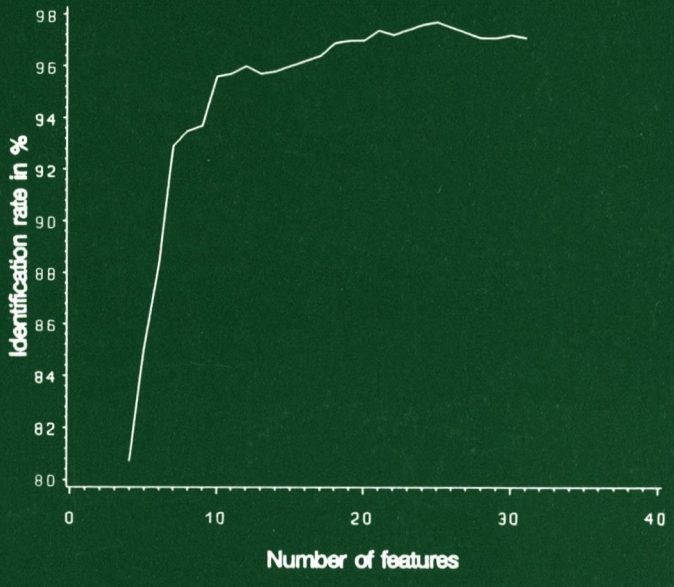
Landbrugsministeriet

Statens Planteavlsforsøg

Weed Seed Identification by shape and Texture Analysis of Microscope Images

Ph.D. dissertation



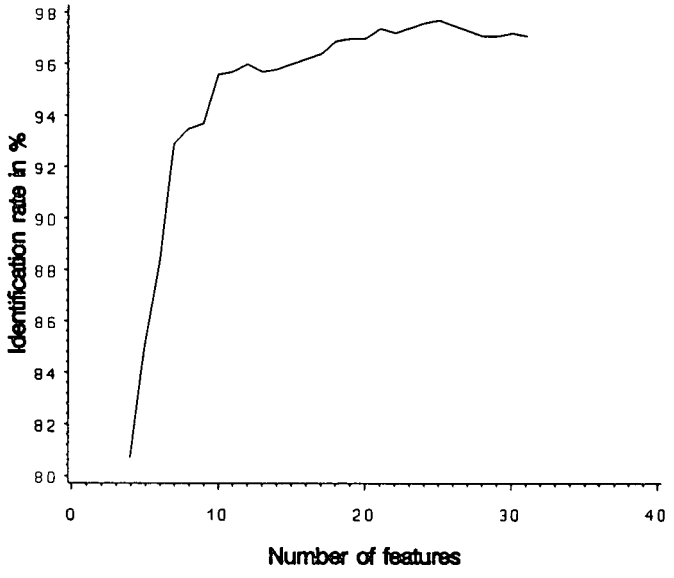
Poul Erik Herløv Petersen
The Danish Institute of Plant and Soil Science
Department of Biometry and Informatics
DK-2800 Lyngby

Tidsskrift for Planteavls Specialserie

Beretning nr. S 2198 - 1992

Weed Seed Identification by Shape and Texture Analysis of Microscope Images

Ph.D. dissertation



Poul Erik Herløv Petersen

The Royal Veterinary and Agricultural University
Department of Mathematics and Physics
Copenhagen

The Danish Institute of Plant and Soil Science
Dept. of Biometry and Informatics

1992

Preface

This thesis is along with the paper of Petersen (1991) and the manuscript of Petersen and Krutz (1991) a partial fulfillment of my Ph.D. study at the Royal Veterinary and Agricultural University, Denmark. The work in image analysis was mainly carried out at the Department of Biometry and Informatics at the Danish Institute of Plant and Soil Science, and the project was financially supported by the Danish Research Academy. The Danish Botanical Garden delivered the weed seeds used in this project.

The main advisor in image analysis and statistics was Professor Mats Rudemo, Department of Mathematics and Physics, and the local advisor at the Department of Biometry and Informatics was Lic. Tech. Klaus Juel Olsen. Lectures and research in Plant Anatomy took place at the university under the directions of Associate Professor Ole Lyshede. I am very indepted to all three advisors for careful and creative participation and support during this project. Special thanks goes to Professor Gary W. Krutz for his efforts and care during my four month stay at Purdue University, USA.

