Expediting Gathering and Labeling of Data from Zebrafish Models of Tumor Progression and Metastasis Using Bespoke Software

Samling och Märkning av Data från Zebrafiskmodeller av Tumörprogression med Hjälp av Skräddarsydd Programvara

Adam Ivarsson

Supervisor: Peter Dalenius
Examiner: Rita Kovordanyi
Upphovsrätt

Copyright
The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances. The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility. According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement. For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: http://www.ep.liu.se/.

© Adam Ivarsson
Expediting Gathering and Labeling of Data from Zebrafish Models of Tumor Progression and Metastasis Using Bespoke Software

Adam Ivarsson
Linköping, Sweden
adam.m.ivarsson@gmail.com

ABSTRACT
In this paper I describe a set of algorithms used to partly automate the labeling and preparation of images of zebrafish embryos used as models of tumor progression and metastasis. These algorithms show promise for saving time for researchers using zebrafish in this way.

Author Keywords
Computer-vision; Algorithms; Zebrafish; Cancer;

INTRODUCTION
Many pharmaceutical drugs¹ are, if taken in the wrong doses or under the wrong circumstances, dangerous. For example Warfarin, an anticoagulant, can have its effects heightened to dangerous levels if a patient is simultaneously taking certain other drugs [1]. This means that before a drug is employed, its effects on humans, both good and bad, have to be studied carefully. However from the outset it is not ethically viable to test how dangerous a potential drug is on human subjects; first the safety for the test subjects have to be raised to an acceptable level. One way to do this is to first test the drug on model organisms.

A model organism is a non-human species which, from a biological standpoint, is well understood. Some of these model organisms have biology that after exhaustive study have been deemed to correlate in part to the biology of humans. This means that if the differences between model and humans are known then the model can be used as a surrogate for some types of drugs; the effects of the drug on the model animal allows you infer information on the effects the same drug will have on a human. One such model organism is the zebrafish².

Zebrafish models have been used in the research of certain types of cancer and the drugs that treat them and have been proven fit for this purpose [8, 5, 9]. The method studied in this paper involve certain strains of zebrafish embryos being injected with cancer-cells and/or fibroblasts to study the progression of any tumors created as well as how drugs affect both the progression and the embryo. To study the progression of the cancer in the embryos as well as the embryos themselves, high-resolution pictures are taken with a microscope before, during and after the trial period. These images are then prepared, labeled and studied. The preparation step involves generating a binary mask of the cancer metastases³ in the image, whilst the labeling of the image consists of the measurements taken from the mask including overall spread, size and count of metastases.

The preparation and labeling of images for this type of zebrafish-study is done using different software at different institutions. The software used is generally not made for cancer nor this type of zebrafish study specifically meaning that the labeling and preparation of the images is labor and time intensive⁴. In theory this step could be expedited if the software were tailor-made for this type of study, even if it only automated or expedited a small subset of the required actions to turn the raw images to useful data.

If this was done successfully resources could be diverted to other parts of the involved institutions and companies which in turn could speed up the time it takes for certain drugs to be approved.

Issues
Issues that this article will try to answer:

Which steps of the image preparation can be automated and to what degree?
When the image comes straight from the microscope, several steps has to be taken to make the image fit for labeling. Depending on the image there can be many required steps from changing the contrast of the image to manually painting out parts of the image the study isn’t interested in.

Which steps of the image labeling can be automated and to what degree?
After the image has been properly prepared the data still has to be labeled/extracted from the image. This can range from individually counting the number of cancer cells in certain studies to calculating the range of the metastasis of the tumor.

How do you minimize the time spent on steps that cannot be automated to an acceptable⁵ degree for the persons performing these steps?

¹Which from here on will be referred to by the more general term “drugs”.
²A small freshwater fish, also a common aquarium fish.
³A secondary site of cancer caused by parts of the primary site, i.e the primary tumor, breaking of and settling elsewhere.
⁴We could not find any such software existing in the public sphere. See Background and related work.
⁵Acceptable being defined as the algorithm and developed correction facilities still saving time.
For example, can a step be partly automated and then only require minimal input from a human to become acceptable?

RELATED WORK
Zebrafish models have now been used extensively for over a decade and during that time multiple software solutions to gather, prepare and label data has been created. In fact, so much work has been done there was even a survey paper [6] exploring the current research of the subject at the time of its writing (2013). However, the only piece of research relating specifically to zebrafish that in the end was applied to the problems studied in this article was “A high-throughput analysis method to detect regions of interest and quantify zebrafish embryo images” [10], whose algorithm for finding the zebrafish embryo was used as an outline to the fish-finding algorithm developed in this paper. Most other attempts to automate further research of zebrafish either focus on the behavior or specific cells of the fish.

