

Identifying Individual Clown Fish

Xiao Liu



Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2013

Abstract

The Fish4Knowledge project has collected millions of tropical reef fish images using fixed underwater cameras. One interesting and common species that we considered in this project was the clown fish which is also known as the *Amphiprion clarkii*, which we have many image resources. Because of the resident characteristic of clown fishes, we are likely to detect the presence of the same individual for many time in a period, which leads to the requirement for identifying the true the number of individuals in a given camera fish detections.

A large amount of previous work has been done in fish detection and extraction from various backgrounds, and a huge number of fish images are stored in the database from which we collected clown fish observations. From the observation of different individuals, despite that each species has a unique colour, shape and patterns on their body, we can see that there are some subtle differences between individuals, mainly in the placement and size of stripes and the appearance colour. To address this identification problem, we present presents some new features to represent fishes. In order to reduce the dimensionality and explore some useful features, 4 feature extraction and selection techniques were adopted. Then we used K-means to solve the classification problem to identify individuals and estimate the true number. Results show that based on the original feature vector containing 3091 attributes the correlation-based feature selection method (CFS) selects a feature subset which has the best clustering performance and it can always estimate a reasonable number of individuals for all experiment datasets, which the number is approximate with the true class numbers.

Acknowledgements

I would like to thank my supervisor Robert Fisher for his patient help and valuable suggestions. He gave me a lot of support during these months. It is my honer to be a student of him. I am also grateful to all members of the Fish4Knowledge group, especially Phenix Xuan Huang, for providing the data and help.

I also would like to thank Micheal Rovatsos who helped me a lot in the progression to my dissertation last year.

Last but not least, I want to thank my family for their endless support and encouragement.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Xiao Liu)

Contents

List of Figures	vi
List of Tables	viii
1 Introduction	1
1.1 Motivation	1
1.2 Project Objectives	2
1.3 Outline	3
2 Background	4
2.1 Literature Review	4
2.1.1 Fish detection	4
2.1.2 Fish recognition	5
2.1.3 Literature Summary	8
2.2 Data Description	8
2.2.1 Data Collection	8
2.2.2 Ground Truth	9
3 Fish Description and Feature Selection	13
3.1 Features	13
3.1.1 Color Ratio	13
3.1.2 Length Ratio	19
3.1.3 White stripes area ratio	21
3.2 Dimensionality Reduction and Feature Selection	22
3.2.1 Principal Component Analysis	22
3.2.2 Feature Selection	25
3.3 Summary	32

4	Classification	34
4.1	Classifier	34
4.1.1	K-means	35
4.1.2	Distance metric	37
4.1.3	Implementation	37
4.2	Model selection	37
4.2.1	Akaike information criteria (AIC) and Bayesian information criteria (BIC)	38
4.2.2	Clustering evaluation indices	39
4.2.3	Model selection algorithm	41
4.2.4	Model selection result	42
4.3	Cluster Validation	45
4.3.1	Result of evaluation	48
4.4	Summary	49
5	Evaluation	50
5.1	New Datasets	50
5.1.1	New Fish Datasets	50
5.1.2	Features for New Fish Datasets	51
5.2	Feature Visualisation	51
5.2.1	Isomap	52
5.2.2	Feature visualisation	52
5.3	Evaluation	56
5.3.1	Model Selection Results	56
5.3.2	Clustering Performance	57
5.3.3	Summary	65
6	Conclusion and Future Work	66
6.1	Conclusion	66
6.2	Future Work	67
A	The description of 19 feature types	68
	Bibliography	70

List of Figures

2.1	<i>The site map of NPP-3</i>	9
2.2	<i>An example of fish image and the orientation result</i>	9
3.1	<i>Comparison of two clown fish</i>	14
3.2	<i>The illumination affects the colour of the fish reflected</i>	14
3.3	<i>Fish image and different parts of the fish; (a) is the orientated fish image; (b) is the part of extracted white stripes; (c) is the chromatic body of the fish; (d) the top part of the chromatic body; (e) the bottom part of the chromatic body; (f) the front part of the chromatic body; (g) the posterior part of the chromatic body</i>	15
3.4	<i>Histogram of the color ratio between chromatic body and white stripes</i>	16
3.5	<i>Histogram of the color ratio of the top part and the bottom part of the chromatic body</i>	17
3.6	<i>Color ratio of the front part and the posterior part of the chromatic body</i>	18
3.7	<i>Average RGB colour ratio between the chromatic body and the white stripes</i>	19
3.8	<i>Average RGB colour ratio between the top part and the bottom part of the chromatic body</i>	20
3.9	<i>Average RGB colour ratio between the front part and the posterior part of the chromatic body</i>	20
3.10	<i>Length ratio calculation</i>	21
3.11	<i>Area ratio</i>	22
3.12	<i>plot of PCA-w</i>	23
3.13	<i>plot of PCA-sep</i>	24
3.14	<i>The flow chart of two feature selection models; (a)is the filter selection model; (b) is the wrapper selection model</i>	26

3.15	<i>Increasing tendency of the clustering accuracy with the growth of dimensionality of CFS</i>	29
3.16	<i>Changing of the clustering accuracy with increasing number of features</i>	32
4.1	<i>The procedure of the K-means algorithm</i>	36
4.2	<i>Model Selection Results for PCA-w Features (4 fish individuals) . . .</i>	42
4.3	<i>Model Selection Results for PCA-sep Features (4 fish individuals) . .</i>	43
4.4	<i>Model Selection Results for CFS Features (4 fish individuals)</i>	43
4.5	<i>Model Selection Results for SPEC Features (4 fish individuals)</i>	44
4.6	<i>Clustering performances of the dataset with 4 fish individuals</i>	49
5.1	<i>Feature Visualisation for 4 Fish Individuals' Dataset</i>	53
5.2	<i>Feature Visualisation for 3 Fish Individuals' Dataset</i>	54
5.3	<i>Feature Visualisation for 7 Fish Individuals' Dataset</i>	55
5.4	<i>Model Selection Results for PCA-w Feature (3 fish individuals)</i>	57
5.5	<i>Model Selection Results for PCA-sep Feature (3 fish individuals) . . .</i>	58
5.6	<i>Model Selection Results for CFS Feature (3 fish individuals)</i>	58
5.7	<i>Model Selection Results for SPEC Feature (3 fish individuals)</i>	59
5.8	<i>Model Selection Results for PCA-w Feature (7 fish individuals)</i>	59
5.9	<i>Model Selection Results for PCA-sep Feature (7 fish individuals) . . .</i>	60
5.10	<i>Model Selection Results for CFS Feature (7 fish individuals)</i>	60
5.11	<i>Model Selection Results for SPEC Feature (7 fish individuals)</i>	61
5.12	<i>Clustering performances of the dataset with 3 fish individuals</i>	62
5.13	<i>Comparison of the fish individual 1 and 6</i>	63
5.14	<i>Clustering performances of the dataset with 7 fish individuals</i>	64

List of Tables

2.1	<i>Data Information</i>	10
2.2	<i>Information of the four different fish individuals</i>	11
2.3	<i>Split the dataset into 5 subsets according to the trajectory number of each observation</i>	12
3.1	<i>Percentage of variation of PCA-w using different number of eigenvectors</i>	23
3.2	<i>Percentage of variation of PCA-sep using different number of eigenvectors</i>	24
3.3	<i>Feature type No. and number of principle components in each type</i>	24
3.4	<i>Four feature subsets</i>	32
4.1	<i>Model selection results for the 4 fish individuals' dataset</i>	44
4.2	<i>The contingency matrix</i>	46
4.3	<i>Clustering validation measurements</i>	46
4.4	<i>Clustering performances of four features using selected model (4 fish individuals)</i>	48
4.5	<i>The contingency matrix of CFS feature(4 fish individuals)</i>	48
5.1	<i>Information about the additional three different fish individuals</i>	51
5.2	<i>Information of the additional three different fish individuals</i>	51
5.3	<i>Model selection results for the 3 fish individuals' dataset</i>	56
5.4	<i>Model selection results for the 7 fish individuals' dataset</i>	56
5.5	<i>Clustering performances of four features using selected model (3 fish individuals)</i>	61
5.6	<i>The contingency matrix of CFS feature (3 fish individuals)</i>	61
5.7	<i>Clustering performances of four features using selected model (7 fish individuals)</i>	63
5.8	<i>The contingency matrix of CFS feature (7 fish individuals)</i>	64

5.9	<i>Model selection results for the 3 datasets</i>	65
5.10	<i>Clustering performance of two datasets using the CFS feature</i>	65

Chapter 1

Introduction

1.1 Motivation

The environment has been considerably affected by human activities. Some phenomena of environmental effects have started to make people concerned up, for example, global warming, glaciers melting and endangered biological species. Researchers have begun to do some research to help people monitor environment changes. The Fish4Knowledge¹ project is trying to develop a system to study marine ecosystems by observing marine animals, which is essential for monitoring environmental effects. Besides that, this project can also help people to achieve commercial purposes such as fish farms and fisheries management. There are many ways to obtain data about fish such as using sonar to collect acoustic information or diving underwater to record fish photos. Computer vision is a good way for biologists to analysis fish data. Fish4Knowledge provides a professional system for capture, storage, analysis of undersea videos. Undersea video cameras that are embedded in different locations are used for recording the biological data of marine animals.

The main purpose of this project is to develop a method to distinguish different fish identities in a set of multiple observations and estimate the number of individuals. Because of the residential property of clown fish, we are likely to observe one individual many times in a video captured by a fixed camera, which means that thousands of fish observations may only be the image of several fish individuals. We present solutions to separate observations into individual groups. Statistics of individual numbers can help researchers to investigate the population size of species. The variation of population size observed in a period indicates some knowledge, for instance, of the increasing

¹<http://homepages.inf.ed.ac.uk/rbf/Fish4Knowledge/overview.htm>

population of a species. Sometimes, the sudden decrease of a population may be because the underwater environment of the area around the camera suffers from pollution or other changes.

1.2 Project Objectives

The main objective of this project was to present a method to identify different fishes and estimate the number of individuals. The sub-objectives are shown below:

- Extract distinctive features from fish images and create image pattern for fish.
- Group same individuals by using clustering methods
- Estimate individual numbers
- Evaluate experimental performance

Numerous works have been done in the classification of different species and the estimation of the number of species in a given large amount of fish detections. The discrimination between different species is easier, most of them have significantly different patterns, colors, length and shapes. For example, the clown fish have two distinct white stripes on the orange body and yellow fins while the *Dascyllus aruanus*, known as the three-stripe damsel, are white with three vertical black bars.

With regard to the same species, intuitively, it is more difficult to identify different individuals among a set of observations of the same fish species, and it is hard to estimate the number of individuals because of subtle distinctions between individuals. We need to use different types of features to represent fish. A lot of features can be used in this project, and one aim is to find some distinctive types from existing features. In this task, some new reliable features should also be designed. However, not all features may be relevant for clustering. In this case, selecting a feature set from the original feature vector may not only lead to better clustering the performance, but also reduce the computational cost.

The classifier used in this project is unsupervised clustering, which is a good way to find "natural" groupings of different individuals by clustering "similar" individuals together. In this project, the k-means algorithm is used as the clustering method. As cluster analysis is an unsupervised learning technique, and we do not know the number of clusters previously, we use model selection to solve this problem.

Some previous works have been done to solve the problem of fish detection and extraction from various backgrounds. This project uses the dataset of extracted fish images which features are extracted based on.

1.3 Outline

This thesis contains six chapters, each of which focuses on one aspect of the project, roughly including the description of data, design of the classification algorithm, the performance of experiments and the evaluation.

Chapter 2 reviews previous literature in which important techniques and methods related to this project were discussed. The detailed information of the dataset are also introduced in this chapter.

Chapter 3 introduces features used to represent fish including new features developed for this project. The feature selection method was also introduced to reduce the dimensionality.

Chapter 4 focus on the classifier used in this project. The clustering algorithm, which is used as the classification method in this project, is introduced in this chapter. Model selection techniques are also discussed for choosing the best feature subset and the best clustering model.

Chapter 5 provides the detailed experiment procedure and algorithm of this project. The evaluation of experiments is presented.

Chapter 6 concludes this dissertation based on the results of the experiments and gives some propositions and prospectives of future work

Chapter 2

Background

This chapter gives an introduction to relevant works and techniques that relate to the fish recognition and classification. The first section introduces the background and some previous works. The limitation of these works are also discussed. The second section focuses on the description of the data used in this project.

2.1 Literature Review

This section presents some previous techniques which can help to solve the problem of fish individual classification. Most works mainly achieved the task of fish recognition, which classify different fish species. In order to recognise the species of an observed fish, accurate fish detection and good separation from background are fundamental tasks. Researches tried to extract valid features to represent fishes and adopt many techniques to do classification and recognition tasks. Some related works have presented performances of different methods.

2.1.1 Fish detection

Fish detection is a fundamental step before recognising fish species or classification of individuals. A good fish detection result provides valid recourses of fish data. The paper [33] introduced some useful fish detection methods they used in practice, which can well handle various effects generated from the undersea unconstrained environment such as the lighting changes, frequent changes of fish scale and appearance because of fast movement of fish, various backgrounds of videos and sometimes the presence of murky water. Two mixture models algorithms were implemented to deal

with the problem of various undersea environment issues in this work; the well-known Adaptive Gaussian Mixture Model (AGMM) and the Adaptive Poisson Mixture Model (APMM). They also adopted other two algorithms, the Wave-back (WB) algorithm and the Intrinsic Model (IM). The former algorithm is a frequency-decomposition technique which considered the unstable movement of background elements, and the latter method performs well in dealing with videos with various lighting conditions. In order to separate the fish from the background accurately, the background scene should be modelled and must also be updated with the change of environmental conditions. A post-processing filtering step was carried out to improve the detection performance, which uses a quality score to evaluate the detection object. Objects with high score were regarded as detected fish objects. Finally, this work adopted common metrics, the detection rate and false alarm rate to evaluate performance of the detection results.

The experiment test on five videos and gave the detection result for each. From the results, it is obvious that the results of the algorithms implementing post-processing filtering were better than those without using this method. A sensible improvement in both detection rate (enhanced by 15% on average) and false alarm rate (reduced by about 10% on average) was achieved. And in most cases, the APMM and IM got better detection result. The detection rate using APMM algorithm with post-processing filter can averagely reach 84% and the false alarm rate can be lower than 10%. For the IM algorithm, the average detection rate was 88% and false alarm rate was 8%. Ideal detection performance makes the fish recognition and classification tasks more accurate and reasonable.

2.1.2 Fish recognition

Fish recognition and classification is a tough problem because of unstable undersea environments, various fish appearances and different illumination levels. Many works have been done to solve this problem. Classification of fish individuals is more difficult than fish species classification because there are significant differences between different fish species. So, more distinct features should be extracted to represent fish in order to distinguish different fish individuals of the same species. While, both tasks applied similar classification methods. The following papers presented different methods to deal with this problem. [18, 37] set up experiments under a limited environments, and the latter two solutions aims to deal with the fish recognition problem in an unrestricted natural environment which have various background and changing illumination.

N.J.C. Strachan

The research[37] had tried to solve the fish species recognition problem by using computer vision techniques. This paper introduced three shape-related features for discriminating between the different species of fish, the invariant moments, the optimisation of mismatch and the shape descriptors. In order to get clear fish silhouettes and perfect fish shapes from images, they took photographs of fishes in a constrained environment. 50 points of the silhouettes were collected and processed to create three shape features. Invariant moments are invariant to changes in scale, rotation and translation, and six invariant moments of the different fish shapes were computed. This method was not time consuming. The optimisation method processed slowly, because each pair of the 50 silhouette points from fish shapes should be compared. 11 geometric shape descriptors were measured from the fish shape. This method is very fast. The experiment was applied on the dataset containing as many as 30 images of 7 different type of fishes. The discriminant analysis was adapted to evaluate the result of using three techniques, which showed that the shape descriptor was the most reliable (reach 90% sorting result) and fastest method. The moment invariant is not time consuming, but the performance is worse (73%). The worst way is using the optimisation (63%) which is not only computationally intensive also inaccuracy in computing the mismatch factor for two fish.

R. Larsen

In [18], Shape and texture based classification of fish species, shape and texture features are mainly extracted to represent three fish species, and the species classification task were performed based on the combination of these two types of features. Fish data were in the form of images which are captured under standardized illumination environment and invariable background. In order to describe the shape and texture features of fishes, a series of corresponding landmarks of fishes were marked up to create the active appearance model (AAM). Instead of manually labelling the landmarks, they can be placed according to the minimum description length (MDL) principle. Fisher discriminant analysis was used to calculate the score for each pair of classes, which can evaluate the discriminant of different features. The result of experiments showed that the combination of texture and shape features was most discriminating, which the resubstitution rate reached 76% by using a dataset containing 108 fish images of 3 types of fishes.

Sampinato

Both of the two introduced research project described above were restricted to

constrained environments. However, in most practical applications, the environment around the fish is unstable and variable. Sampinato in [34] presents an automatic fish classification system that can help biologists to understand fish behavior in the real natural undersea environment. As images are all in 2D views, a fish with different pose presents different appearances. Affine transformation is a useful technique to solve the limitation of 2D image, where affine fish images of different views can represent the 3D fish shape. After doing affine transformations for each fish, two type of features are mainly extracted to represent fish images, the shape feature and the texture feature. Image moments, spatial Gabor filtering and the co-occurrence matrix were three texture feature extraction techniques used in this paper. The shape features were extracted by using the Curvature Scale Space transform and the Fourier descriptors. A dataset containing 360 fish images of 10 species were tested and the average accuracy reached 92%. The limitation of this work was that the shape and texture features were not enough to represent fishes. So, more useful features should be implemented.

Phoenix X. Huang

Compared with previously introduced three research projects, Phoenix in [16] investigated novel techniques to solve the problem of fish species classification in an unrestricted natural undersea environment. The SVM method is a popular classifier which was originally used for the binary classification. Some variations of SVM, such as a one-vs-one strategy, had been proposed for multi-class classification. However, this method ignored the correlation within classes. Another classifier, the hierarchical classification tree method, was adopted to replace the flat classifier. The drawback of hierarchical classification is that the classification error will accumulate with the increasing depth of the hierarchical tree. In this paper, they used the Balance-Guaranteed Optimised Tree (BGOT) to deal with this drawback. For improving the recognition result, some pre-processing was undertaken. The main purpose of the pre-processing steps are making all fish have the same orientation, which was achieved with 95% accuracy. After that, 66 types of features of different parts of fish body, such as tail, head, top, bottom and the whole fish, are extracted including colour, shape, texture properties. The experiment used a dataset with 3179 fish images of 10 species, and applied a 6-fold cross validation procedure. The average recall (AR), the average precision (AP) and the accuracy over count (AC) were three popular measurements to evaluate the result. The experiment result showed that the BGOT method had the best performance in all metrics compared with the Ada-boost method and the flat SVM method, which achieved 90% AR, 91.7% AP and 95% AC.

2.1.3 Literature Summary

The papers introduced above provided an overview of related works in fish detection and species recognition and classification. The paper that introduced fish detection techniques provided a fundamental process for data collection of our project. Four papers which proposed useful methods for fish recognition and species classification were introduced and compared. The former two research were restricted to constrained environments, and the latter two attempted to solve the recognition and classification problem caused by changing illumination and variable backgrounds. However, all these works tried to classify different fish species. What we are trying to do in this project is to recognise different fish individuals, which is more difficult. In our project, more discriminating features are extracted.

2.2 Data Description

2.2.1 Data Collection

This section describes the dataset used in the experiments. The Fish4Knowledge project started from October 2010. It has developed a full prototype of the data collection system. 9 cameras are installed at 4 different locations, from which nearly 100K hours of videos have been recorded and many million of images of tropical reef fish have been collected using these fixed cameras. All videos are uploaded to the website ¹. The first stage of processing the video is fish detection and tracking. 35 ground truth species are labeled and verified by the marine biologists. All detection and tracking results are recorded in an SQL database. NPP-3, LanYu, HuBihu and NMMBA are the four sites around Taiwan Island.