It is my hope with this thesis to reach other plant scientists with interest in image analysis for description and recognition of biological objects. The background for the present investigation was mainly the wish to build up theoretical and practical knowledge in image analysis for use in future agricultural applications. The weed seed identification seemed ideal for that purpose, because the seeds show significant differences in many aspects (*i.e.*, size, shape and texture), and because there is a need for automatic identification. For building up general knowledge the project became mainly methodological and not specifically directed towards a single practical implementation. Therefore, scientists from other fields, I hope, might also find useful information in this thesis.

The reader may not need to read the full thesis. Some sections may be found of minor relevance for particular groups of readers, or the readers may possess inadequate background knowledge in certain subjects. For choosing chapters of interest the organization of the text will be briefly described.

The first chapter deals with previous research in image analysis applied to seeds, mainly grains, and the second presents a botanical and anatomical description of seeds, to provide background information for the following investigation. The third chapter deals with video microscopy in relation to certain problems faced in the seed project. These three chapters have an introductory character compared with chapter four, which contains descriptions of the selected seeds, measurements of detected error sources in the equipment, and descriptions of the image analyses related to images of weed seeds. The final evaluation of the analyses is presented in chapter five with classification results from the individual analysis as well as different types of combinations. In chapter six the experiences and results are discussed together with remaining difficulties for future constructions of automatic weed seed identification systems.

Summary

The main objective of this investigation is to evaluate image analysis as a tool for automatic identification of weed seeds. For this reason a number of different analyses were applied to the digitized images of weed seeds.

Forty plant species of commonly occurring weeds of which some were closely related with similar looking seeds were selected. Samples of 25 seeds per species were taken, and images acquired by a black and white CCD camera mounted on a stereo microscope, one seed per image. The images were segmented by thresholding followed by a minor smoothing of the seed contour before the analyses were applied.

A relatively large variation in the recognizing power of the separate analyses were found. Thus, the analyses of the seed shape showed between 26.2% and 77.0% correct identification, and texture analyses showed correspondingly between 31.7% and 61.3% correct identification in average of all species. These figures refer to groups of closely related features (between 3 and 11 features per analysis).

Identification of the plant family in stead of species increased these rates 15.3% on average (all analyses). Certain species showed relatively high confusions with species within the same family, but others were confused with species of foreign families. Estimated processing times of the separate analyses showed large variation too.

Combining the analyses of seed size, shape and texture greatly improved the identification rate. The highest identification result obtained in this study was 97.7% on average for all species. This was based on 25 features selected by a stepwise procedure.

The image analysis seems to be highly reliable, thus encouraging the efforts for developing an automatic identification system. Certain practical difficulties of constructing such a system are discussed in the text.

Resume

Summary in Danish

Hovedformålet med denne undersøgelse er at belyse, hvorvidt billedanalyse er egnet til automatisk identifikation af ukrudtsfrø. En række forskellige metoder til beskrivelse af digitale billeder bliver anvendt og vurderet i denne sammenhæng.

Der blev udtaget 25 frø fra hver af 40 ukrudtsarter, som repræsenterede et bredt udsnit af praktisk forekommende arter og samtidig indeholdt en række nært beslægtede arter. Billeder blev optaget med et CCD-kamera, der var monteret på et stereomikroskop. Det digitale billede viste et frø pr billede. På billedet udskiltes frø fra baggrund ved en simpel segmenteringsproces efterfulgt af en skånsom udjævningsprocedure. Herefter kunne billedanalyserne udføres.

Der blev fundet stor variation i styrken af de enkelte analyser. Således viser analyser af karakteristika hos frøkonturen mellem 26.2% og 77.0% korrekt identifikation i gennemsnit af alle frø, og for analyser af frøteksturen blev opnået gennemsnitlige identifikationsgrader på mellem 31.7% og 61.3%. Antallet af karakteristika i hver analyse varierede fra 3 til 11.