There is therefore reason to develop a custom solution from the ground up for the problems stated in this paper, as there are no published solutions nor any commercially available that can solve them.

BACKGROUND
The solution presented in this paper was developed for a specific company working the type of zebrafish research that was described earlier. This company currently employs 2 people working full time with the research, divvying up time between the image analysis discussed here and other related research.

For the questions in this study, the relevant part of their workflow is as follows:

1. Images of the zebrafish embryos are taken in a microscope over a period of days 3+1 (including day 0). Images are taken with the same magnification within each test to ensure the final measurements are consistent. The actual measurement used is not important (e.g. pixels or %), only how it compared to other images taken previously, either compared to the same fish or compared to the total recorded on a day over a set of fish.

2. Images are analyzed one by one, day by day
(a) Images are prepared with an image-editing program, this is done by manually selecting filters to turn non-cancer parts of the image black, this is a multiple step process and the settings for the filters have to be manually entered and differ from image to image.
(b) Pixels are manually marked or erased where they have been wrongfully left marked or unmarked by the filters until the only non-black pixels are those where cancer is seen.
(c) The number of non-black pixels are acquired using a function in the program.
(d) The number of metastases is manually counted.
(e) The “width” of the spread of the cancer manually calculated.
(f) The data and its calculated meta-data is recorded.

3. For each day, compile all data into a summary (often an average) for the day. The total time for only step 1 and 2 were measured prior to the study by the company itself to determine whether the production of a quicker solution was necessary. It was calculated that, for each image, once the employee was “in beat”, only about 30 seconds were needed to give accurate results. That meant that with a potential of 10 test-series running in parallel per week, with 22 fish each and a picture taken every day for 4 days, slightly more than seven work hours would be required for those two steps alone each week.

Step 2a is intended to be replaced by a cancer masking algorithm, taking in the raw image and generating a binary representation of the cancer. Step 2b-2e and 3 is intended to be replaced by a labeling (and) data gathering algorithm which will both measure the data and output a compiled version.

METHOD
Software development
The software solution was developed in several stages, some of them done concurrently when viable.

Baseline measurement phase
First a set of baseline measurements had to be taken to help determine how the solution developed compared to the old software and workflow. Here it was measured that, for a skilled image analyst working undisturbed, about 1 image per 30 seconds could be analyzed using their old workflow and software.

Development of image preparation algorithm
This step was done in parallel to the development of the cancer masking algorithm (CMA), this was done due to the fact that the CMA depended largely on how the image was prepared. This step also includes generating a “fish” mask helping the cancer-masking algorithm determine the relevant parts of the image, improving accuracy.

Development of cancer masking algorithm
Using the measurements taken in the baseline measurement phase the algorithm for creating a binary mask of the metastases was developed. Here the algorithm was first tested on images deemed difficult by the employees. Once an acceptable level had been reached as deemed by the employees, it was tested on images not used during development, if the performance here was deemed insufficient, these images were added to the training set and the process repeated. These images are added to the set to make sure that the cases where the algorithm failed are covered in future testing.

This is backed up by the data collected during the Baseline measurement phase described in the method of this study.

A binary image wherein pixels marked as white make up the mask, i.e the pixels containing the fish in this case.
Development of labeling algorithms
The development of the labeling algorithm turned out to be
trivial from a programming sense when the performance of
the CMA was sufficient. This involved iterating over the
pixels of the image and counting the pixels marked, from
which meta-data such as average cancer location was col-
lected. Here the issues instead lied within finding what met-
data about the image was interesting, again an iterative pro-
cess was used wherein a constant channel of feedback and
suggestions was open to the employees.

Development of correction facilities
The algorithms are assumed to be imperfect regardless of
time allowed for their development, therefore a set of correc-
tion facilities(a computer program) had to be implemented
wherein an employee can correct any mistakes by the algo-
ритм. These facilities where also iteratively improved by the
employees using the facilities and giving feedback. When the
correction facilities had reached what the employees deemed
a usable level in terms of efficiency and ease-of-use, the soft-
ware solution itself started to automatically gather data on
time spent on correcting mistakes, and from this point this
data was used in conjunction with the feedback from the em-
ployees to improve the algorithms.

Overall improvement phase
Here measurements were taken on how long the employees
took to prepare and label the images, as well as how long it
took them to compile the data. The solution could then be
iteratively improved upon by comparing the new measure-
ments to previous ones; using this comparison as a guide to
further hone the solution.