In this project, we use one of common fishes, *Amphiprion clarkii* known as clown fish as the experiment objective because of its large observation number and its more characteristic body feature compared with other fish species. Clown fish is a "resident" species, meaning that this species has a limited home range that is approximate 1-3 meters. So we are likely to see one individual many times in a video. There are about 1030 thousand detected clown fishes in the database, but most are wrong detection. We have to pick out correct clown fish detections manually. From previous works, among all labeled species, 4049 clown fish detections are correctly labeled in the dataset. Some clown fish examples are shown in (figure: note the differences in color and

¹<http://homepages.inf.ed.ac.uk/rbf/Fish4Knowledge/resources.htm#>

texture as well as different stripes). All these observations are from the site of NPP-3. NPP-3 is a southeast harbor of Taiwan where four cameras are fixed, shown in Figure 2.1. The estimated distances are: D-A: 2 m; A-B: 3 m; A-C: 7 m



Figure 2.1: *The site map of NPP-3*



(a) The original image (b) The binary image (c) Orientation result

Figure 2.2: *An example of fish image and the orientation result*

From the dataset, we get original images and binary masks of detected fishes. Since the fish are freely swimming, they always have different appearances in observations. Pre-processing procedures have been done to make the orientation of fishes standardised. The new format of each fish image is that fish are always at the image's centre and are facing to the right. Figure 2.2a shows an origin frame of fish image. The figure 2.2b is the binary image result of detected fish. Figure 2.2c presents the orientation result.

2.2.2 Ground Truth

This project is mainly to find out how many individuals are we seeing in a large number of fish detections, so we should label each fish manually. In order to give the

groundtruth of the dataset, each fish image is appended with some information including the site, the camera number, the date and time, the frame id and the trajectory number. Therefore, fish from the same trajectory should be the same individual. Fish from different cameras are very likely to be different individuals, especially when the cameras are far from each other. Because the home range of the clown fish is limited with 1-3 meters, if one individual appears around one camera, it will not swim away. Fish from the same camera may be the same individual but may also be different individuals, because more than one fish individual may have a home range within the view of that camera.




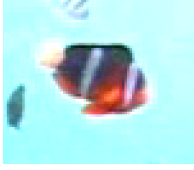
Table 2.1: *Data Information*

Information of a fish detection frame	Description
Site	The site that fishes are captured from (there are 4 sites which have been described in section 2.2.1)
Video No.	Each video has a video number
Camera No.	The No. of the nine cameras that are fixed in different locations
Date and time	The time when the video was recorded
Trajectory No.	An identifier that tells which fish are from the same tracking

According to the information described in above Table 2.1, we built a dataset with a total of 478 fish observations of four clearly labeled fish individuals. We created the ground truth for these four individuals according to their video captured location and date. Table 2.2 shows some observation examples of these four fishes and some detailed information.

As we described in the previous section, there are 9 fixed cameras in 4 sites. So, the three videos of fish individual 1, 2 and 3 were captured by the same camera No. 2 fixed in NPP-3. The video of fish individual 4 was captured by the camera No.3 fixed in NPP-3. Most labelling work was done intuitively for those fish observations that are captured by the same camera. We watched every video that contains clown fish individuals, and compared the colour, size and some details of fish observations

Table 2.2: Information of the four different fish individuals

Fish No.	Individual 1	Individual 2	Individual 3	Individual 4
Fish Observation				
Number of Observations	112	110	132	124
Video Captured Site	NPP-3	NPP-3	NPP-3	NPP-3
Camera No.	2	2	2	3
Video Captured Date	2010-11-08	2010-09-25	2010-08-22	2010-08-20

to determine whether observations from different trajectories or videos are different or same.

The fish individuals 1, 2 and 3 were all captured by the No.2 camera, and the individual 4 was by the No.3 camera. So, it is easy to determine that the fish 4 is a single individual. Individual 2 obviously has a different appearance with the other two. It is difficult to confirm that the first and the third are two different individuals. Comparing all observation images of these three individuals, we can see that individual 1 was slightly brighter than 3 whose chromatic body is reddish, and the thin area between its chromatic body and stripes of individual 1 is blackish. So, we determined that they are different individuals.

In order to classify different individuals of clown fish, we created the training set and the testing set, both of which contains images of the four individuals. We do feature selection based on the training samples. We reduce the dimensionality of the test set using selected features. Then, we use half of the samples of the test set to select the best model, and finally evaluate the clustering model's performance on the remaining samples of test set. As fishes from the same trajectory are captured within a short time frame and they have small variance which may affect the training and testing performance, we made the training and testing sets contain observations from different trajectories, meaning that we separate fish detections from same trajectory into different sets. The main idea of creating training and testing set is that we equally separate

these 478 fish detections into five sets, and each time we use one set to do testing and other four sets to do training. The final performance evaluation will be the average value of five testing results. The separation should consider the trajectory number of each detection. So, our idea is that for each trajectory, we split fish observations from the track into four sets, and then we got five generally average sets (see Table 2.3). While we say it "generally average" is because that the number of some trajectories can not be averagely separated into 5 sets. What we did is randomly split them.

Table 2.3: *Split the dataset into 5 subsets according to the trajectory number of each observation*

Individual No.	Set 1	Set 2	Set 3	Set 4	Set 5
Individual 1	23	24	23	20	22
Individual 2	25	24	25	29	27
Individual 3	20	26	27	31	28
Individual 4	24	22	23	27	28
Total Number	92	96	98	97	95

The table above shows how we separated the 478 observations of 4 fish individuals into 5 subsets containing 92, 96, 98, 97 and 95 observations in each. In each round, we pick out one set as the test set and the others as the training set. The final result is obtained from combination of the five intermediate performance evaluation.

Chapter 3

Fish Description and Feature Selection

This chapter introduces features used in the experiment. Some previous work has been done to extract some useful features. In [23] there are generally four types of features to build the pattern of fish images, colour features, texture features, boundary features and complex moments. In this project, some more intuitive colour and structure related features are introduced, and the HOG feature descriptor is also used to describe gradient information of images. The detail of new features are discussed in section 3.1. We combine all extracted properties into a feature vector, resulting in a high-dimensional representation of samples. Section 3.2 explores dimensionality reduction and feature selection methods.

3.1 Features

Some previous works have been done to adjust all fish into the same direction, where the head of fishes are facing to the left[16]. After rotation, features are extracted based on these uniform fish images. In this project, we introduced some new features to represent fishes.

3.1.1 Color Ratio

Depending on species, clown fishes are overall yellow, orange, or a reddish or blackish colour with white stripes or patches. We can separately analyse the colour of the colour body and white stripes.

Fish individuals have various colours and the colour is also unevenly distributed over fish body. For example, In Figure 3.1b, the fish is uniform orange, while the

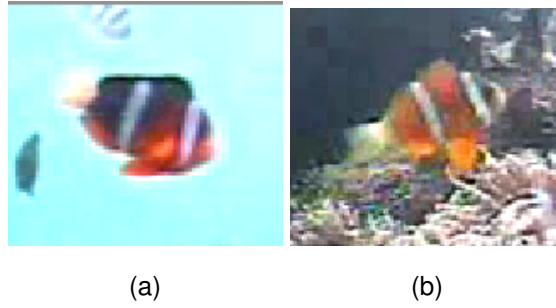


Figure 3.1: Comparison of two clown fish

second fish in Figure 3.1a shows blackish on the top part of body and reddish on the bottom part. So the RGB colour ratio of particular parts of fish body may encode some distinctions between different fish individuals. Eq 3.1 defines the colour of objects reflected in specific conditions, the colour value is based on the illumination and the colour of environment as well as the real colour of objects.

$$L = I \cdot W \cdot C \quad (3.1)$$

where I is the illumination of images, W denotes the water colour, and C is the fish colour. In RGB colour spaces, there are three additive primary colour channels: red green and blue.

$$L_r = I \cdot W_r \cdot C_r; L_g = I \cdot W_g \cdot C_g; L_b = I \cdot W_b \cdot C_b \quad (3.2)$$

Usually, the same fish individual captured in various lighting conditions looks very different. The Figure 3.2 provides an example. These two observations are the same fish individual, while the first detection in Figure 3.2a of the fish looks darker and the second in Figure 3.2b looks brighter.

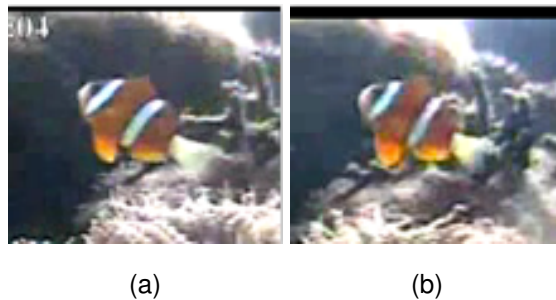


Figure 3.2: The illumination affects the colour of the fish reflected

So we prefer to use colour ratio Eq 3.3 between different fish part which can remove

the effect of illumination and environment colour, meaning that the contrast between real colour of fish body is only taken into account. After division, only the real colour ratio of different parts of fish body remains. L_{p1} and L_{p2} represent the colour of two part of fish respectively.

$$\frac{L_{p1}}{L_{p2}} = \frac{I \cdot W \cdot C_{p1}}{I \cdot W \cdot C_{p2}} = \frac{C_{p1}}{C_{p2}} \quad (3.3)$$

The colour of fish is most reflected on the chromatic part except white stripes. So we aim to separate the chromatic fish body and white stripes. The separation method is to find a threshold of the colour histogram of the fish image, which can separate the white area and the colour area. Figure 3.3 (a) to (c) show a separation example. After separation, we analysis the colour ratio between chromatic fish body and white stripes, the colour ratio between top and bottom part of chromatic body, and the colour ratio between front and posterior part of chromatic body. We described details of these three type of colour ratio.

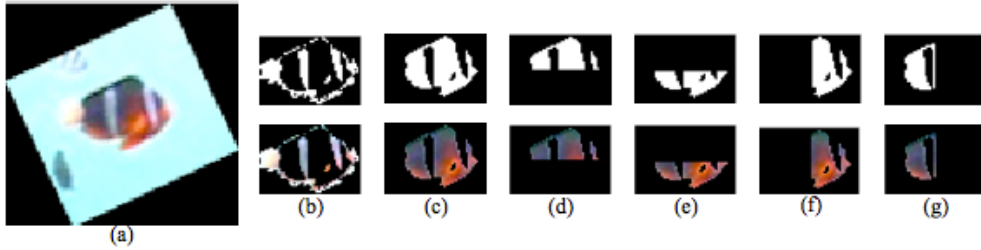
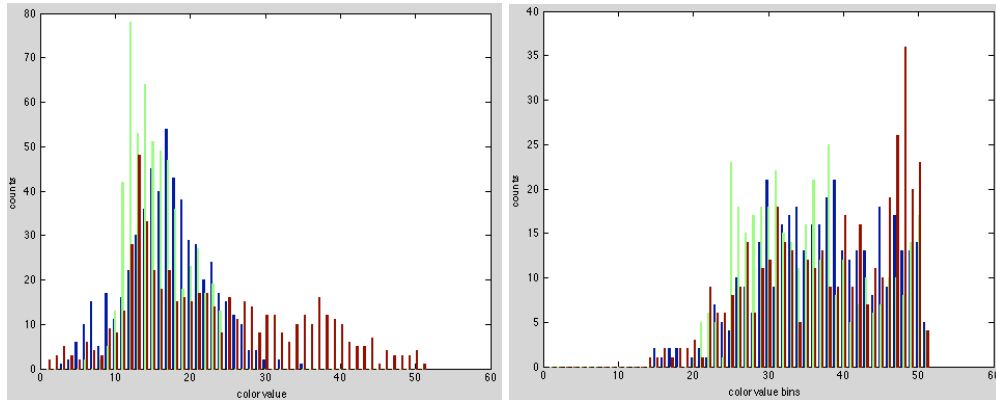


Figure 3.3: *Fish image and different parts of the fish; (a) is the orientated fish image; (b) is the part of extracted white stripes; (c) is the chromatic body of the fish; (d) the top part of the chromatic body; (e) the bottom part of the chromatic body; (f) the front part of the chromatic body; (g) the posterior part of the chromatic body*

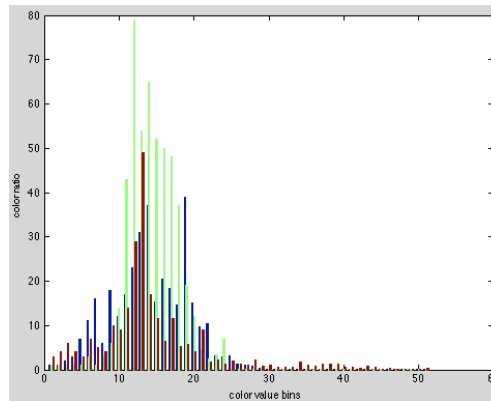
- 1 The colour ratio between chromatic body and white stripes. White stripes reflect more colour of the environment and illumination, so the division procedure helps remove that colour, then we can get the more actual body colour. Firstly, we created RGB colour histograms for the white stripes and the chromatic fish body respectively. Then, for each bin of the histogram, we calculated the ratio of these two body parts. Finally, three colour ratio vectors were generated and then were combined into a single vector. Figure 3.3(b) and (c) shows the two parts of the fish, stripes part and chromatic body part. Figure 3.4a presents the

colour histogram for the chromatic body of the fish, Figure 3.4b shows the colour histogram for the white stripes, and Figure 3.4c shows the ratio histogram of two corresponding bins coming from the previous two histograms.



(a) color histogram of the chromatic body

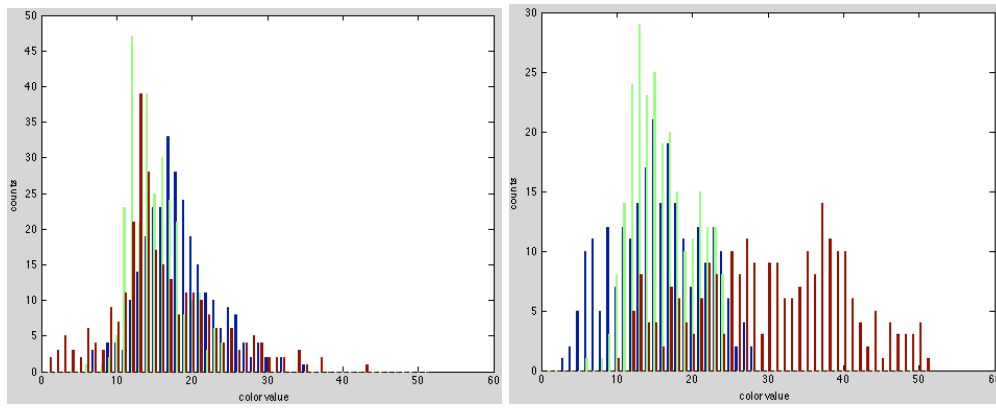
(b) color histogram of the stripes



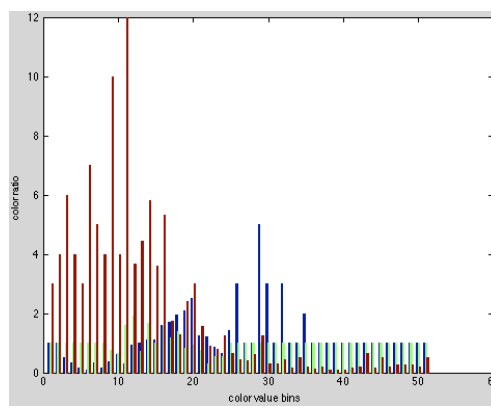
(c) colour ratio between white stripes and chromatic body

Figure 3.4: *Histogram of the color ratio between chromatic body and white stripes*

- 2 Similarly, a the colour ratio of the top part and the bottom part of the chromatic body is firstly created using a RGB colour histograms for the top part of body and the bottom part of body respectively, and then calculating the colour ratio of each bin of the histogram for the fish. Finally, we got a colour ratio vector combining histograms of three colour channels. Figure 3.5a shows the colour histogram for top part of chromatic body, Figure 3.5b shows the colour histogram for bottom part of chromatic body, and Figure 3.5c shows the ratio histogram of two corresponding bins coming from previous two histograms.
- 3 The colour ratio between the front part and the posterior part of the chromatic



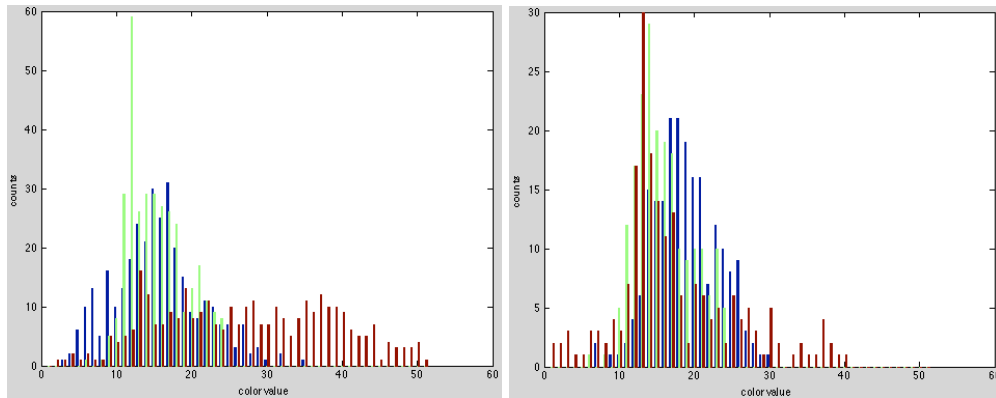
(a) color histogram of the top part of the chromatic body (b) color histogram of the bottom part of the chromatic body



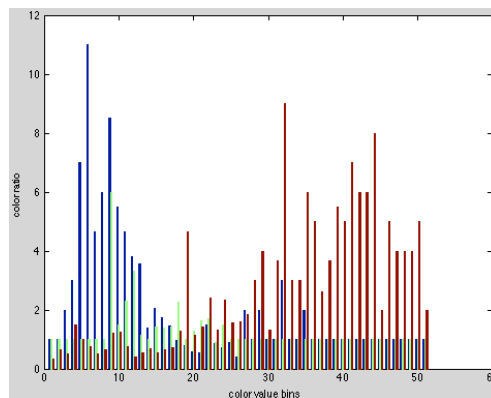
(c) colour ratio between top and bottom

Figure 3.5: Histogram of the color ratio of the top part and the bottom part of the chromatic body

body is used to create RGB colour histograms for the front part of the chromatic fish body and the posterior part of chromatic body respectively. Then, for each bin of histograms, we calculated the ratio of these two body parts of the fish. Finally, the three colour ratio vectors were generated and then were combined into a single vector. Figure 3.6a shows the colour histogram for front part of chromatic body, Figure 3.6b shows the colour histogram for posterior part of chromatic body, and Figure 3.6c shows the ratio histogram of two corresponding bins coming from the previous two histograms.



(a) color histogram for front part of chromatic body (b) color histogram for posterior part of chromatic body



(c) Histogram of the colour ratio of the front part and the posterior part of the chromatic body

Figure 3.6: Color ratio of the front part and the posterior part of the chromatic body

We can also calculate the average RGB colour value of different parts of the fish body and then compute the average colour ratio of two different parts. Figure 3.7 to

Figure 3.9 show the average RGB colour ratio comparison between four fish individuals. Four example images of these individuals are shown. Figure 3.7 shows the average colour ratio of chromatic body and white stripes of four individuals over three colour channels. It is obvious that this feature can separate four different individuals well. Figure 3.8 shows the average colour ratio of the top part and the bottom part of the chromatic body over RGB channels. The red ratio can performs better than green and blue ratio Figure 3.9 shows the average colour ratio of the front and the posterior part of the chromatic body over three channels. This feature can not separate different fish individuals well.

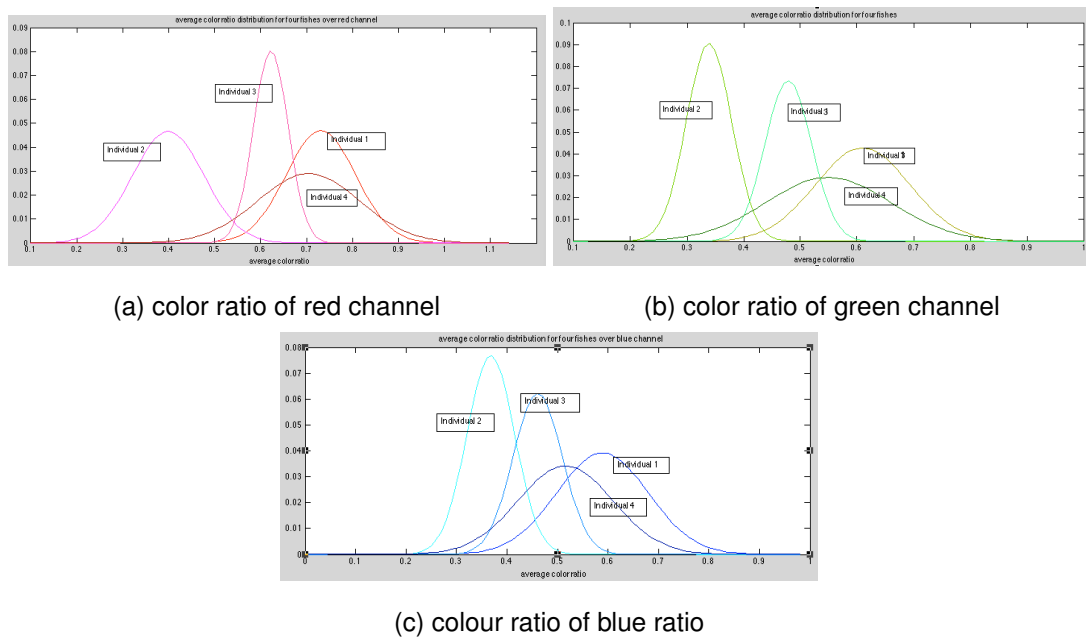


Figure 3.7: Average RGB colour ratio between the chromatic body and the white stripes

3.1.2 Length Ratio

Different clown fish individuals have different length of stripes. Some clown fishes? stripes are wider, while some are narrower. So different individuals show distinctive stripes and body length ratio. We introduced a method to calculate the ratio.

