Technical Report

UCAM-CL-TR-464
ISSN 1476-2986

Number 464

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Feature representation for the automatic analysis of fluorescence in-situ hybridization images

Boaz Lerner, William Clocksin, Seema Dhanjal, Maj Hultén, Christipher Bishop

May 1999

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

Fast and accurate analysis of fluorescence in-situ hybridization (FISH) images will depend mainly upon two components: a classifier to discriminate between artifacts and valid signal data, and well discriminating features to represent the signals. Our previous work has focused on the first component. To investigate the second component, we evaluate candidate feature sets by illustrating the probability density functions and scatter plots for the features. This analysis provides insight into dependencies between features, indicates the relative importance of members of a feature set, and helps in identifying sources of potential classification errors. The analysis recommends several intensity and hue-based features for representing FISH signals. The recommendation is confirmed by the probability of misclassification using a two-layer neural network (NN), and also by a feature selection technique making use of a class separability criterion. Represented by these intensity and hue-based features, 90% of valid signals and artifacts are correctly classified using the NN.

1 Introduction

Fluorescence in-situ hybridization (FISH) allows the detection of specific DNA sequences in intact cells and chromosomes. It enables selective staining of various sequences in interphase nuclei and therefore the detection, analysis

and quantification of specific numerical and structural chromosomal abnormalities within these nuclei. FISH is a widespread and diversely applied technology. Among the fields of biology in which FISH is employed include karyotype analysis, gene mapping, DNA replication and recombination, clinical diagnosis and monitoring of disease, radiation dosimetry, gene transcription and expression and the study of chromatin organization and structure [1]. This diversity of applications is complemented by a similar diversity in the types of probes, detection systems and multi-waveband epifluorescence filter cubes that are currently being employed for FISH.

Digital microscopy in FISH allows the application of image analysis techniques for automation of time consuming tasks, such as dot counting. Dot counting, the enumeration of signals (also called dots or spots) within the nuclei, is considered as one of the most important applications of FISH. One approach to dot counting relies on an auto-focusing mechanism to select the 'clearest' image for the analysis [2, 3]. However, basing dot counting on auto-focusing can have some shortcomings [4]. First, special instrumentation is required to focus the stage under computer control. Second, as image acquisition depends upon finding a single 'optimal' image, it can fail if the mechanism focuses on artifacts such as debris or background fluorescence, or if the field of view is empty. Manual inspection for discarding such images is sometimes inevitable. Moreover, even if the 'clearest' image is found, it can only represent a section of a three-dimensional image, where signals in other

sections, which are above or below that section, are out of focus. Finally, automatic focusing is found to be both time-consuming and a source for a large percentage of the total error rate of the analysis.

Recently, it has been proposed [4] to base FISH dot counting on images that are sampled at a fixed focal plane. This method is motivated by the assumption that nuclei are approximately uniformly distributed in the sample, so that translations at a fixed focal plane will provide a statistically equivalent sample as projections through different focal planes. Images can be captured by any scanning method of the slide, and the microscope stage can stop for collecting images arbitrarily, even at random. Randomly-captured images in a fixed focal plane 'intersect' nuclei on the slide at random sections, which are equivalent to those encountered by the auto-focusing mechanism. This method enables most of the shortcomings of auto-focusing to be overcome, since it shortens the length of image acquisition and requires no special instrumentation. However, the system needs to acquire sufficient analysable images and to exploit most of the information contained within these images in order to enable dot counting. It may have to deal with more unfocused nuclei and signals, and so its ability to distinguish between focused and unfocused signals should be better than that of a system employing an auto-focusing mechanism. Therefore, the proposed system is based on extracting characteristics that discriminate between focused and unfocused signals, and on a highly-accurate classifier, trained using large numbers of examples of

the two classes. A two-layer perceptron neural network (NN), trained by a scaled conjugate gradient algorithm, was found [4] to be a highly-accurate classifier of FISH signals into real (valid) signals and artifacts of two fluorophores. In the present paper we aim to find well-discriminating feature representations of FISH signals to ensure an efficient and accurate classification. Together, the two components provide a complete framework for automatic signal classification in FISH images.

Section 2 of the paper describes the procedure we use to acquire FISH images, while Section 3 depicts a methodology for multi-spectral FISH image analysis. Sections 4 and 6, respectively, describe the applications of feature extraction and feature selection to signal measurements, which are introduced in Section 5. Section 7 presents an NN-based classifier of signals into valid and artifact signals of two colours. An evaluation of feature representations for FISH signals by visual analysis of scatter plots and probability density functions, as well as by a scatter criterion and the classifier probability of misclassification, is given in Section 8, whereas conclusions for the paper are discussed in Section 9.

2 Biological materials and methods

2.1 Slide preparation

The interphase nuclei preparations from amniotic fluid were made using the method by Klinger *et al.* [5] with minor modifications. 1-2ml of amniotic fluid was centrifuged and the cell pellet washed in PBS warmed to 37°C. The cells were resuspended in 75mM Potassium Chloride (KCl) and put directly on to slides coated with APES (Sigma) and incubated at 37°C for 15 minutes. Evaporation of PBS was compensated with filtered distilled water. Excess fluid was carefully removed and replaced with 100ml of 3% Carnoys fixative, 70% 75mM KCl at room temperature for 5 minutes. The excess fluid was carefully removed and 5 drops of fresh fixative were dropped on to the cell area. Slides were briefly dried on a 60°C hotplate, and then either used immediately for hybridization or dehydrated through an alcohol series and stored at -20°C until required.

