REPORT

Adnane HAMID, Zhuofan YU

1. INTRODUCTION

Biology has inspired innovation in computer vision, a particularly well known one is the convolutional neural network. Histological image analysis, the study of microscopic images from biological tissues, and research on cell segmentation has been going on since 1960 [6]. It has also been an active field of study in recent years with papers published and Kaggle competitions on the subject [7]. This may be linked to an increasing general interest in medical imagery, as well as increasing capabilities in microscopic imaging and computation capabilities, the latter partly characterized by the emerging use of neural networks in histology. In the following, we study histological images of fly embryos from the species drosophila melanogaster, widely used in biological research in genetics, physiology, microbial pathogenesis, and the life history evolution. As of 2017, eight Nobel prizes had been awarded for research using Drosophila [4].

We propose an automatic classification method to identify the stages of development of embryos in the dataset. Given the small dataset, we put a strong focus on image processing and feature extraction. It includes the implementation of the methods described in Jia-Yun Pan et al [1] and Abdolhoseini et al [8] and we present a novel improvement for the latter.

We would like to thank Dr. Ahmed Mosallam, the internship supervisor one of us has, for proposing the subject and overseeing our work.

2. LITERATURE AND STATE OF THE ART

Our approach was to start with a broad literature review of the research made on drosophila embryos age estimation and more broadly on histological image processing and feature extraction.

Pure biological research based on observation and study of the evolution of cell embryos and their anatomy is useful to understand where our data is situated in the general embryo evolution chronology. It is also useful to understand the features that characterize the embryo's evolution. In that regard, the O'Farrell Lab's website provides a clip of the development phase they call mid-blastula transition which includes the stages corresponding to our dataset [10]. Their imaging method is also the same as for our dataset which makes understanding and comparison easier. Also, the website Flymove documents the evolution of the embryo's anatomy from start to end.

Alongside that, image processing techniques are required to extract relevant features. One such process is cell segmentation, which has been subject to research in histology since 1960 and continues to be an active field of research today [6]. A variety of methods can be used to achieve segmentation, examples are intensity thresholding, blob detection, morphological filtering. Most recent methods are variations or combinations of existing methods or use learning methods such as neural networks. Since our dataset has less than 100 images and is not labeled for segmentation, we did not take the machine learning route. We took interest in the cell segmentation method described

in Abdolhoseini et al 2019 [8]. This was notably because their method yielded good results on the Kc167 dataset which consists of drosophila embryo cells as well as other histological datasets.

Apart from cell segmentation, other features may be extracted. For example, Jia-Yu Pan et al [1] extract features derived from independent component analysis (ICA), which give information on the structures that characterize the embryo's development. Another example of features are the Fourier coefficients used in Myaskinova et al.

As a side note, before starting our work, we made a summary of two papers on age and stage of development estimation based on histological images. This was done as a mean to familiarize ourselves with the general methodology in the field before proposing our method. We will give a brief description of those to conclude this paragraph on literature review and provide a two-page summary of both papers in the project's folder.

In a few words, Myaskinova et al involves using gene expression images as data [2]. After some image preprocessing, Fourier transform coefficients of the image are extracted as features. A support vector regressor is then trained to classify the images based on their stage of development.

The work in Chen et al is part of another class of histological analysis research that uses epigenomic matrices as input, in which rows represent genes, columns represent cells and entries correspond to gene expression values [9]. Such information is used to extract knowledge on the stages of evolution of the cells and the role that each gene has at each stage of evolution. Their method relies on affinity propagation clustering enhanced with linear optimization on elastic principal graphs.

3. METHODS AND ALGORITHMS

3.1 DATASET

To avoid confusion, we precise that there are only two entities in each image: background and nuclei which we also refer to as cells.

Our dataset composes of 70 microscope images files where each image file contains two layers. The first layer of each is a microscopic image of the embryos stained for DNA, where bright regions represent the presence of nuclei. The second layer represent the gene expression of the Krüppel gene. These 70 image files are acquired and labeled by a biologist and distribute unbalanced in 7 different developmental stages (Tab. 1). Each image is of 1024×2048 pixels and 256 grayscale.

Stage	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Number	1	7	12	16	10	20	4

Table 1. Distribution of the image files in terms of developmental stage

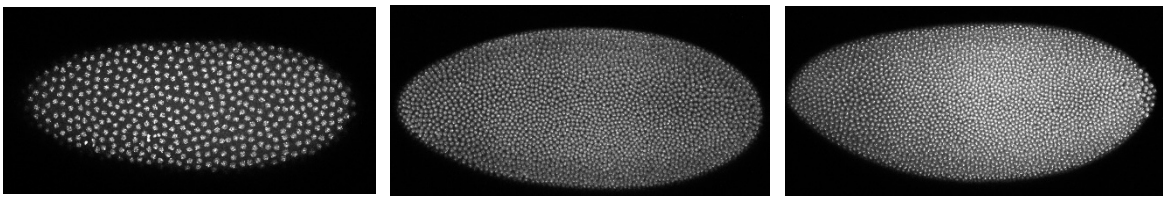


Figure -1. Typical microscopic images of drosophila embryos. Stage 1 (left) Stage 4 (center) Stage7 (right)

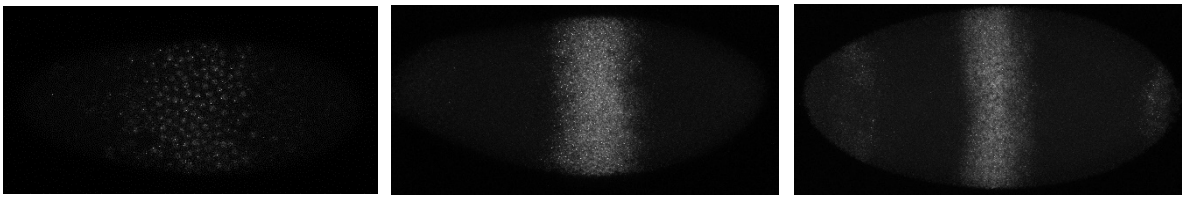


Figure -2. Typical gene expression of drosophila embryos. Stage 1 (left) Stage 4 (center) Stage 7 (right)

Some typical images of embryos and gene expression in different developmental stages are shown respectively in Fig. 1 and Fig. 2. Embryos are of different shapes and sizes, there is luminosity imbalance between images, some images are also of lesser quality with blurred limits between cells. Background brightness may strongly vary locally as well as between different regions of the image. Brightness inside the cells also differs within the cells and between cells.

3.2 BIOLOGICAL STUDY OF THE DATASET

General knowledge on embryo development suggests it can be divided into 17 phases of development from its creation to its full development [5]. Identifying where our dataset is situated in this chronology allows us to understand the biological processes going on. This understanding then translates into features we look to extract from the images for later classification of the stages.

The number of nuclei doubles after each division and there should be around 50 nuclei at the center of the embryo at the end of the 8^{th} mitosis, while the majority of nuclei should populate the periphery and number around 200 [5]. The fact that there is a continuous migration of the cells from the center towards the periphery [5] means that the number of nuclei at the center might not double at each mitosis but have a smaller increase instead. The nuclei counting indicate that embryos stage 1 of our dataset has more than 200 nuclei in the center. With 50 nuclei in the center after the 8^{th} mitosis, this means we have a maximum of 200 nuclei after the 10^{th} mitosis. Thus, stage embryos are at least after the 11^{th} mitosis. We also know that our data is from before invagination before which we have 13 mitosis happening. Also, stages 2 to 4 have double the cells of stage 1 and stages 5 to 7 have double the cells of stages 2 to 4. Meaning we have only 3 levels of density in our dataset, and the only 3 levels possible for stage 1 are the 11^{th} to the 13^{th} mitosis. Thus stage 1, stages 2 to 4 and stages 5 to 7 are after the 11^{th}, 12^{th} and 13^{th} mitosis respectively. In the grand scheme of the embryo's development, this corresponds to stages 4 to 5 of embryonic development, also known as the blastoderm stage [5].