As fish images are all rotated into a horizontal direction, stripes are all vertically shown in images. We can extract the middle pixel row vector of the stripes binary image according to the area centroid the the binary fish image. Figure 3.10 shows the binary image of the fish stripes, the red point in the centre is the position the area

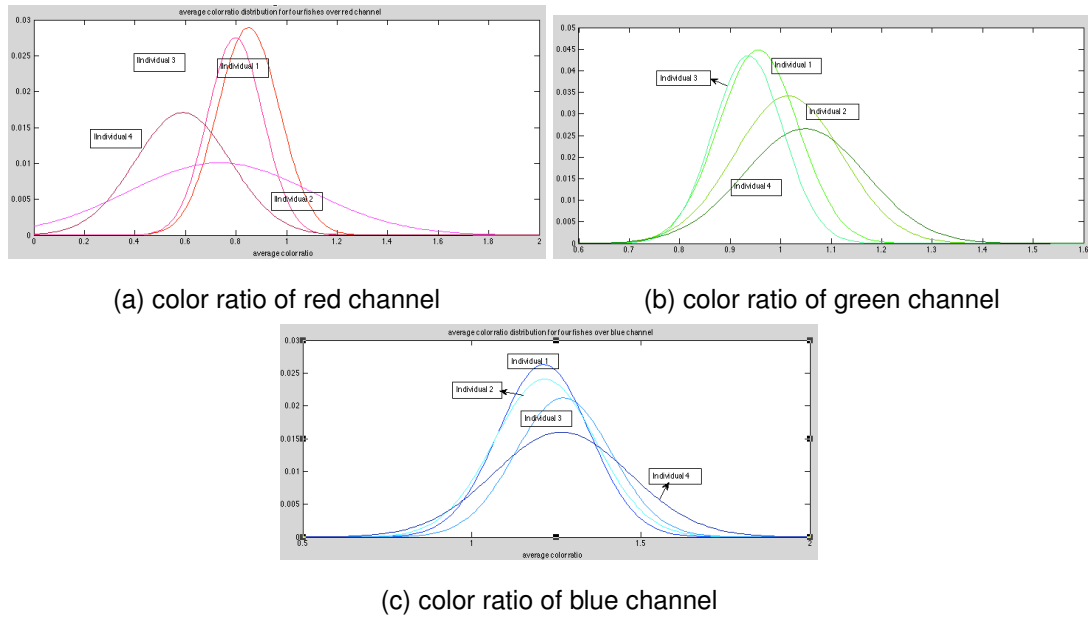


Figure 3.8: Average RGB colour ratio between the top part and the bottom part of the chromatic body

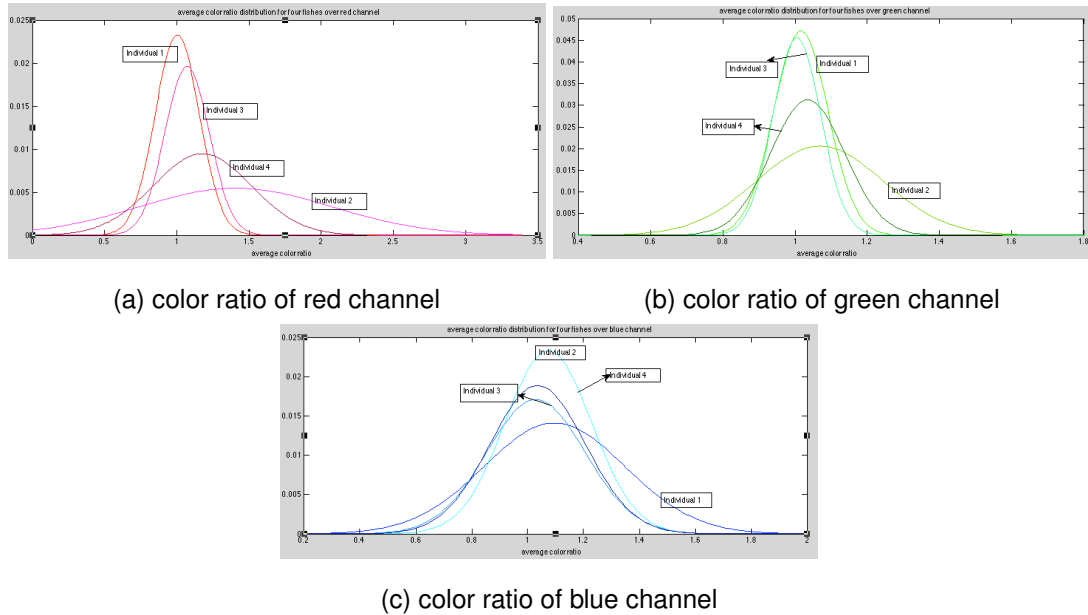


Figure 3.9: Average RGB colour ratio between the front part and the posterior part of the chromatic body

centroid of fish body, and the parallel blue line is the middle row vector of the image. White pixels in the row vector represent the part of stripes and the black pixels represent the part of chromatic body. By counting the number of white pixels, we can obtain the length of white stripes.

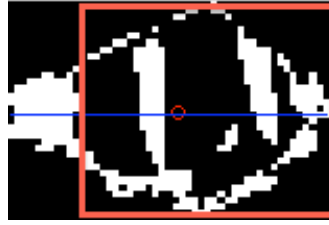


Figure 3.10: Length ratio calculation

While the waving of the fish tail makes the length of the tail unstable, meaning that images of the same fish individual display unequal length of tails. Considering this, we only analysis the stripes binary image without the part of the tail, which is the red box showed in this figure.

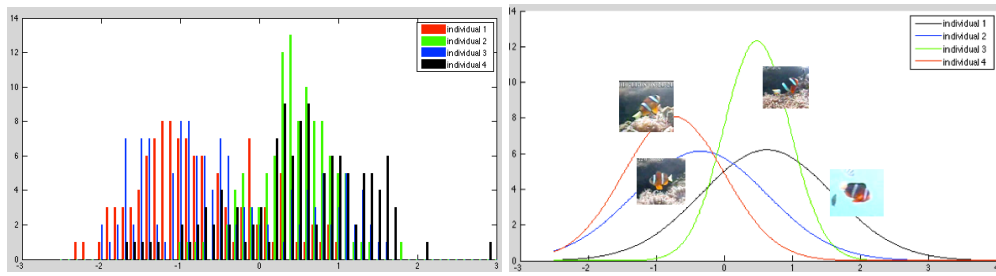
3.1.3 White stripes area ratio

The percentage of white stripes' area in the area (Eq 3.4) of the whole fish body can be a useful feature to distinguish different fish individuals. Figure 3.3 (b) Figure 3.10 show the extracted area of the white stripes. For the same reason which we described in 3.1.2, the area of tail changes significantly according to different pose of fish. So, we will remove the tail from the image and calculate the area ratio between white stripes and the whole fish body without considering the tail area.

$$AreaRate = \frac{area_w - area_t}{area_f - area_t} \quad (3.4)$$

Where the $area_w$ is the area of white stripes, the $area_f$ is the area of the whole fish body and the $area_t$ is the area of tail.

Figure 3.11a is the comparison histogram of stripes area ratio of four fish individuals. Figure 3.11b is the corresponding gaussian distribution graph. The graphs shows that this feature can separate some individuals, but it can not make all fish individuals separable. For instance, the fish individual 1 (red line) and fish individual 4 (black line) are separable, while the gaussian distribution for fish individual 1 and fish individual 3 (blue line) are almost overlapped.



(a) the histogram of the area ratio for the four fishes
(b) Gaussian distribution of the area ratio for the four fishes

Figure 3.11: *Area ratio*

3.2 Dimensionality Reduction and Feature Selection

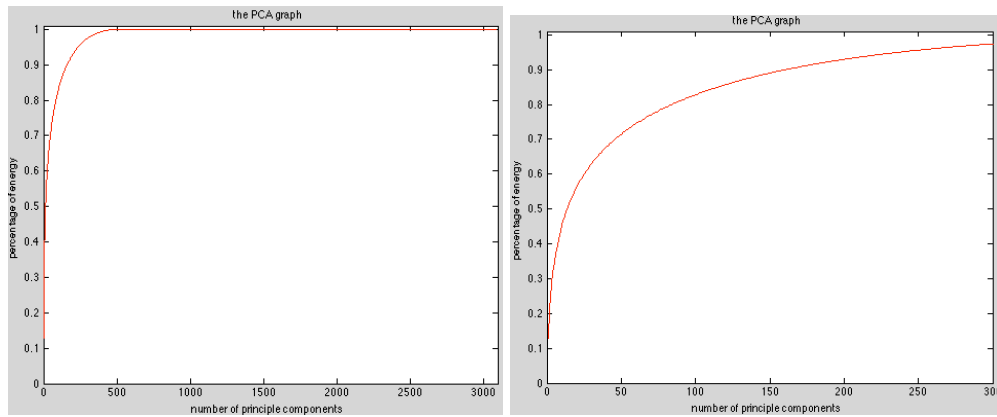
After extraction of all features, the feature vector for each sample is combined. Each image contains 3091 attributes. Not every feature contains useful information for classification and high dimensionality is time consuming. We should reduce the dimensionality of feature vectors. There are two general approaches for performing dimensionality reduction: feature extraction and feature selection[5]. The differences between these two methods is that feature extraction generates a transformation matrix to project existing features into a lower dimensional space, like the PCA, and feature selection selects a feature subset by using evaluation criteria instead of a transformation. In section 3.2.1, we described the process of doing dimensionality reduction by which a lower dimensional feature set is generated. PCA (Principal Component Analysis)[24] is an efficient dimensionality reduction technique which can be used for unsupervised learning problem. In section 3.2.2, we introduce the sequential forward selection technique which can select a lower dimensional feature set. Spectral feature selection[41] which is a filter-based unsupervised feature selection method and the Correlation-based feature selection[12, 13] which is a wrapper model are used in our work.

3.2.1 Principal Component Analysis

From previous works, we get a feature vector for each fish sample with 3091 attributes which is very high dimensional. In order to reduce the dimensionality, we applied Principal Component Analysis[24]. PCA is a useful statistical technique for feature dimensionality reduction, the main idea of which is to project the original feature space onto a new one.

There are two ways to extract principle components. One way is to do PCA on the integrated feature vector which contains 3091 attributes. Another way is trying to do PCA on each feature family respectively. There are totally 19 feature types, including colour features, texture features, shape features, image moments and so on. The detail of the feature types can be found in Appendix A.

1. We do the PCA on the data set based on the whole feature vector with 3091 attributes. Figures 3.12 show the number of eigenvectors and the cumulated energy of each number. The table shows the explicit dimensionality of features and corresponding percentage of variations. We applied the top 80 principle components to reduce the dimensionality, which cover nearly 80% energy of variance. Table 3.1 shows the number of eigenvectors and the cumulated energy of each number.



(a) The plot of cumulated variation of Eigenvectors
(b) The plot of cumulated variation of top 300 Eigenvectors

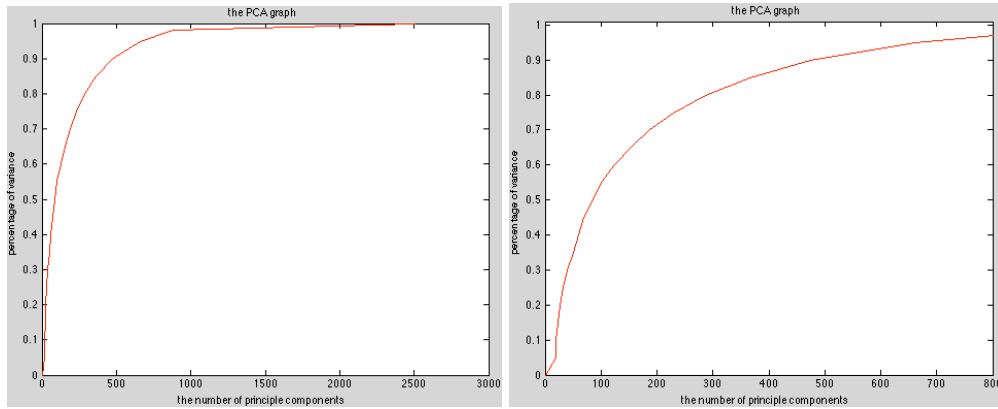
Figure 3.12: plot of PCA-w

Table 3.1: Percentage of variation of PCA-w using different number of eigenvectors

Energy	65%	70%	75%	80%	85%	90%	95%	98%
Number of Eigenvectors	34	46	62	84	115	161	237	322

2. Instead of doing PCA on the whole feature space, we do PCA on different feature families separately. The Figure 3.13 shows the number of eigenvectors and the cumulated energy of each number. Table 3.2 shows the energy of eigenvalues and their corresponding explicit number of principle components. In this project, we used the total 287 eigenvectors which cover 80% of variance of each feature family to reduce

the dimensionality.



(a) The plot of cumulated variation of (b) The plot of cumulated variation of top
Eigenvectors 300 Eigenvectors

Figure 3.13: plot of PCA-sep

Table 3.2: Percentage of variation of PCA-sep using different number of eigenvectors

Energy	65%	70%	75%	80%	85%	90%	95%	98%
Number of Eigenvectors	153	187	229	287	366	475	660	879

Table 3.3: Feature type No. and number of principle components in each type

Feature Type No.	1	2	3	4	5	6	7	8	9	10
Number of Eigenvectors	35	34	13	9	1	1	3	9	7	38
Feature Type No.	11	12	13	14	15	16	17	18	19	
Number of Eigenvectors	8	8	7	2	18	47	42	4	1	

While using this method, some less global principle components will be extracted. For the instance of 80% of variation, the table 3.3 shows how many principle components are extracted in each feature family. The No. of feature type corresponds to the feature type shown in Appendix A. Principle components only reflect local variance in each feature type.

3.2.2 Feature Selection

Feature selection[11] which is also known as attribute selection is a technique to enhance classification efficiency by selecting a feature subset. A lower feature subset can also reduce the computational efficiency. In practice, some attributes may be of low relevance to the classes, and fewer features means lower dimensionality. So picking out those ideal features which can distinguish samples from different classes may help to improve the generalisation capabilities and reduce the complexity and clustering time.

Feature subset selection mainly requires the following two factors:

1. Search strategy to select attributes from all features
2. Evaluation criteria to evaluate selected feature subsets

In this section, we introduced two feature selection methods to select lower dimensional feature subset based on the original feature space containing 3091 attributes.

3.2.2.1 Search Strategy

There are a large number of search strategies, but generally there are three types:

- Exponential search algorithms evaluate a number of subsets that grows exponentially with the dimensionality of the search space. Exhaustive search is one of the representative algorithms.
- Sequential search algorithms which we used in experiments search for the feature subset by adding or removing features sequentially, but have a tendency to become trapped in local minima. Sequential forward selection is an example.
- Randomized searching algorithms.

In this project, the sequential forward selection (SFS) algorithm[4] is adopted to select a new feature subset. SFS starts with an empty subset and then sequentially adds a new attribute which combining with already selected attributes would achieve the best classification accuracy or get the highest score according to evaluation criterion to the existing features that have already been selected. There are two types of termination criteria. One is using a fixed number of attributes as the termination criterion, which means that when the the number of attributes in the feature subset achieves this number, the selection process stops. Another method is that when the score of feature

subset does not improve any more, the algorithm stops.

The procedure of SFS is:

- Start with an empty feature set: $F = \emptyset$;
- Evaluate remaining attributes according to evaluation criteria;
- Select the best attribute according to $f^+ = \arg \max_f [J(F_k + f)]$
- Combine the existing feature set F_k and the best attribute f^+ to a new feature subset F_{k+1}
- Repeat above steps until we achieve the termination criterion

3.2.2.2 Evaluation Criteria

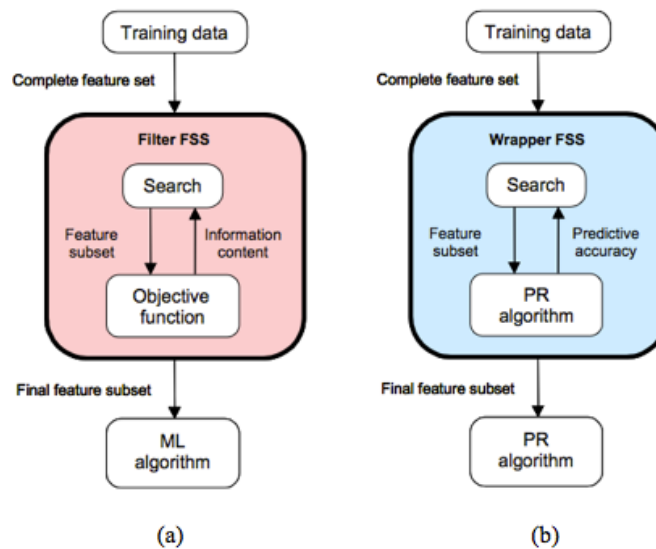


Figure 3.14: The flow chart of two feature selection models; (a) is the filter selection model; (b) is the wrapper selection model

The evaluation function evaluates the feature subset combined with a new attribute and returns a measure of its performance or "goodness" based on some criteria, using search strategies to pick out a suitable new attribute and add it into the existing feature subset. Filters and wrappers are two main evaluation models applied in feature selection. Figure 3.14 (a) shows the flow chart of filter model, and Figure 3.14 (b) shows the wrapper model. The difference between these two models is that the filter model evaluates the score of each attribute according to some criteria like information content, correlation of features, statistical dependence or information theoretic measures of feature subsets, then the attributes with the best score should be selected and added in to existing feature subset. The wrapper model evaluates selected feature subsets by utilising pattern recognition algorithms to measuring their clustering accu-

racy, clusters' interclass and intraclass distances or separability of data instead of using objective functions. Comparing these two models, the filter model is faster and more general but may not have a good performance, while the wrapper model is accurate but slow in execution. Most unsupervised feature selection algorithms are wrappers. In this project we adopted a wrapper model and a filter model.

1. Correlation-based feature selection

Correlation-based feature selection (CFS)[12, 13] is an example of wrapper model. It is a heuristic method for evaluating the correlation between feature subset and class labels and the correlation between attributes. So we also have to know the class label of each sample. The main idea is that a good feature subset contains attributes which are highly correlated with the true class labels, yet attributes should be uncorrelated with each other. So this algorithm applies a measurement to each feature. Feature subsets with high correlation value are preferred to selected. The following equation is used to represent the heuristic merit[10] of feature subset with d dimensional features:

$$Merit_s = \frac{dr_{cf}^-}{\sqrt{d + d(d-1)r_{ff}^-}} \quad (3.5)$$

where r_{cf}^- is the average value of the correlation between feature subset and the class labels and r_{ff}^- is the average value of the correlation between features. The Information Gain[27] is used to compute the correlation, which is defined as

$$InformationGain = H(Y) - H(Y|X) \quad (3.6)$$

Where $H(Y)$ is the entropy of Y and $H(Y|X)$ is the entropy of Y after observing X

Correlation-based feature selection use sequential forward selection algorithm with a termination criterion that the merit of feature subsets will not improve any more.

