AUTOMATIC CLASSIFICATION OF COPEPOD SPECIES USING DEEP LEARNING

AN UNDERGRADUATE THESIS

Presented to the Faculty of

Bachelor of Science in Computer Science

MSU - Iligan Institute of Technology

Iligan City

In Partial Fulfillment of the

Requirements for the Degree

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

Franco Jigo C. Pacana

May 28, 2019

**Abstract**

**Acknowledgement**

**Contents**

**List of Tables**

**List of Figures**

CHAPTER 1

**Introduction**

**1.1 Research Description**

Invertebrates are animals without a vertebral column also known as backbone or spine. They occur in great numbers living in both fresh and marine waters of lakes, swamps, marshes, rivers and oceans. Familiar examples are crabs, lobsters and their kin, snails, clams, octopuses and their kin, starfish etc. Many fresh and marine water invertebrates, however, including the copepods which constitute the secondary producers of the marine environments and a fundamental step in the trophodynamics of the oceans, are so tiny that they need special attention to collect and observe them. Thus, making them unnoticeable by casual visitors to aquatic habitats (Boehler, 2012). Many studies have been conducted in almost all aspects of copepods from its population, abundance, morphology, taxonomy, diet, diversity, ecology etc. but most of these studies employed conventional techniques especially in identification and classification where manual process is used and expertise is the primary requirement. To address this problem, scientists have started using Image processing as a tool for automatic identification of these species and Artificial Neural Networks for classification.



Figure 1: Copepods under a compound microscope

The digital image processing according to Gonzales and Woods refers to processing digital images by means of a digital computer. Furthermore, it is also a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. Nowadays, it is among rapidly growing technologies and act as a core research within engineering and computer science disciplines.

Image processing along with a system called Artificial Neural Network which is inspired by brain’s structure and function are commonly used in today’s researches. The ANN is considered a nonlinear statistical data modeling tool where the complex relationships between inputs and outputs are modeled or patterns are found (<https://www.investopedia.com/terms/a/artificial-neural-networks-ann.asp>). Many advancements in science and technology including biology, computer vision, speech recognition, machine translation, social network filtering, video games, medical diagnosis and artificial intelligence esp. in robotics use this model to perform specific tasks such as clustering, classification, pattern recognition, etc.(Gill, 2017).

**1.4 Background of the Study**

**1.4.1 Importance of studying copepods**

Copepods are microscopic crustaceans which ranges 200μm to 2mm in total length. They can be found in a large number approximately 60,000 individuals per cubic meter of water. Studying the community structure and abundance of the copepods in relation to their environment is important to evaluate their contribution to mangrove trophodynamics and coastal fisheries. They act as a linking factor between phytoplankton which are the primary producers of the aquatic environments and main food of the copepods and organisms of higher trophic level. They also act as a bioindicator for changes in water quality because their distribution and abundance can be affected by both abiotic and biotic factors such as salinity, temperature, food quantity and quality. Thus, copepods are one of most studied species in both marine and freshwater ecosystem (Leow, 2015). The routine in identification of copepods is very technical, needs taxonomical expertise to do, and takes so much time and effort (Leow, 2015). Hence, a need to develop an advanced method using new technologies today to automate the identification and classification of these samples.

**1.4.2 The Conventional way of Identification and Classification of Copepods**

The identification and classification of copepods which is summarized as: Collection from sampling sites, Preservation, Sorting, Examination and preparation which requires information of their morphology can be very time consuming and may require taxonomic expertise which is not readily available for undergraduate students and even some graduate biologist which does not specialize in copepods. Specific requirements for identification also include the Body shape to characterize the genera and Appendages such as fifth legs for species level. Thus, image processing tools of copepods are very useful for error-less digital recognition and may save up time and energy.

**1.4.3 Latest Advancement in Copepod Research**

Although there is an existing technique such as ZOOSCAN digital imaging system which uses image processing and yields semi-automatic recognition system (Grosjean et. al, 2004) for zooplankton, copepods were only covered in a few categories from the entire zooplankton community (Plourde et. al, 2008). Another technique which uses diffraction patterns as a tool for identification was also conducted by various researchers such as Zavala-Hamz et. al in 1996, Castro-Longoria et. al in 2001, Alvarez-Borrego et. al in 2001, and Castro-Longoria et. al in 2003 but it only caters calanoid copepods. The latest advancement in copepod recognition is in 2015 where Lee Kien Leow and his colleagues used image processing and artificial neural network to produce a computer software where the automatic recognition takes place of eight species of copepods but his technique uses only the conventional square grid lattice in image sampling using MATLAB’s Image processing toolbox R2013a.

**1.4.4 Neural Network for copepod Classification**

Classification methods for image identification systems have been used such as neural network, structural, fuzzy, and transform based techniques for many biological specimens but not with copepods. Artificial Neural Networks have shown promising results in classifying various specimens of insects (Wang et. al, 2012), dinoflagellates by Culverhouse in 1996, metazoans and protozoans by Ginoris et. al in 2007, and many more.

**1.5 Statement of the Problem**

The current problem that Biologists face in copepod research is the conventional way of identification. It would take a lot of time and effort to identify and classify species of copepods without expertise.

**1.6 Research Objectives**

**1.6.1 General Objectives:**

The study aims to utilize image processing techniques in recognition and classification of a copepod sample down to species level.

**1.6.2 Specific Objectives:**

1. Design an application which caters automatic identification of copepods down to species level using image processing.

2. Use ANN algorithm as a tool for classifying the copepods.

4. Use performance evaluation schemes to evaluate the system and compare manual identification versus an automated one.

**1.7 Scope and Limitations of the Research**

The research will focus in developing an application which will identify the copepod species for the users. The number of copepod species will limit based on the copepod species used in previous studies which is Eight species for efficient system evaluation. The system will utilize OpenCV toolbox for image processing and a Convolutional Neural Network using Tensorflow framework.

**1.8 Significances of the Study**

This study will help transition biologists from manual identification to automated process. It will also lessen the workload for any researchers and students which currently studies copepod.

CHAPTER 2

**Review of Related Literature**

**2.1 Future Perspective of real time plankton-imaging-system**

One research studies the future perspective of automating the image analysis of plankton in which the researchers said that the major requirement to achieve any progress in addressing biological diversity in ocean plankton is the high-resolution sensors for imaging field-collected and *in situ* specimens in a non-invasive manner (Culverhouse,2006). They further explained that a large distributed Database in the form of high-resolution 3D rotatable plankton images must be created to attain automatic categorization of species.

The research proposes several advancements in plankton-imaging-system. First, to establish a standard of taxonomic quality images of specimens that will be validated for use in automatic categorisation machines and for other scientific researches. Second, to identify ‘holes’ in taxonomic expertise in basic technology developers. Hence, through validation of the database contents, experts for the taxonomy of specific groups would become more ‘visible’ and approachable to the wider community. Lastly, to disseminate images of unknown species to experts across the world, through the World Wide Web. This would provide the widest possible access to expert taxonomic opinions (Culverhouse,2006).

