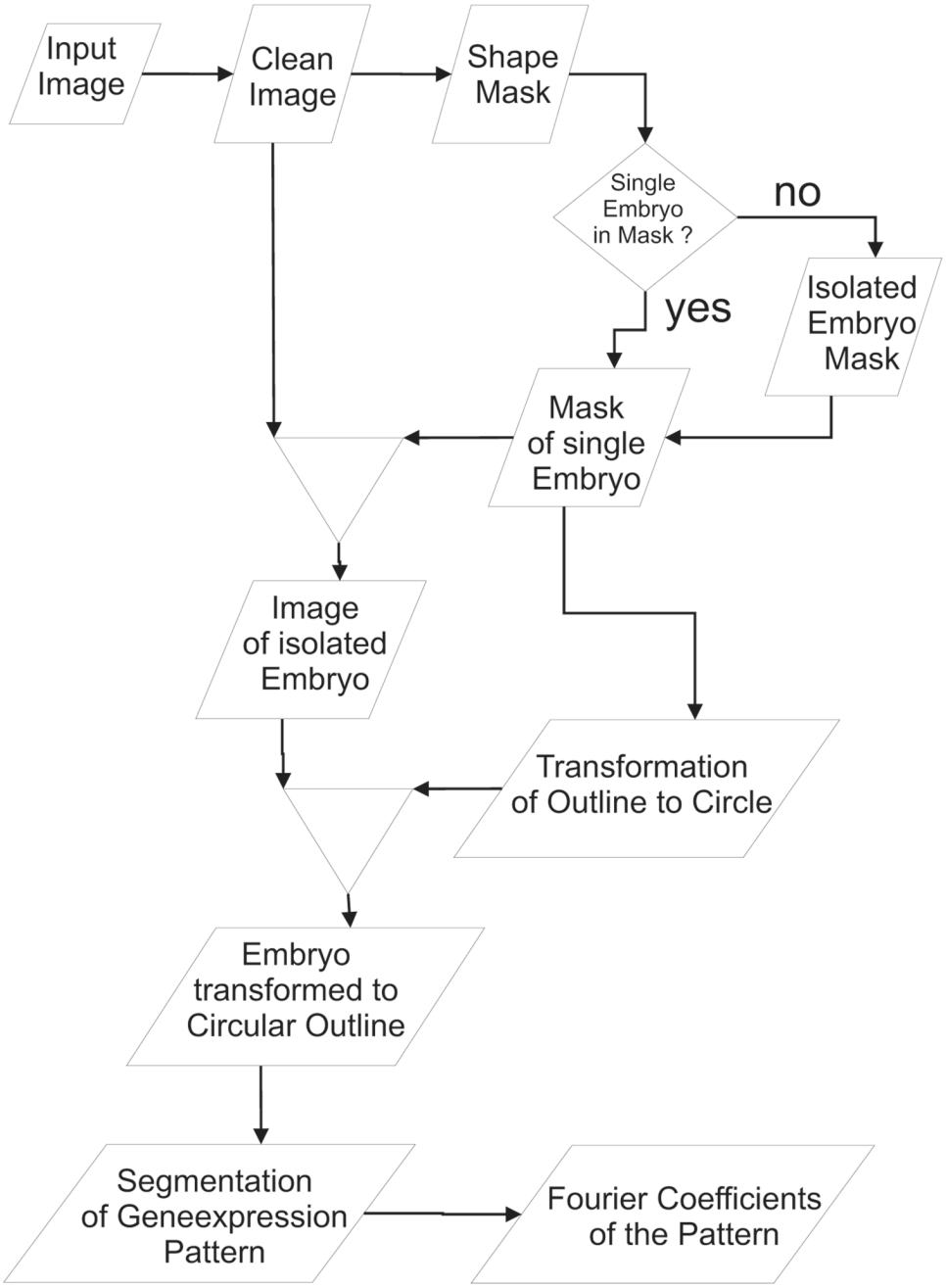

Fig. 1.
Extraction of the Fourier coefficients of the expression pattern