The CFS is implemented by adopting the the algorithm (see Algorithm 1). We applied the CFS algorithm on the original feature space. Inputting the dataset with 3091 dimensional features, a feature subset with 177 attributes is generated by using CFS. The dimensionality of new feature subset should also be reduced, because it is also high dimensional. We evaluate the clustering result for datasets

Algorithm 1: Correlated-based Feature Selection Algorithm

- Main Algorithm;**Data:** Dataset with N samples;

Training Set : 4/5 of N samples (A);

Test Set: 1/5 of N samples (B)

Result: *Feature_Subset* and *Evaluation_Subset***begin**

```

  % Using Training Set to do feature selection;

```

```

  % Randomly generate M sets of training sets(T) and validation sets(V) both

```

```

  % of which contains half of training samples (1/2 of A);

```

```

for round  $\leftarrow$  1 to M do

```

```

  Feature_Subset(F(found))  $\leftarrow$  Sequential Forward

```

```

  Selection(T(round));

```

```

  Evaluation_SubsetF(found)  $\leftarrow$  Result

```

```

  Evaluation(T(round),V(round));

```

```

  Generated M sets of feature subsets and pick out the best attributes as the
  final feature subset;

```

;

- Sequential Forward Selection Algorithm;**Data:** Training set (T)**Result:** *new_feature_subset***for** loop1 \leftarrow 1 **to** number of attributes **do**

```

  for loop2  $\leftarrow$  1 to number of remaining attributes do

```

```

    temporary_feature_subset(loop2)  $\leftarrow$ 

```

```

    feature_subset + one_attribute(loop2);

```

```

    evaluation_result(loop2)  $\leftarrow$  evaluate( temporary_feature_subset)

```

```

if the best evaluation_result of temporary_feature_subset decreased then

```

```

  break loop2;

```

```

else

```

```

  find the new_attribute(best_loop2) which makes the

```

```

  temporary_feature_subset get the best evaluation result;

```

```

  feature_subset  $\leftarrow$  feature_subset + one_new_attribute(best_loop2);

```

```

  new_feature_subset = feature_subset

```

with different number of features. As the Figure 3.15 shows below, the clustering accuracy has an increasing tendency with the growth of dimensionality. The growth becomes mild when the dimensionality reaches about 50. So we picked out the top 50 attributes from the 177 dimensional CFS selected feature vector as the new feature subset.

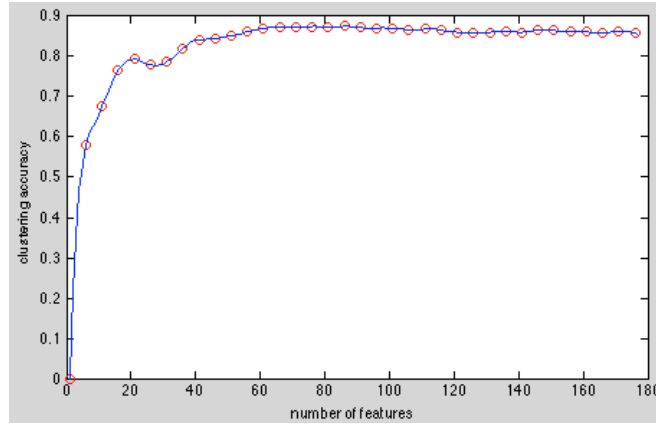


Figure 3.15: Increasing tendency of the clustering accuracy with the growth of dimensionality of CFS

2. Spectral Feature Selection

Without labels for unsupervised learning problems, lower dimensional features can not be selected by evaluating features's correlation with class. In this case, unsupervised feature selection exploits data variance and separability to evaluate feature relevance. Spectral Feature Selection[41] which belongs to a filter model is one of the unsupervised feature selection techniques estimating the feature relevance by estimating feature consistency with the spectrum of the graph induced from the similarity matrix of instances. The spectrum indicates the separability of samples. The Radial-based function [equation][3] between two instances can be used to calculate the similarity matrix.

$$S_{ij} = e^{-\frac{\|x_1 - x_2\|^2}{(2\delta)^2}} \quad (3.7)$$

Then an undirected graph $G(V, E)$ can be constructed from the similarity matrix where V denotes the vertex set and E denotes the edge set. Given graph G , we construct the adjacency matrix W , defined as $W(i, j) = w_{ij}$, and the degree matrix D , defined as $D(i, j) = d_i$ if $i = j$ and 0 otherwise where $d_i = \sum_{k=1}^n w_{ik}$. The reason why spectrum of graph helps to reflect the feature relevance is that

samples close to each other have similar feature values according to the graph structure, meaning that the graph structure shows that which features are more relevant to the target concept. In order to rank features, some ranking functions should be evaluated by using the Laplacian matrix. The Laplacian matrix L and the normalised Laplacian matrix are represented as:

$$L = D - W; \mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \quad (3.8)$$

The evaluation criterion we used here is calculated in the following way:

$$\varphi(f_i) = \frac{\hat{f}_i^T \gamma(\mathcal{L}) \hat{f}_i}{1 - (\hat{f}_i^T \xi_1)^2} = \frac{\sum_{d=2}^D \alpha_j^2 \gamma(\lambda_j)}{\sum_{d=2}^D \alpha_j^2} \quad (3.9)$$

$$\hat{f}_i = (D^{\frac{1}{2}} f_i) \cdot \left\| (D^{\frac{1}{2}} f_i) \right\|^{-1} \quad (3.10)$$

f_i is the i th feature vector of feature space. (λ_j, ξ_j) is the eigensystem of Laplacian matrix; $\alpha_j = \cos \theta_j$, where θ_j is the angle between \hat{f}_i and ξ_j ; and γ is an increasing function used to rescale the eigenvalues of \mathcal{L} for denoising. The top eigenvectors of \mathcal{L} are the optimal soft cluster indicators of the data [39].

We aim to reduce the original dimensionality by removing features which may affect the clustering accuracy. According to the Algorithm 2, the k-means algorithm was used as the evaluation method to evaluate the clustering accuracy of feature subset in order to select an efficient dimensionality. The pattern recognition algorithm was ran for 100 times, and the blue line in the figure connected all the average accuracy value corresponding to each feature dimensionality. The figure 3.16 below shows the clustering accuracy using datasets with increasing number of features. It is obvious that feature set with the feature dimensionality from 0 to 500 obtained increasing clustering accuracy, while after 800, the clustering accuracy decreased gently. So, the algorithm stops when the dimensionality reaches 500. The top 500 weighted attributes were selected to create a new feature subset.

Algorithm 2: Spectral Feature Selection Algorithm

Data: Training Dataset (A)**Result:** *Feature_Subset* with top feature weight*Using Training Set to do feature selection;***begin**

- Construct similarity matrix S from data set using RBF function;
 - Construct undirected graph G from S ;
 - Construct adjacency matrix W from S ;
 - Construct degree matrix D form W ;
 - Build Laplacian matrix according to Eq. 3.8;
 - Do evaluation for each feature vector f_i according to Eq. 3.9;
 - Rank features according to the feature evaluation result.;
- % now the Training Dataset (A) is reorder according to the rank features;*
% Using Training Set to do feature selection;

for $loop2 \leftarrow 1$ **to** *Number_of_Attributes* **do** Initialise the *number_of_trials* = 100; Initialise *Evaluate_Result*[100]; Randomly generate M sets of training sets(T) and validation sets(V)
 both of which contains half of training samples (1/2 of A); **for** $loop1 \leftarrow 1$ **to** *number_of_trials* **do** | *Temporary_FeatureSubset* = T (dimensional from 1 to $loop2$); | *Evaluate_Result(loop1)* = *Evaluation(Temporary_FeatureSubset)*; *Average_Evaluate_Result* = *Average(Evaluate_Result[100])*; **if** *Average_Evaluate_Result* of *Temporary_FeatureSubset* change
 smoothly or decrease **then** | break $loop2$; *Feature_Subset* = *Temporary_FeatureSubset* ;

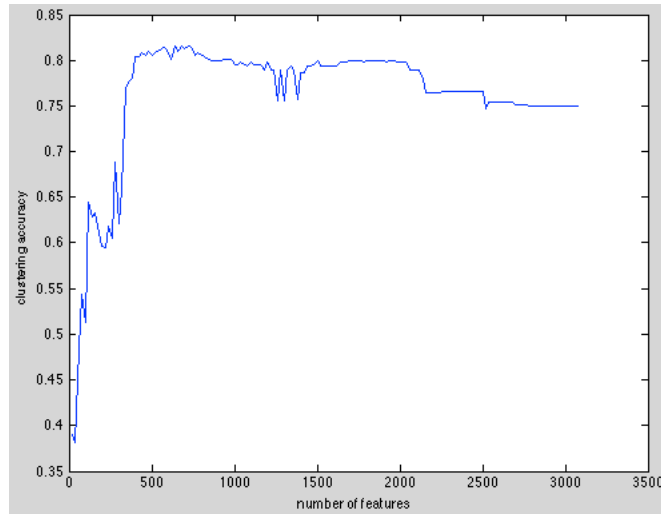


Figure 3.16: Changing of the clustering accuracy with increasing number of features

3.3 Summary

Table 3.4: Four feature subsets

Abbreviation	description	Dimensionality
PCA-w	Do PCA on the whole feature vector	84
PCA-sep	Do PCA on each feature families respectively	287
SPEC	Spectral feature selection result	500
CFS	Correlation-based feature selection result	50

In this chapter, we introduced several new features in section 3.1. Combining with features presented in some previous works, a high-dimensional feature vector with 3091 attributes was generated. Then, some feature extraction and selection methods are used to reduce the dimensionality, which were introduced in section 3.2. PCA is one of the popular feature extraction methods. We did PCA in two ways, one of which is based on the whole feature vector, and another way is that we did PCA over each feature family respectively and then combined lower dimensional features of each feature family together. Among all feature selection techniques, there are two different models, the filter model and the wrapper model. Spectral feature selection (SPEC) that we used in this project is one of the filter model. This method is a kind of unsupervised feature selection technique, which does not need to consider the groundtruth of the dataset. A example of wrapper model is the correlation-based feature selection (CFS),

which needs to use the groundtruth of the dataset to analysis the correlation between features and classes. Both SPEC and CFS were implemented by the sequential forward selection algorithm. Table 3.4 shows details of all feature subsets we used in following experiments.

Chapter 4

Classification

This chapter introduces the classifier used in this project. The main objective of this project is to classify different clown fish individuals and learn the number of individuals using previous extracted features. To measure the classification performance, the dataset was separated into a training set and a test set. As the number of classes is not known previously, in this project, unsupervised learning was used. K-means clustering[15, 22] is one of the simple and effective partition clustering analysis methods. Determining the number of individuals is by estimating the number of clusters. Model selection helps to select the appropriate parameters for learning algorithms, meaning that it can help the dataset to select k using k-means. In section 4.2, several model selection methods were introduced. Using selected models by different features, we applied K-means to cluster all data samples in each feature space. Then we used some clustering validation methods to evaluate the performance of clustering, which were introduced in section 4.3.

4.1 Classifier

K-means clustering algorithm is an algorithm to classify or group N observations (data points or samples) based on a D dimensional features into K clusters. In this project, we have a large number of fish observations, and we need to classify all observations into a number of clusters in each of which we attempt to make sure that most are the same fish individuals.

4.1.1 K-means

The basic idea of K-means clustering is to iteratively locate K centroids which minimize the overall intra-cluster distance until the result converges which means that the K centroids are stable and all samples will not move to another group anymore. First of all, we should determine the number of clusters K , and initialize the location of the K centroids of these clusters. In most cases, the initial position of K centroids are randomly located, meaning that we can take any random K samples as the initial centroids. Then the k-means algorithm will do the following steps until convergence.

Initialization: (Figures 4.1a)

- Determine K = number of clusters;
 - Randomly choose K samples as initial centroids;
 - We denote the data point by x_n , $n = 1, \dots, N$ and centroids by m_k : $k = 1, \dots, K$.
- Iterate following steps until convergence:**

Step 1. compute the distance of each data point to each centroid.

- For each sample X_n , we compute the distance for X_n to all m_k achieving a $1 \times K$ distance vector
- For all N samples, we compute the distance vectors getting a $N \times K$ distance matrix D .

Step 2. Assign each data point to the closest cluster based on the minimum distance to centroids (Figures 4.1b)

- Assign each data point to the cluster with the nearest centroid, meaning that find minimum value for each row of the distance matrix.

$$\hat{k}^{(n)} = \arg \max_k \{d(m^{(k)}, x^{(n)})\} \quad (4.1)$$

-We get an assignment matrix R according to the distance matrix D . The element of matrix R given by Eq 4.2 meaning that set $r_k^{(n)}$ to one if the sample x_n is closest to mean k , otherwise $r_k^{(n)}$ is zero.

$$r_k^{(n)} = \begin{cases} 1 & \text{if } \hat{k}^{(n)} = k \\ 0 & \text{if } \hat{k}^{(n)} \neq k \end{cases} \quad (4.2)$$

-Now we get new clusters

Step 3. determine the new position of centroids according to the new groups (Figures 4.1c)

- According to the new clusters, recompute the new centroid of each cluster with new sample assignment (Eq 4.3). $R^{(k)}$ is the total responsibility of mean k.

$$m^{(k)} = \frac{\sum_{n=1}^N r_k^{(n)} x^{(n)}}{R^{(k)}} \quad (4.3)$$

$$R^{(k)} = \sum_{n=1}^N r_k^{(n)} \quad (4.4)$$

- the centroid of each cluster is just the mean vector of all samples in this cluster.

Step 4. determine whether convergence is achieved

- Yes: Algorithm End. Return with cluster assignment for each data point and centroids of clusters (Figures 4.1d).

- No: Repeat the above steps 1 to 3

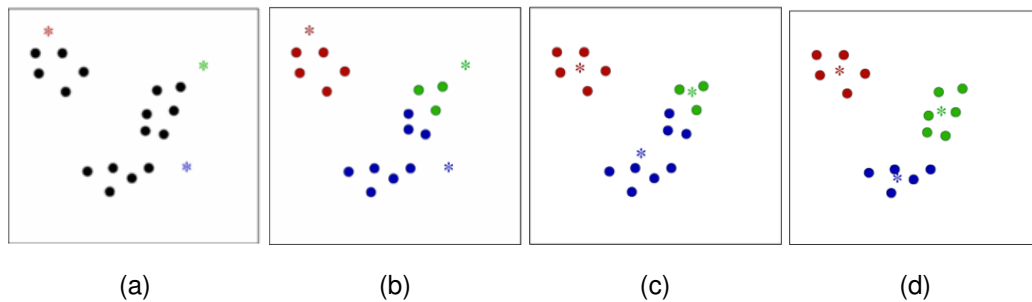


Figure 4.1: *The procedure of the K-means algorithm*

When the following two conditions are satisfied, the algorithm can be viewed as converged:

1. The overall distances between samples to their assigned cluster centroids begins to decreased.
2. There are only finitely many partitions of the training examples into K clusters.

4.1.2 Distance metric

Many distance metrics can be used to calculate the distance between samples and centroids in K-means algorithm. Squared Euclidean distance is one of the commonly used metric. Given two vectors of attributes, $A\{a_1, a_2, \dots, a_D\}$ and $B\{b_1, b_2, \dots, b_D\}$, where D is the dimensionality of the feature. The squared Euclidean distance between A and B is given by:

$$d(A, B) = (a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_D - b_D)^2 = \sum_{d=1}^D (a_d - b_d)^2 \quad (4.5)$$

4.1.3 Implementation

Matlab[®]'s built-in implementation was adopted in experiments.

- Distance Metric: Squared Euclidean distance.
- Cluster centroids initialization method: Selecting k observations from dataset at random.
- Iteration times: the algorithm was run for a maximum of 100 iterations.

4.2 Model selection

The model selection problem for clustering[2] is to select the number of clusters. Using the K-means algorithm with different choices of K will result in different statistical models for data[14, 26]. A model selection technique helps to choose the best model which in K-means is the best choice of K to fit the training data. There are a lot of model selection methods to solve the K selection problem for clustering. All of these methods apply a measure to evaluate the model.

In practice, we start with a set of candidate models, which in K-means is a set of different candidate K numbers. In this project, We tried K from 2 to L , where L is 30 in this project. A set of candidate models is generated, $M_{k_2}, M_{k_3}, \dots, M_{k_L}$. Then by evaluating these models using a specific measure, we can get a set of measurements, $S_{k_2}, S_{k_3}, \dots, S_{k_L}$, according to which the best model or K number can be selected. In some criterion the best model corresponds to the minimum value of the measure and in others the maximum.

4.2.1 Akaike information criteria (AIC) and Bayesian information criteria (BIC)

The AIC-based[1, 30] and BIC-based criterion[17, 31] techniques are both information-theoretic measures. The basic idea of these two penalty methods is trading off distortion of the model against model complexity, meaning that complex models will be penalised more severely. Complex models have large number of parameters.

The K-means algorithm aims to find the best location of K centroids which maximises the log-likelihood of the data. The likelihood reflects the conformity of the model to the observations. The log-likelihood monotonically increases with the increase in model complexity, because the model becomes more fit to the dataset. If $K = N$ (N is the number of samples), each cluster contains a singleton instance. Although the log-likelihood is maximum, the model with N parameters overfits to the data. AIC (Eq. 4.6) and BIC (Eq 4.7) provide measures that combines two elements: the distortion which is a measure of how much samples deviate from the centroid of their clusters, and a measure of model complexity which means the number of parameters used in a model or the number of K tried by K-means.

$$AIC = -2L(K) + 2K \quad (4.6)$$

$$BIC = -2L(K) + K \ln N \quad (4.7)$$

where $-L(K)$ is the negative maximum log-likelihood of the data for K clusters. The log-likelihood is a measure of distortion. The K is the number of parameters to be estimated. This part is a measure of model complexity in the AIC. In BIC, $K \ln N$ is used as the penalty of the model complexity, where N is the number of samples. The log-likelihood for the K-means is defined as (Eq. 4.8):

$$L(K) = -(1/2)RSS_{min}(K) \quad (4.8)$$

where the $RSS_{min}(K)$ denotes the objective function in K-means and the goal is to minimise it. The RSS means the residual sum of squares, which is a measure of how well the centroids represent the members of their clusters. It monotonically decreases with the increase of K. So, the log-likelihood increases with the increasing of K. It is defended by Eq. 4.9 to Eq. 4.11

$$RSS = \sum_{k=1}^K RSS_k \quad (4.9)$$

$$RSS_k = \sum_{\vec{x} \in \omega} |\vec{x} - \vec{\mu}(\omega_k)|^2 \quad (4.10)$$

$$\vec{\mu}(\omega_k) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x} \quad (4.11)$$

The optimal fitted model is identified by the minimum value of AIC or BIC. Both AIC and BIC present the same goodness-of-fit term, but the difference is that the penalty term of BIC is more stringent than that of AIC. So, BIC tends to favour models with fewer parameters than AIC.

4.2.2 Clustering evaluation indices

Some clustering quality evaluations can also be used to help select the best cluster number. We can run the K-means with a set of K and calculate the clustering quality for each clustering result. Then the value of K achieving the best clustering quality is determined as the cluster number.

1. Dunns Index

We applied one of the useful clustering evaluation indices, the Dunns Index, in this project. The Dunns Index[6, 7] which was proposed in 1974 by Dunn is a clustering quality evaluation method, which can also be used to estimate the cluster numbers. The Dunns index is defined as:

$$C = \frac{\min_{k \neq k'} d_{kk'}}{\max_{1 \leq k \leq K} D_k} \quad (4.12)$$

where $d_{kk'}$ is the minimum distance between two samples which come from the different clusters.

$$d_{kk'} = \min_{i \in I_k; j \in I_{k'}} \left\| X_i^{\{k\}} - X_j^{\{k'\}} \right\| \quad (4.13)$$

D_k is the largest distance between two samples which come from the same cluster.

$$D_k = \max_{i, j \in I_k; i \neq j} \left\| X_i^{\{k\}} - X_j^{\{k\}} \right\| \quad (4.14)$$

2. Silhouette Index

Silhouette index[20, 29] refers to measuring the validation of clusters of data. The average silhouette width of the data can be used as a criterion for determining the number of clusters. The basic idea of silhouette is that a good clustering make the sample similar with samples from the same cluster and different from

points in other clusters. So the silhouette width for each sample should be valued, and the definition is:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4.15)$$

where $a(i)$ is the mean distance of sample i to other samples of the same cluster with i ,

$$a(i) = \frac{1}{n_k - 1} \sum_{i' \in I_k, i' \neq i} d(x_i, x_{i'}) \quad (4.16)$$

and $b(i)$ is the minimum of the average distances between sample i and all the other samples from each other cluster.

$$b(i) = \min_{k' \neq k} \frac{1}{n_{k'}} \sum_{i' \in I_{k'}} d(x_i, x_{i'}) \quad (4.17)$$

Then, we can get the average silhouette width of samples in each cluster S_k .

$$S_k = \frac{1}{n_k} \sum_{i \in I_k} s(i) \quad (4.18)$$

Finally, the global silhouette index for the clustering can be calculated by averaging all cluster silhouette mean widths:

$$C = \frac{1}{K} \sum_{k=1}^K S_k \quad (4.19)$$

3. Scatter metric

Using clustering methods, we can separate samples into clusters. In order to estimate the best cluster number, we adopted the scatter metric[32] to evaluate the cluster separability. Given a data set, the quality of clusters can be evaluated by the within cluster scatter S_w and the between cluster scatter S_b which are defined as:

$$S_b = \sum_{k=1}^K \pi_k (\mu_k - \mu)(\mu_k - \mu)^T \quad (4.20)$$

$$S_w = \sum_{k=1}^K \pi_k E \{ (X - \mu_k)(X - \mu_k)^T \mid \omega_k \} = \sum_{k=1}^K \pi_k \Sigma_k \quad (4.21)$$

$$\mu = E \{ X \} = \sum_{k=1}^K \pi_k \mu_k \quad (4.22)$$

where π_k is the probability of cluster k in the data set. S_b denotes the between cluster scatter and S_w denotes the within cluster scatter. An ideal feature subset

should make the data points within the same cluster as close as possible and make the data points between clusters as far as possible. In order to represent the separability of samples, many measures can be obtained using S_w and S_b . We adopted a scatter related measure, which is defined as J_3 (Eq 4.23):

$$J_3 = \frac{tr(S_b)}{tr(S_w)} \quad (4.23)$$

We use J_3 criteria to evaluate models. The main idea is that the best clustering model makes the samples most separable. In practice, we applied the K-means algorithm for the dataset. By trying a set of K numbers, we generate a set of candidate clustering models and corresponding cluster assignments for samples. According to the cluster assignments of each model, we can get the within group scatter and between group scatter from which we can compute the separability of data for each model. The best clustering model group samples into k clusters, which can maximise the J_3 criteria.