All of the stated proposed advancements were made entirely for all plankton species in the ocean but not particularly for copepods.

**2.2 ZOOSCAN digital imaging systems**

ZOOSCAN is a system for identifying and counting plankton net samples. It describes image processing and semi-automatic recognition using various machine learning methods of plankton species.

The system showed an accuracy level 85% in classifying taxa of plankton species and a faster than manual sample handling thanks to the new combined algorithm known as discriminant vector forest (Grosjean,2004).

Although the system showed promising results, copepods were not entirely studied in this research.

**2.3 Diffraction patterns as a tool for recognition**

Diffraction patters were used for 3 species of Calanoid copepods. The images were digitized, binarized and edited to remove debris present and to simulate different degrees of segmentation. A total of 28 segmented images of the three species were generated to get their digital fourier transform of diffraction pattern. The method has a lot of potential but it needs more research to develop an automatic system for plankton identification (Zavala-Hamz,1996).

**2.4 Circular Harmonic Filters**

CHF’s are group of filters which most widely used method of accomplishing rotation invariance. The filters have several advantages. First, the correlation plane is invariant to rotation, and second, they are proven sensitive to noise. The main disadvantage of the use of CHF’s is that they lose their power of discrimination if the expansion center, is not previously selected (Zavala-Hamz,1997).

Researchers found out that the symmetry of genus Acartia permitted discrimination to the species and sex levels, while the asymmetry of the genus Calanus permitted discrimination only to the generic level. The differences among organisms of different sex of the genus Calanus could not be detected by these particular CHF’s.

The researchers suggested that more research should be carried out to implement an automated optodigital system to identify and count marine plankton.

**2.5 Novel Study in Copepod Automatic copepod identification and Classification**

Another study about the Identification and Classification of copepods by L.K. Leow et.al in 2015 uses Image processing and neural network. The researchers used Matlab’s image processing software. They used eight species of copepods namely *Acartia spinicauda, Bestiolina similis, Oithona aruensis, Oithona dissimilis, Oithona simplex, Parvocalanus crassirostris, Tortanus barbatus and Tortanus forcipatus*. The researcher used 240 samples which were then divided into three sets, the training set (168 samples, or 70% of samples), validation set (36 samples, 15%) and testing set (36 samples, 15%). The data from the training set were used for network training; the validation set for measuring network generalization and terminating training before overfitting; and the testing set for independent measure of network performance during and after training. The overall approach demonstrated not only a fast and automated technique for copepod identification and classification but also an accuracy rate of 93.13%. The performance evaluation of the system was evaluated using MSE or the Mean Square Error and Confusion matrices. The other 160 independent samples (20 samples from each species) were used for system performance evaluation. The trained network was simulated using the testing data as input and the output was then compared to the predicted data and recorded in a confusion matrix.

His approach demonstrated an overall classification accuracy of 93.13% (100% for A. spinicauda, B. similis and O. aruensis, 95% for T. barbatus, 90% for O. dissimilis and P. crassirostris, 85% for O. similis and T. forcipatus).

Although his work showed a promising result, it still needs improvements especially in feature extraction method used and the selected features used.

**2.6 Artificial Neural Networks**

There are many types of neural networks such as Hopfiled Neural Network,Radial Basis Function Neural Network, Probabilistic Neural Network, Convolution Neural Network,Fuzzy Neural Network but one of the most famous used for image segmentation is the Feed Forward Neural Network (Z. Shi and L. He, 2010). Shi and He also noted in their study (“Application of Neural Networks in Medical Image Processing”) that the said network is less sensitive to the selection of the training sets than the Maximum Likelihood classifier.

Neural networks have been utilized in many fields of science especially in image or samples of species’ detection, recognition and classification. It is used in Insect classification by J. Wang et.al in their study “A new automatic identification system of insect images at the order level”. It was also utilized in other species or groups organisms such as Macroinvertebrates by S. Kiranyaz et.al in their study “Classification and retrieval on macroinvertebrate image databases using evolutionary RBF neural networks”; Algae by P. Coltelli et.al in their study “Water monitoring: automated and real time identification and classification of algae using digital microscopy” in 2014; Fishes in the study “Fish recognition based on robust features extraction from size and shape measurements using

neural network” by MK Alsmadi et.al in 2010 and other groups of organisms such as protozoa and metazoa (Y.P. Ginoris et.al, 2007), dinaflagellates (PF Culverhouse et.al, 1996), etc. However only Leow and his colleagues have use neural network for copepod classification.

In Leow’s study, A two-layer (hidden and output layer) feed-forward network was trained using a back-propagation algorithm which is based on ten neurons at the hidden later and eight neurons at the output layer. They used a total of 240 sample images for training set with 30 for each class. They obtained seven selected features of each species which is used as input data presented to the input nodes of the network from the training set, whereas eight desired output classes were defined by the target data. The results showed 93% correct classification from the confusion matrix of all 240 samples in the training, validation, and testing sets.

CHAPTER 3

**Theoretical Framework**

This chapter will discuss the basic concepts and framework that is used in implementing Image Processing and Deep learning for copepod Identification and Classification.

**3.1 Noise reduction on images**

Signal processing devices which is categorized to analog and digital are vulnerable to noise. This noise often can be, random or white noise with an even level of frequency distribution, or frequency dependent noise introduced by a device's mechanism or signal processing [algorithms](https://en.wikipedia.org/wiki/Algorithm).

In images, noise is often acquired when one shoots a photo at night and/or with high ISO and with wide exposure latitude. To remove noise and attain a clear image of the copepod species, a filter must be applied to the image (Attila, n.d).

Although noise must be removed from the images, the famous Convolutional Neural Network which is also used in this paper as a comparison to the standard simple Feed Forward Neural network utilize noise during training to so that the Network can still classify objects even if it has Noise. In fact, “Adding noise to an underconstrained neural network model with a small training dataset can have a regularizing effect and reduce overfitting” (Brownlee,2018).

**3.2 Edge Detection**

Edge detection is the process of identifying and locating sharp discontinuities in an image. The discontinuities are abrupt changes in pixel intensity which characterizes boundaries of objects in a scene (Maini, 2009).

Edge detection methods can be grouped in to two main categories. First, Gradient which detects edges of the image by searching for the maximum and minimum in the first derivative of the image. Second, Laplacian methods which looks for zero crossing in the second derivative of the image (Igbinosa, 2013).

Most used edge detection techniques are the Sobel operator, Robert Cross operator, Prewitt detection, and Canny operator. The Sobel operator is a pixel-based edge detection which can detect edges by calculating partial derivatives in 3 x 3 neighborhood. The Robert Cross operator performs a simple and quick 2-D spatial gradient measurement on an image and it also consists of a pair of 2x2 convolution kernel. Prewitt detection is similar to the Sobel operator and it is used for detecting vertical and horizontal edges in images (Gonzales, 2002). Canny edge detector is the most rigorously defined operator and is widely used dues the following reasons:

1. Good Detection – low error rate
2. Good Spatial localization
3. Good Response rate

**3.3 Segmentation**

Image segmentation technique is used to partition an image into meaningful parts having similar features and properties. The main aim of segmentation is simplification i.e. representing an image into meaningful and easily analyzable way (Kaur, 2014).