Developed algorithms
Image preparation algorithms
The first step of preparing the image is isolating its red chan-
nel since the cancer and background both exhibit a red fluo-
rescence. The next step is generating a fish-mask indicating
where in the image the fish is. This mask helps increase the
accuracy of the cancer-masking algorithm. When displayed
within the correction facility this mask also helped employ-
ees with the error-correction of the algorithm both in terms
of efficiency and accuracy. My method shares some steps as
those described in [10], the biggest difference being the order
of operations, my use of a tuned Canny edge detector rather
than just a Sobel operation as well as using a contour detec-
tion algorithm [2] with a template window size of 10 pix-
els and a search window size of 21 pixels.

A. The image is first downsampled 4 times, from 2560x1920
to 160x120 pixels using a Gaussian pyramid wherein ev-
ery new image is the result of a Gaussian average of the
preceding image. This type of down-sampling helps pre-
serving the areas of high contrast at the edge of the fish.

B. The image is then normalized, this is done by re-scaling
the colors of the image so that the darkest color sits at 0
and its brightest sits at 255, its maximum.

C. Then a truncating threshold operation is done wherein all
values of the image which are higher than a predefined
value are set to that value. Currently the predefined value
used sits around half the images maximum value, this
“cuts off” the cancer in the image which is not useful for
identifying the fish in the image.

D. The image is then normalized again to utilize the values
freed by the previous threshold operation.

E. The image is denoised using the Non-local means denoi-
sing algorithm [2] with a template window size of 10 pix-
els and a search window size of 21 pixels.

F. Then another threshold operation is performed wherein
all pixel values which are below a predefined value are
set to 0, in this algorithm that predefined value sits at 1/25
of the images maximum value. This helps solidify the
darker parts of the image which now should be the fish.

G. At this point a Canny edge detector [3] is applied to the
image. Here its first threshold is to 0 and its second
threshold to

\[26.1 * e^{0.001+9.43*\text{contrast}} + 10\]

with contrast being defined as the average magnitude of
the 2d vectors formed by the normalized x and y gradient
of a Sobel operation on the image as it is in figure 1F. A
10 pixel black edge is then added to the image to ensure
the proper functioning of the contour detection algorithm
since many of the fish used in this workflow are not fully
in frame.

H. The next step is to “solidify” the edges created by step
G. This is done by two morphological transformations to
the edges, both with a 3x3 pixel kernel. First a dilation
transformation is performed followed by 4 morphological
“close” operations. This results in the edges of the fish
almost always being solid.

I. The last step is to fill the voids of the fish. First the
contour(s) of the fish are found by using Suzukis edge
finding algorithm [7] which are then simplified using the
Douglas-Peucker algorithm [4] with \( \epsilon = \frac{P}{100} \) where P is
the perimeter length of the contour. Then all the outer-
most perimeters are filled on a mask of the original size
of the image before step A had downsampled it. Note that
these perimeters need to be offset by -10 pixels in both di-
rections due to the the 10 pixel border added in step G.

Image labeling algorithms
A. Again the first step is to isolate the red channel of the
image, the resulting image being seen in figure 2A.

B. The image is then masked using the previously generated
fish mask, result displayed in figure 2B.

C. To then get the mask of the cancer a tuned threshold op-
eration is performed. Here everything over \( Q \) is set to 255
in the mask, and everything below it is set to 0 were Q is
defined as:

\[Q = 64 + x*Ic - y*Cst - z*Ib\]

\(x, y \) and \( z \) were the result of external mathematical analy-
sis wherein \( Q \) was manually set. \( Ic \) is the image contrast
Figure 1: Stages in the fish masking algorithm

as defined in the fish finding algorithm step G, Cst is the second Canny threshold as defined in the fish finding algorithm step G and Ib is the image brightness defined as the average brightness of the image. The result is displayed in figure 2C.

D. If the cancer mask is deemed insufficient by the employee they use the correctional facilities to provide additional data to the image, see Developed correction facilities.

E. Metadata is then collected and stored, including settings used by the algorithm as well as the amount of cancer marked.

Note that figure 2D is not the result of the algorithm but a manually made image to better illustrate what parts of the image was marked as cancer (sections outlined in white).

Developed correction facilities

The algorithms are in themselves not enough for acceptable results, for these reasons facilities facilitating the use and correction of the algorithms were developed in parallel to the algorithms. The developed facility comes in the form of a computer program, which lets the employees work on images on a set-by-set basis with no loading between images. The facility has the following features.

A. The ability to manually select the settings for the cancer-thresholding algorithm, intended as the fastest tool in the case that the algorithm fails to correctly label the cancer.

B. The ability to manually select regions to be marked or unmarked as being cancer, intended to be used when manual tuning of the cancer labeling algorithm is not sufficient for good results.