4.2.3 Model selection algorithm

The model selection algorithm was designed for this project, and the code for pseudo model selection is given as below:

Algorithm 3: Model Selection Algorithm

Data: Test Set (B)

Result: BestK

Initialize the Criteria_value[29, M] matrix (29 is the number of k was tested, and M means the number of trails);

begin

for $i \leftarrow 1$ **to** M **do**

 Initialize a set of predefined cluster numbers $k = 2 : 30$;

for $k \leftarrow 2$ **to** 30 **do**

$Cluster_assignment_result = \text{K-mean algorithm}(B)$;

$Criteria_value(k, i) = \text{model selection criteria algorithms (AIC, BIC,$

 Scatter metric, Clustering Indices);

For each criteria, select the k with the best average criteria value as the suitable cluster number;

 BestK = k with the best *criteria_value*

4.2.4 Model selection result

This section presents the model selection results for the dataset of four fish individuals. The AIC, BIC, Dunns Index and Silhouette index all select the K number with minimum median, and the J_3 scatter metric selects the K with maximum median. We applied model selection on four feature sets, PCA-w, PCA-sep, SPEC and CFS. Figure 4.2 to figure 4.5 show the model selection results of these four feature spaces. The red dash line in the blue box denotes the median of 100 values of each K. The best K is selected according to the maximum or minimum median value of all K numbers. The green asterisk in each graph shows the selected best K and the best value of the measurements. The table 4.1 summarise the selected models of all feature sets using different selection methods.

1. Model selection results using PCA-w features (Figure 4.2)
2. Model selection results using PCA-sep features (Figure 4.3)
3. Model selection results using PCA-CFS features (Figure 4.4)
4. Model selection results using SPEC features (Figure 4.5)

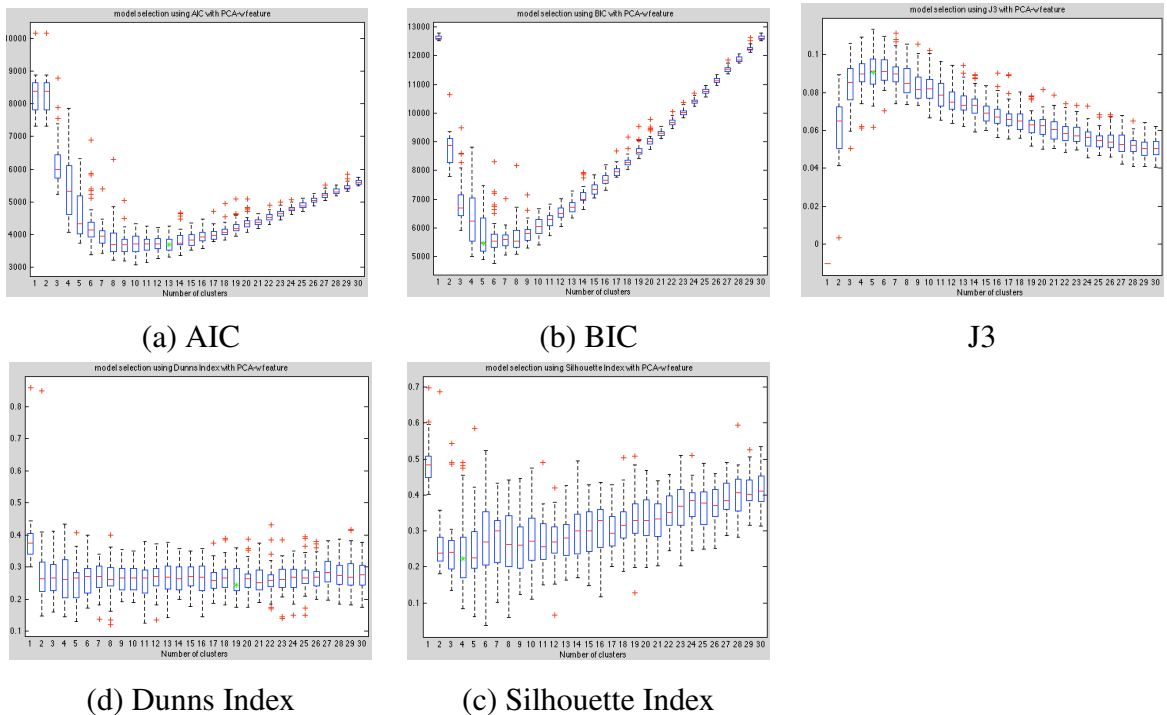


Figure 4.2: Model Selection Results for PCA-w Features (4 fish individuals)

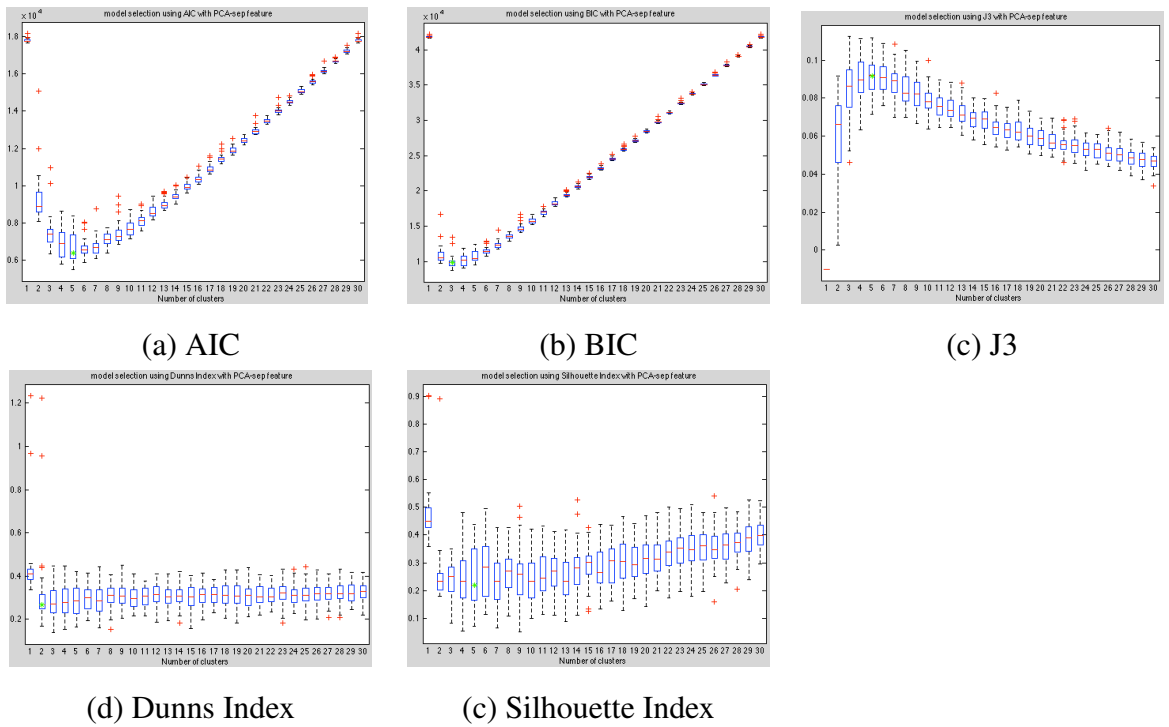


Figure 4.3: Model Selection Results for PCA-sep Features (4 fish individuals)

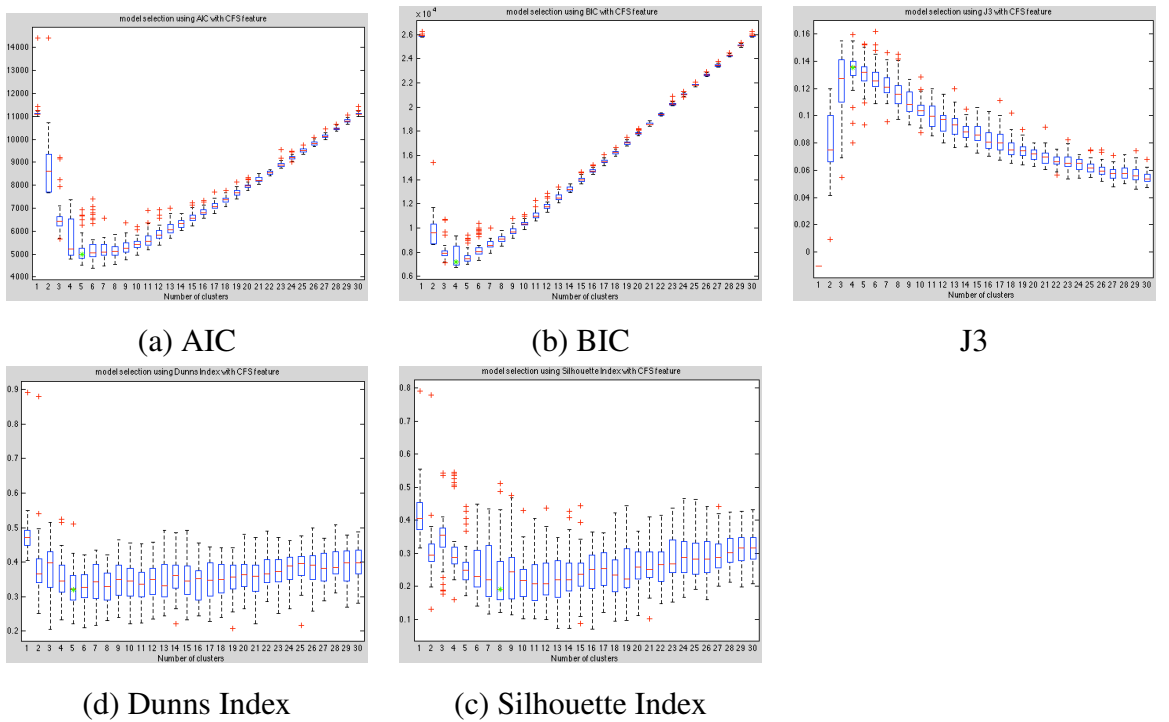


Figure 4.4: Model Selection Results for CFS Features (4 fish individuals)

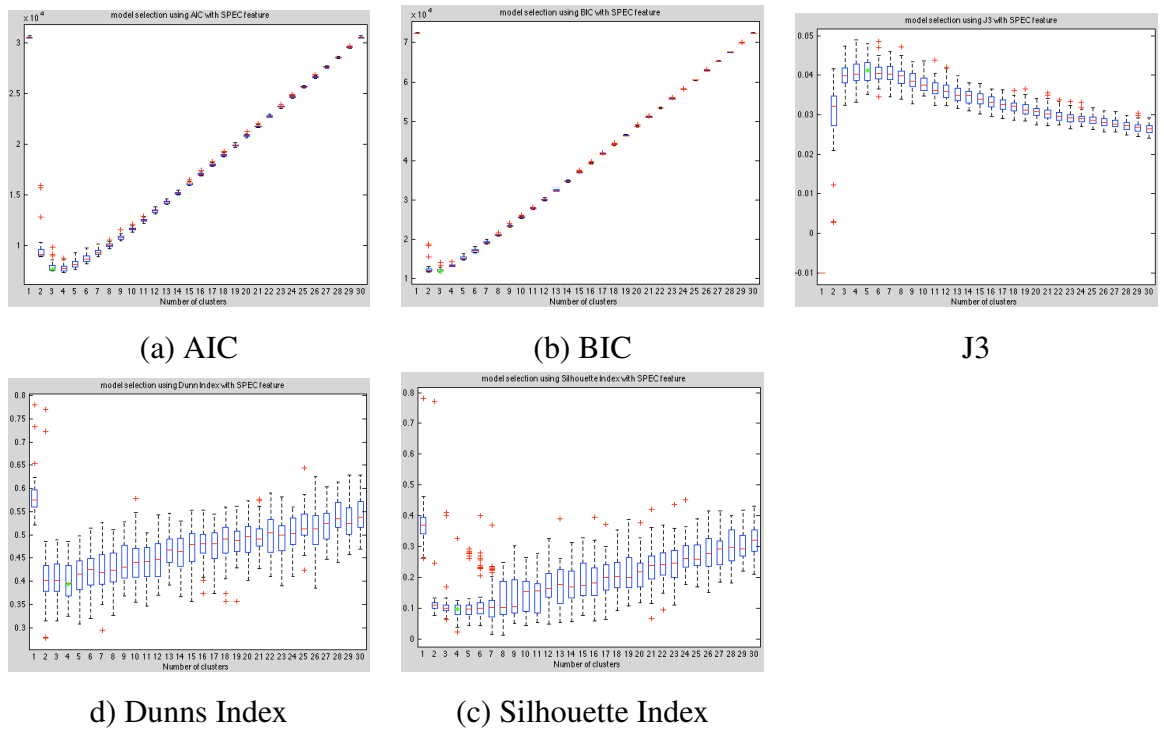


Figure 4.5: Model Selection Results for SPEC Features (4 fish individuals)

Table 4.1: Model selection results for the 4 fish individuals' dataset

Features	Model Selection Strategies					Model determination
	AIC	BIC	J_3	Dunns Index	Silhouette Index	
PCA-w	13	5	5	19	4	5
PCA-sep	5	3	5	2	5	5
CFS	5	4	4	5	8	5
SPEC	3	3	5	4	4	4

From the figures and the table, we can see that the PCA-sep and SPEC got approximate cluster numbers (K), where the true class number is 4. The BIC and J_3 selection methods can get approximate K number for all feature sets. Both AIC and Dunns Index estimated a large cluster number using PCA-w feature set. The Dunns Index and Silhouette Index are not good model selection methods, because the values do not show an obvious increasing tendency after the best model.

Then we can use the voting method to determine the model selected by each feature, which means that we determine the K number as the model selected by most model selection criteria. For example, in table 4.1, two model selection methods determine the K as 5 for the PCA-w feature set, so we choose 5 as the selected model by the PCA-w feature. If two models have the same vote, we choose the larger one, because a larger model can get better performance in clustering. For example, the SPEC selected two models, K is 3 or 4, then we prefer to use 4. The last column of this table shows the final determination of the K .

4.3 Cluster Validation

Cluster validation is a core task to measure the goodness of the clustering results. Given the true class labels (ground truth), we can analyse the clustering performance by comparing class labels with cluster results. There are many measures have been proposed in previous works[9]. In this project, we applied some external criteria of clustering quality. These measures can be divided into three categories according to different sources, which are statistics-based measures, information-theoretic measures and classification-based measures.

Most measures are computed based on the contingency matrix[25] which reflects the relationship between the true class labels and the cluster results. Table 4.2 shows the structure of the contingency matrix. Given a data set D with n samples, assume that we know the true class labels of data $Class = \{Class_1, Class_2, \dots, Class_{K'}\}$, where K' denotes the true class numbers of data. After clustering the data using k-means, we have the predicted partition of data $Cluster = \{Cluster_1, Cluster_2, \dots, Cluster_K\}$, where K is the estimated number of clusters. The contingency matrix G reflects the overlapped information between true classes and estimated clusters. The elements of G , $n_{i,j}$, is the number of data points in cluster i from class j . The following table gives the detail of contingency matrix.

A good clustering performance will get a contingency matrix in which the sum of

Table 4.2: The contingency matrix

	Class 1	Class 2	...	Class K'	Sum
Cluster 1	n_{11}	n_{12}	...	$n_{1K'}$	$n_{1.}$
Cluster 2	n_{21}	n_{22}	...	$n_{2K'}$	$n_{2.}$
...
Cluster K	n_{K1}	n_{K2}	...	$n_{KK'}$	$n_{K.}$
Sum	$n_{.1}$	$n_{.2}$...	$n_{.K'}$	n

values on the diagonal is as large as possible. According to the information reflected by the contingency matrix, cluster validity measures can be completed. In this project, we used five evaluation measurements to evaluate the performance of clustering. We briefly introduce these measures and the detailed computation of these measures can be found in table 4.3.

Table 4.3: Clustering validation measurements

Measures	Notation	Computation	Range
Entropy	E	Eq 4.24	$[0, \log K']$
Mutual Information	MI	Eq 4.25	$[0, \log K']$
Purity	P	Eq 4.26	$(0, 1]$
Rand Index	R	Eq 4.27	$(0, 1]$
F-measure	F	Eq 4.28	$(0, 1]$

Note: find the definition of n from table 4.2, and $p_{ij} = n_{ij}/n, p_i = n_{i.}/n, p_j = n_{.j}/n$

1. Information-Theoretic Measures

The following two measures are widely used methods for the evaluation of clustering results. All are calculated based on information theory.

- Entropy (E):

Entropy is an important and basic concept in information theory. It can measure the quality of clustering by calculating the entropy of samples in each cluster[40]. Entropy is a negative measures, meaning that a better clustering performance has a smaller entropy value.

$$-\sum_i p_i \left(\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i} \right) \quad (4.24)$$

- Mutual Information:

Mutual information[38] can be used to evaluate the cluster quality by measuring the amount of information cluster labels can tell about class labels. The value of mutual information can be increased by increasing the number of clusters, meaning that it can trade off quality of the clustering against the number of clusters. In order to penalise a large number of clusters, a normalisation process is adopted. The Mutual Information value is a number from 0 to $\log K'$. After normalisation, we can get a value with in the range from 0 to 1.

$$\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j} \quad (4.25)$$

2. Statistics-Based Measures

- Purity (P):

Purity measures the percentage of correctly assigned samples in the data set. The "correctly assigns" is that samples are those of the class most frequent in the cluster. But there is a defect with the purity measure, which is that large cluster numbers may lead to a high purity result. So purity cannot penalize the result with large number of clusters.

$$\sum_i p_i (\max_j \frac{p_{ij}}{p_i}) \quad (4.26)$$

- Rand index:

Rand index[28] gives the percentage of decisions that are correct which is the sum of true positive decisions and true negative decisions. True positive decisions are the number of sample pairs that come from same clusters and also from same classes, and true negative decisions are the number of sample pairs that come from different clusters and also from different classes.

$$\left(\binom{n}{2} - \sum_i \binom{n_i}{2} - \sum_j \binom{n_{.j}}{2} + 2 \sum_{ij} \binom{n_{ij}}{2} \right) / \binom{n}{2} \quad (4.27)$$

3. Classification-Based Measures

- F-measure:

F-measure[36] evaluates clustering results by combining the precision and recall concepts from information retrieval. The detailed computation is shown in the following table.

$$\sum_j p_j \max_i \left(2 \frac{p_{ij} p_j}{p_i p_j} / \left(\frac{p_{ij}}{p_i} + \frac{p_{ij}}{p_j} \right) \right) \quad (4.28)$$

4.3.1 Result of evaluation

This section presents the clustering results for the dataset containing four fish individuals. According to the selected models by each feature, we applied K-means on the test sets in the four feature spaces respectively using the selected models and compare the performances of the four feature sets. Figure 4.6 shows the comparison of the performances of four features in 5 measures. In the figure, the red dash line in the blue box denotes the median, and the green asterisk is the position of the mean. Both values show that the CFS feature did the best performance. Table 4.4 summarise the average measurements of each feature. The results show CFS is the clearly better feature set. Then we use the contingency matrix (Table 4.5) to show the clustering result of CFS feature. From this table we can see that about 95% samples of the individual 2 were grouped into the same cluster and 93% of the individual 4 were grouped together. But both the samples of individual 1 and 3 were separated into 2 different clusters.