Image segmentation can be categorized in to two types, one is Discontinuity detection-based approach where an image is segmented into regions via discontinuity. Two, Similarity detection-based approach where an image I segmented based on similarity.

Image segmentation techniques are as follow:

1. Thresholding methods

- This method divides the pixels with respect to their intensity and comes in three types. The Global thresholding which is done using any appropriate threshold value and the value T is constant for the whole image. Variable thresholding which is done by varying the value T across the image. Multiple thresholding has more than one threshold values i.e T0 and T1.

b. Edge Based Segmentation Method

- based on the rapid change of intensity value in an image because a single intensity value does not provide good information about edges.

c. Region Based Segmentation Method

-are methods which segments the image into various regions having similar characteristics.

d. Clustering Based Segmentation Method

- are methods which segment the image into clusters having pixels with similar characteristics.

e. Watershed Based Methods

- uses the concept of topological interpretation. It’s like Any grayscale image can be viewed as a topographic surface where high intensity denotes peaks and hills while low intensity denotes valleys.

f. Partial Differential Equation Based Segmentation Method

- These are appropriate for time critical applications.

g. Artificial Neural Network Based Segmentation Method

- simulate the learning strategies of human brain for the purpose of decision making.

**3.4 Feature Extraction for plankton images**

One of the most important method in Artificial Intelligence is Feature Extraction. It consists methods to extract the most relevant features of an image and assign it into a label. In image classification, the crucial step is to analyze the properties of image features and to organize the numerical features into classes. In other words, an image is classed according to its contents (Medjahed, 2015).

Many techniques of Feature Extraction come in major categories such as Color features, Texture features and Shape features.

**3.5 Feature Selection**

Feature selection is different from Feature Extraction. The latter creates new features from function of the original features, while the former returns a subset of the features.

Three major selection methods are Filter, Wrapper, and Embedded methods. The Filter methods apply statistical measure in order to assign a scoring to each available feature. Each feature is then ranked based on their scores and selected to be kept or to be removed. The wrapper method considers selection as a search problem, where different combinations are prepared, evaluated and compared to other combinations. While the Embedded methods finds the best feature for the model while the model is being created. The most common type of embedded feature selection methods are regularization methods.

**3.6 Artificial Neural Network training**

The ANN is a model inspired by the human brain and its functions. The brain’s ability to learn new things and adapt to the changes in the environment is the key to this Network. It is a computational learning system that uses a network of functions to understand and translate a data input of one form into a desired output, usually in another form. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules (Kukreja, 2016).

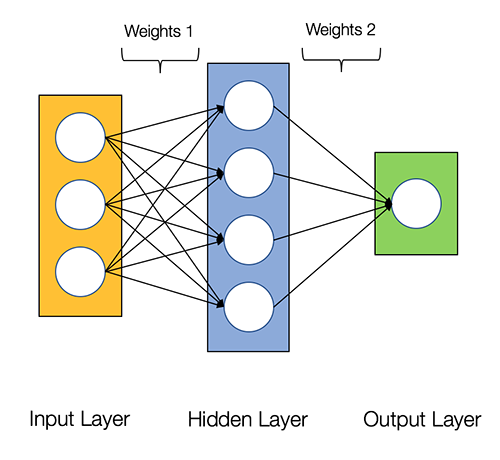
The following are the ANN’s strengths and limitations:

1. Adaptive learning
2. Faster than Normal Brain speed in processing information
3. Parallel operation
4. Fault Tolerance
5. Self-organization
6. ANN’s output is not clear compared to a normal program.
7. ANN can be used for applications that has unclear data
8. ANN cannot be utilized when the input and Output are known already.

The basic Neural Network consist of the following components:

1. An input layer, x
2. An arbitrary number of hidden layers which usually depends on the task at hand
3. An output layer, y
4. A set of weights and biases between each layer, W and b
5. A choice of activation function for each hidden layer, **σ.**
   * 1. **Feed Forward Neural Network Architechture**

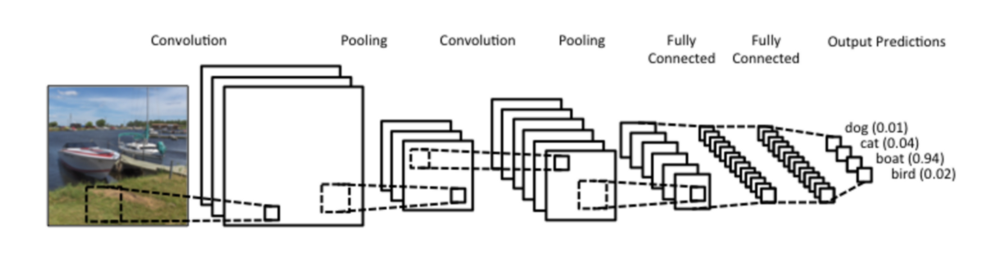
Feed Forward Neural Networks are network of models known as Multi-lasyered Network of Neurons (MLN). These type of models are called feedforward because the information only travels forward in the neural network, through the input nodes then through the hidden layers and finally the output nodes (Kumar,2019).



**Figure 2.1 Standard Architecture of a Feedforward Neural Network**

* + 1. **Convolutional Neural Network**

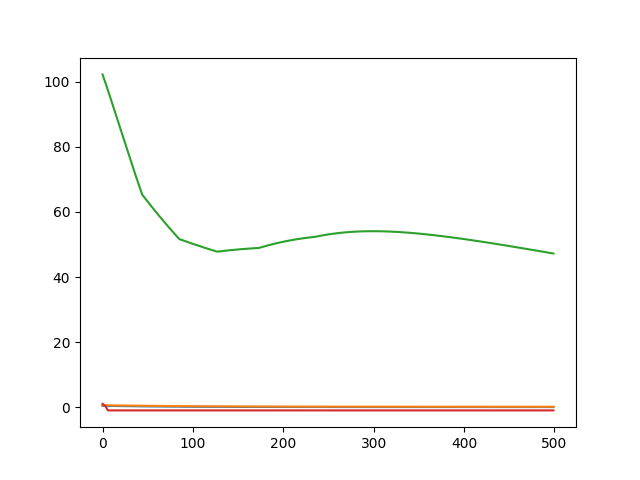
The CNN is similar to a Multilayer Perceptron Network. The main differences between the two are that what the network learns, how they are build and the purpose they are mostly used for. CNN was inspired from biological processes in which they resemble the visual cortex present in an animal. They are widely used in computer vision due to large numbers of researches which produce high performance models in various test cases (Gandhi,2018).