With this information we researched the biological mechanisms taking place during this phase of the embryo's evolution. Obviously, the number of cells allow to make a clear separation between the groups of classes we just mentioned. Furthermore, there is an ongoing of formation of the polar buds on the right which has already started after the 11^{th} mitosis [5]. Additionally, there is a progressive elongation and angular distortion of the cells that happen specifically after the 13^{th} mitosis [8]. Finally, the different steps of the mitosis process happen right after the 12^{th} mitosis and before the 13^{th}, corresponding to stages 2 to 4. This is characterized by an increase in the size of the nuclei shortly after the 12^{th} mitosis, followed by a contraction right before cell division with the 13^{th} mitosis. Regarding the Krüppel gene expression, it is characterized by a vertical band located in the center of the embryo. After the end of the 13^{th}

mitosis (stages 5 to 7), it starts expressing at the poles, with a higher in the overall embryo, with a reduction of the band's thickness in the center [12].

After careful observation of the dataset, it seems impossible to distinguish between stages 2 to 4 using the above information. The same goes for stages 5 to 7. This indicates a probable issue with the labelling of the stages, with a strong possibility of stages 2 to 4 corresponding to one same stage and the same goes for stages 5 to 7. Therefore, while we present the classification results later in this report, we believe they are not significant and a relabeling is required for a proper evaluation of the age estimation. Therefore, we put a stronger focus than initially intended on feature extraction, with the aim of presenting indications that it is working as intended and delivers us the correct information.

3.3 PREPROCESSING OPERATIONS

With observations, we list the following aspects to be improved and the corresponding solutions:

1. Luminosity imbalance

   Due to the influence of some factors (ambient brightness for example) during the acquisition process, there is a non-ignorable luminance difference between embryos images. Since the luminosity imbalance also appears in the embryos of the same stage, we can deduce that the luminosity imbalance is not related to the age of the embryo. There is also luminosity imbalance between cells in the same embryo. These luminosity imbalances can stop us from correctly distinguishing the cell and may mislead the PCA and ICA process. Therefore, we apply the sliding histogram equalization[2] to reduce the luminosity imbalance in embryo and between embryos. In order to eliminate the influence of the luminosity, we unify the feature matrix in the later process and make each latent gene expression pattern have the same amplitude. #

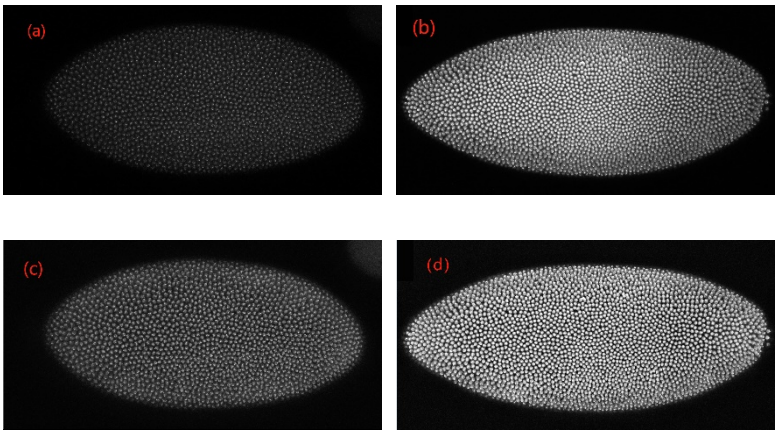


Figure -3. Both (a) and (b) are original embryo images in stage 5. Filtered image (c) and (d)

2. Discrepancy of the embryos' shape and noise in the background.

   Embryos are neither well centralized nor having the same shape. Some embryos are plump while the others are skinny, and some of them may be close to one side of the image and far from another. What is worse, in many microscope samples there are visible stains in the background. The biased positions, the different shapes and the blurs can both mislead the PCA-ICA algorithm into believing that they are some important features. So before

applying the feature extraction algorithm, we need to firstly remove the background and standardize the embryos' shape.

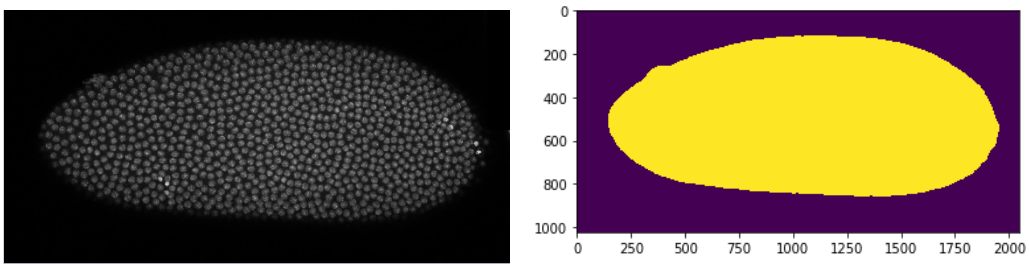


Figure -4. Original embryo (left) and its extracted contour (right)

The first step is to extract the contour of the embryo and then remove the background.

To achieve this:

1. We first apply a global threshold at 20 which only removes background pixels on the entire dataset.
2. We apply a median blur with a large filter of dimension 51*51, this blur puts at zero any pixel for which the filter has a majority of zero pixels. Then all non-zero pixels are set to 255. This has the effect of giving a rough estimate of the embryo's shape which is precise enough for PCA/ICA.
3. As can be seen in the top right of Figure 3, b, parts of other embryos might appear in the image, these are removed simply by keeping the biggest connected component of the image which is the central embryo.
4. An example of resulting image can be seen in figure 4.

Then we need to consider the centralization and the standardization process. Since the first embryo in the dataset has an oval shape and is not used for model training, we define its contour as the standard contour, which means all other embryos will be reshaped according to the standard contour. We first crop the original image into a rectangle that just fits the embryo and rescale horizontally the embryo according to the standard embryo. An example of this centralization and standardization is shown in Fig 5.

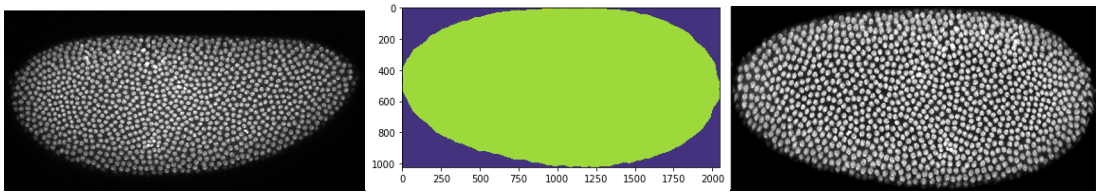


Figure -5. An example of the original embryo (left) standard contour (center) standardized embryo(right)

3.4 EXTRACTION OF CHARACTERISTIC FEATURES

There are two paths to extract features from the embryos' images. The first path is more intuitive, which is realized by recognizing and calculating cells in an embryo and their size. The number of cells appears to be greatly different

between certain different stages. The second path is more mathematic, which is realized by projecting the original image into a small feature vector [1] . These two paths will be introduced in this chapter.

3.4.1 Cell count calculation

See appendix 1 to visualize the algorithm's steps.

Here we present our methodology and results for cell counting. It is a means to obtain the cell density rather than the total number of cells in the embryo. Therefore, we perform the counting not on the entire embryo but in a fixed area at its center to take advantage of the higher quality of the image there compared to the embryo's limits.