2.2 Hybridization

Target areas were marked on the slides using a diamond tipped scribe. Target DNA was denatured by immersing in 70% formamide:30% 2xSSC at 73°C for 5 minutes. 10 μL of probe mix containing spectrum orange LSI 21 and spectrum green LSI 13 (Vysis UK) was applied to the target area and a coverslip placed over the probe solution. Coverslips were sealed using

rubber cement and slides placed in a pre-warmed humidified container in a 37°C incubator for 16 hours. Coverslips were removed and slides washed in 0.4xSSc/0.3%NP-40 solution at 73°C for 2 minutes. Slides were then placed in 2xSSC/0.1% NP-40 solution at room temperature for 1 minute. When completely dried 10 μL of DAPI II counterstain (Vysis UK) was applied to the target area and sealed under a coverslip.

2.3 Fluorescence microscopy

Slides were screened under a Zeiss axioplan epifluorescence microscope using x100 objective. Signals were viewed using appropriate filters and images acquired using a CCD camera and SmartCapture software (Vysis UK). Slides were scanned by starting in the upper left corner of the coverslip and moving from top to bottom. Images were captured by stopping at random intervals. Red and green signals were seen on blue DAPI stained nuclei, corresponding to chromosomes 21 and 13 respectively. The focus and colour ratios were adjusted for the first captured image from each slide, and then kept at those values for all the following images from that particular slide. A total of 400 images were collected from five slides and stored in TIFF format.

3 Multi-spectral FISH image analysis

3.1 Motivation

Colour systems such as colour television and colour photography usually follow the human visual system and describe and synthesize colour images using the three primary colours– red, green and blue (RGB). Similarly, a tricolour digital image can be considered as a two-dimensional image having three intensity levels (red, green and blue) at each pixel [6]. By analysing each of the three colour channels of the RGB image separately and in various combinations, pre-processing and segmentation of multi-spectral images can be facilitated.

In FISH preparation, multiple probes, labelled by different fluorophores, are frequently combined. In the current study, for instance, chromosomes 13 and 21 are detected as green and red signals respectively, whereas the nuclei are indicated by blue. Although the position in the image and the characteristics of the fluorophores are of importance, much previous analysis [2, 3, 7] usually converts colour information into a gray-intensity scale, and FISH image analysis is then based on brightness contrast. However, difficulties encountered during the analysis of intensity-based FISH images can be avoided if colour information is maintained and used [8]. Nuclei can be analysed using the blue channel of the RGB image, whereas the signals can be analysed using the red and green channels. In our case, for example,

red and green signals are analysed separately using the red and green channels of the RGB image, respectively. Multi-spectral FISH image analysis is beneficial not only to facilitate pre-processing and segmentation, but also to yield colour-based features that contribute to an efficient signal classification. Finally, the benefit of using multi-spectral analysis is expected to increase with the number of fluorophores involved in the analysis.

3.2 Colour specification

Using the RGB colour format, which is the most basic quantitative description of a colour image, we represent colour by the scaled (usually between zero to one) red, green and blue intensities of each image pixel. In the HSI (hue, saturation, intensity) format, which is more suitable for approximating human colour judgements, the colour of a pixel is represented by its hue and saturation, whereas the intensity indicates the pixel overall brightness, regardless of its colour. If we wish to compare the two formats geometrically, we would describe, using the RGB format, a pixel as a point within a cube, whose three edges coincide with the red, green and blue axes of the RGB space. In three of the cube corners– (1,0,0), (0,1,0) and (0,0,1) one primary colour reaches its maximum value, say 1, and the other two colours reach their minimum values, say 0. Two other corners of the cube– (0,0,0) and (1,1,1) represent, respectively points of no brightness at all ('black') and

of full brightness of all the three primary colours ('white'). Geometrically speaking, the HSI space can be considered as a cylinder whose vertical axis represents the intensity from black at the bottom to white at the top. Saturation and hue are expressed within the cylinder bottom, as the radius of a point from the origin of the colour circle and the angle between this radius and the x-axis, respectively [6].

In this work, RGB colour is recorded during the acquisition stage because pre-processing and nuclei and signal segmentations are performed more easily using this colour format than using the conventional conversion of the image to gray-level scale. However, as intensities of red and green signals, each measured in its own channel, are very similar to each other, the RGB format is not suitable for discriminating between signals of different colours. By contrast, signals of different fluorophores represented by the hue parameter of the HSI colour format can be easily resolved due to their different hues. Therefore, we use the HSI format when measuring signal features.

To convert RGB to HSI format, we use a coordinate system in which the RGB cube is rotated so that its main diagonal (1,1,1) lies along the z-axis (the vertical axis of the HSI cylinder) and its R-axis lies in the xz-plane [6]. We then convert to cylindrical coordinates and following some normalization we obtain the equations for the required conversion as [9],

$$H = \arctan 2(3^{1/2}(G - B), (2R - G - B)), \tag{1}$$

$$S = 1 - 3(min(r, g, b)) \tag{2}$$

and

$$I = (R + G + B)/3, \tag{3}$$

where $r = R/(R + G + B)$, $g = G/(R + G + B)$ and $b = B/(R + G + B)$, and R, G and B are the intensities in the three channels, respectively.

3.3 Colour image segmentation

The first step of processing the image is to perform segmentation on each of the three channels of the RGB image. Global thresholds are applied to segment objects, which are then employed as candidates for nuclei and red and green signals. Finding thresholds is straightforward compared with thresholding an intensity image since only red (green, blue) objects are found in the red (green, blue) channel, and intensities of objects (nuclei and signals) and background pixels in each of the channels are well separated. Noise elimination and boundary smoothing of nuclei, and spatial correlation between nuclei and signals complete the segmentation. A signal whose area is larger than 5% of the area of the corresponding nucleus is rejected as 'background fluorescence'. Since our interest in this work is signal classification, we allow the system to accept signals of nuclei of irregular shape or which are part of a cluster. For dot counting, such nuclei, as well as unfocused nuclei, can be

rejected.