**Figure 2.1 Standard CNN Architechture**

**3.7 System Evaluation**

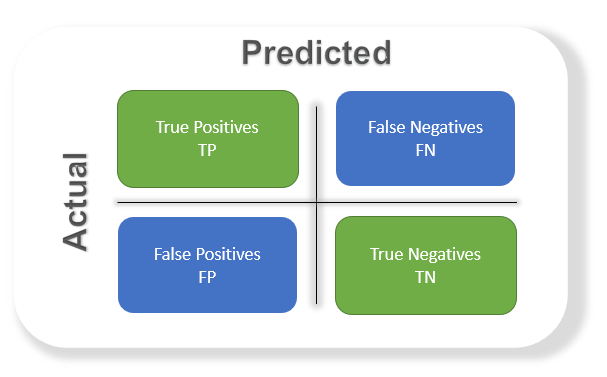
It is normal to evaluate the System in every research to measure the efficiency of that system. The MSE or Mean Square Error is one of the validation techniques used to evaluation the result of the neural network. The mean(average) magnitude of the squares of the error*:*i.e., the distance between the model's estimate of your test values and the actual test value (Rostampour, 2013).



**Figure 3.3 Mean Square Error**

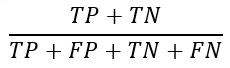
The Confusion matrix is an evaluation technique for the classification algorithm. It can give you the idea of where and how is your classification algorithm right and what errors does it have.

The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix (Brownlee, 2016).



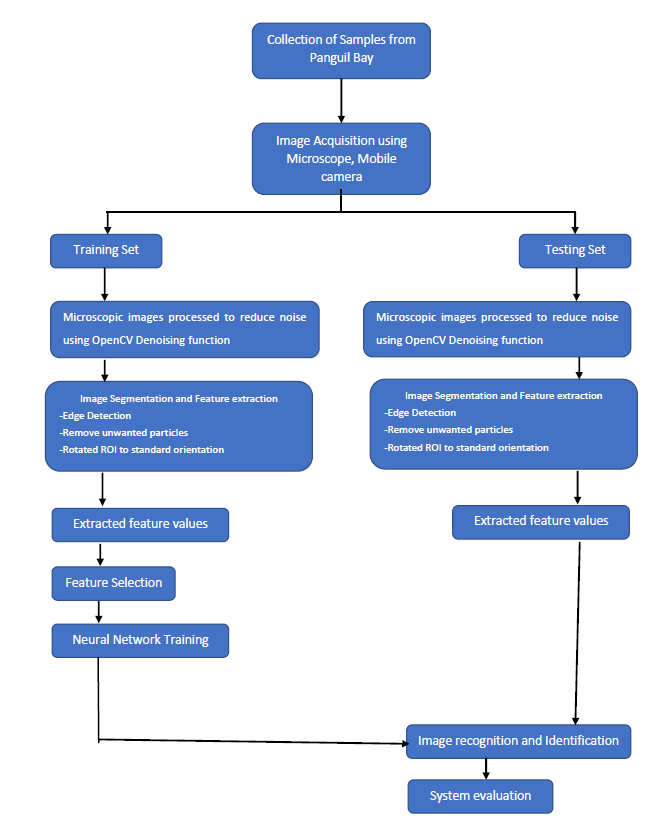
**Figure 3.4 Confusion Matrix**

The Accuracy is the most commonly used metric to judge a model and is actually not a clear indicator of the performance. The worse happens when classes are imbalanced.



Chapter 4

**Methodology**



**Figure 4.1** Methodology Flowchart

**4.1 Sample Collection and Image Acquisition**

Five genera of marine copepods commonly encountered in mangrove waters were examined: Acartia (A. spinicauda), Bestiolina (B. similis), Oithona (O. aruensis, O. dissimilis and O. simplex), Parvocalanus (P. crassirostris) and Tortanus (T. barbatus and T. forcipatus). Copepods were sampled from four stations from the upper estuary in the Panguil bay to near shore waters of Marandin Lala Lanao Del Norte. Horizontal plankton tows (0.5-1 m depth) using paired 45 cm-diameter bongo nets (180 μm) were made and collected plankton were preserved in buffered 10% formaldehyde. In the laboratory, collected copepods were sieved through stacked Endecott sieves of 1,000 μm, 500 μm, 250 μm and 125 μm mesh sizes, and the sieved fractions were preserved in 80% alcohol in individual vials for a long-term preservation.

**4.2 Image acquisition**

Specimens of copepod will be randomly pipetted onto a microscope slide from the preserved samples and each identified to species level under a compound microscope.



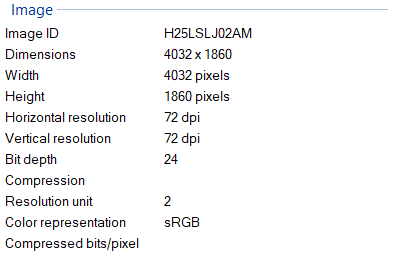
# Figure 4.1 SB-BC-200 Microscope

To enable the dorsal aspect of the identified copepod to be imaged, often the copepod body had to be rotated. Body rotation could be easily achieved by first placing two short nylon fishing lines (0.36 mm diameter) on either side of the specimen and gently moving a cover slip placed over them by using the tip of the index finger. The desired view of the copepod body will be acquired by a Mobile camera – Samsung A50 with triple camera on its back - the main 25MP PDAF f/1.7 snapper is joined by an 8MP fixed-focus, f/2.2 ultra-wide and a 5MP, fixed-focus, f/2.2 depth sensor.



**Figure 4.2 Samsung A50**

The images taken from the mobile camera has the following specification and properties.



**Figure 4.3 Image Specifications**

**4.3 Image Processing**

**4.3.1 Image Pre-processing**

The Images were taken from the Microscope as a raw image with a Circular frame because of the Microscope’s Ocular Lens. It is cropped out using a manual Copping algorithm in OpenCV. Images were then converted to binary using Otsu’s binarization method with proper threshold. Images were next converted to 2D grayscale image using OpenCV cvtColor() function. Then noise were removed using morphologyEx() function.

**4.3.2 Image Segmentation**

Image segmentation was done using the marker-based watershed segmentation algorithm by OpenCV where the regions with one color, background region, or regions with another color, or unsure regions were labelled/marked with 0 beforehand. After doing so, apply the watershed algorithm then the markers will be updated and the boundaries will have a value of -1.

The image were converted to binary image with appropriate threshold and the borders will be cleared. The holes that occurred during the conversion from grayscale to binary were filled using imfill() in OpenCV. Particles that are less than 50000 pixels will be excluded to ensure that only copepods will be segmented in the image.

The orientation represented by the angle between the x-axis and the major axis of the ellipse that has the same second-moments as the region of interest (ROI) will be obtained using Contour properties function in OpenCV. Image rotation was done manually.

**4.3.3 Feature Extraction**

The ROI of the copepod were cropped by getting the coordinates of the boundary of copepods.

Features were extracted from the shape descriptors represented by the binary images of the ROI using Contour properties function in OpenCV. The measurement to be taken are area, perimeter, aspect-ratio, equivalent diameter, extent, convex area, solidity, major axis length, minor axis length and eccentricity.