Table 4.4: *Clustering performances of four features using selected model (4 fish individuals)*

Features	Selected Model	Evaluation Measurements				
		P	F	R	E	MI
PCA-w	5	0.824	0.786	0.576	0.467	0.916
PCA-sep	5	0.837	0.80	0.597	0.453	0.93
CFS	5	0.907	0.868	0.743	0.26	1.12
SPEC	4	0.754	0.738	0.539	0.583	0.80

Table 4.5: *The contingency matrix of CFS feature(4 fish individuals)*

	Individual 1	Individual 2	Individual 3	Individual 4
Cluster 1	21	0	0	1
Cluster 2	0	26	0	0
Cluster 3	5	0	20	0
Cluster 4	0	0	0	25
Cluster 5	0	1	6	1

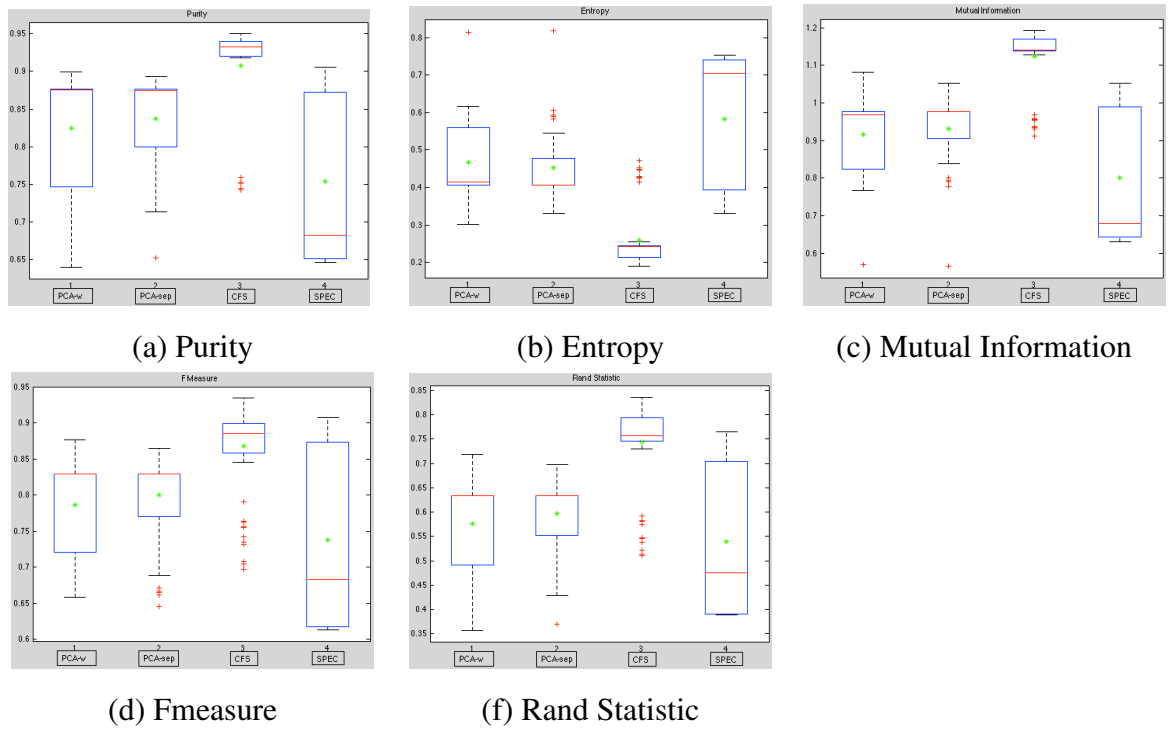


Figure 4.6: Clustering performances of the dataset with 4 fish individuals

4.4 Summary

In this chapter, we introduced the K-means algorithm, model selection strategies and clustering evaluation metrics. K-means is a popular clustering method which we used to classify different fish individuals. Using a test set with selected features, we implemented selected model selection techniques to select the best cluster number. The final model, the K number, was selected by voting by all model selection strategies. The result shows that PCA-w, PCA-sep and CFS all selected K as 5, and the SPEC select 4 which is the true class number (there are total 4 fish individuals in the dataset). We applied the K-means algorithm on each feature set using the selected model. Finally, we introduced several clustering validation methods to evaluate the clustering performance. From the results, we found that the CFS had the best performance, with purity 90.7%, F-measure 86.8% and Rand Statistic 74.3%.

Chapter 5

Evaluation

This chapter introduces some new fish individuals. We use the new data to create two new datasets which can be used to test the performance of selected features and model selection methods. In section 5.1, the new datasets are described. We also present a visualisation of four selected features using Isomap, which was introduced in section 5.2. Section 5.3 presents the results of model selection and clustering performance on the two new datasets.

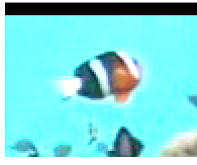

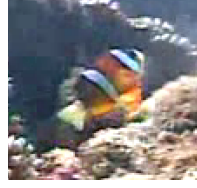
5.1 New Datasets

5.1.1 New Fish Datasets

We used the dataset containing with 4 fish individuals to do feature selection, and then tested the classification and estimation performance on them. However, the final purpose of this experiment is applying this method to a new set of fish observations which are not labeled to identify different individuals and estimate the individual number. In order to test the validity and reliability of selected features on new dataset, we picked out additional three fish individuals with a total of 307 observations from the dataset using the same strategy which we introduced in section 2.2.2. The detailed information and observation examples of these 3 new fish individuals are listed in the table 5.1.

We also combined the seven individuals together into a single dataset. The whole dataset totally contains 785 fish observations of 7 fish individuals. So, now we have 2 new fish datasets containing 3 and 7 fish individuals respectively.

Table 5.1: Information about the additional three different fish individuals

Fish No.	Individual 5	Individual 6	Individual 7
Fish Observation			
Number of Observations	84	122	101
Video Captured Site	NPP-3	NPP-3	NPP-3
Camera No.	3	2	2
Video Captured Date	2010-08-20	2010-09-25	2010-11-24

5.1.2 Features for New Fish Datasets

We applied feature extraction and feature selection based on the 4 fish individuals' dataset, which was described in section 2.2. For the new datasets, we also use PCA to extract lower dimensional features, which can generate two PCA feature spaces, the PCA-w and PCA-sep. The other two feature subsets with lower dimensional attributes were generated by using the attributes selected by CFS and SPEC based on the dataset with 4 fish individuals. The four feature sets are all extracted or selected based on the original feature vector containing 3091 attributes, which are detailedly introduced in section 3.2. The following table 5.2 summarises the dimensionality of each kind of feature for all the three datasets.

Table 5.2: Information of the additional three different fish individuals

DataSet	Number of observations	Dimensionality for each feature subset			
		PCA-w	PCA-sep	SPEC	CFS
4 Fish Dataset	478	84	287	500	50
3 Fish Dataset	307	74	265	500	50
7 Fish Dataset	785	116	327	500	50

5.2 Feature Visualisation

This section introduces a visualisation method, the Isomap, to evaluate to what extent fish individuals could be visually distinguished in a 2-D projection space of fish sam-

ples. According to the visualisation graphs, we can determine which feature space is more effective for classification. As we introduced in section 3.2, after dimensionality reduction using PCA, spectral feature selection and correlation-based feature selection, we have four feature spaces to represent the fish dataset (table 3.4). Then we did the visualisation process on the three datasets containing 4, 3 and 7 fish individuals respectively which we introduced in table 2.2 and table 5.1 and presents the visualisation result using graphs. We also applied the K-means algorithm to evaluation the performance of each feature subset of three fish datasets.

5.2.1 Isomap

Isomap[21] is an extension of MDS[19], aiming at capturing the nonlinear structure in the data. Multidimensional scaling (MDS) provides a geometrical representation of samples reflecting the spatial distribution of all samples. using a dissimilarity measure. The basic idea is to seek a projection of data points by iteratively minimising the approximation error. In metric multidimensional scaling, the function which allows to transform the dissimilarity into distances are calculated by minimising the following function, called the metric stress function in the context of MDS:

$$S = \left[\frac{\sum_{i=1}^n \sum_{j>i}^n (D_{ij} - \|\bar{x}_i - \bar{x}_j\|)^2}{\sum_{i=1}^n \sum_{j>i}^n D_{ij}^2} \right]^{\frac{1}{2}} \quad (5.1)$$

$D_{i,j}$ is the dissimilarity between two samples. S contains a set of new projections. We can visualize the data points in a 2D space using the top two projections.

For the Isomap method, the Euclidean distances between data points are replaced by geodesic distances which can be done by constructing a neighbourhood graph. Then computing the length of a graph shortest path distances between data points in the graph and finally applying classical MDS to the graph distance matrix to construct a low dimensional embedding. We find the 2D dimensional embedding of data points to visualize the spatial distribution of samples, where points nearby are mapped nearby in the 2D dimensional space and points far away are mapped far away.

5.2.2 Feature visualisation

We did visualisation on the three datasets with different features. The table 5.2 gives the detail of the three datasets and the four features. The visual 2D projection results obtained for each feature are shown below.

1. Dataset with 4 Fish Individuals

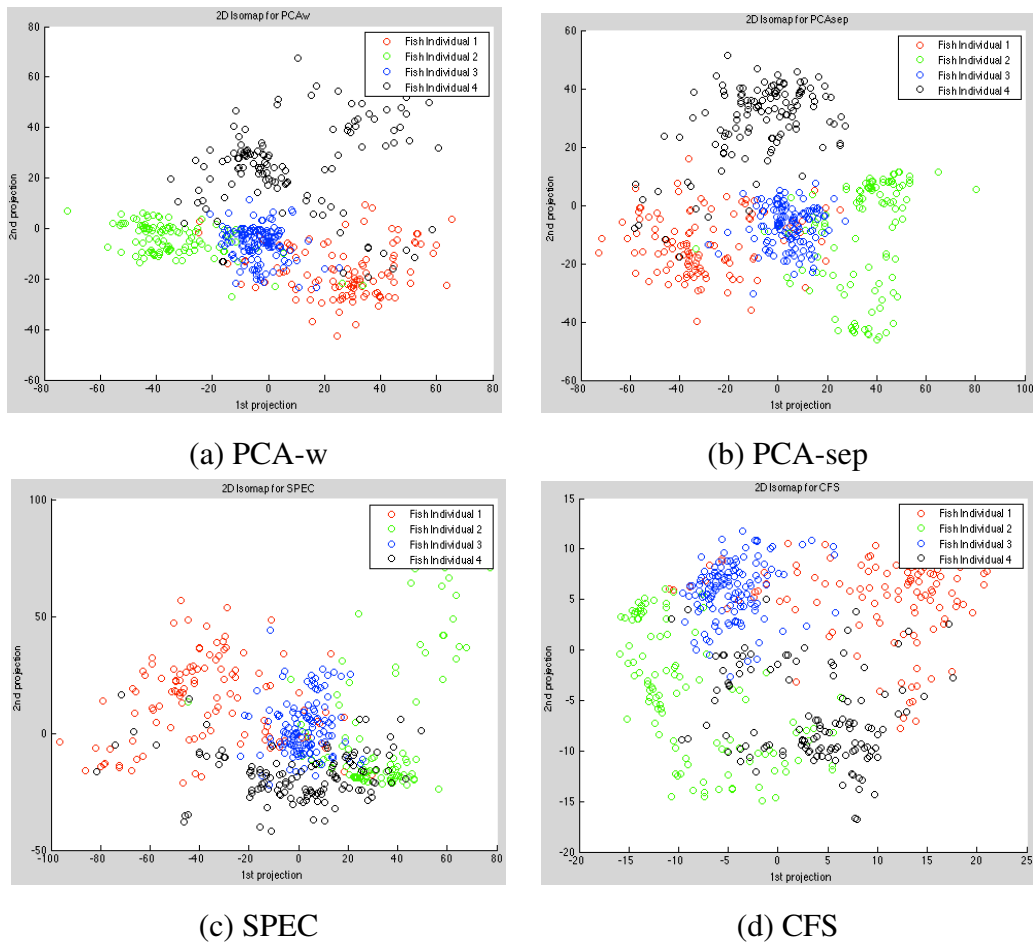
Visualization

Figure 5.1: Feature Visualisation for 4 Fish Individuals' Dataset

Figure 5.1 shows the four features' visualisation of the dataset with 4 fish individuals. Figure 5.1(a) shows the visualisation of PCA-w feature. It is obvious that all fish individuals were well separated from each other. Especially for the individual 2 and 3, most samples were grouped together, while the data points of fish individual 4 were scattered, which means that these samples were tended to be grouped into two clusters. Figure 5.1(b) shows the result of the PCA-sep feature space. Among all features, it had the best performance, where the four fish individuals were clearly separated into 4 groups except for a little bit dispersion of the samples within the individual 2. The separation of the same individual may be because of the different illumination of fish samples. Figure 5.1(c) is the visualisation result of SPEC feature. This graph shows that this feature was the most unreliable one, in which samples were scattered and the four individuals

can not be clearly separated. Most samples of individual 1 and 4 were overlapping. Figure 5.1(d) gives the 2D projection of CFS feature space. Despite that the samples within each individual were scattered, the distance of intra-classes is large, meaning that different individuals were well separated.

2. Dataset with 3 Fish Individuals

Visualization

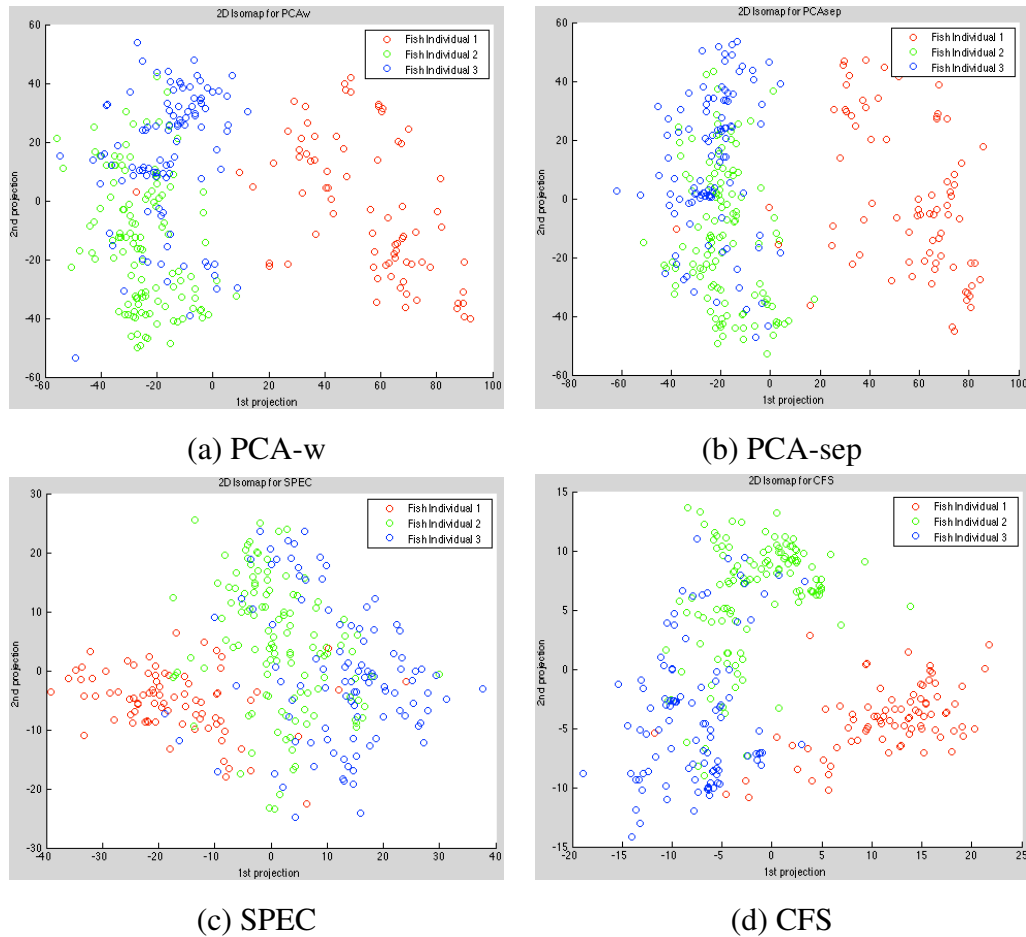


Figure 5.2: Feature Visualisation for 3 Fish Individuals' Dataset

Figure 5.2 shows the four features' visualisation of the dataset with 3 fish individuals. The SPEC feature space and CFS feature space use attributes selected by using the dataset with 4 fish individuals. From all graphs in the figure 5.2, we can see that the samples of individual 1 which is represented by red circles are far away from data points of other two fish individuals, which means that the individual 1 can be easily separated from the other two individuals. But the data points of individual 1 in PCA-w and PCA-sep 2D projection figures ((a) and (b))

are scattered. Samples of individual 2 and 3 show some overlap in all feature spaces.

3. Dataset with 7 Fish Individuals

Visualization

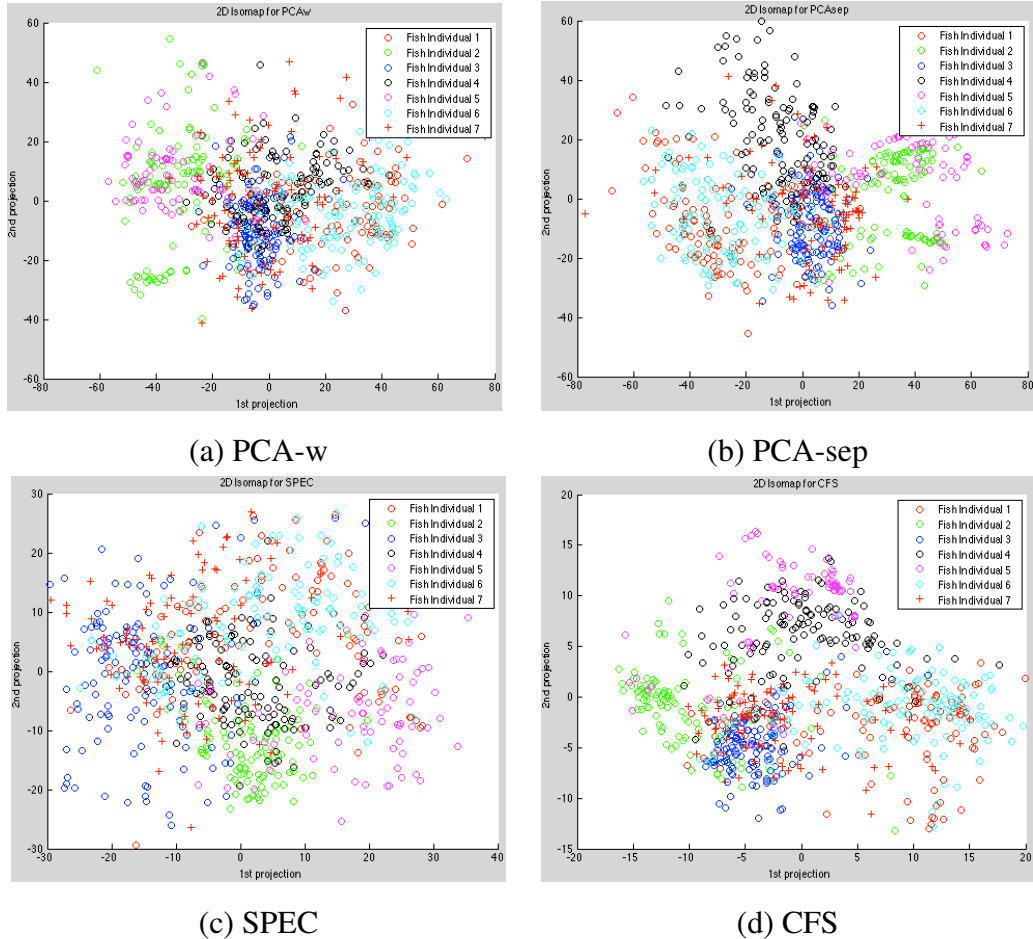


Figure 5.3: Feature Visualisation for 7 Fish Individuals' Dataset

Figure 5.3 shows four features' 2D projection of the dataset with 7 fish individuals. It shows that all features can not separate these 7 individuals clearly, because it is difficult to separate all classes clearly in a 2 dimensional space. The SPEC feature set is the worst intuitively, because all data points of the same individual are scattered. Despite that the overlap of samples presents between different individuals in PCA-w and PCA-sep feature space, the data points of the same fish individual are compact. Data points of some particular individuals show dispersion in the 2D projection space of CFS feature.

5.3 Evaluation

This section shows some figures to describe the model selection results and clustering performances of two new datasets with the four feature spaces.

5.3.1 Model Selection Results

Using the same model selection strategies described in section 4.2.4, we presents the selection results below.

1. Model selection results for the dataset containing 3 fish individuals

Table 5.3 summarises the model selection results of all feature spaces. Figure 5.4 to figure5.7 plot the selection results of all strategies for all features.