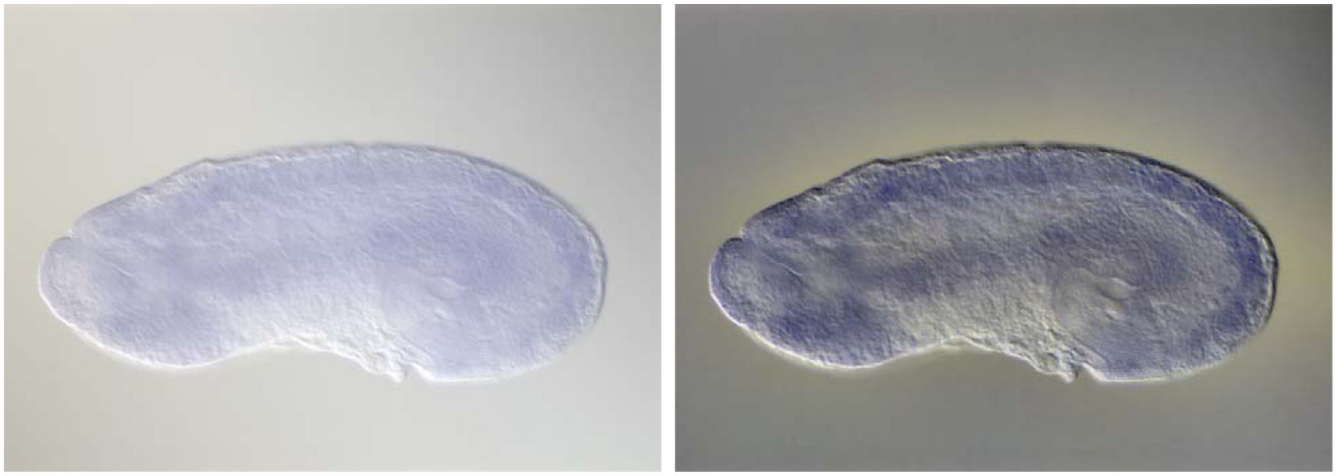

Fig. 2.
Preprocessing of the original image (left) to the clean image (right)