4 Spectral feature extraction

Following signal segmentation, the next step is to characterize signals by sets of pixel intensities. A set (signal) can include one or many members (pixels). Since the content and dimension of each set can vary dramatically from signal to signal, raw data (intensities) are not considered discriminating enough to act as features for classification. It is therefore necessary to determine a more discriminating and compact representation of the data. One representation can be derived by measuring a set of features of the data. This representation can be further refined by applying feature selection to the set in order to select a subset of features which maximizes a separability criterion. In addition, another representation can be obtained by feature extraction that maps the intensities into more effective features. A simple, yet very effective, method for doing this, is to transform the data linearly using principal component analysis (PCA) [10, pp. 400–403]. In PCA, we represent samples of a distribution (e.g. intensities) with a small subset of linear combinations of the samples. These linear combinations are formed by the projection of the samples onto principal axes, which maximize the data variance. We use the mean-square error as a criterion to evaluate the effectiveness of the subset. Let $\boldsymbol{X} = f(\boldsymbol{Y})$ be a linear mapping of a random vector $\boldsymbol{Y}$, $\boldsymbol{Y} \in R^d$, $\boldsymbol{X} \in R^m$

and $m < d$. The approximation,

$$\hat{\boldsymbol{Y}} = \sum_{i=1}^{m} \boldsymbol{x}_i \phi_i, \tag{4}$$

with the minimum mean-square error,

$$\varepsilon = E\{(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T (\boldsymbol{Y} - \hat{\boldsymbol{Y}})\}, \tag{5}$$

is obtained when ϕ_i $(\forall i = 1, m)$ satisfy

$$\Sigma_Y \phi_i = \lambda_i \phi_i. \tag{6}$$

The m most effective principal axes ϕ_i are those eigenvectors associated with the m largest eigenvalues λ_i $(\lambda_1 \geq \lambda_2 \geq \ldots \lambda_m \geq \ldots \lambda_d)$ of the covariance matrix of the mixture density Σ_Y. $\boldsymbol{x}_i = \phi_i^T \boldsymbol{Y}$ are the projected values of $\boldsymbol{Y}$ on ϕ_i.

PCA is applied here to signal intensities. Red and green signals have intensity components also in the blue channel of the RGB image, as the signal is part of a nucleus. To find projections which are capable of discriminating between red and green signals and artifacts, we apply the PCA only to the intensities in the red and green channels. We would expect that the principal axes of red and green real signals will coincide with the R and G-axes, since the signals have intensities in only one channel (R or G). Overlap between signals of different fluorophores, or between a signal and background fluorescence due to other fluorophores, is frequent in multi-spectral FISH images. This overlap leads to artifacts in which principal axes are expected to be

between the R and G axes. Using the eigenvectors as features can therefore improve the ability to distinguish between real signals and artifacts of different colours.

Other research [9] has shown that PCA-based features are very effective, yet computationally demanding, for colour image segmentation. In an effort to find a projection with the same characteristics as PCA but which is less costly to compute, the three eigenvectors of typical colour images were examined. Projection of the image along the eigenvector which corresponds to the largest eigenvalue captures most of the variance (information) contained in the image. For typical colour images, this eigenvector has a value of around $(1/3, 1/3, 1/3)^T$, which remains similar across different images [9]. Therefore, projecting a typical colour image onto its first eigenvector is equivalent to computing

$$I_1 = (R + G + B)/3, \tag{7}$$

which is the intensity of the image. The other two eigenvectors are typically $(1/2, 0, -1/2)^T$ (or $(-1/2, 0, 1/2)^T$) and $(-1/4, 1/2, -1/4)^T$ and are also similar across different images. Likewise, the projections of the image onto these eigenvectors are equivalent to the computations of the intensities,

$$I_2 = (R - B) \; or \; (B - R) \tag{8}$$

and

$$I_3 = (2G - R - B)/4. \tag{9}$$

Therefore, the computation of these three intensities provides effective colour features such as those computed by a PCA but with only little computation.

5 Signal measurement

The next step is to measure a set of features for each of the segmented signals. These include area (a size measure) and eccentricity (a shape measure), which have been previously suggested [2]. In addition, we measure a number of spectral features. We compute, at the specific colour plane, three RGB intensity-based measurements: the total and average intensities and the intensity standard deviation. We also compute four HSI hue-based measurements: maximum hue, average hue, hue standard deviation, and Delta hue. Delta Hue is the difference between the maximum and average hue normalized by the average hue. This last feature has been added to the set because it was observed that the difference between values of the average and maximum hue for real signals is usually near zero, whereas for some kinds of artifacts (e.g. overlap of two fluorophores) this difference is substantially large. Two additional features of the set are the two coordinates of the eigenvector corresponding to the largest eigenvalue of the red and green intensity components of the signal. The last feature is the average intensity, I_1 (Eq. 7), which in other colour image interpretation tasks [9], is found to be superior to the other intensities (Eqs. 8, 9).

Table 1: The set of features studied in the work. Numbers are used in the rest of the paper to identify the features. Texture indicates standard deviation of intensity (5) or hue (8). Eig. 1, 2 are abbreviations for the two coordinates of the eigenvector corresponding to the largest eigenvalue of the red and green intensity components of the signal.

Number	Feature	Number	Feature
1	Area	7	Average Hue
2	Eccentricity	8	Hue Texture
3	Total Intensity	9	Delta Hue
4	Average Intensity	10	Eig. 1
5	Texture	11	Eig. 2
6	Maximum Hue	12	Average Intensity (I_1)

Table 1 lists and numbers the twelve features to facilitate their identification in the rest of the paper.

6 Feature selection

Once a sufficient set of features is measured, we can use each one or even all of them to classify signals into 'reals' (valid signals) and 'artifacts'. However, the 'best' single feature may not be sufficiently discriminating for an accurate classification. For example, there is no reason to believe that the area feature

alone can discriminate between red and green real signals. On the other hand, because of the 'curse of dimensionality' [11, pp. 7–9], classification based on whole or most of the set may be complex, costly to compute, and inaccurate. Moreover, some of the features can be found to contribute very little to the classification accuracy.