3.4.1.1 Cell counting methodology

The method essentially rests on the application of an adaptive mean thresholding, followed by the watershed algorithm on the distance map to separate touching cells before counting. Application of median filters can be done after and before each step for different purposes, mainly to remove noise and small connected components.

3.4.1.2 Cell counting results

We evaluated the method by observing the counted cells and whether or not touching cells were correctly separated by watershed. The method works well on different stages, with different levels of luminosity. Cell separation is less efficient on some images, typically blurry images or where cells are touching a lot. But even on those images, since the number of cells grows exponentially, we can still tell how many cell divisions occurred even with an error on the cell counting.

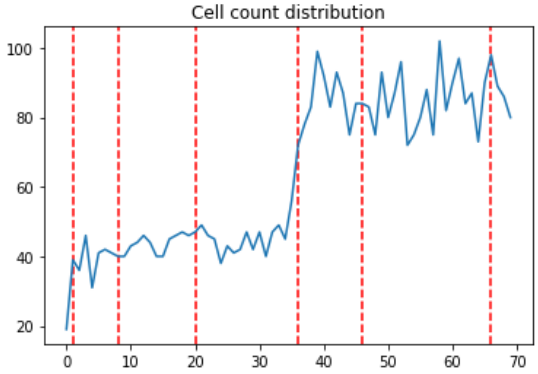


Figure -6. Cell count distribution, y axis represents the cell count and x axis the image number, red lines separate the 7 different stages.

Developmental stage	AVG number of cells
S01	19
S02	39
S03	43
S04	45
S05	85
S06	85
S07	88

Table -2. Average number of cells for 40000 pixels at each stage.

On average, stages 2 to 4 and 5 to 7 have the same number of cells. There is a factor of two between the two groups of stages, indicating the occurrence of a cell division.

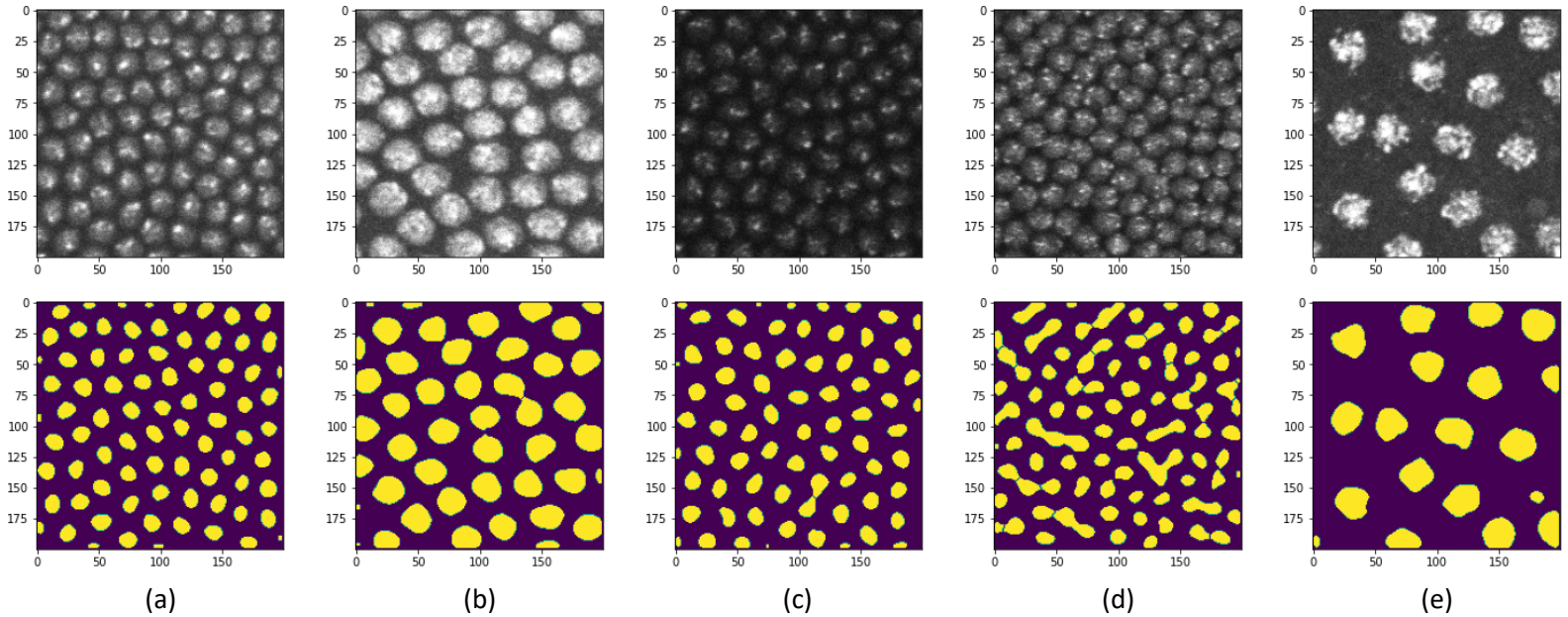


Figure -7. Top images are originals and bottom one are counted connect components *a.* 80 cells correct *b.* 45 cells, correct *c.* 72 cells, correct *d.* 84 cells, Incorrect: 10 cells were not separated and thus not counted *e.* 19 cells, Inexact because the small bright region in the bottom right was counted as a cell.

3.4.2 Nuclei segmentation and background removal

In the following sections we describe the methodologies we have tried and led us to our final method for this project. The goal of segmenting the nuclei is to obtain an average value of the cell size. This information might be useful to determine how far the embryo has progressed in the cell division process since the cells grow progressively after each division.

3.4.2.1 Initial attempts at cell segmentation

We initially tried several methods of cell segmentation:

1. Otsu thresholding and multi-Otsu thresholding
2. Local mean thresholding
3. Canny edge detection

The Otsu thresholding consists in fining an automatic threshold for the given image based on the histogram. We tried a local Otsu thresholding as well as a local one. We also tested removing the embryo's background from the threshold(s) calculation. For the multi-Otsu thresholding, we also tried both local and global thresholding, testing several thresholds from 2 to a maximum of 4 for computational reasons.

The local mean thresholding was performed with a mask with a size that guarantees the presence of at least a couple cells inside the mask. It consists in putting pixels above the mean of the mask to 255 and the rest to zero.

For both methods, several preprocessing methods were tried, including different methods of luminosity balancing such as histogram equalization and CLAHE. Background removal using the Top hat morphological transform was also

attempted to facilitate cell segmentation. All OpenCV's classic blurring methods were tested with different sizes to remove noise. Both methods included postprocessing operations to remove noise, based on blurring or connected component filtering based on size.

However, it turned out that these methods did not give satisfying results:

First, such methods are understandably very sensible to luminosity variations. This makes it difficult to have a method based on these techniques that would work on the entire dataset, given the luminosity variations. Moreover, luminosity variations inside the images also cause problems for these methods.

Second, for the local mean thresholding, there is no reason we can think of that pixels with brightness greater than the local mean would be part of cells.

Finally, looking at the global and local regions, it is difficult to derive thresholds from their histogram, thus it is understandable that both Otsu and multi-Otsu thresholding would not be successful.

3.4.2.2 Implementing a research-based method for segmentation [8]

See appendix 1 to visualize the algorithm's steps.

We implemented Abdolhosseini et al's method for cell segmentation. Explanations on the method can be found in the paper and we provide our implementation of the method in python. The novelty of this algorithm being the definition of local regions using the watershed algorithm [8].

We found little success with this method on our dataset. The segmentation did not fit the cells and the separation between cells was mostly incorrect, separating a cell in many in most cases. We tested the addition of preprocessing steps not mentioned in the method, such as blurring, top hat background removal and luminosity equalization, but without significant results.