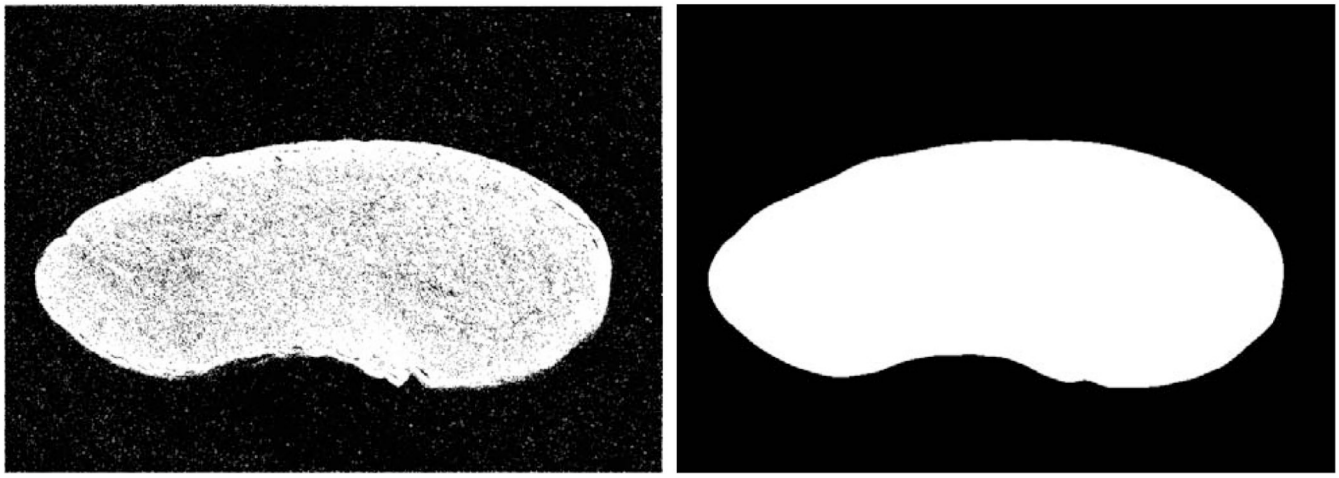

Fig. 3.
Shape Segmentation

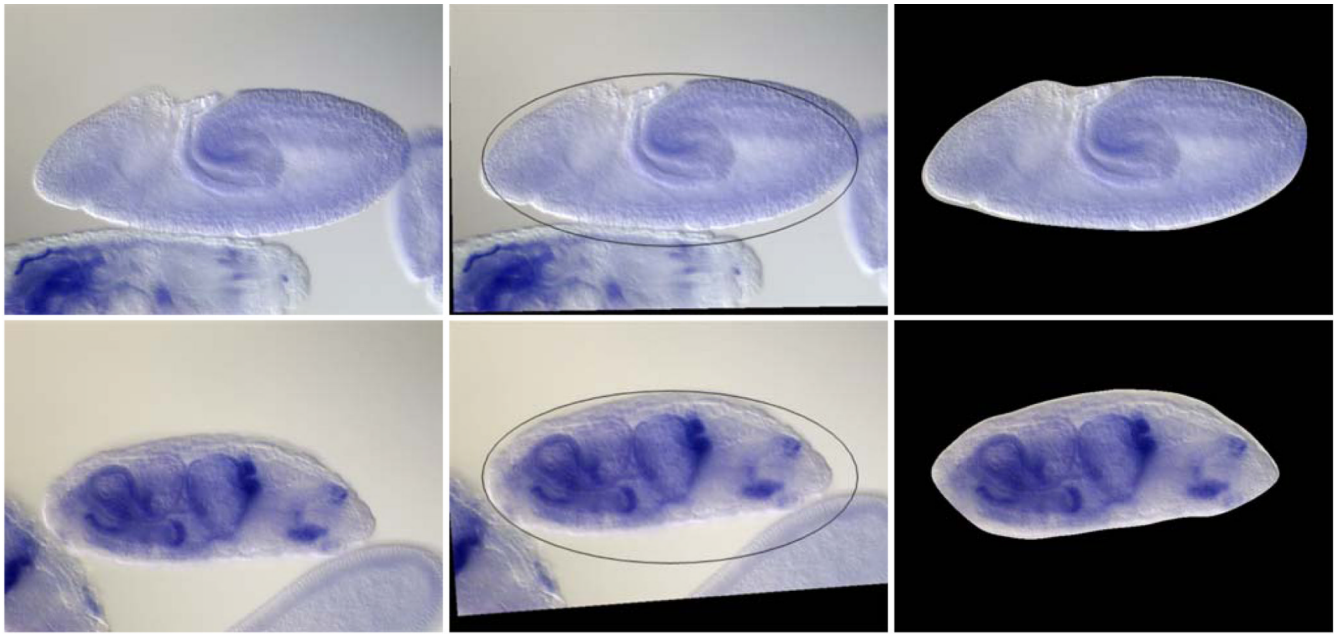

Fig. 4.
Processing of touching embryos; the original images (left), after rigid registration of the outline on a centered ellipse (middle) and the extracted embryo image from the gradient vector field snakes (right)