C. The ability and requirement to select the area of interest for the data-gathering algorithm, this is needed due to the common occurrence of false-positives outside the main area of cancer, see Results.
D. The ability to quickly jump between images in the same set.

E. The ability to export the generated data in CSV format.

See Appendix 1 for an labeled image of the program.

RESULTS

After the overall improvement stage was finished, all employees used to the old workflow as well as two other people who were not used to the workflow were timed labeling 20 images. All subjects were first allowed to label a set of images with guidance to familiarize them with the workflow being timed. The results are displayed in table 1.

The accuracy of the cancer-masking algorithm was also tested after the overall improvement stage by the employee most familiar with the new workflow and developed correction facilities. Here, 250 images were analyzed and timed. Out of those 250 images, 169 images had a cancer mask accurate enough to require less than a single second for the subject to correct it, 70 required between one and ten seconds, and eleven required more than ten seconds.

DISCUSSION

Validity

Internal validity
To avoid internal bias the results presented are merely factual, any conclusion drawn about the results are clearly separated into the conclusion section.

One source that might cause a lack of internal validity is the choice of images used in testing; due to the scope of the paper only a small amount of images was used in the tests (the average size of one days images of one set of embryos). It was however discovered throughout development that the difficulty between sets of images varied little, with the only spikes occurring in images deemed as exceptions by the employees, for example images showing contaminants suspended in the solution the fish are kept in such as fibers.

External validity
It’s important to remember that the results presented here was generated with a specific microscope, with a specific light-box and a specific type of prepared embryo. If these conditions are met, the methods described in this paper should perform close to as it is indicated in this paper. The scope does however become even narrower when the low number of employees used during testing are considered; but due to them having varying levels of experience, I believe their results are a good indication of the results that would be gotten in a more thorough study.

**Ethical considerations**

There are two ethical considerations that feel especially important in this paper.

**Automation**

The first consideration lies in the fact that the techniques and methods developed in this paper reduces the required work-hours for analyzing and labeling these types of images. Currently it only reduces the time required at the company that this paper collaborated with (Bioreperia), but in theory it could save time at many more companies. This initially seems like a good thing, but many will also quickly think about the effects these types of time-savings have on the employees at such companies rather than the effect it will have to the companies: will any of the employees be made redundant by the algorithm? The answer is likely no. The employees that are qualified for analyzing these types of images are the same ones that are not only required to spot when the algorithm fails to be accurate, but also to correct it when it happens. This type of qualification means that they are likely to be the same persons preparing the test-series themselves; caring for the embryos, handling the injection and the preparation of the cells. This means the throughput will increase without an increase in cost, therefore making any layoffs unnecessary for a company which wants to grow. When it comes to non-profit organizations the case for this type of automation being a good thing increases even further. For the same investment, more drugs to combat cancer can be developed which is an enormous gain for society as a whole.

**Animal testing**

In any study involving animal testing the question of the ethics of using those animals is a consideration which should not be taken lightly. In this paper, the animals (or more correctly, the embryos) were already part of another study, and no embryos were used solely for the purpose of this paper. The results of the paper could however have consequences for those types of studies in making them easier to do, and therefore it still has to be considered. However, this type of zebrafish study (using the workflow described earlier) is almost universally agreed to be ethically justifiable; the age of the embryos means that they are not even considered animals in terms of animal testing in Sweden (where this paper was written). This in tandem with the proven usefulness of already conducted studies using zebrafish [8, 5, 9] means that the continued use of zebrafish embryos seems justifiable, which in turn would mean streamlining those types of studies, as this study tried to do, is also justifiable.

**Conclusion**

The algorithms and facilities developed in this paper do save time for the subjects used during testing, however with such a small testing population no guarantees can be made that the results presented here can be generalized even to other people working in the same field with the same type of workflow. To guarantee that, a larger study with a bigger and more diverse population would have to be included. I do however believe that the results would generalize if such a study would be performed due to both the recorded accuracy of the cancer masking algorithm as well as the overall time-saving shown regardless of the subjects familiarity with both the old and new workflow.

The results also indicate that a large subset of images can be both prepared and labeled automatically with only a small subset of images requiring time-consuming human intervention. Using the correction-facilities for handling this human intervention resulted in time being saved even in the cases where human intervention was a necessity.

**REFERENCES**


1 Appendix 1

Figure 1: Developed correction facility – Images in current set (A), Tools for selecting area-of-interest and adding and removing automatically generated cancer selections (B), Tool for exporting generated data in CSV format (C), Outline of fish-mask (D), Selected area of interest (E), Manually and/or automatically marked cancer (F), Sliders for manual tuning of cancer-masking threshold (G).