1. Contour Area - gives the area of the region.
2. Contour Perimeter - gives the perimeter of the region.
3. Contour Aspect-Ratio - ratio of width to height.
4. Equivalent Diameter - equivalent diameter of the circle with same as area as that of region.
5. Convex Area - gives the area of the convex hull.
6. Solidity - contour area / convex hull area.
7. Major Axis length - gives the length of major axis
8. Minor Axis length - gives the length of minor axis
9. Eccentricity - gives the eccentricity of ellipse

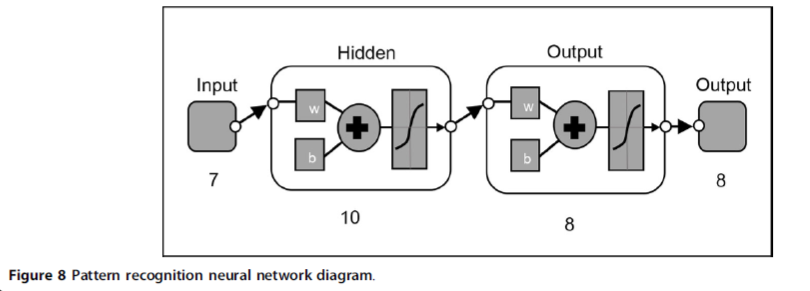
**4.4 Feature Selection**

To avoid overfitting in the Neural Network training and to increase performance, not all the 10 extracted features were used. The extracted features were evaluated to make sure that only significant features will be selected to classify the copepods into their respective taxa.  Recursive Feature Elimination with Logistic Regression was used to aid the selection of the most useful top 6 features. In order to visualize how well a selected feature clustered the specimens in the training set into the eight classes (species), 2D and 3D scatter plots were graphed with different combinations of features as the axes.

**4.5 Artificial Neural Network Training**

**4.5.1 Forward Feed Neural Network Training**

A Feed Forward Neural Network was used as the pattern recognition tool to classify the extracted features values into the nine classes (species). The architecture of the ANN is a two-layer feed-forward network with sigmoid hidden (ten nodes) and output (nine nodes) neurons and the network were trained with scaled conjugate gradient backpropagation. A total of 270 sample images were used in the training set with 30 images from each class. The input data for the input nodes of the network have six selected features of each specimen from the training set, whereas the target data defined eight desired output classes. The 240 samples will be divided into three sets, the training set (168 samples, or 70% of samples), validation set (36 samples, 15%) and testing set (36 samples, 15%). The data from the training set was used for network training; the validation set for measuring network generalization and terminating training before overfitting; and the testing set for independent measure of network performance during and after training. The performance of the network training was evaluated using Mean Square Error (MSE) and confusion matrices. The training stopped when the MSE of the samples in the validation set started to increase indicating that the network generalization stopped improving. The network was trained several times to get the trained network with best performance. Another 160 independent samples (20 samples for each species) was used for system performance evaluation. The trained network was simulated using the testing data as input and the output was then compared to the predicted data and recorded in a confusion matrix.



**Figure 4.4** Pattern recognition neural network diagram

**4.5.2 Convolutional Neural Network Training**

The Convolutional Neural Network was implemented through the following steps.

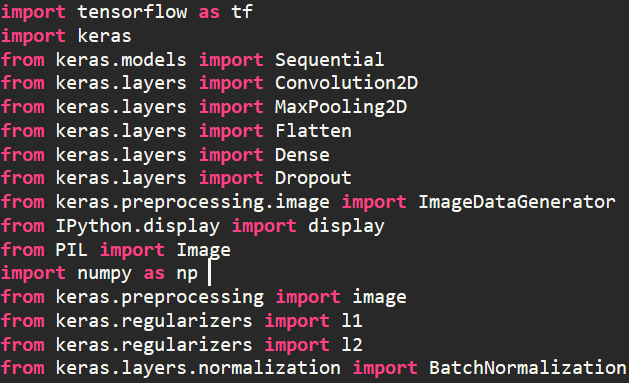
1. **Importing the libraries and splitting the Dataset**

**TensorFlow** is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications

**Keras** is one of the leading high-level neural networks APIs. It is written in Python and supports multiple back-end neural network computation engines.

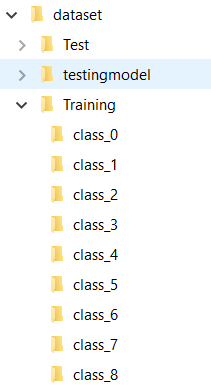
**Python Imaging Library** (abbreviated as **PIL**) (in newer versions known as Pillow) is a [free](https://en.wikipedia.org/wiki/Free_and_open_source_software) [library](https://en.wikipedia.org/wiki/Library_(computing)) for the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)) that adds support for opening, manipulating, and saving many different image file formats.

**NumPy** is a library for the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)), adding support for large, multi-dimensional [arrays](https://en.wikipedia.org/wiki/Array_data_structure) and [matrices](https://en.wikipedia.org/wiki/Matrix_(math)), along with a large collection of [high-level](https://en.wikipedia.org/wiki/High-level_programming_language) [mathematical](https://en.wikipedia.org/wiki/Mathematics) [functions](https://en.wikipedia.org/wiki/Function_(mathematics)) to operate on these arrays.



**Figure 4.5 Libraries**

In our case, the dataset is already split into three parts. The training set has 30 images each of the class/species while the test set has 20 images of each class. The dataset was stored and arrange in the following manner. Another folder was made ‘testing\_model’ for the testing of the CNN performance later after the training.



**Figure 4.6 Dataset Directory**

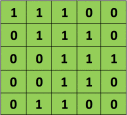
1. **Building the CNN**

There are three primary parts in implementation of the CNN.

1. Convolution
2. Polling
3. Flattening

The primary purpose of **Convolution** is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data.

Since every image can be considered as a matrix of pixel values. Consider a 5 x 5 image whose pixel values are only 0 and 1:



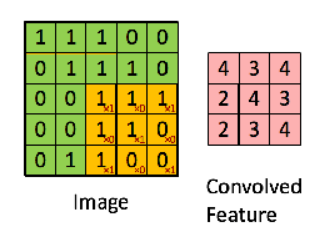
**Figure 4.7 5x5 matrix**

Also, consider another 3 x 3 matrix as shown below:



**Figure 4.8 3x3 matrix**

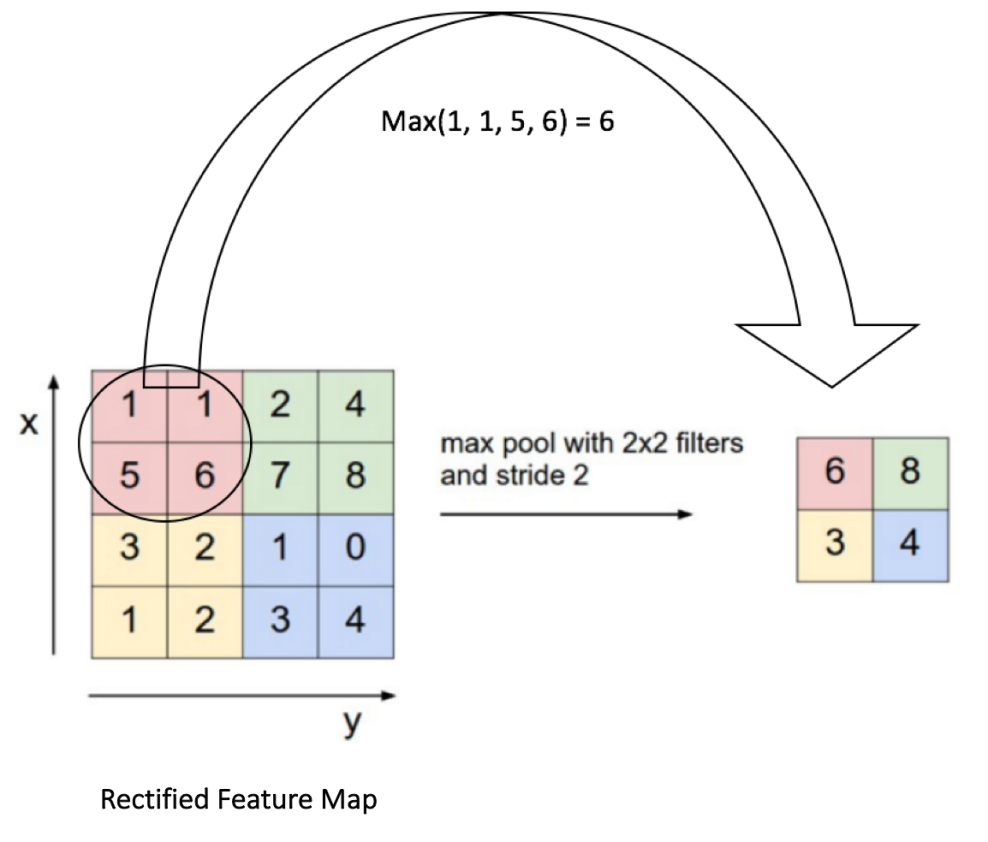
The Convolution of the 5 x 5 image and the 3 x 3 matrix can be computed as shown in the animation in **Figure 4.9** below:



**Figure 4.9 Feature map**

The obtained matrix is also known as the feature map. An additional operation called ReLU is used after every Convolution operation.

Pooling reduces the dimensionality of each feature map but retains the most important information. In case of Max Pooling, we define a spatial neighborhood (for example, a 2×2 window) and take the largest element from the rectified feature map within that window. Instead of taking the largest element we could also take the average (Average Pooling) or sum of all elements in that window. In practice, Max Pooling has been shown to work better.



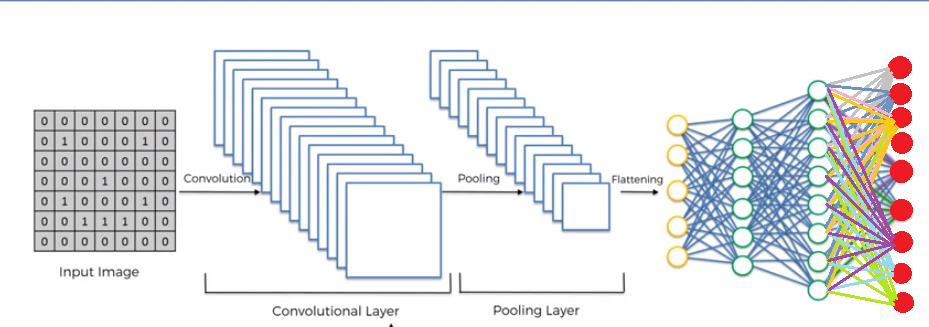
**Figure 4.10 Rectified Feature Map**

**Flattening** is where the matrix is converted into a linear array so that to input it into the nodes of our neural network.

1. **Full Connection**

Here we have made 2 layer neural network with a **Softmax function** as an activation function for the last layer as we need to find the probability of the object being an object belonging to what species.

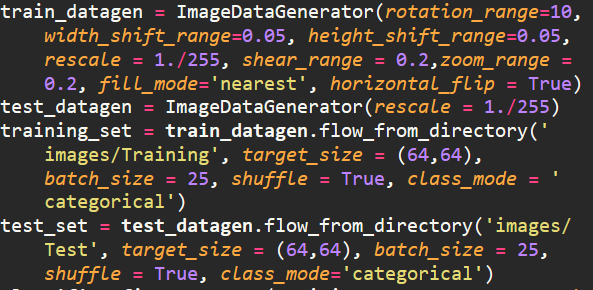
So our network is just like the diagram below but with 9 outputs representing the 9 species of Copepods.



**Figure 4.11 Fully Connected CNN**

1. **Data Augmentation**

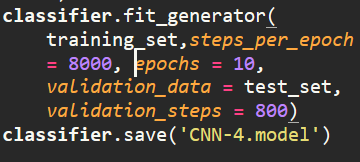
To compensate for a small number of images , Data augmentation was used so that we can reduce overfitting on models, where we increase the amount of trainingdata using information only in our training data.



**Figure 4.12 Data Augmentation**

1. **Training the Network**

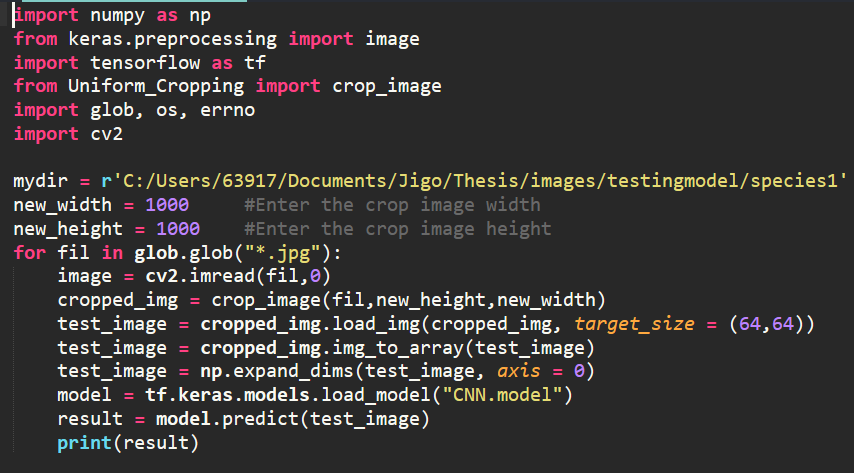
We trained the network through using tensorflow-gpu installed through Anaconda.



**Figure 4.13 Training the CNN**

1. **Testing**

The trained network will be used to predict 20 independent images from each species. The images the folder will be iterated and each image will be cropped to fit the desired size for the CNN model.



Chapter 5

**Results and Discussion**

* 1. **The dataset**

The images for CNN were quite noisy due to salt water and other unwanted particles. In contrast, noise was suppressed as much as Possible for the Forward Feed Neural Network. Aside from that, images taken using the Samsung A50 has blurry areas such as the antennae and the lower appendages of the Copepods.