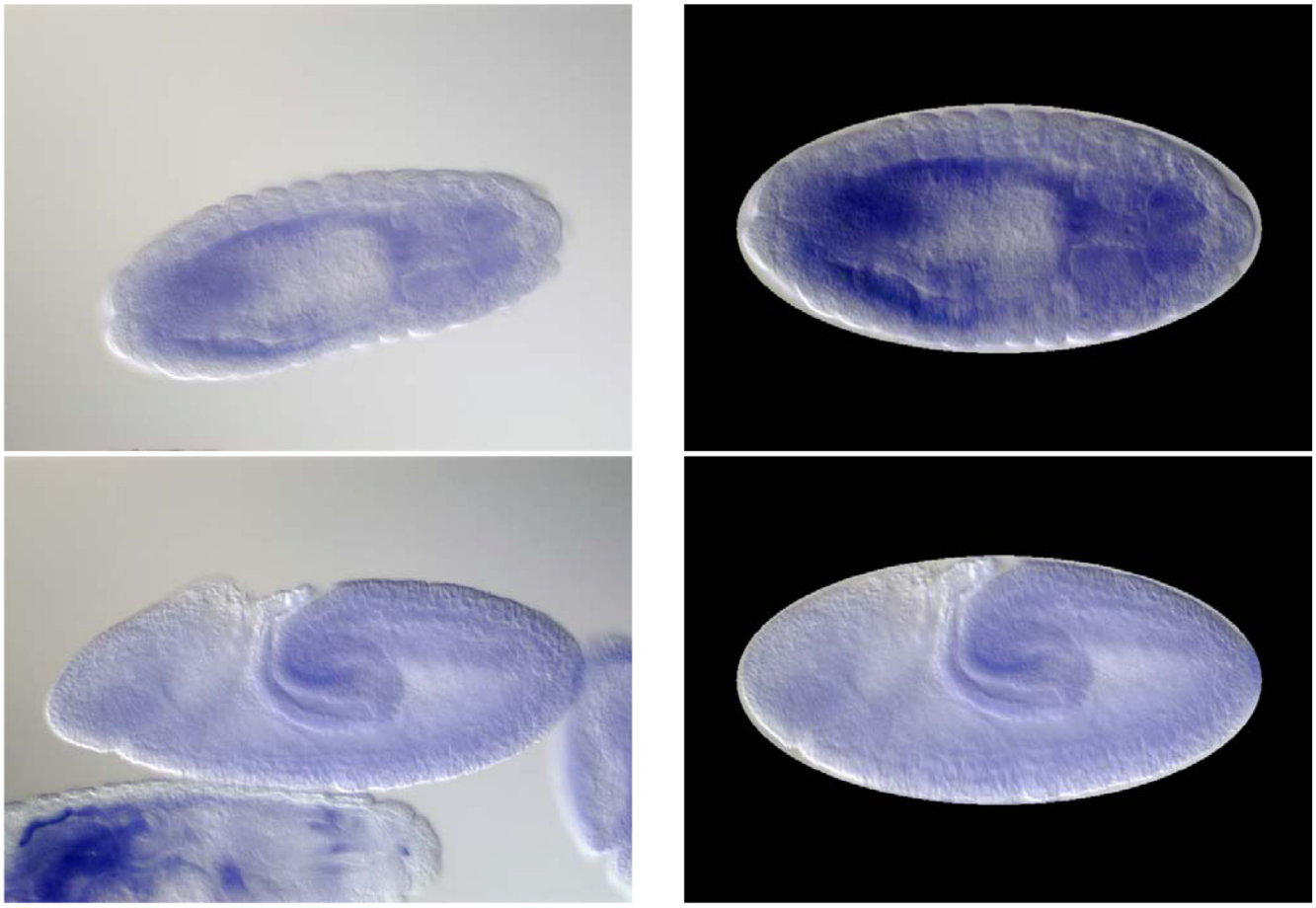

Fig. 5.
Masked embryo images (left) and the result of the registration (right)