The purpose of feature selection is to select a (small) subset of the feature set that yields an accurate classification in minimal computational cost. In practical problems and for a not very large feature set, we can search among all the possible feature subsets and evaluate each one of them using a criterion of class separability. The subset that achieves the highest value of the criterion is then selected to represent the patterns to the classifier.

The criterion of separability that is considered here, called J_1, is based on the within-class scatter matrix [10, pp. 446-447],

$$S_w = \sum_{i=1}^{L} P_i E\{(X - M_i)(X - M_i)^T | \omega_i\} = \sum_{i=1}^{L} P_i \Sigma_i, \tag{10}$$

and the between-class scatter matrix,

$$S_b = \sum_{i=1}^{L} P_i (M_i - M_0)(M_i - M_0)^T, \tag{11}$$

where

$$M_0 = E\{X\} = \sum_{i=1}^{L} P_i M_i \tag{12}$$

is the mean pattern of the mixture distribution. $X|\omega_i$ are patterns of class ω_i $(i = 1, L)$ with mean M_i, covariance matrix Σ_i and a priori probability P_i. The criterion

$$J_1 = tr(S_w^{-1} S_b), \tag{13}$$

where $tr(A)$ is the trace of matrix A, is expected to be larger when the between-class scatter matrix is larger and/or the within-class scatter matrix is smaller.

7 Signal classification

In [4] we demonstrated the feasibility of automatic signal classification of randomly-captured FISH images. The signal classification method for the current application is briefly summarised as follows. Signals are classified into four classes– 'real red', 'artifact red', 'real green' and 'artifact green'. Within the 'artifact' classes we expect to find mostly unfocused and overlap signals, and signals which are the result of background fluorescence. These signals will have patterns with different values of features than those of real signals, and hence will be classified as artifacts. Labels for the patterns, as belonging to one of the four classes, are needed to train the classifier, and they are obtained by an expert cytogeneticist using a custom-built graphical user interface for labelling FISH images [12].

Before performing each classification experiment, outliers (around 3% or

less of the data) are automatically removed from the data and the features are then normalized to zero mean and unit variance. Patterns are divided randomly into training and test sets and classification into one of the four classes is implemented using cross-validation [11, pp. 374–375]. The classifier is a two-layer perceptron NN [11, Ch. 4] trained by the scaled conjugate gradient algorithm [11, pp. 282–285]. A validation set which is drawn from the training set assures that the classifier is not over-trained. It also allows the selection of a minimal network configuration based on only a few hidden units. Both factors ensure rapid training and improved generalization.

The three classification strategies in [4] are examined also here. In the first, called the 'monolithic strategy', patterns are classified into the four classes using a single NN. In the second, termed the 'independent strategy', patterns are classified into 'red' and 'green' classes using the 'colour network' and independently by a second network, the 'real network', into reals and artifacts. Classification of a pattern into one of the four classes is achieved by a common decision of both networks. In the third strategy, called 'combined', patterns are first classified into 'red' and 'green' classes using the 'colour network' and then based on the results of this network they are classified by two other networks, the 'real-red network' and the 'real-green network', into reals and artifacts of the two colours.

8 Experiments and results

A few experiments to study feature sets to represent FISH signals are performed. Before the experiments, we created a database of 400 FISH images, which were randomly-captured from five slides. Following nuclei segmentation, the system identified 944 objects within these images as nuclei, of which 613 also contained signals. Following signal segmentation, 3,144 objects within the above nuclei were identified as potential signals and features were measured for them. Based on labels provided by expert inspection (Section 7), 1,145 of the signals were considered as 'reals' (among them 551 were red) and 1,999 as 'artifacts' (among them 1,224 were red).

Features are first analysed visually using probability density functions (pdfs). Figure 1 shows examples of histogram estimates of one-dimensional conditional pdfs for three features– area, average intensity and average hue. In the first three examples, red signals– reals and artifacts are compared, whereas in the last example real signals– red and green are evaluated. From the first example (Fig. 1a), we can see that the value of the area parameter of real signals is much more confined than that of artifacts. However, overlap between the two distributions for most values of area implies that classification of reals and artifacts could not be based on the area feature alone. The two distributions for the average intensity (Fig. 1b) indicate less overlap between distributions of reals and artifacts, where values of artifact intensi-

ties are usually lower. Average hue is found in Fig. 1c and Fig. 1d to be a well-discriminative feature for distinguishing between red and green signals and even between reals and artifacts. Similar graphs to Fig. 1a-c are also derived for green signals. In summary, the large extent of overlap between distributions in these estimates demonstrates some of the difficulty in classifying signals into reals and artifacts of two colours. However, as estimates for other features reveal a higher degree of overlap between the distributions, the pdfs of Fig. 1b-d suggest the superiority of the corresponding features for classification.

To extend the evaluation of single features for signal classification we perform additional experiments using two classification criteria. In the first experiment, feature selection (Section 6) is applied to the set of twelve features. Criterion J_1 (Eq. 13) is computed for each and every feature to give an indication for the amount of class separability the feature provides. In the second experiment, a second classification criterion– the probability of misclassification is employed by the 'monolithic' strategy (Section 7) to classify signals, represented using each of the features. For each feature representation, the optimal configuration of the NN classifier is determined by a validation set, and training continues for 200 epochs. Table 2 shows, for each of the features, the value of J_1, the rank of the feature among all the features according to J_1 ('highest' is 'best'), the configuration of the classifier and the probability of success on the training and test sets. We can draw two main