**Figure 5.1 CNN Dataset**

The Dataset for Feed Forward Neural Network were preprocessed using OpenCV. The initial Image for FFNN is show in Figure 5.2. Notice the Ocular lens of the Microscope is very conspicuous in the image so it is copped out just like in Figure 5.3.



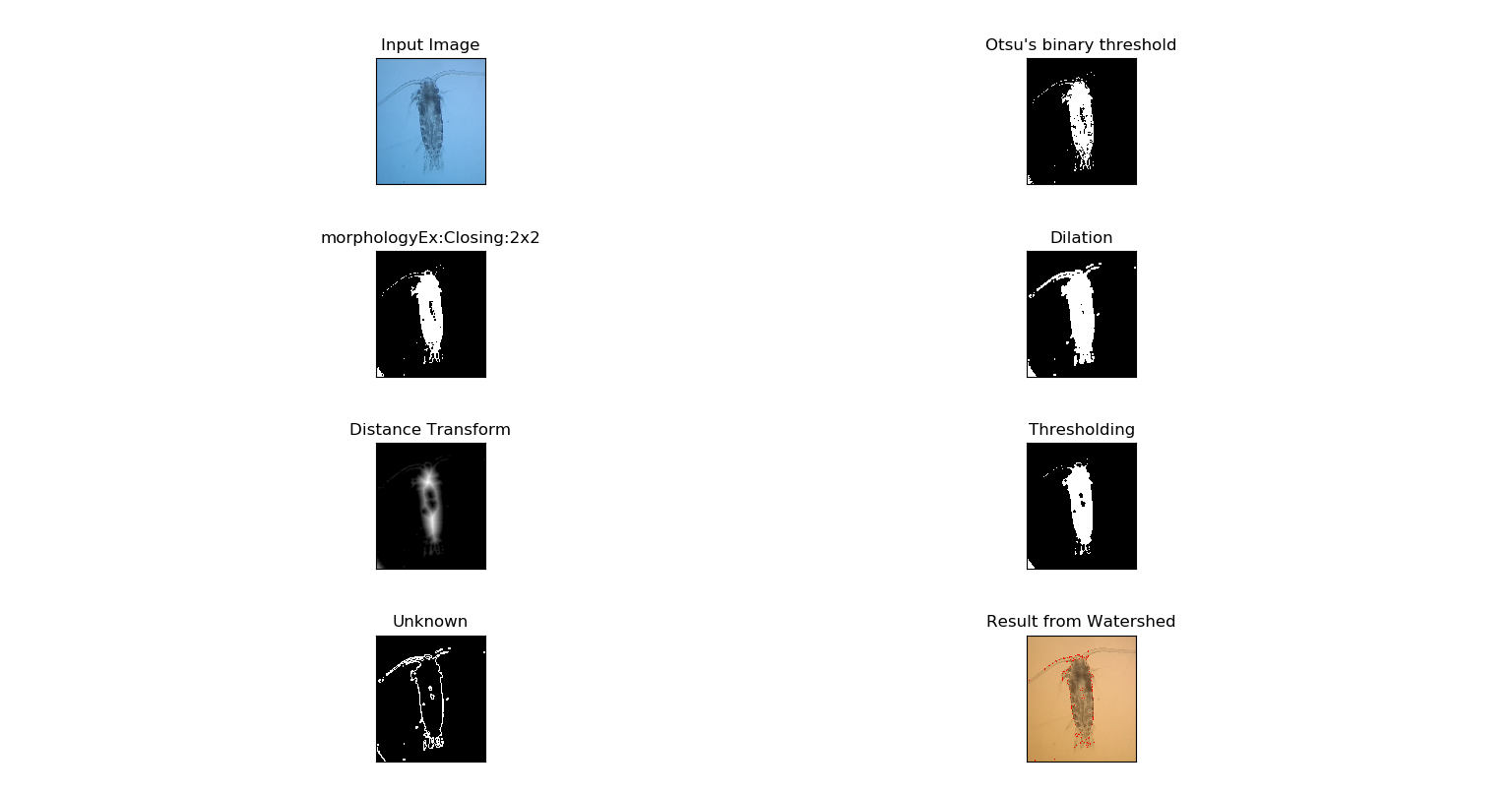
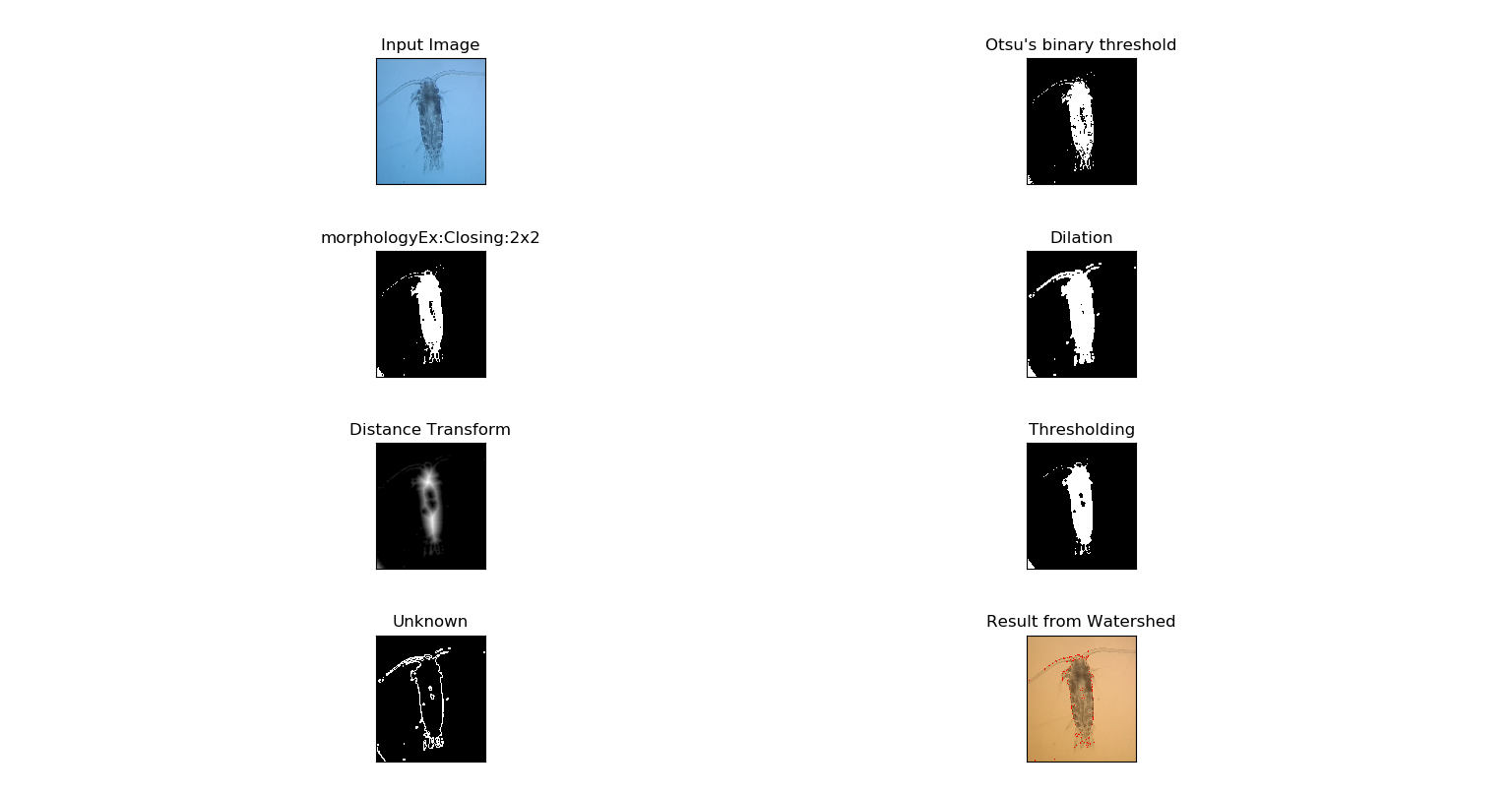
**Figure 5.2 Raw Image of Species 3 under a Microscope**