Table 5.3: *Model selection results for the 3 fish individuals' dataset*

Features	Model Selection Strategies					Model Determination
	AIC	BIC	J_3	Dunns Index	Silhouette Index	
PCA-w	8	5	4	3	4	4
PCA-sep	3	3	3	4	4	3
CFS	3	3	4	3	7	3
SPEC	2	2	4	3	5	2

2. Model selection result for the dataset containing 7 fish individuals

Table 5.4 summarises the model selection results of all feature spaces. Figure 5.8 to figure 5.11 plot the selection results of all strategies for all features.

Table 5.4: *Model selection results for the 7 fish individuals' dataset*

Features	Model Selection Strategies					Model Determination
	AIC	BIC	J_3	Dunns Index	Silhouette Index	
PCA-w	16	10	4	4	5	4
PCA-sep	10	4	4	4	5	4
CFS	12	8	4	8	9	8
SPEC	7	4	4	2	7	7

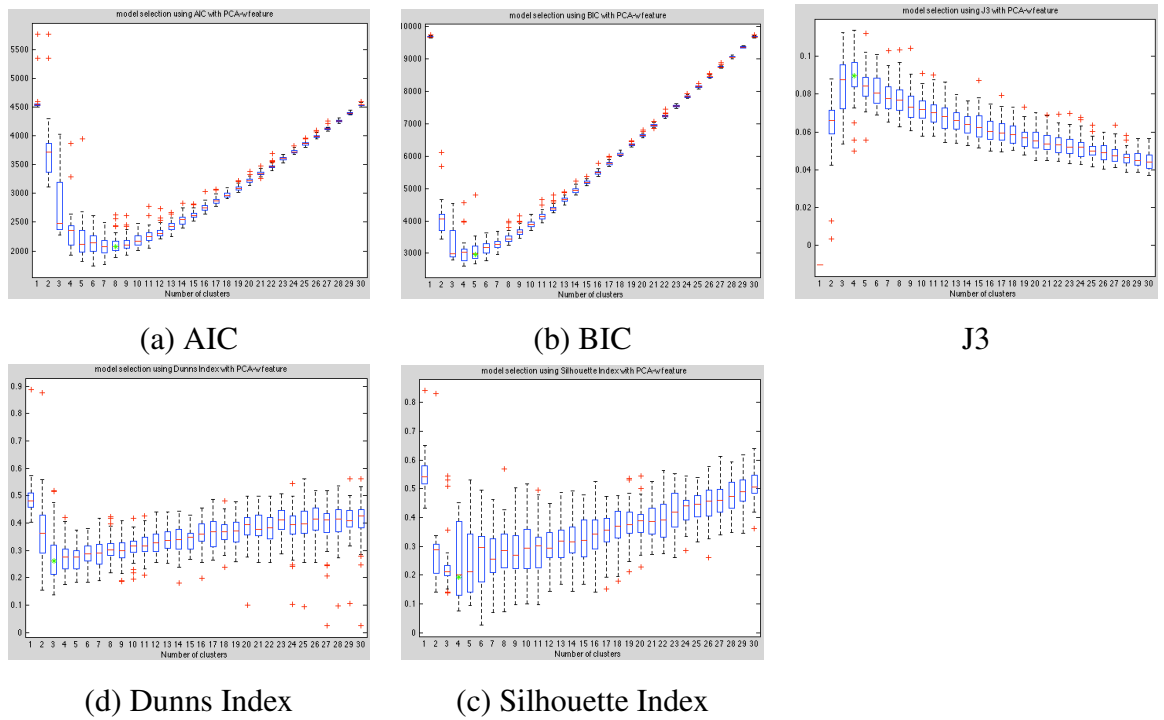


Figure 5.4: Model Selection Results for PCA-w Feature (3 fish individuals)

5.3.2 Clustering Performance

This section evaluates the clustering performance on the two new datasets in the four different feature spaces according to the selected models in last section.

1. Clustering performance of the dataset containing 3 fish individuals

We present the clustering results for the dataset containing three fish individuals. According to the selected model for each feature set, we applied K-means on a half of randomly selected samples of dataset in the four feature spaces respectively using selected models and compare the performances of the four features. The K-means algorithm ran for 100 trials, and the final evaluation results were the average value. Figure 5.12 shows the comparison of the performances of the four features sets on 5 measures. In the figure, the red dash line in blue box denotes the median, and the green asterisk is the position of mean. Comparing both values of all measurements, it shows that the CFS feature had the best performance. Table 5.5 summarises the average measurements of each feature. Then we use the contingency matrix (Table 5.6) to show the clustering result of CFS feature.

2. Clustering performance of the dataset containing 7 fish individuals

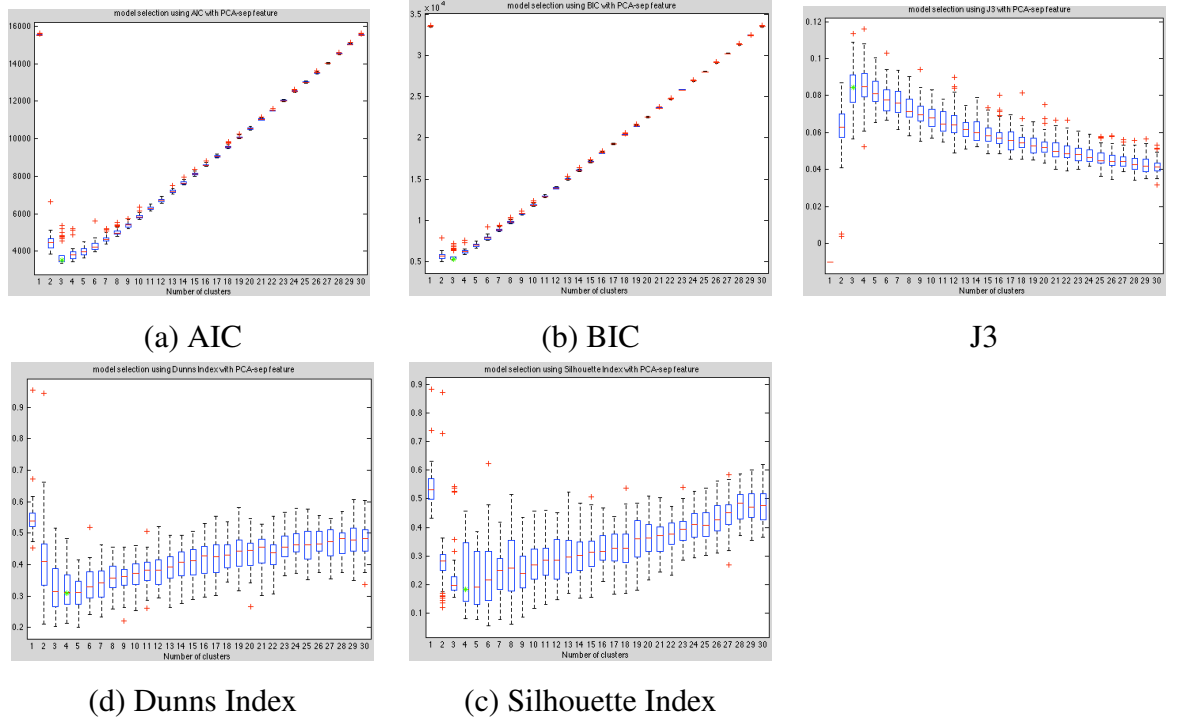


Figure 5.5: Model Selection Results for PCA-sep Feature (3 fish individuals)

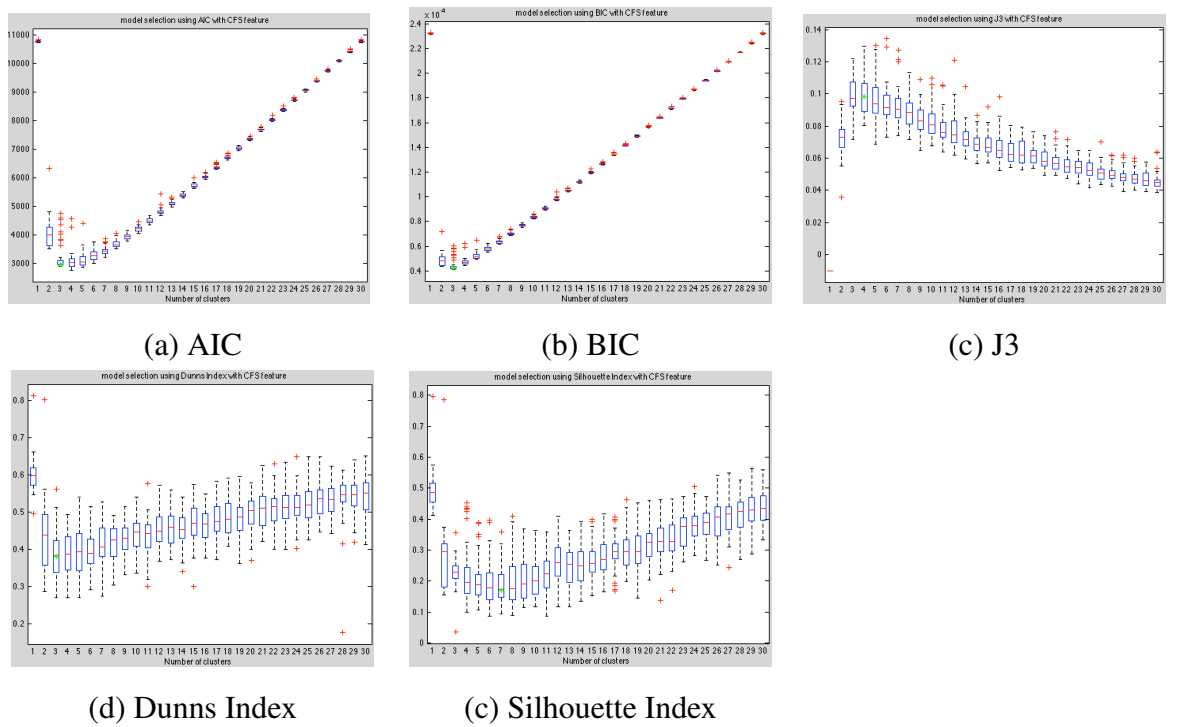


Figure 5.6: Model Selection Results for CFS Feature (3 fish individuals)

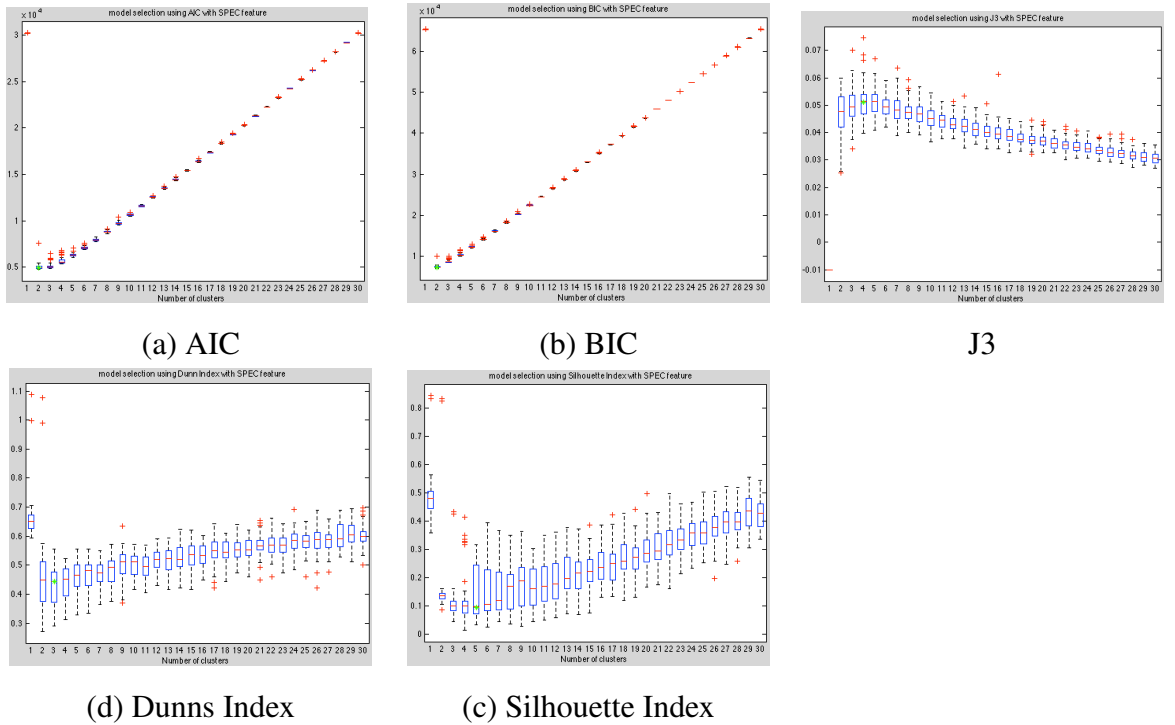


Figure 5.7: Model Selection Results for SPEC Feature (3 fish individuals)

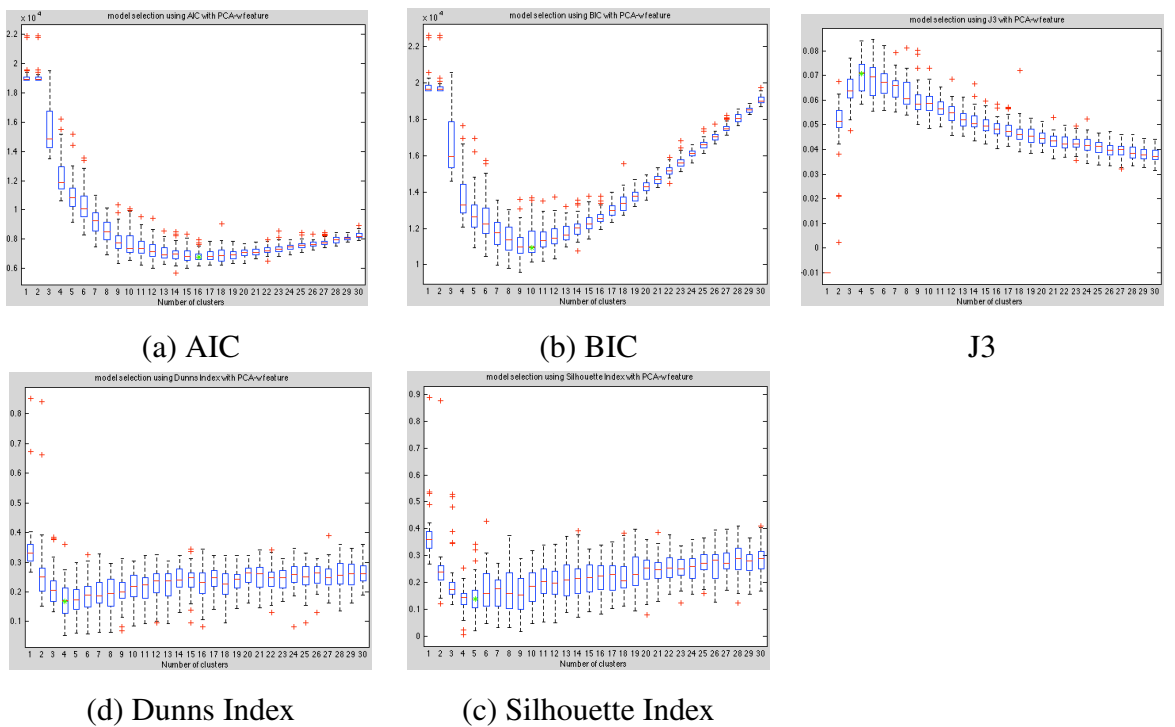


Figure 5.8: Model Selection Results for PCA-w Feature (7 fish individuals)

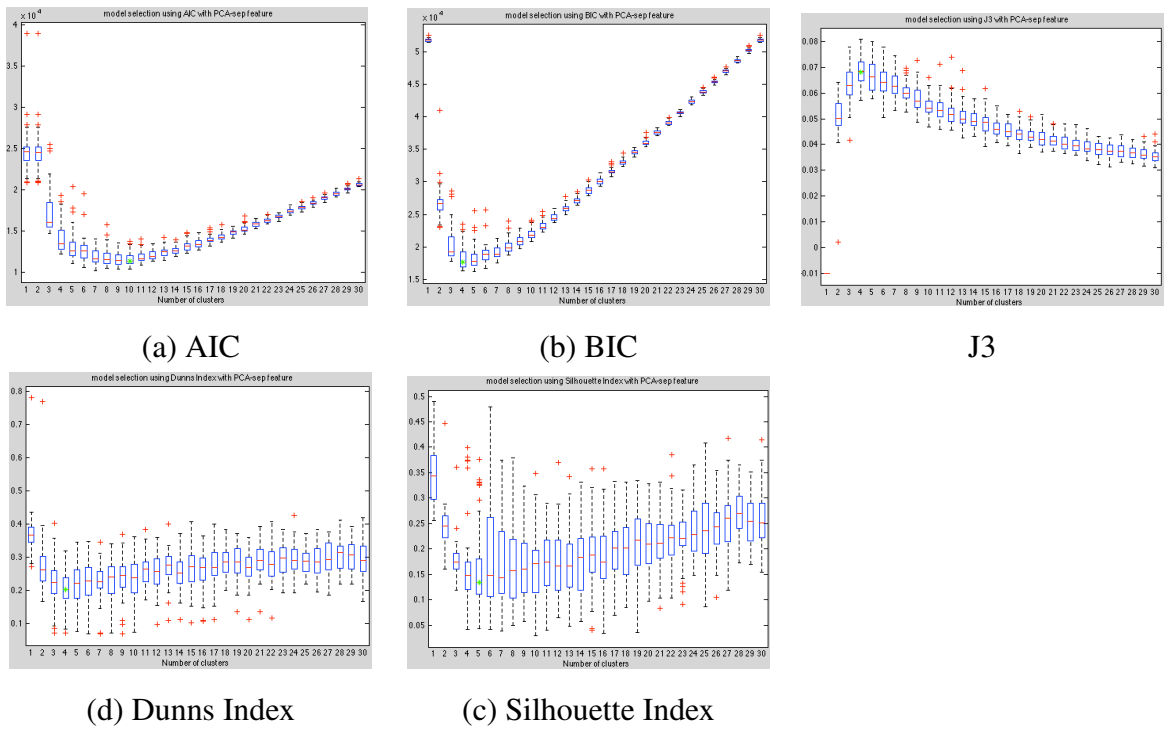


Figure 5.9: Model Selection Results for PCA-sep Feature (7 fish individuals)

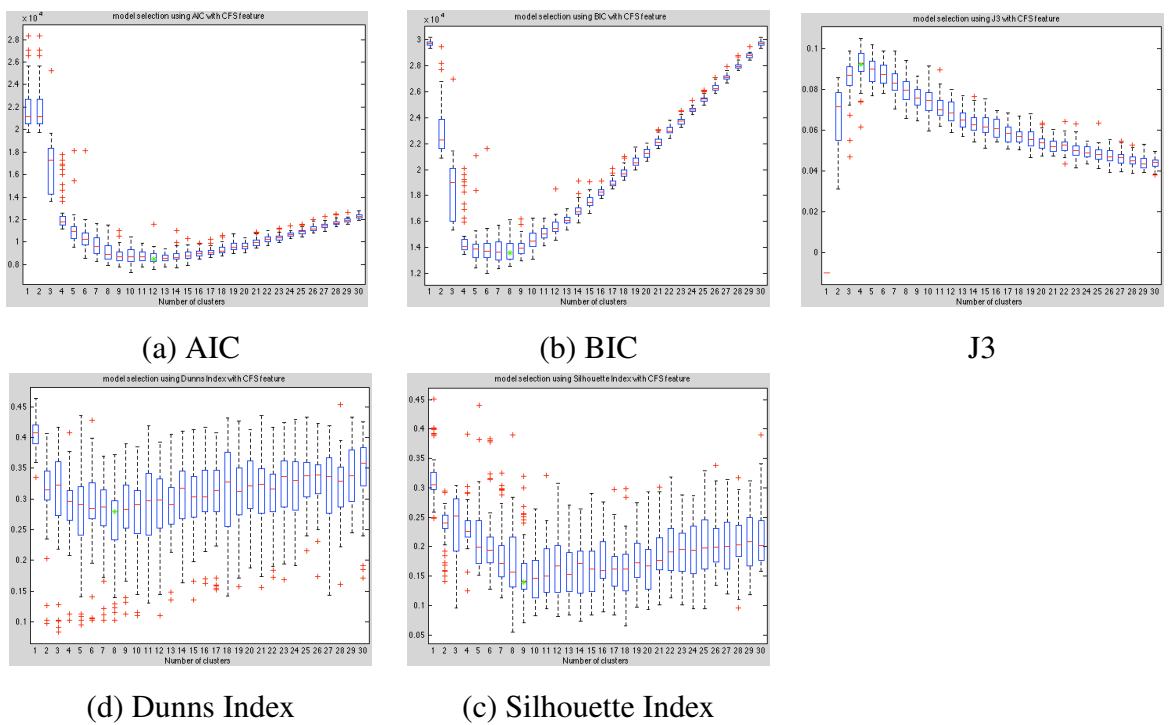


Figure 5.10: Model Selection Results for CFS Feature (7 fish individuals)