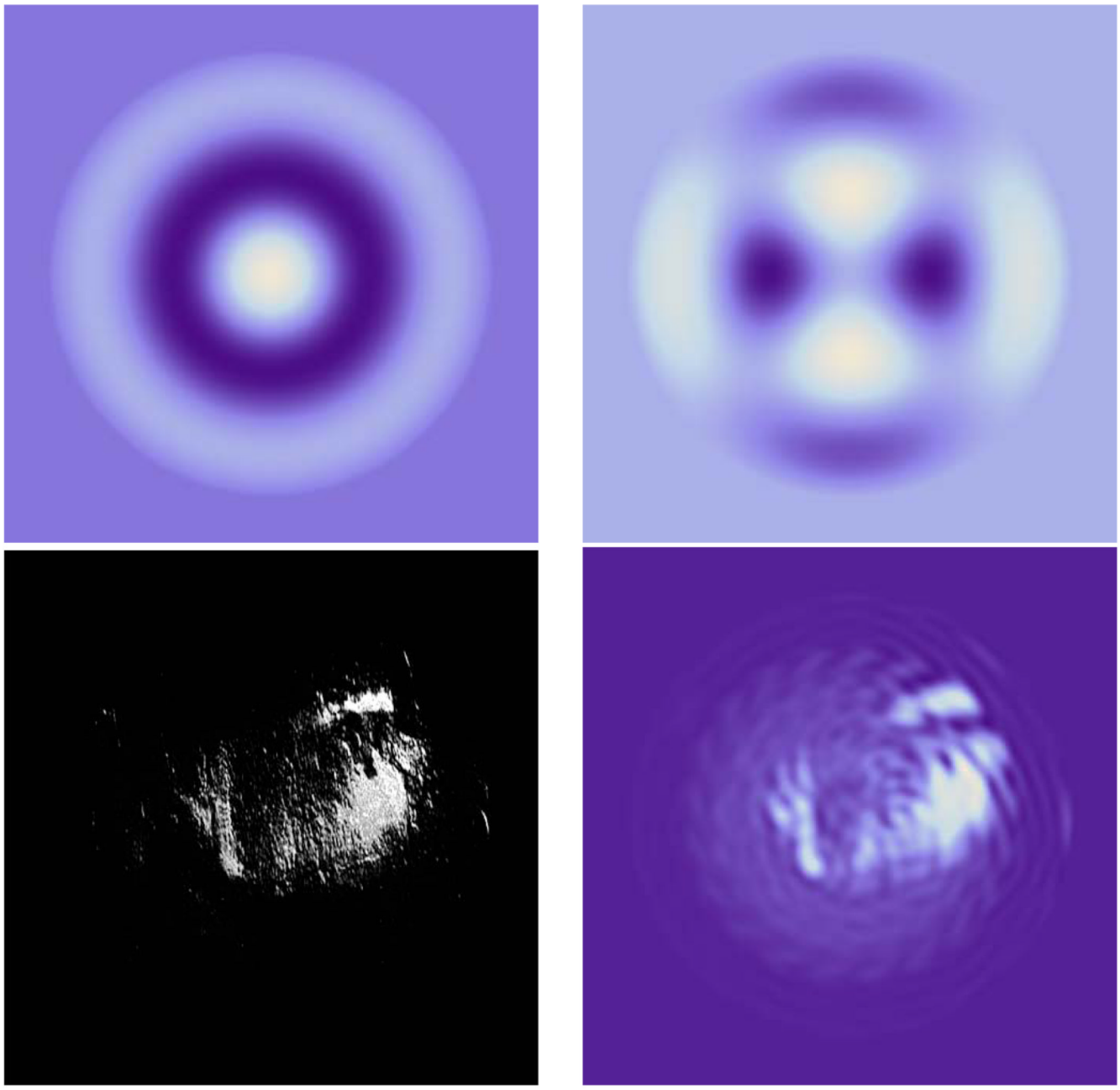

Fig. 6.
Examples of two Bessel-Eigenfunctions with $k = 0$, $j = 3$ (upper left) and $k = 2$, $j = 2$ (upper right). Circle fitted GEP (lower left) and the reconstructed GEP from the computed coefficients (lower right).

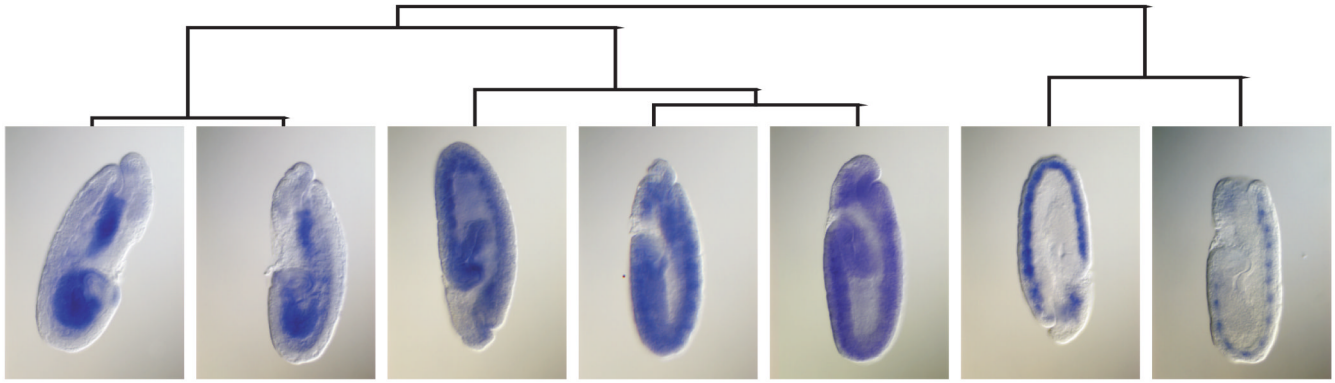

Fig. 7.
Hierarchical clustering on a set of 7 images that demonstrates that visually similar pictures are grouped together.