However, in the paper we can see that the dataset was successful for many datasets, one of which is the Kc167 dataset which consists of drosophila cells [13]. Yet the Kc167 does not contain images of embryos, but of cells obtained through a different method. In this dataset, cells are much sparser than for our dataset, also it does not have a number of issues including complicating variations of brightness. These differences mean to us that it is possible they had a good result for their datasets but that did not transfer to our dataset. Then we can propose an explanation as to why it does not work as well. Although Abdolhosseini et al's method have an innovative way of defining local regions with watershed and employ several processing methods all the way to the results, it essentially is a local Otsu thresholding method at its core. And as we said in the above, multi-Otsu thresholding and Otsu thresholding methods, even when applied locally were not successful and robust throughout the dataset.

3.4.2.3 Our novel cell segmentation method inspired by Abdolhosseini et al

See appendix 2 to visualize the algorithm's steps.

Our failure with the approaches we used so far led us to think about a different way to segment the cells. We considered how a human would proceed to segment the cells. Let us take the example of an image with only two cells. One way is

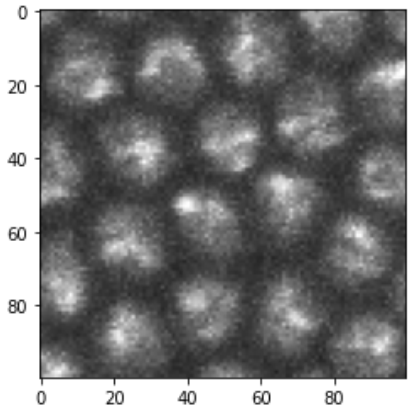
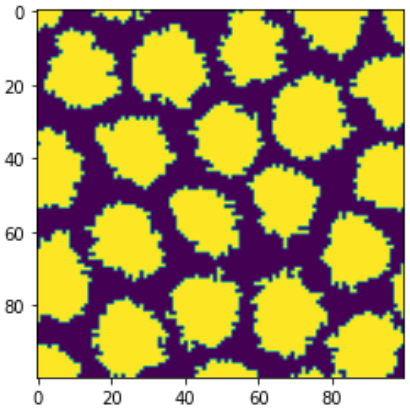
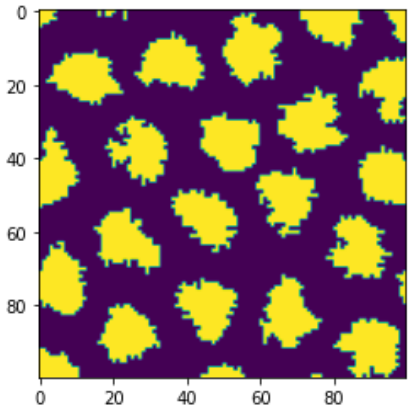
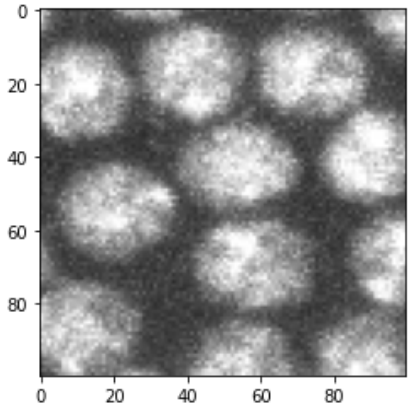
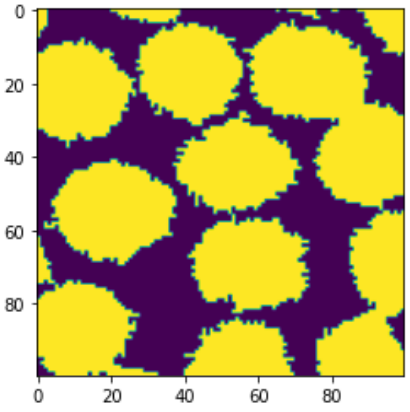
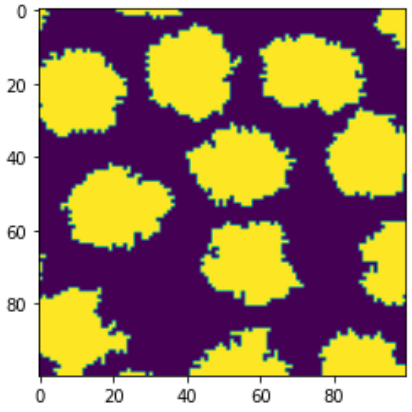
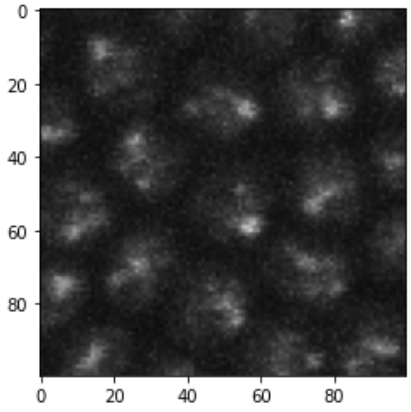
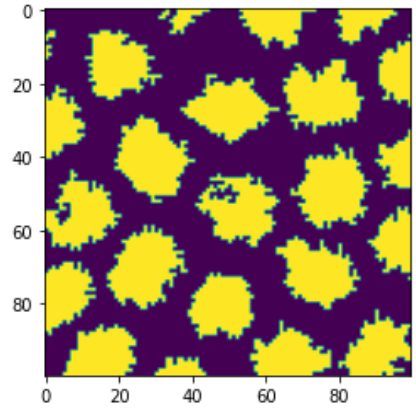
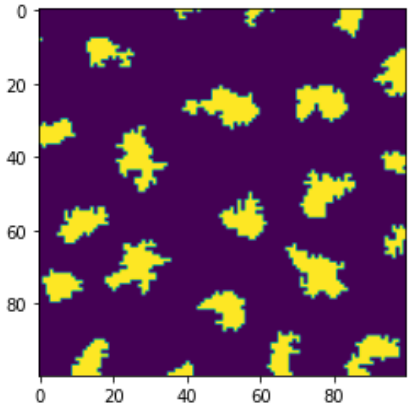
to look at bright circular shapes and to consider those as roughly the shape of cells. We would have identified roughly the shapes of the two cells, but some bright areas directly around the cells cannot be identified for sure as cell or background. Then looking at the brightness in the space located at equal distance between the two cells gives us an idea of the maximum brightness of the background. Then we would mark the shapes of the cells with precision by removing any regions that has brightness at the same level or below the maximum brightness in the space between the two cells. This is the intuition behind our method that we explain below.

To execute such a method, we need a way of defining rough zones containing the cells and the spaces situated between cells which we will use to define the threshold. For this matter we reuse Abdolhosseini et al's idea of using the watershed algorithm to define local regions. The expected result is that the watershed lines would be between cells and thus would be located exclusively in the background. Then for each local region, we would gather the pixels located on the watershed lines drawing the boundaries of this region. We would then use the maximum of these pixels' values to threshold the background. We expect that we will replace the maximum by a statistically significant value, less sensible to noise such as the upper value of the 95% confidence gaussian interval.

Abdolhosseini et al's method starts directly with an Otsu thresholding. However, this often results in cells separated in more than one connected component. This is problematic because this will cause watershed lines between these connected components to go over the cell region belonging to the original cell. Yet we want the watershed lines to be exclusively located in the background to extract a background threshold from them. Therefore, we first apply a median blur with a mask of size 11*11, roughly the size of a cell. This prevents the Otsu thresholding from dividing one cell into many connected components. This will also result in more components containing more than one cell. The problem with this is that large connected components render the computed threshold for that component less local. Yet this is undesirable because of the luminosity differences between regions. Still, it is a good enough trade off since having watershed lines go over cell regions would render the defined thresholds useless. After the Otsu thresholding, we apply the watershed algorithm, using the bright regions as zeros, to obtain the watershed lines and local regions, just like in Abdolhosseini et al [8]. However, we then found that the Otsu thresholding would reduce some cells to a very small component. The problem is that the watershed lines issued from that component are then more likely to go through the cell's region. Therefore, we replaced the Otsu thresholding method with local mean thresholding with a mask of 51*51 to typically includes a couple of cells.