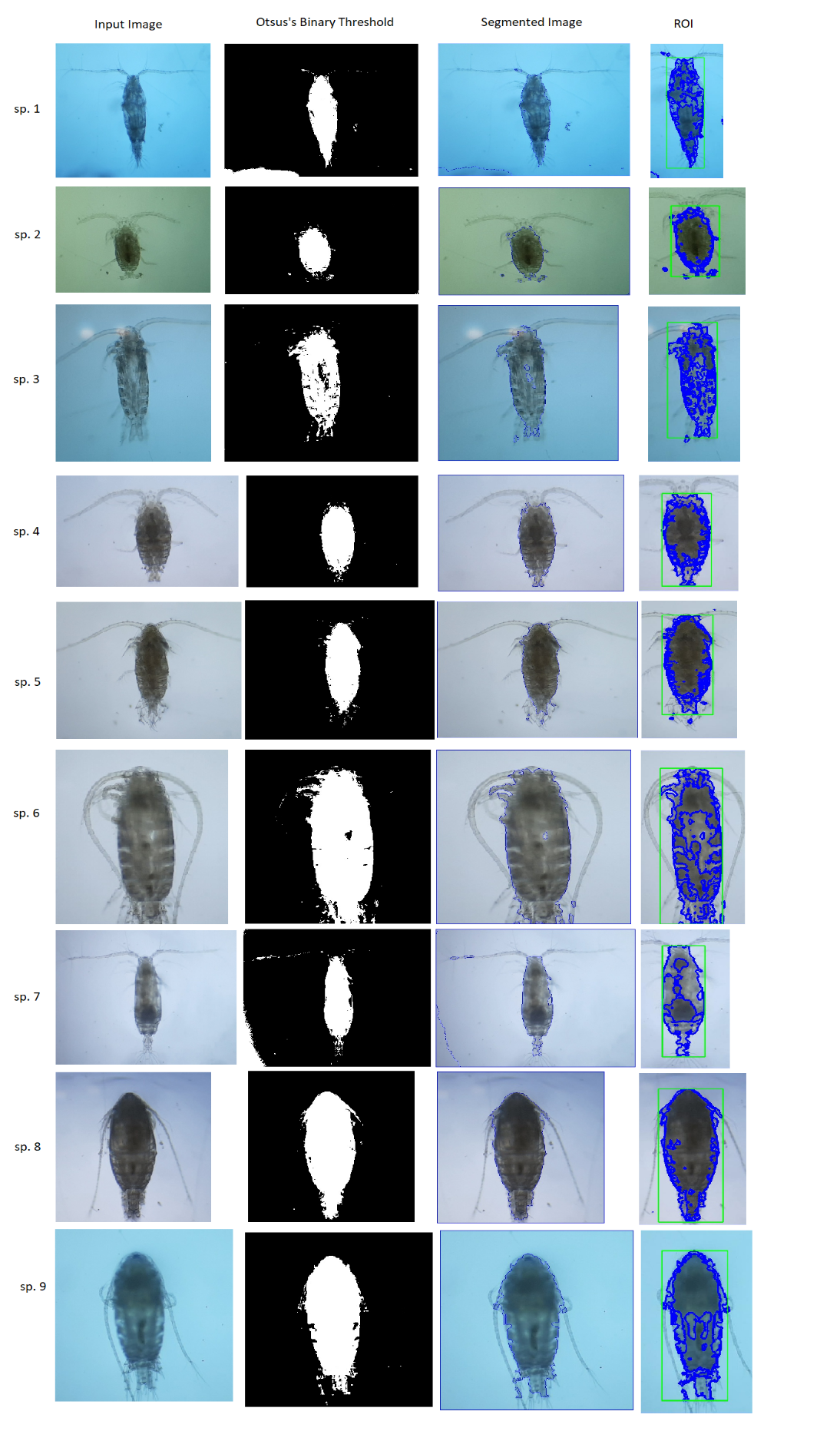
**Figure 5.3 Cropped Copepod Image**

**5.2 Preprocessing of Data**

The cropped images were then subjected for segmentation using OpenCV’s Watershed Algorithm. Species 3 watershed Summary of results are shown in Figure 5.4.



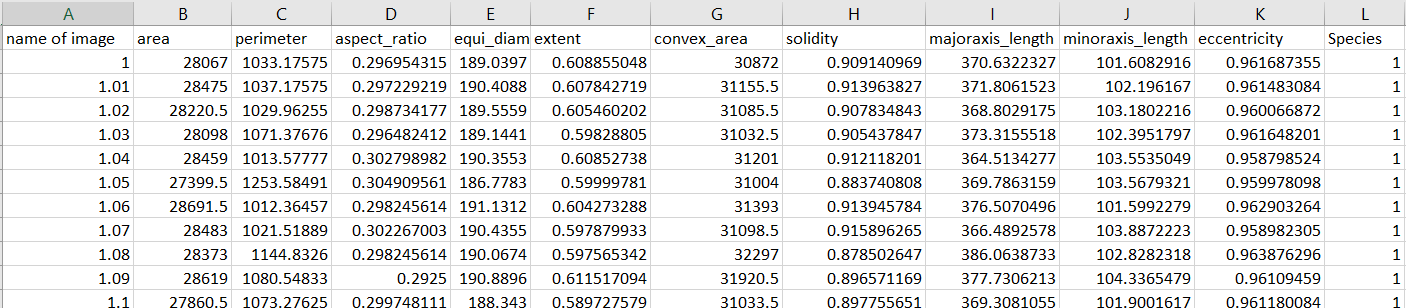
**Figure 5.4 Summary of Watershed Algorithm**



**Figure 5.5 Preprocessed Images of 9 Species**

* 1. **Feature Extraction**

A total of 10 features were extracted from each image via openCV’s contour feature extraction and then Record on Microsoft Excel and converted to CSV file.



**Figure 5.6 Extracted Features**

* 1. **Feature Selection**

The extracted feature were subjected for selection through Scikit learn’s Recursive Feature Elimination with Logistic Regression.