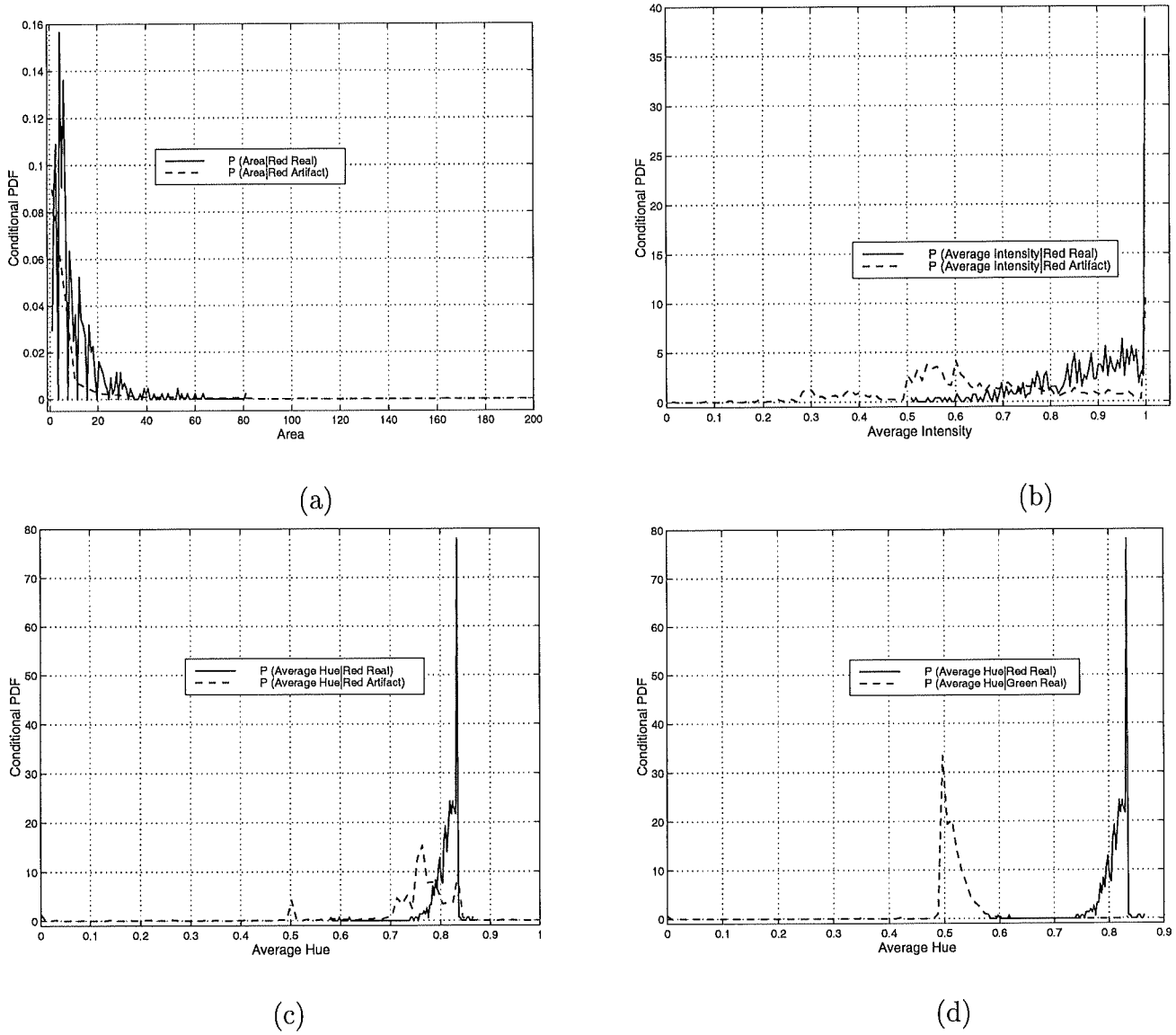

Figure 1: Histogram estimates of the one-dimensional conditional pdfs for red signals– reals vs. artifacts (a,b,c) and for reals– red vs. green (d). Density functions are plotted for the features: (a) area, (b) average intensity and (c) and (d) average hue.

conclusions from Table 2. First, there is a general agreement regarding class separability between criterion J_1 and the probability of success. The three features with the highest values of J_1 (average hue (7), maximum hue (6) and Eig. 2 (11)) are those with the highest probabilities of success. For the rest of the features, there is a weaker agreement, probably since the difference between two probabilities of success, or two values of J_1, is sometimes marginal. The second conclusion is that in order to achieve signal classification which is sufficiently accurate, we would need more than a single feature. Classification based on most of the single-feature representations failed since the representations could not discriminate between signals of two colours. Only representation by the maximum or average hue could yield correct classification of red and green signals, and in some of the cases even of reals and artifacts. Table 2 also strengthen results shown in Fig. 1 about the relative importance of features 1 (area), 4 (average intensity) and 7 (average hue).

Visual analysis of scatter plots for the features may help in evaluating the impact of adding more features to the classification process. Figure 2 shows scatter plots for four pairs of features. To facilitate the analysis, only 200 patterns of each of the four classes are randomly selected and presented. In Fig. 2a, the average hue reveals good colour discrimination, whereas the area is not able to resolve overlap between reals and artifacts. The linear dependency of the average intensity on the hue feature in Fig. 2b, and the 'tendency' of points on the graph toward points of the other colour require

Table 2: Evaluation of single features for signal classification by two criteria. The table includes values of J_1, ranks according to J_1, configurations of the 'monolithic' NN classifiers and the corresponding success rates on the training and test sets in classifying the signals into the four classes. Feature numbers are defined by Table 1.