At this point, Abdolhosseini et al's method applies some post processing, and most importantly does local thresholding on each local region, using Otsu's method. Instead, we will use the idea of using the watershed lines we described above to obtain a threshold. For each region, we gather the pixels which define of watershed lines, at the region's border. This is essentially a sample of background pixels. The idea is then to extract a statistically significant threshold, such as any value below that in the local region would be considered as background. We tested the maximum, the 95% Gaussian confidence interval and the 95% percentile and the latter yielded the best results amongst the three. Finally post processing is applied in the form removing connected components smaller than a minimum size set as 100.

3.4.2.4 Cell segmentation methods' result comparison

Image	Our segmentation	Abdolhosseini et al
		
		
		

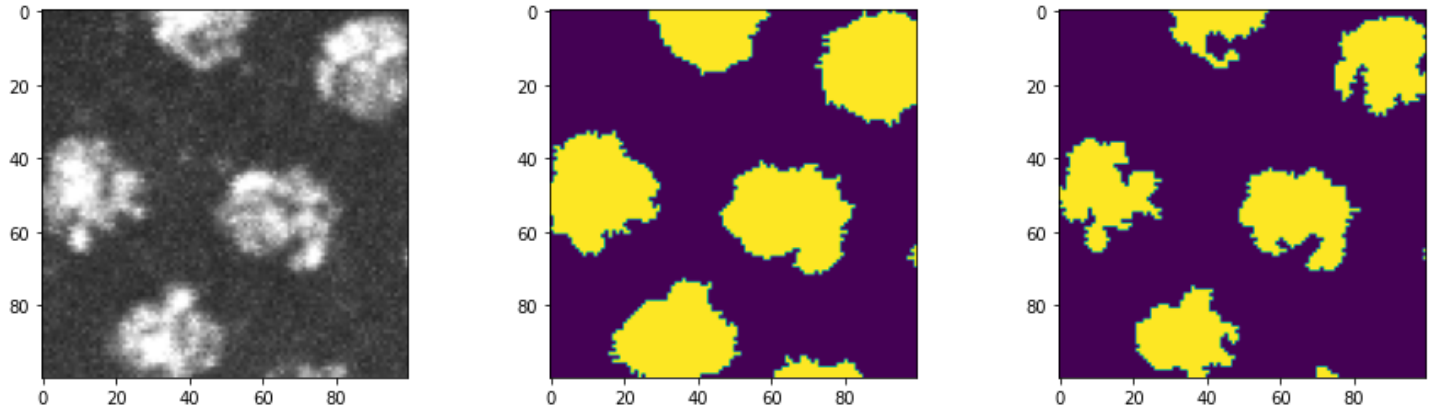


Figure -8. These figures support our observation throughout the dataset that our method gives better results than Abdolhosseini et al's. On some images, the result is almost perfect segmentation. On other images, with typically many touching cells, the result can be improved but is still better than Abdolhosseini et al's method and the methods we had tested initially.

3.4.2.5 Computing the cell size feature

Once the cells segmented, we calculate the cell area on a 200*200 window of the image. To do so, we remove connected components touching the borders of the window which can be incomplete cells. Then we compute the number of cells on the same image after separating touching cells by watershed. Components smaller than a minimum size are not counted in the total.

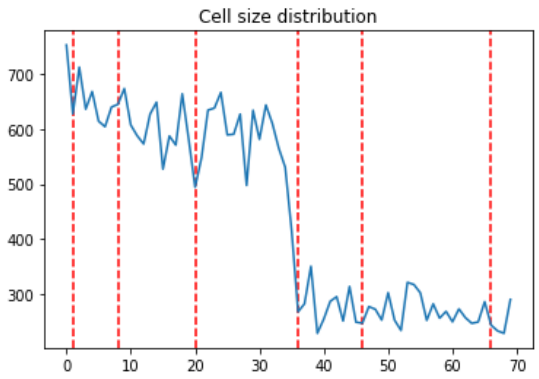


Figure -9. Cell count distribution, y axis represents the cell count and x axis the image number, red lines separate the 7 different stages.

Developmental stage	Average cell size in pixels
S01	753
S02	643
S03	608
S04	579
S05	278
S06	270
S07	249

Table -3. Average number cell size at each stage.

In conclusion, there is no significant difference of cell size between stages 2 to 4. There is also no difference between stages 5 to 7. Although the segmentation is still to be improved, this goes along our observation that cell size does not change between images of those stages. There is however a difference between stages 2 to 4 and 5 to 7, since a cell division occurs between the two groups of stages.

3.4.3 PCA and ICA projection

In this section, we will talk about the features' extraction and aggregation. We use PCA and ICA for 2 main purposes: 1) discover the potential biological patterns 2) build independent feature space. By projecting the original dataset on the feature space, we can summarize an image into a vector of several dimensions which is a suitable input of a SVC model.

Through observation, we can find out that as the embryo grows, there are significant changes inside the embryo. Before diving into details of PCA-ICA projection, we firstly introduce some notations:

$D = [I_1, \dots, I_M]$ Images of embryos where each image I_i(I =1,…,M) is viewed as a 1-by-N vector of $N = 1024 \times 2048$ pixels

$X = [X_1, \dots, X_{M'}]$ Matrix of M' eigen-image where $M' \leqslant M$

$S = [S_1, \dots, S_{M'}]$ Matrix of independent patterns

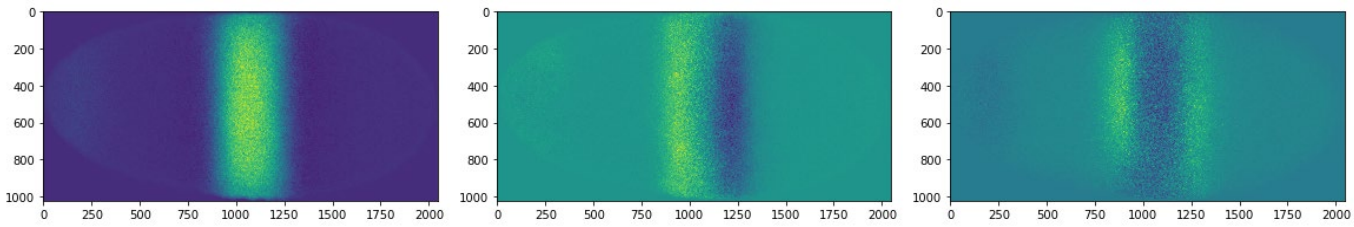
$I_{LGE} = [I_1', \dots, I_M']$ Matrix of projected images

Firstly, we apply PCA (Principle Component Analyze) to compute the eigenvectors of the covariance matrix of the data (DD^T). Each eigenvector can be visualized as a 1024 x 2048 image, which is called eigen-image. In our project, we select only a reduced number M' of eigen-images, those associated with the highest eigenvalues. We maintain 90% of the "total energy" (sum of squared eigenvalues) to conserve the most important information. However, each single Eigen-image fails to describe the biological features. To mine the latent biological features, we use ICA (Independent Component Analyze) which transforms Eigen-images into pattern images. The top 3 pattern images of embryos are shown in Fig.5 and an interesting gene expression pattern is shown in Fig.6. This gene expression pattern has a bright zone on the left which indicates the brightness on the left zone is a biological feature for separating the early stages and the late stages. The vertical band located in the middle of the embryo is another important biological feature for the stage classification. The model of ICA transform is as follow:

$$\boldsymbol{X = B\,S}$$

Where X the M'-by-N Eigen-image matrix, B an unknown M'-by-M' matrix and S and M'-by-N matrix where each row corresponds to a pattern and can be visualized as a 1024 by 2048 image.