Ved identifikation af frøets familie i stedet for art opnåedes en gennemsnitlig stigning på 15.3% (alle analyser). Nogle arter viste stor sandsynlighed for forveksling med arter fra samme familie, mens andre blev hyppigt forvekslet med arter fra fremmede familier. De proces-tider, som blev målt for udførelsen af de enkelte analyser, udviste ligeledes stor variation.

Ved kombination af analyser som beskriver både størrelse, kontur og tekstur kunne opnås en betydelig forbedret genkendelsesprocent. Den højste identifikationsprocent, som blev opnået i denne undersøgelse, var 97.7% ved anvendelse af 25 karakteristika udvalgt ved en trinvis udvælgelsesprocedure.

Selve billedanalysen må således siges at være meget pålidelig, hvilket må opmuntre til at fortsætte opbygningen af et automatisk identifikationssystem. En række praktiske problemer i forbindelse med konstruktionen af et sådan system bliver diskuteret i teksten.

Contents

1. Introduction

Image analysis is a computer technique used for calculating features of objects in a digitized image, such as area, length, perimeter etc. Thus, visual information is processed and analyzed to extract certain features which can be used for taking decisions. These decisions could be related to some quality criteria, such as grading classes of a product or simply accept/reject decisions. With these properties image analysis could be a valuable tool for agriculture, horticulture and the food processing industry, as food is often visually inspected on its way through the chain: Production - processing - packaging - distribution. Today agricultural production has risen to a level where yield increases might increase the environmental damage, and therefore the importance of research in food quality is growing. Consequently, research in crop or food quality is a major challenge for agricultural applications of image analysis.

Some applications of image analysis to plant science have been reviewed by Price and Osborne (1990). Previous work on quality evaluation of fruits has involved determination of size, shape, colour and blemish area. These are all quality features, which can be visually recognized and used for automatic sorting and grading of the product. Other research fields mentioned include quantitative plant disease assessment, seed description and classification, measurements of plant cover, plant growth, root length and soil porosity, eliminating imperfect plantlets in plant tissue culture, robotic fruit harvesting and many cytological applications.

Often it must be realized that for practical use, the vision system has to be combined with some machinery. For example, in robotic fruit harvesting the vision system has to control the robot so the fruit can be picked and delivered. Also, in vision based fruit sorting systems new sorting machinery has to be developed, and in development of an automatic system for weed seed identification new machinery which presents the seed to the vision system will play an important part. However, in the present study only development of the imaging system is considered.

Identification of weed seeds may be regarded as an important factor for determination of crop seed quality. In purity determinations of the crop seeds, samples of seeds are divided into fractions of pure seed, other crop seeds, weed seeds and inert matter. Standards for contents of non-pure seeds have been set up in different national and international organizations. Purity analysis is important for seed trading. According to the Danish rules the seed selling firm has to give guarantees on the quality, at minimum the EEC standards (Klitgård, 1989). If the guaranteed standard is not met, the firm has to pay compensation. Certification issuing and control of quality are performed by the national seed testing stations. In the year 1988/89 approximately 27000 examinations of purity were made at the Danish Seed Testing Station. Furthermore, the percentage by weight was specified in 6113 samples tested for *Agropyron repens*, 2255 for *Rumex spp.* and 8185 for single species of other weed seeds. The numbers of the weed seeds and other crop seeds were examined in 13018 number count tests, where test samples are of size 25000 seeds weight (max. 1 kg) (Jensen, 1989).

This identification work is very labour intensive, and at the Danish Seed Testing Station it is based on seasonal employment of trained personnel which may cause difficulties in the future. Therefore, automation of the identification work is an important issue for seed testing.

Another area for large scale identification of weed seeds is agricultural weed research, where the seed banks in the soil are investigated. The seed bank reflects in some way the composition of the weed flora on the land from many years,

and, therefore, registration of changes in the seed bank is useful for evaluating the effect of different cultivation and weed control programs. Thus, the use of seed bank data to forecast the weed problems is an important perspective for future weed control. In the light of the present development of new herbicides with a more selective effect, and the wish to reduce herbicide application, a prognosis of where the weed problem will occur, how severe it will be, and which kind of species, will be valuable as an advisory tool.