Feature Number	J_1	Rank	NN Configuration	Training (%)	Test (%)
1	0.0048	10	1:1:4	45.9	45.9
2	0.1058	7	1:2:4	46.5	46.6
3	0.0048	11	1:2:4	48.2	48.2
4	0.3237	5	1:3:4	46.4	46.3
5	0.1801	6	1:2:4	42.8	42.5
6	1.1398	2	1:7:4	65.8	65.4
7	1.1649	1	1:4:4	70.6	70.6
8	0.0013	12	1:9:4	44.9	44.5
9	0.0061	9	1:10:4	47.8	47.5
10	0.3534	4	1:2:4	41.3	40.8
11	0.6690	3	1:6:4	53.8	53.7
12	0.0704	8	1:3:4	49.2	49.0

some explanations. Equation 1 shows that the hue of a signal non-linearly depends upon the intensity according to an arctangent function. However, those intensity values which we find in the RGB channels of FISH images, fall onto the linear section of the arctangent function. Therefore, in our case, hue changes linearly with the average intensity. In addition, since artifacts are mostly the result of unfocused signals, their intensities are weaker than those of real signals. As the intensity of such an artifact signal, say red, decreases, it means that its main component, say R, decreases while the other two components remain the same (G=0 and B=1 in this example). Based on Eq. 1, the outcome is that the signal 'changes' its colour toward the other colour as its intensity decreases. 'Shift' of colour is observed for both red and green signals, and having very weak intensities, signals of the two classes may have almost similar hues (Fig. 2b).

This also explains the dependency in Fig. 2c of the average intensity (I_1) on the maximum hue. In this case, however, lines have different slopes as the average intensity (I_1) depends on the intensity according to Eq. 7. The two major clusters in Fig. 2c are mostly due to real signals. These signals have almost fixed values of I_1, since all their three colour components are fixed (for red signals, e.g., R=1, G=0, B=1). Other artifacts that are caused by overlap of signals of different colours (or signals and fluorescence background of the other colour) create some anomaly in the graph, which was not seen in Fig. 2b. These artifacts have an additional intensity of the

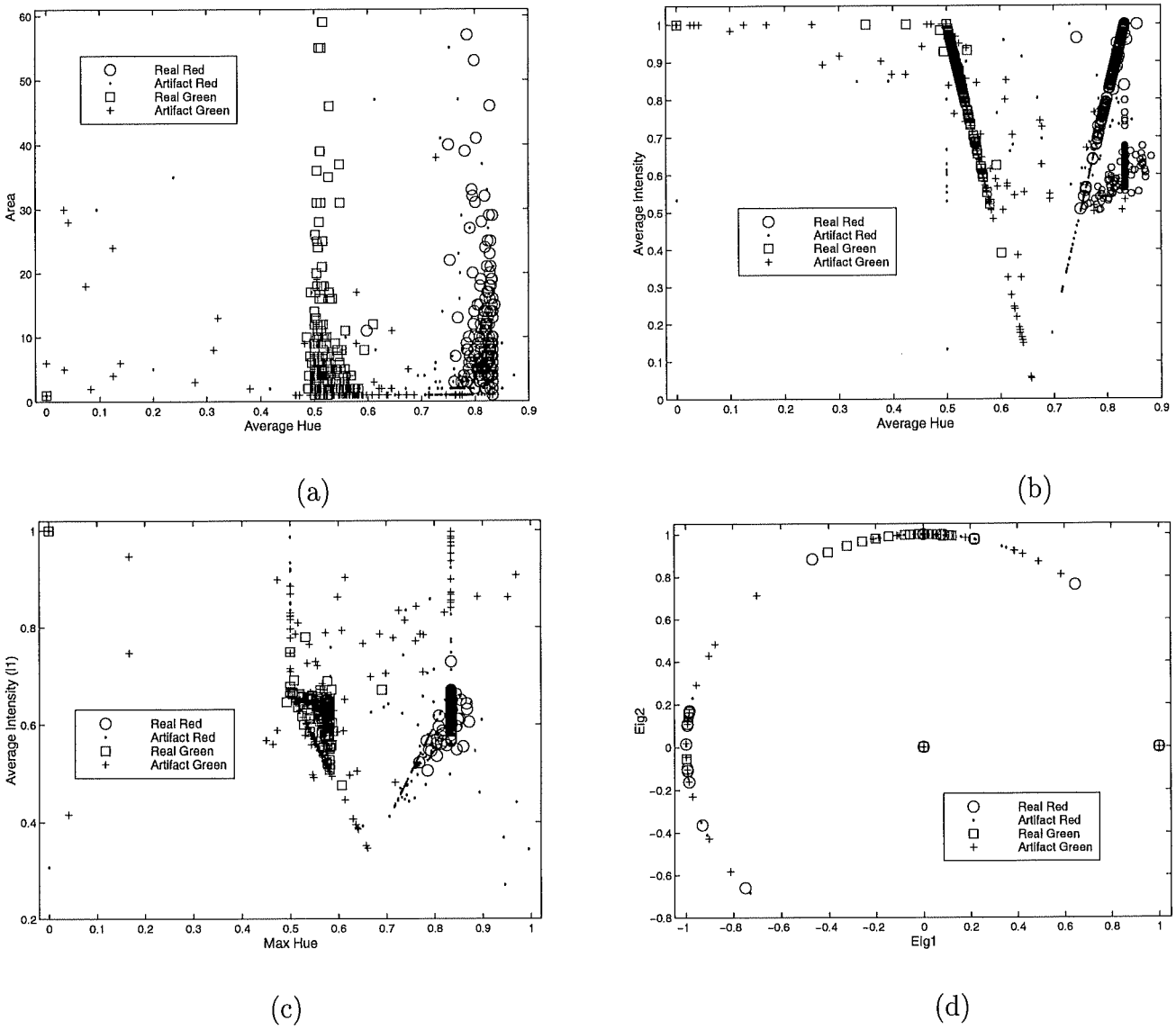

Figure 2: Scatter plots for four pairs of features: (a) area vs. average hue, (b) average intensity (RGB) vs. average hue, (c) average intensity (I_1) vs. maximum hue and (d) the two coordinates of the eigenvector corresponding to the largest eigenvalue (Eig. 1, 2).

occluded signal that increases the average intensity (I_1) according to Eq. 7 (but not the average intensity (RGB) in Fig. 2b). However, as the intensity of the top signal is much stronger than that of the occluded signal, hue is almost unchanged, although the intensity of the signal increases. These artifacts are responsible for the two almost 'x=constant'-lines in Fig. 2c. In addition, within these two lines we can find points that 'swap classes'. This interesting phenomenon can be again explained by Eq. 1. When one intensity component, say G, of a mixed-colour signal, is larger than the second component, say R, and the two are large, there is an agreement between visual analysis and analysis based on Eq. 1 about the signal hue, say green. Therefore, both the system and the expert cytogeneticist will agree on the signal hue. However, when the two components are small (for $G > R$) or close to 0.5 (for $R > G$), the expert will still judge the hue by the top (stronger) signal, but as Eq. 1 predicts, hue will 'shift' toward the colour of the occluded (weak) signal, and the system will decide on that latter colour.