By projecting the original image on patterns, we get a M'-by-1 feature vector for each image. In Fig. 5, we show 3 distributions of the projection values of all 70 images under their pattern image. Red dashed is the separation of two adjacent stages. We can see that some features can distinguish stages. For example, with the first feature we can well separate the early stages (1-4) and later stages (5-7). We can also find out some biological patterns more easily by directly observing those images. In Fig 6 we can see a bright zone on the left side which is an important feature. Until now, we successfully reduce the original 1024 x 2048 matrix to a M'-by-1 vector which can be the input of the classifier.



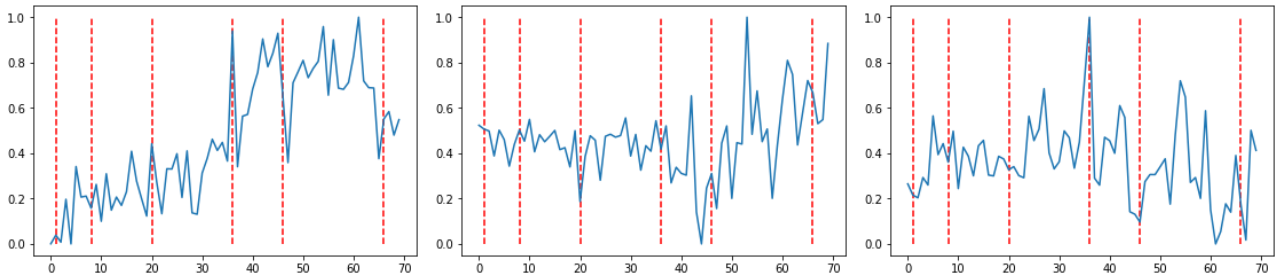


Figure -10. Top 3 independent patterns extracted by PCA-ICA (top) and the distribution of their projection values (bottom)

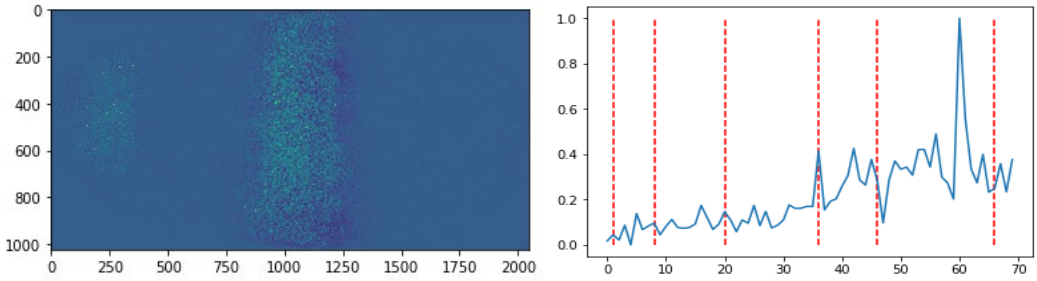


Figure -11. Interesting gene expression pattern (left) with bright zone on the left and its distribution of projection values (right)

3.5 TRAINING & TESTING SET AND CLASSIFIER

Since our dataset has only 70 images which distributed not evenly in 7 stages, the division of training and testing set is limited. There are only 1 image in stage 1 while 20 images in stage 6, which can apparently influence the final classification. In order to make the classification results take into account all stages as much as possible, we delete the only image in stage 1 and divide the whole remained data into a training set and a testing set according to a ratio of 8 to 2 and setting the stratify parameter to True.

By following the observation of a relative experiment[1] , we use the linear support vector classifier(SVC) for the age classification. The choice of the linear kernel is a trade-off of processing time and classification accuracy. Our parameter "cost" is set as 0.5 to improve the generalization ability of our classifier.

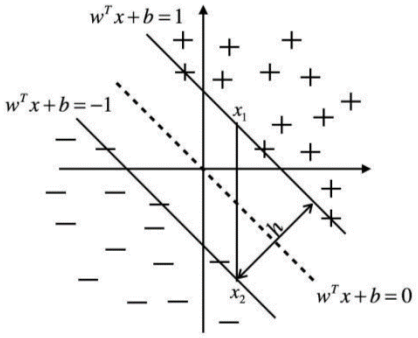


Figure -12. Explanation of the idea of a 2-D SMC[3] . The parameter "cost" is represented as the distance h in this image.

In order to exclude the influence of random factors, we take the previous process for 300 times and take the average as the final result, which means separating randomly the training and testing set with consideration of the stratification for 300 times.

4. RESULT

We have seen in the previous paragraph that the first component of the PCA-ICA can be used to separate stages 2 to 4 from stages 5 to 7. Now we also want to see whether our PCA-ICA projection values can recognize some other biological features. We design another SVM classifier to see if an image contains a "head" feature (Figure 6, left zone) or not. We manually label the images by observing whether a gene expression image has the "head" feature and then train the SVM classifier with 80% images in the dataset. The same idea as the previous section, we use the other 20% images for the test and repeat 5000 times to have a more stable result. Without fine tuning, base on the PCA-ICA projection values, we have the following confusion matrix:

	Positive (with head)	Negative (Without head)
True	43%	1%
False	56%	99%

The result shows that we have 43% possibility to correctly detect the images with "head" feature and almost never classify an image without "head" as containing the "head". This result supports that the PCA-ICA projection values can have a positive influence on our project.

The confusion matrix of 300 iterations is shown in Tab. 2. Each row is corresponding to a labeled stage and each column is a classified stage. The number is represented in percentage and the higher the value, the deeper the color. We can see that most of the embryos in stage 2 – 4 are well classified while the embryos with the labels between 5 – 7 are more likely to be classified as stage 6. Most of the classification errors are within 1 stage, which represents a good classification ability. According to our biological study of the dataset, it makes more sense to use 3 developmental stages instead of 7, so we can assume that stage 5, 6 and 7 are biological similar and can be summarized as one single stage.

		Classification Age					
		2	**3**	**4**	**5**	**6**	**7**
Label	**2**	86	14	0	0	0	0
	3	0	77	22	0	0	0
	4	0	12	83	0	4	0
	5	0	0	0	5	94	0
	6	0	0	5	4	87	2
	7	0	0	0	0	100	0

Table -4. Confusion matrix. Each row is corresponding to a labeled stage and each column is a classified stage. The number is represented in percentage.

5. DISCUSSION

Our drosophila age prediction system is still far away from perfect. We have summarized the following points which may probably improve the general classification accuracy:

1. Lack of information on the labelling of the stages: We do not know how the original images are labeled which makes us unable to tell the difference between some adjacent stages. Furthermore, as explained in our biological study, nothing seems to separate stages 2 to 4 and stages 5 to 7. We believe a revision of the labelling could open doors to a better classification.

2. Larger dataset: Having more data would allow to test the robustness of our feature extraction methodology. Typically, our method relies on the fact that the large mask median blur would prevent the initial thresholding from cutting any cell in two. The fact that it works on our dataset of 70 images does not guarantee success on additional data. A larger dataset also would benefit the training of the final classifier.

3. We do not have a quantitative way to measure the accuracy of the segmentation on our dataset du to the absence of ground truth. The same goes for the cell counting.

This project allowed us to get a lot more familiar with cv2, skimage and image processing in general. We found it to be a challenging and ultimately good and formatting project because we would very often face frustrating problems for which we had no immediate solution which we feel has forced us to think more creatively.