A survey of Roberts (1981) summarizes the difficulties connected with predictive use of seed bank data for weed control. Firstly, the number of soil samples should be relatively high because of uneven distribution of the seeds, and also the time of the year when the samples are taken is important due to major new inputs of seed. Secondly, the number of seeds giving rise to seedlings is dependent on soil type, soil disturbance, soil moisture, temperature and seed dormancy. These requirements for germination vary among species. The age of the seed will influence the percentage of germination differently from species to species, and even during optimal conditions the percentage of germinating seeds will be dependent on the species. Therefore, it was assumed that the expected emergence, or at least the relative distribution among species, could be estimated from seed bank data combined with a knowledge of the germination pattern in the region and knowledge of the percentage germination of each species related to the cultivation method used. However, the use of seed bank data for determination of future vegetation is a perspective that will increase the need for seed identification at a low cost.

Previous work

Most studies on seeds by image analysis have been concentrated on varieties and species of cereal grains, but other cultivated species and weed seeds have been included to a lesser extent. The purpose of the investigations is varied. Some works take a taxonomic approach, where image analysis is studied as a descriptive tool for the seeds, but most try to assess the classification capability (discriminating power) of image analysis. Although, there seems to be a clear connection between the two approaches, certain differences also appear. For taxonomy the classification results of the single seed are not important, the imaging system is not going to recognize the seed. Instead, the taxonomic approach concentrates on descriptive morphological values of sample means in order to separate the seeds as a group (variety or species). The discriminatory approach classifies single seeds in the samples.

The following review describes the research of image analysis on seeds. It starts with describing the work of the taxonomic approach, and proceeds with the work on discrimination of crop species, separation of wheat from non-wheat, classification of grains into grading classes, and discriminating wheat varieties. Practical objectives, features used, seeds used and major results are mentioned in order to evaluate the work.

Measurements of area, perimeter, length, width, shape factor ($4 \cdot \pi \cdot \text{area}/\text{perimeter}^2$) and aspect ratio (width/length) were studied by Draper and Travis (1984) on seeds of barley, wheat, lettuce and five species of weeds. Values of means and standard errors are presented to compare differences among species. It is suggested that the shape factor might be used in taxonomy. A larger study including seven crop species and 42 species of weed seeds, which are likely to occur in samples of the selected crop species, was carried out by Travis and Draper (1985). Separation of sample means using 95% confidence regions of the means was demonstrated in a bivariate plot of shape factor and seed length. It showed that sample means of most species could be distinguished from each other. In Keefe and Draper (1986) new morphological measurements of the wheat kernel were developed. The wheat kernel was placed crease side down, and in that position different axis lengths and angles were determined on the image. The following features were measured: 1) area, 2) seed length, 3) seed height, 4) per-

imeter, 5) germ height, 6) germ length, 7) the tangent of the germ angle, 8) high point, 9) brush height, and 10) the tangent of the dorsal angle. From these other derived measurements were calculated: 1) shape factor, 2) aspect ratio, 3) relative germ height, 4) relative germ length, 5) relative brush height and 6) horizontal axis. Seeds of five wheat varieties were compared pairwise, with the purpose of identifying bulk lots. For randomly selected seeds it was possible to separate all pairs except one. Later, Keefe and Draper (1988) presented a mobile camera gantry designed for automatic presentation of the seeds to the camera.

Keefe (1990) investigated the univariate distributions of some measurements on wheat seeds from the data in Keefe and Draper (1986). The majority of the distributions appeared to be multimodal due to different position of origin in the inflorescence. Consequently, it was claimed that the normality assumption in multivariate methods of wheat seed identification (discriminant analysis) is invalid.

A very early attempt to classify single seeds by image analysis was reported by Segerlind and Weinberg (1972). They investigated 7 grain species (corn, oats, wheat, barley, rye, soybean and pea) and in addition two varieties of wheat, barley and oat. For kernel identification amplitudes of the first ten Fourier coefficients were used. Discrimination among the seven species was 94.3 % correct, and between the two varieties 77.8 % correct on average of the three species.