Finally, in Fig. 2d, the two coordinates of the eigenvector corresponding to the largest eigenvalue are plotted against each other. Real signals have only one colour intensity component, either R or G, and therefore, are projected onto either (1,0) (or (-1,0)) or onto (0,1) (or (0,-1)). Colour-mixed artifacts are projected in between. As PCA (Section 4) cannot be applied to single-pixel signals, these signals are projected artificially on (0,0).

Based on the complete visual analysis, we find that average intensity,

maximum hue, average hue, average intensity (I_1), the two coordinates of the eigenvector of the largest eigenvalue and (slightly) the area each provides reasonable discrimination capability between the four classes. Therefore, these features are utilized to combine a few combinations of feature sets to be tested in another experiment by the three classification strategies (Section 7). Input and output dimensions for each of the NN classifiers are set by the feature space dimension and the number of classes, respectively. The number of hidden units is determined such that the network has the highest generalization capability. This is achieved by evaluating networks of different numbers of hidden units on an independent validation set drawn from the training set [4]. The network which has the lowest error measured on the validation set is selected for training. Training of each of the networks, in each of the experiments reported here, is continued for 200 epochs and using three random network initializations. The results are averaged over these initializations using the cross-validation (CV-5) technique. Each of the strategies, classifying signals represented by each of the feature sets, is checked using its own optimal configuration and the results are shown in Table 3. Table 3 reveals that unseen signals, represented by different combinations of features, can be classified as reals or artifacts of two colours with accuracies higher than 80%. In addition, the 'combined' strategy is found to be the best among the three classification strategies, even when inferior feature sets are employed (see e.g., features 1, 4, 12). Finally, feature sets

consisting of the area (1), average intensity (4), average hue (7), the two coordinates of the eigenvector (10 and 11) and the average intensity (I_1) (12) are found in Table 3 to provide the best representations for the signals. We examined, although not shown here, the probabilities of success of the 'colour' and 'real' networks (Section 7), which are responsible for the results of the 'independent' and 'combined' classifiers in Table 3. The examination reveals that the hue-based features (6, 7) are crucial for separating signals of the two colours while the intensity-based features (4, 12) are essential for separating real signals from artifacts.

We also apply feature selection, using criterion J_1, to evaluate feature sets chosen from the features of Table 1. However, before this application, we remove the two coordinates of the eigenvector corresponding to the largest eigenvalues (features 10 and 11) from the data. This is done since coordinates of single-pixel signals are determined artificially to be (0,0), as the PCA cannot be applied to single-element vectors (as explained for Fig. 2d). Therefore, many of the artifacts and some of the real (single-pixel) signals get those values, thereby contributing a bias of the J_1 criterion toward these features. Since only ten features are included in the reduced feature set, we can allow exhaustive search for the 'best' (according to criterion J_1) subset of, say three features. This search is done quickly since it involves the evaluation of only 120 subsets. Values for J_1 and ranks of the ten combinations of three features with the highest values of class separability criterion (J_1)

Table 3: Accuracy of the three strategies in classifying FISH signals, represented by different combinations of features. Features are defined by their numbers according to Table 1. Reported are the number of hidden units in the NN classifier (hid.) and the per cent probabilities of success on the training (Tr.) and test (Tst.) sets for each of the strategies– 'monolithic', 'independent' and 'combined'. The two values under hid. for the 'independent' and 'combined' classifiers are the numbers of hidden units in the 'color' and 'real' networks, respectively, which make up these two strategies.

Feature	'monolithic'			'independent'			'combined'		
Combination	hid.	Tr.	Tst.	hid.	Tr.	Tst.	hid.	Tr.	Tst.
4, 7, 12	7	79.0	78.2	3, 11	78.5	77.7	3, 11	81.9	81.4
4, 5, 6	16	79.0	77.3	12, 7	79.1	77.3	12, 7	82.4	81.3
6, 9, 12	21	80.4	79.3	8, 8	79.9	79.0	8, 8	83.8	83.4
1, 4, 12	9	56.8	54.9	1, 4	55.7	54.6	1, 4	89.5	89.0
1, 4, 7	15	84.3	83.0	1, 14	82.1	81.5	1, 14	88.3	87.5
1, 4, 7, 9	20	85.3	83.4	13, 13	85.8	83.9	13, 13	88.9	88.1
1, 4, 7, 10, 11	19	86.9	84.1	9, 6	86.2	83.8	9, 6	89.4	87.9
1, 4, 7, 12	20	88.4	86.3	12, 7	88.2	86.3	12, 7	89.9	89.2
1, 4, 6, 7, 10–12	15	89.0	87.4	7, 12	89.3	87.0	7, 12	90.5	89.1

are given in Table 4. The table also presents values for the combinations of three features of Table 3. Table 5 shows the percentage of times each of the ten single features appears in the 30 'best' (according to J_1) combinations. Both Tables 4 and 5 demonstrate the superiority, regarding criterion J_1, of the features: average intensity (4), maximum hue (6) and average hue (7). In each of the ten 'best' combinations, the average intensity and either the maximum or average hue are selected.