6. FUTURE WORK AND POSSIBLE IMPROVEMENTS

We believe we could improve our segmentation method by choosing a possibly more sensible local threshold than the 95% percentile. Ways of doing that may include the incorporation of the cell's brightness into the threshold, looking into ways of adapting the chosen percentile. We could also adapt the zone from which we select background samples, for instance, larger components may have a larger and more widespread watershed border so we could divide it into local subsections where each would have its own threshold.

Also, there are interactive background removal methods where the user selects portions of the image as the background and other portions as the foreground. There are similarities with our method of segmentation as it can be seen as an automated way of selecting portions of the background and deriving thresholds from that. Therefore, we could in the future take inspiration from those methods and how they use the user selected portions of the image to perform segmentation.

Ultimately, we feel there is a lot of potential significant improvement to our segmentation method. The idea we present is still at its initial stage, yet the results were very promising. We have little doubt that improvements in the same direction would yield much better results, and this idea could also be applied for segmentation on other fields.

7. REFERENCE

[1] . Jia-Yu Pan et al. *Automatic Mining of Fruit Fly Embryo Images.* Carnegie Mellon University.

[2] . E. Myasnikova et al. *Determination of the developmental age of a drosophila embryo from confocal images of its segmentation gene expression patterns*. St. Petersburg State Polytechnic University

[3] . 土豆磊, *SVM*、*SVC*、*SVR 三者的区别*,https://zhuanlan.zhihu.com/p/37702043

[4] . Wikipedia. *Drosophila melanogaster*. https://en.wikipedia.org/wiki/Drosophila_melanogaster

[5] . Weiigman et al. *FlyMove – a new way to look at development of Drosophila*. Universität Münster https://www.sdbonline.org/sites/fly/aimain/2stages.htm

[6] . Meijering: *Cell segmentation: 50 years down the road.* http://helios.mi.parisdescartes.fr/~lomn/Cours/CV/BME/HistoPatho/LongPapers/CellSegmentationAreview2012.pdf

[7] . 2018 Data Science Bowl: https://www.kaggle.com/c/data-science-bowl-2018

[8] . Abdolhoseini et al: *Segmentation of Heavily Clustered Nuclei from Histopathological Images.* https://www.nature.com/articles/s41598-019-38813-2#Sec2

[9] . Chen et al : *Single cell trajectory inference.* https://www.nature.com/articles/s41467-019-09670-4

[10] . O'Farell Lab website's mid-blastula transition clip: http://biochemistry2.ucsf.edu/labs/ofarrell/movies.html

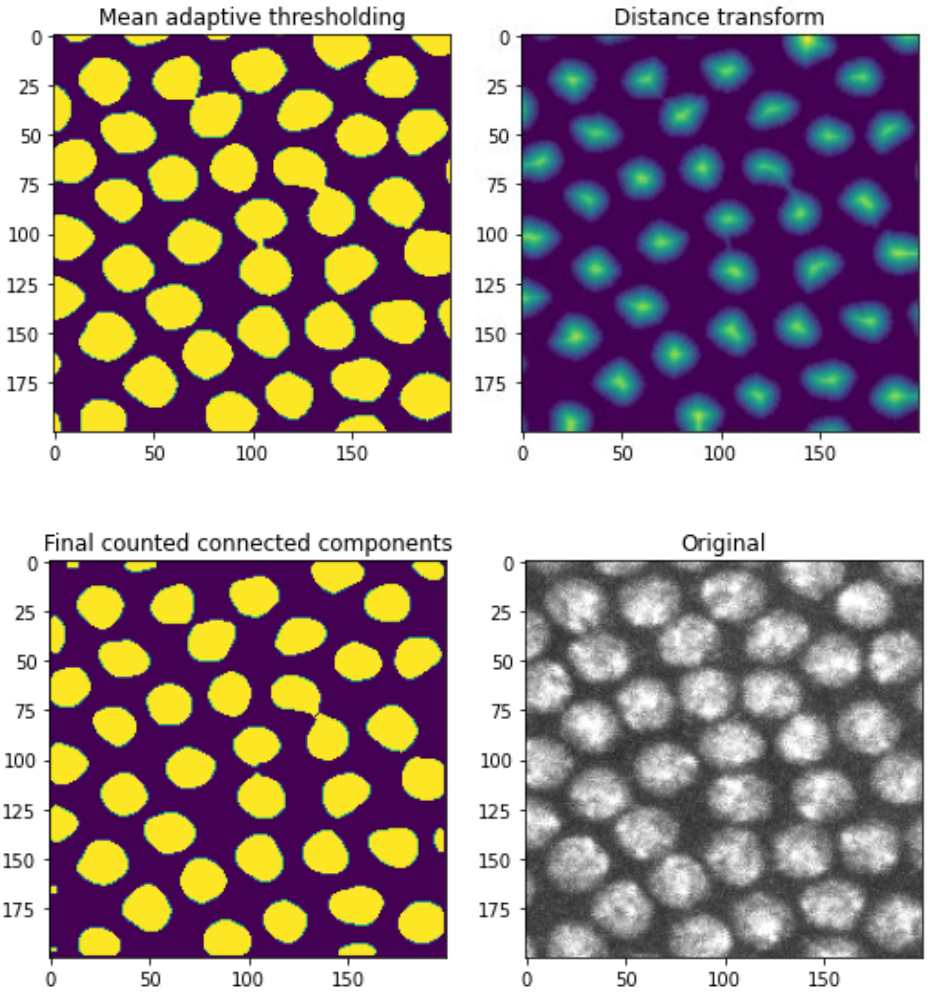
[11] . Kong et al : *Robust Cell Segmentation for Histological Images of Glioblastoma* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5382998/

[12] . Huslampk et al 1990 : *A morphogenetic gradient of hunchback protein organizes the expression of the gap genes Krüppel and knirps in the early Drosophila embryo* https://www.nature.com/articles/346577a0

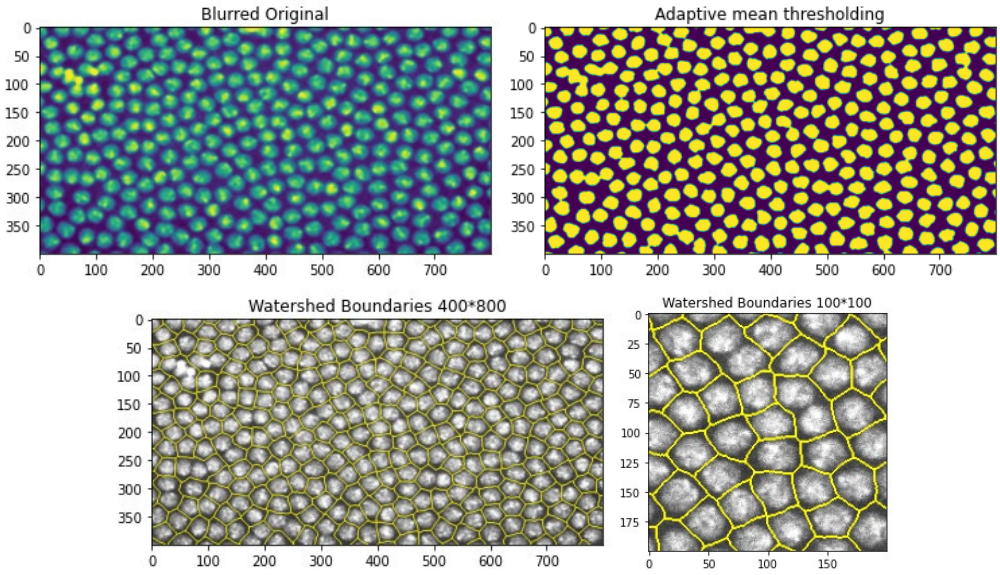
[13] . Kc167 Dataset : https://www.google.com/search?q=kc167+data&rlz=1C1CHBF_frFR865FR865&oq=kc167+data&aqs=chrome.0.69i59j69i60.868j0j7&sourceid=chrome&ie=UTF-8