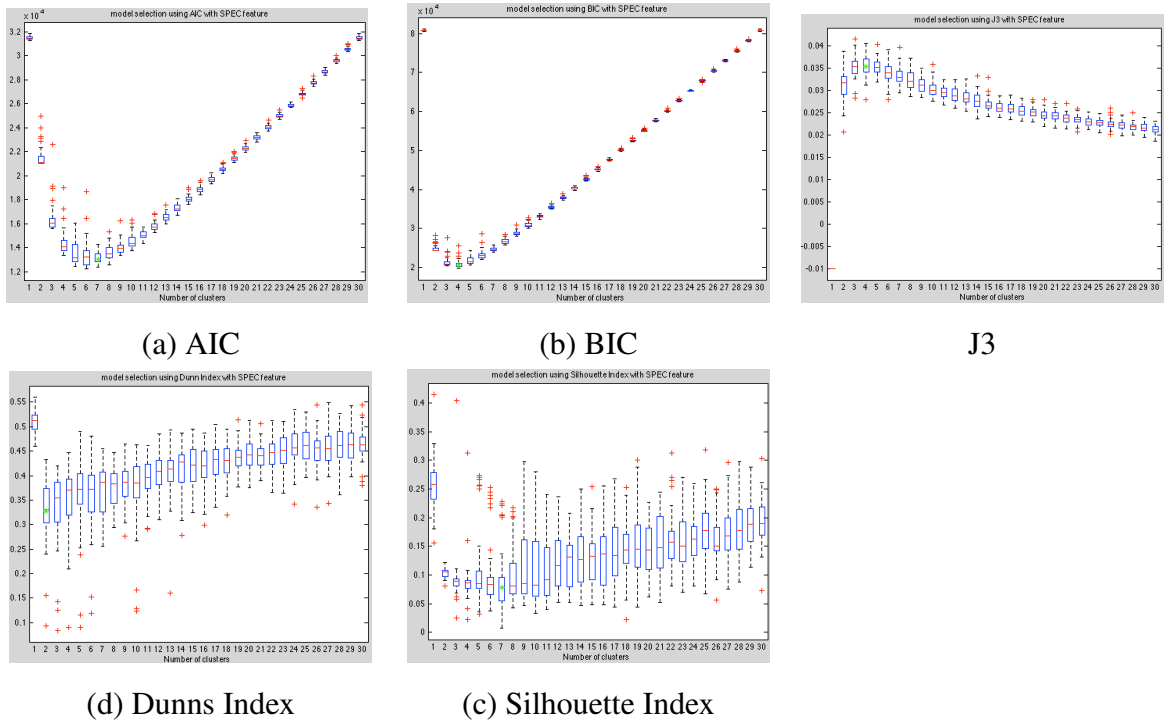


Figure 5.11: Model Selection Results for SPEC Feature (7 fish individuals)

Table 5.5: Clustering performances of four features using selected model (3 fish individuals)

Features	Selected Model	Evaluation Measurements				
		P	F	R	E	MI
PCA-w	4	0.853	0.814	0.562	0.400	0.685
PCA-sep	3	0.829	0.837	0.581	0.443	0.642
CFS	3	0.856	0.864	0.638	0.379	0.705
SPEC	2	0.634	0.698	0.378	0.722	0.362

Table 5.6: The contingency matrix of CFS feature (3 fish individuals)

	Individual 1	Individual 2	Individual 3
Cluster 1	43	11	6
Cluster 2	6	49	0
Cluster 3	0	0	39

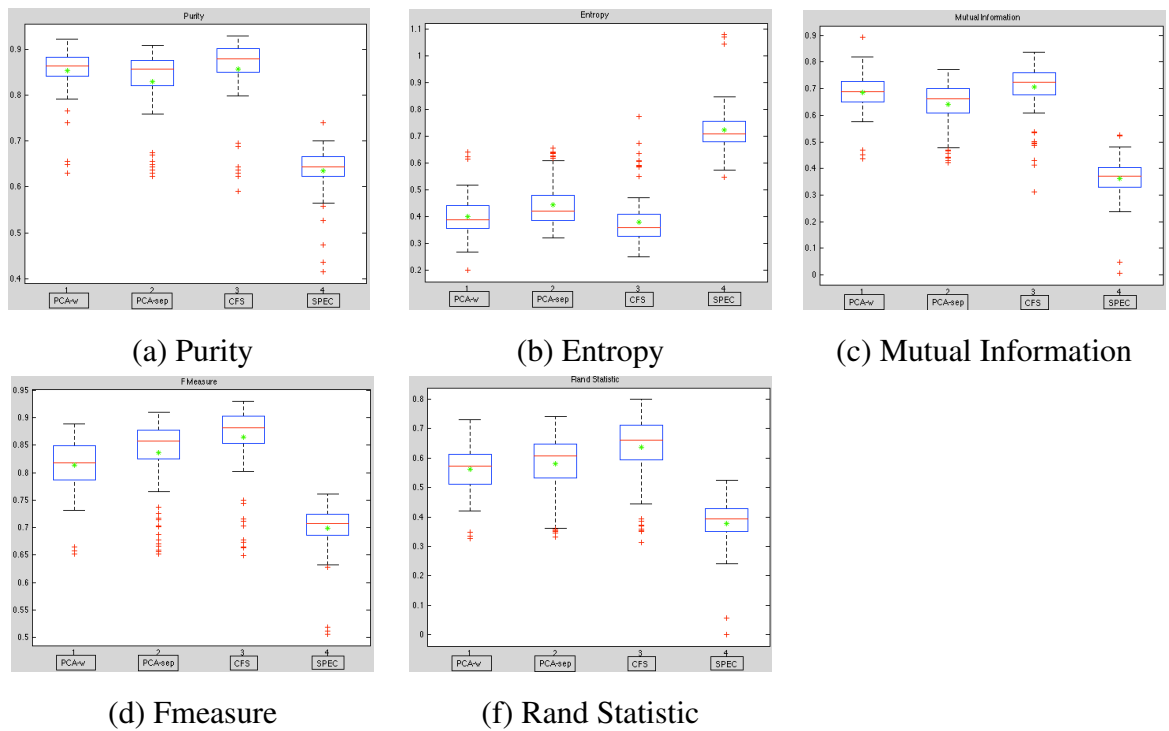


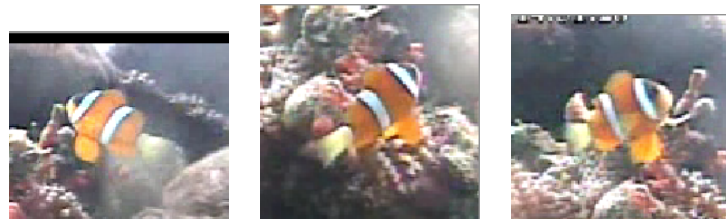
Figure 5.12: Clustering performances of the dataset with 3 fish individuals

We present the clustering results for the dataset containing seven fish individuals. According to the selected model for each feature set, we applied K-means on a half of randomly selected samples of dataset in the four feature spaces respectively using selected models and compare the performances of the four feature sets. The K-means algorithm ran for 100 trials, and the final evaluation results were the average value. Figure 5.14 shows the comparison of the performances of the four feature sets on 5 measures. It also shows that the CFS feature had the best performance. Table 5.7 summarises the average measurements of each feature. We use the contingency matrix (Table 5.8) to show the clustering result of CFS feature. From this matrix, we can see that all samples of the individual 3 were grouped into the same cluster. Most samples of the individuals 2,4 were grouped together. Samples of the individuals 5 and 7 seems scattered. But the individual 1 and the individual 6 were identified as the same individual. The video of fish individual 1 was captured in 25/09/2010, and the the video of fish individual 6 was captured in 08/11/2010. Two videos are captured by the same camera, the camera 2. The reason why we labeled these two fishes as different individuals is that although the captured date of two videos are only 2 month apart, the appearance of observations of these two fishes are slightly different.

The figure 5.13 gives some observation images of these two fishes. We can see that the individual 1 is slightly darker than the individual 6. However, as the groundtruth was given manually, there must be some mistakes in labelling, meaning that the individual 1 and the individual 6 may be the same fish. The appearance differences between these two sets of observations maybe because the same fish was captured under the different illumination condition. So, this experiment may help us to correct the labelling error.



(a) Observations of the the fish individual 1



(b) Observations of the the fish individual 6

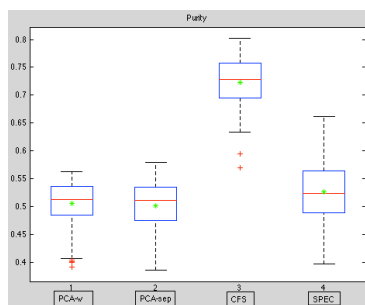
Figure 5.13: Comparison of the fish individual 1 and 6

Table 5.7: Clustering performances of four features using selected model (7 fish individuals)

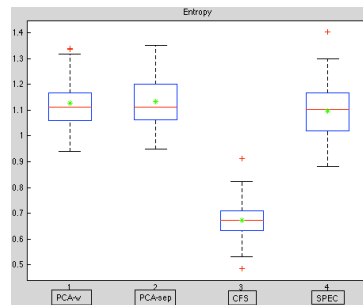
Features	Selected Model	Evaluation Measurements				
		P	F	R	E	MI
PCA-w	4	0.506	0.576	0.350	1.126	0.80
PCA-sep	4	0.501	0.57	0.345	1.134	0.799
CFS	8	0.723	0.712	0.534	0.672	1.260
SPEC	7	0.527	0.544	0.317	1.096	0.837

Table 5.8: The contingency matrix of CFS feature (7 fish individuals)

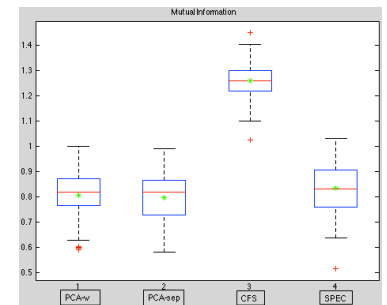
	Ind. 1	Indiv. 2	Indiv. 3	Indiv. 4	Indiv. 5	Indiv. 6	Indiv. 7
Cluster 1	45	0	0	5	0	43	6
Cluster 2	0	42	0	0	12	0	0
Cluster 3	12	0	64	1	2	15	4
Cluster 4	0	0	0	59	0	0	0
Cluster 5	0	0	0	0	30	0	0
Cluster 6	0	0	0	0	0	1	13
Cluster 7	1	6	0	0	1	0	24
Cluster 8	1	0	0	6	0	0	0



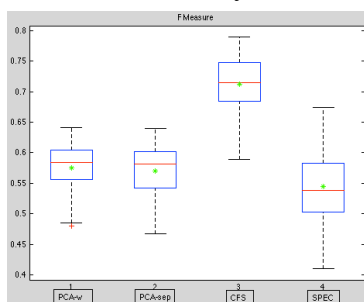
(a) Purity



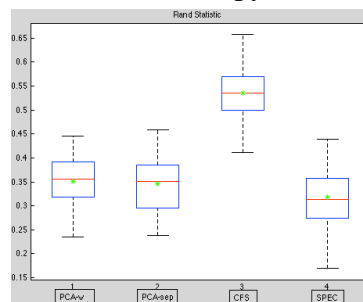
(b) Entropy



(c) Mutual Information



(d) Fmeasure



(e) Rand Statistic

Figure 5.14: Clustering performances of the dataset with 7 fish individuals

5.3.3 Summary

In this chapter, we created two new datasets in order to evaluate the performance of selected features on new data and test how well the model selection strategies will be performed by using different features. One dataset contains 3 different fish individuals which are totally different from the previous four fish individuals. Another dataset combined the dataset with 4 fish individuals and the dataset with 3 fish individuals. We also use the K-means algorithm to do clustering on data and evaluate the performances. Table 5.9 summarises the model selection results for the two datasets using four features. In section 5.3.2, the evaluation results showed that the CFS feature set had the best clustering performance in both datasets. Table 5.10 summarises the results. In addition, the CFS feature set can also select an approximately correct cluster number (the K number).

Table 5.9: *Model selection results for the 3 datasets*

Dataset	Selected model by each feature			
	PCA-w	PCA-sep	CFS	SPEC
3 Fish Dataset	4	3	3	2
7 Fish Dataset	4	4	8	7

Table 5.10: *Clustering performance of two datasets using the CFS feature*

Dataset	Feature	Selected Model	Evaluation Measurements				
			P	F	R	E	MI
3 Indv. Dataset	CFS	3	0.856	0.864	0.638	0.379	0.705
7 Indv. Dataset	CFS	8	0.723	0.712	0.534	0.672	1.260

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The main purpose of this project is to identify individual clown fish and estimate the number of different individuals by taking machine learning approaches. The image data were collected from the underwater cameras which are fixed in 10 different places around Taiwan Island, and most clown fish observations come from the NPP-3 site which is a southeast harbour of Taiwan. We picked out 7 different individuals with 785 images to build the the experiment dataset, in which 4 of them containing with 478 observations were used for training in order to select reliable features. Additional 3 fish individuals with 307 observations and the combination of all 7 fish individuals with 785 observations were then used to do the experiments.

In the feature description chapter, we introduced 4 new features to represent fish images and presented the discrimination of each feature. Combining with features explored in previous works, a final 3091 high dimensional feature vector was created to represent each fish image. In order to reduce the dimensionality of features, feature extraction and selection methods were applied. Feature selection methods are divided into two categories according to whether we know the groundtruth of dataset. If the groundtruth is known previously, the feature selection procedure is supervised, otherwise, it is unsupervised. Both PCA and spectral feature selection are unsupervised feature selection methods. Correlation-based feature selection explore the correlation between features and classes, which is therefore supervised. The 2D projection visualisation shows that PCA-sep did the best separation of different classes, but the CFS method gave features that produced the best clustering. Then, we implemented sequential forward algorithm to as a feature selection procedure. By applying the feature

extraction and selection algorithms on the dataset with 4 fish individuals, we created four new sub-features.

The classification method used in this project was K-means, and the Euclidean distance metric was used. We tried a series of K and find the best K using the clustering model. Several model selection methods were used for selecting the best model, and the suitable model was determined by the voting between all selection strategies. The model selection results shows that the CFS feature set can select an approximate cluster number in all datasets (selected 5 in 4 fish individual dataset, selected 3 in 3 fish individual dataset and select 8 in 7 fish individual dataset). The CFS feature had the best clustering performance in all datasets as well. Despite that the SPEC feature set can select approximate K clusters, the performance of clustering was worse than the CFS feature set. We also found that the PCA-w and PCA-sep feature sets performed well using selected models, but the model was not reasonable when the dataset contains a larger number of different individuals, for example, they all selected 4 as the number of clusters while the true class number is 7.

6.2 Future Work

There are a lot work should be done for further research. The biggest limitation in our work is that we do not have enough clown fish data of a large number of individuals. We only collected 7 fish individuals with 785 observations, which is a small dataset to do experiments. So, more work should be done in the data collection and the groundtruth labelling. Exploring useful features to represent fish is an important task. Simple features were not discriminative enough to perfectly distinguish different fish individuals of the same species. In this project we developed some useful features, and some new reliable features should be explored in the future. We can also investigate some more effective clustering algorithms instead of k-means which is too simple to deal with complex clustering tasks.

Appendix A

The description of 19 feature types

Feature Type.	Size	Name	item
1	510	NormalRG Colour hist	Head/Tail/Top/Bottom/Whole
2	255	H hist in HSV	Head/Tail/Top/Bottom/Whole
3	110	Normal RG Color (re-hist)	Head/Tail/Top/Bottom/Whole
4	55	H hist in HSV (re-hist)	Head/Tail/Top/Bottom/Whole
5	1	CS ratio	
6	1	CS half tail area ratio	
7	12	Density	RGB
8	720	Co-occurrence matrix	Correlation/Energy/Homogeneity InvDiffMoment/ClusterShade ClusterProminen/MaxProbilirity Autocorrelation
9	42	Moment Invariants	Head/Tail/Top/Bottom/Whole Half-head/Half-tail
10	680	HOG	Level 0/Level 1/Level 2/Level 3
11	15	Fourier	
12	160	Gabor	Head/Tail/Top/Bottom/Whole
13	63	Affine Moment Invariants	Head/Tail/Top/Bottom/Whole Half-head / Half-tail
14	2	head and tail Area ratio	Head/Tail
15	153	colour ratio	Chromatic body and white stripes
16	153	colour ratio	Top and bottom parts of the chromatic body

Feature Type.	Size	Name	item
17	153	colour ratio	Front and posterior parts of the chromatic body
18	5	stripes length ratio	
19	1	stripes area ratio	

Bibliography

- [1] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.
- [2] Joachim M Buhmann. Information theoretic model selection in clustering.
- [3] Martin D Buhmann et al. Radial basis functions. *Acta numerica*, 9(1):1–38, 2000.
- [4] Shelley Derksen and H. J. Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282, 1992.
- [5] Geoff Dougherty. *Pattern Recognition and Classification*, volume 1. springer New York, 2013.
- [6] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.
- [7] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.
- [8] Alberto Faro, Daniela Giordano, and Concetto Spampinato. Adaptive background modeling integrated with luminosity sensors and occlusion processing for reliable vehicle detection. *Intelligent Transportation Systems, IEEE Transactions on*, 12(4):1398–1412, 2011.
- [9] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*, volume 20. Siam, 2007.
- [10] Edwin Ernest Ghiselli. *Theory of psychological measurement*, volume 13. McGraw-Hill New York, 1964.
- [11] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [12] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.
- [13] Mark A Hall and Lloyd A Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *FLAIRS Conference*, pages 235–239, 1999.

- [14] Greg Hamerly and Charles Elkan. Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 600–607. ACM, 2002.
- [15] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.
- [16] Phoenix X Huang, Bastiaan J Boom, and Robert B Fisher. Hierarchical classification for live fish recognition.
- [17] P Kontkanen, P Myllymäki, and H Tirri. Comparing bayesian model class selection criteria by discrete finite mixtures. *Information, Statistics and Induction in Science*, pages 364–374, 1996.
- [18] Rasmus Larsen, Hildur Olafsdottir, and Bjarne Kjær Ersbøll. Shape and texture based classification of fish species. In *Image Analysis*, pages 745–749. Springer, 2009.
- [19] Tao Li and Chris Ding. The relationships among various nonnegative matrix factorization methods for clustering. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 362–371. IEEE, 2006.
- [20] R Lleti, M Cruz Ortiz, Luis A Sarabia, and M Sagrario Sánchez. Selecting variables for k -means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 515(1):87–100, 2004.
- [21] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- [22] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.
- [23] Rory McGrath. Learning to recognise fish. *Master's thesis, University of Edinburgh*, 2011.
- [24] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- [25] Karl Pearson. *On the theory of contingency and its relation to association and normal correlation; On the general theory of skew correlation and non-linear regression*. Cambridge University Press, 1904.
- [26] DT Pham, SS Dimov, and CD Nguyen. Selection of k in k -means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1):103–119, 2005.
- [27] John Ross Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

- [28] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- [29] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- [30] Carla M Santos-Pereira and Ana M Pires. Robust clustering method for the detection of outliers: Using aic to select the number of clusters. In *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications*, pages 409–415. Springer, 2013.
- [31] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.
- [32] Ohad Shamir and Naftali Tishby. Stability and model selection in k-means clustering. *Machine learning*, 80(2-3):213–243, 2010.
- [33] Yi-Haur Shiau, Sun-In Lin, Yi-Hsuan Chen, Shi-Wei Lo, and Chaur-Chin Chen. Fish observation, detection, recognition and verification in the real world.
- [34] Concetto Spampinato, Daniela Giordano, Roberto Di Salvo, Yun-Heh Jessica Chen-Burger, Robert Bob Fisher, and Gayathri Nadarajan. Automatic fish classification for underwater species behavior understanding. In *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, pages 45–50. ACM, 2010.
- [35] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.
- [36] Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.
- [37] NJC Strachan, Paul Nesvadba, and Alastair R Allen. Fish species recognition by shape analysis of images. *Pattern Recognition*, 23(5):539–544, 1990.
- [38] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64, 2000.
- [39] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
- [40] Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. *Machine Learning*, 2001.
- [41] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007.