A comparison of the last three tables shows that the average intensity (4) and the average hue (7) (or the maximum hue (6)) are the most discriminating features, and inclusion of these two in a feature set guarantees successful classification of FISH signals. Some agreement between the probability of success and J_1 is shown by the relatively high rank (14) that the subset of three features with the second best classification performance (1, 4, 7) achieves. The comparison however reveals that this is not necessarily the case with other combinations (compare, for example, the probability of success of the two combinations with the highest J_1 (first two rows in Table 3) with other combinations in Table 3). This may suggest that J_1 is not the best criterion to measure class separability of signals represented by the proposed feature sets.

Table 4: Evaluation, using criterion J_1, of combinations of three features for signal classification. The table includes results for the ten combinations with the highest values of J_1 and for other combinations of three features from Table 3. For each combination, the rank among the 120 subsets according to J_1 is given. To be consistent with previous feature numbers, we keep number twelve for the average intensity (I_1), although only ten features are involved in the selection.

Feature Combination	J_1	rank
4, 7, 12	1.7543	1
4, 5, 6	1.6789	2
4, 5, 7	1.6580	3
4, 6, 7	1.6046	4
4, 7, 8	1.6022	5
4, 6, 12	1.5781	6
2, 4, 7	1.5699	7
4, 7, 9	1.5692	8
2, 4, 6	1.5511	9
4, 6, 8	1.540	10
6, 9, 12	1.2218	56
1, 4, 12	0.4342	74
1, 4, 7	1.4958	14

Table 5: The percentage of times in which each of the ten features of Table 1 (excluding features 10 and 11) appears in the 'best' (according to J_1) 30 combinations.

Feature Number	1	2	3	4	5	6	7	8	9	12
(%)	2.2	6.7	3.3	16.7	12.2	17.8	20.0	6.7	5.5	8.9

9 Discussion

This paper has explored suitable feature representations for FISH signals making use of an existing signal classification methodology. For this purpose, a family of features, consisting of measurements of size, shape, intensity, texture and colour, has been examined. In addition, the application of feature extraction to signal intensities has provided features which are capable of improving the accuracy of the classification, mainly due to the identification of artifacts resulting from signal overlap of two types of fluorophores. Moreover, the intensity derived using the HSI format, which can also represent the projection of the image on its principal axes, has been measured also.

A set of twelve measured features has been evaluated by different criteria. Histogram estimates of probability density functions and scatter plots provide preliminary visual insight into dependencies between features, and their relative importance for the classification. Moreover, these tools permit the identification of sources of errors when specific features are used.

Feature selection enables the choice of feature sets of any type and number, which maximizes class separability criterion J_1. The ultimate criterion for evaluating features for classification, however, is the probability of misclassification. Mismatches in selecting optimal feature sets, according to the two different classification criteria, can be partially attributed to the additional feature extraction stage performed by the hidden layer of the NN classifier. In addition, introducing feature selection and feature extraction into the classification process guarantees the selection of an optimal feature set in terms of both discriminative power and dimensionality. This set enables the classifier to utilize the maximum information contained in the data to accomplish high classification accuracy.

Both the qualitative and quantitative analyses have demonstrated the superiority of hue and intensity-based features. When features of the two families are combined together, even a single hue-based feature can separate completely signals of two fluorophores, leaving the task of discriminating real signals from artifacts to an intensity-based feature. Consequently, feature sets consisting of both hue and intensity-based features enable an NN-based hierarchical strategy to classify nearly 90% of the signals as reals or artifacts of two fluorophores.

Acknowledgements: This work is supported by EPSRC contract GR/L51072: *Automatic Analysis of FISH Images.*

References

[1] N. P. Carter. Fluorescence in situ hybridization– state of the art. *Bioimaging*, 4:41–51, 1996.

[2] H. Netten, L. J. van Vliet, H. Vrolijk, W. C. R. Sloos, H. J. Tanke, and I. T. Young. Fluorescent dot counting in interphase cell nuclei. *Bioimaging*, 4:93–106, 1996.

[3] H. Netten, I. T. Young, L. J. van Vliet, H. J. Tanke, H. Vroljik, and W. C. R. Sloos. FISH and chips: Automation of fluorescent dot counting in interphase cell nuclei. *Cytometry*, 28:1–10, 1997.

[4] B. Lerner, W. F. Clocksin, S. Dhanjal, M. A. Hultén, and C. M. Bishop. Automatic signal classification in fluorescence in-situ hybridization images. Technical Report 466, Computer Laboratory, University of Cambridge, May 1999.

[5] K. Klinger, G. Landes, D. Shook, R Harvey, L. Lopez, P. Locke, T. Lerner, R Osathanondh, B Leverone, T. Houseal, K Pavelka, and W. Dackowski. Rapid detection of chromosome aneuploidies in uncultured amniocytes by using fluorescence in situ hybridisation (FISH). *Am. J. Hum. Genet*, 51:55–65, 1992.

[6] K. R. Castleman. *Digital Image Processing.* Prentice-Hall, New Jersey, 1996.

[7] P. M. Nederlof, S. van der Flier, N. P. Verwoerd, J. Vrolijk, A. K. Raap, and H. J. Tanke. Quantification of fluorescence in situ hybridization signals by image cytometry. *Cytometry*, 13:846–852, 1992.

[8] K. R. Castleman and B. S. White. Dot count proportion estimation in FISH specimens. *Bioimaging*, 3:88–93, 1995.

[9] Y. Ohta. *Knowledge-based Interpretation of Outdoor Natural Color Scenes.* Pitman Publishing Limited, London, 1985.

[10] K. Fukunaga. *Introduction to Statistical Pattern Recognition (2nd ed.).* Academic Press, San Diego, 1990.

[11] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[12] B. Lerner, S. Dhanjal, and M. A. Hultén. GELFISH- graphical environment for labelling FISH images. Technical Report 465, Computer Laboratory, University of Cambridge, May 1999.