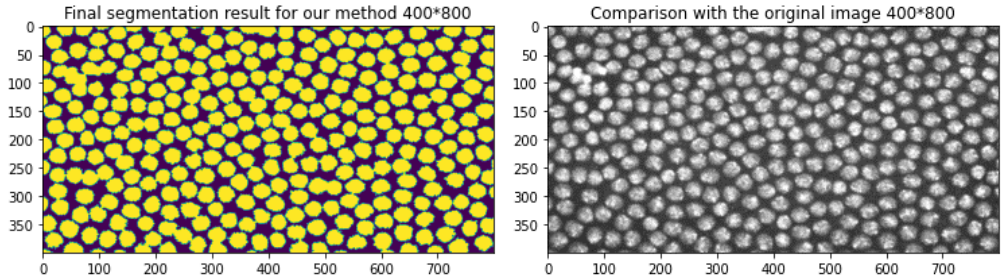
APPENDIX

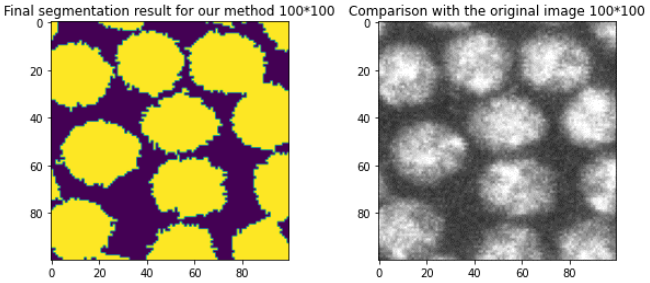
7.1 STEPS FOR CELL COUNTING

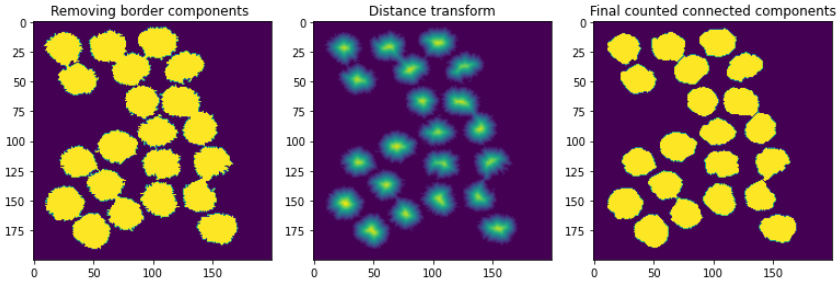


7.2 STEPS FOR SEGMENTATION AND AVERAGE CELL SIZE COMPUTATION: OUR METHOD

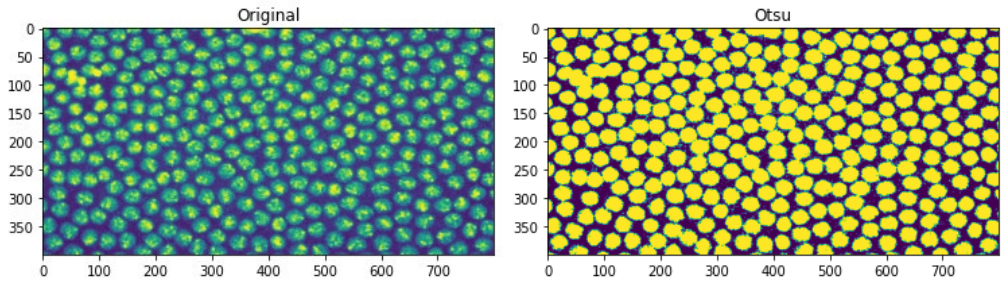


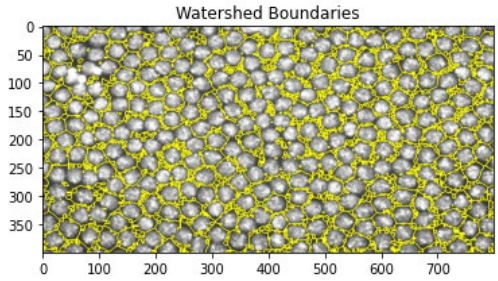


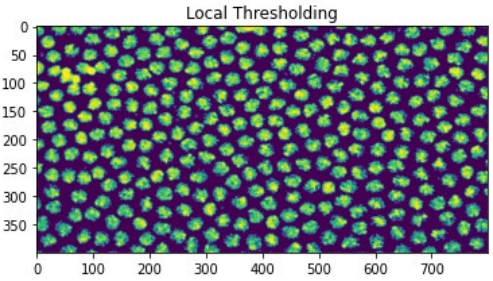


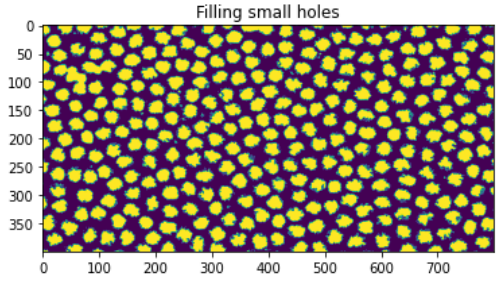


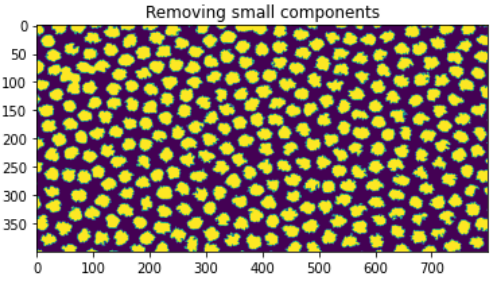
7.3 STEPS FOR SEGMENTATION: ABDOLHOSSEINI ET AL

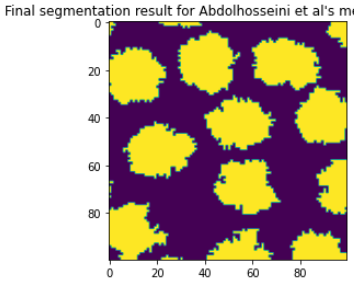


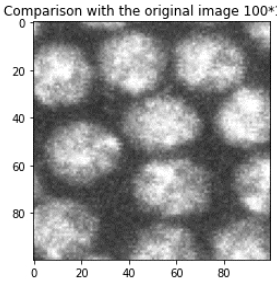

Watershed Boundaries

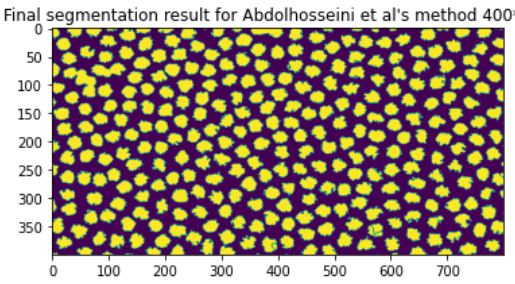

Local Thresholding

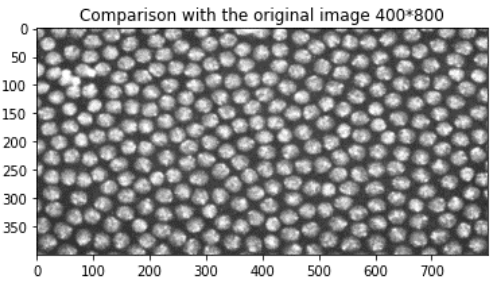

Filling small holes


Removing small components


Final segmentation result for Abdolhosseini et al's method 100*100


Comparison with the original image 100*100


Final segmentation result for Abdolhosseini et al's method 400*800


Comparison with the original image 400*800