Brogan and Edison (1974) used a learning algorithm to increase identification of corn, wheat, soybean, oats, barley and rye. The purpose was to determine admixtures of foreign cereals in a shipment. Extracted features were length, width, depth and area. In general the classification rates were 80-85 % using normal discriminant analysis, but improving the analysis with the learning algorithm the classification rate increased to about 98%. This learning algorithm operated by adjusting the *a priori* probabilities from the knowledge of already identified kernels in a sample.

Sapirstein *et al.* (1987) used normalized moments, lengths, area, compactness (reciprocal of shape factor), principal axis length, width and aspect ratio to characterize seeds of wheat, barley, oat and rye. The purpose was to detect admixtures of other cereals in wheat as an important grading factor in commercial grain inspection. Four descriptors were selected for discriminant analysis showing classification rates between 96.5 and 100 %. There was no confusion between wheat and the other cereals.

Chen *et al.* (1989) tried to improve the identification of 8 crop species and 3 weed species by including features of the height profile from laser scanning. These features were the average first derivative of height profile and height. Other features from the contour image were maximum length, maximum width, and the mean and variance of brush end roundness. The identification was 100 % correct for the 3 weed and of the 4 crop species and between 68 and 94% for the remaining 4 species. Misclassifications were mainly found among varieties of wheat and barley.

For evaluation of wheat in marketing channels, the amount of non-wheat components is measured. For this purpose Zayas *et al.* (1989) present results from wheat (mixtures of 6 varieties), 6 weed species, stones and glass. Wheat from non-wheat was separated using a wheat pattern structure, where the unknown object was matched to certain measurements of a wheat prototype. All wheat kernels were correctly classified, and only 3 out of 386 non-wheat were classified as wheat.

Symon and Fulcher (1988a) investigated the variation within varieties, between varieties and the environmental influence on six oats varieties. The features were kernel length, area and aspect ratio. Variation arising from environment was significant for both size and shape, and, therefore, they concluded that this might exclude a rapid identification of varieties by image analysis. However, the size measurements could replace weight samples in the oat processing industry.

With the purpose of replacing microscopical determinations of the ploidy levels (chromosome numbers) in ryegrass with image analysis

of the seeds Berlage *et al.* (1988) investigated 6 tetraploid and 2 diploid varieties. Correct classification into two ploidy levels was 85% using measurements of perimeter, length, average grey level and two width measurements.

Zayas *et al.* (1986) classified 10 wheat varieties into three grading classes (soft red winter, hard red winter and soft red spring). Features used were length, length ratio, width, length of parabolic segment, and sine and tangent of angle. The classification was performed pairwise between varieties from different grading classes. Classification rate was 83% for mixtures of three varieties of either soft red winter or hard red winter.

Classification of 9 wheat varieties into three grading classes (soft white winter, hard red winter and hard red spring) was studied by Symons and Fulcher (1988b and 1988c). Seven kernel morphological features were used, and correctly classified kernels were above 80% for hard red spring and hard red winter, and soft white winter varieties were totally segregated.

Ten wheat varieties representing six grading classes were investigated by Neuman *et al.* (1989a and 1989b) using colour information (RGB values). Pairwise discrimination between pairs of different grading classes was 88% correct on average. Discrimination among all ten varieties into the six grading classes was 67% correct.

Neuman *et al.* (1987) used kernel thinness ratio, contour length, one normalized central moment and one Fourier descriptor of the contour for classification of wheat kernels. Fourteen varieties from five grading classes were investigated. Classification into grading classes was 100% correct for 9 varieties and between 42 and 75% correct for the other five varieties. Identification of varieties was 56% correct.

The separation of two wheat varieties was studied by Zayas *et al.* (1985). In classification of 240 kernels 235 were correct using nine morphological features. Symon and Fulcher (1988b) investigated wheat variety identification with features obtained from 1) whole kernel image and 2) transverse sections of kernels. From the transverse sections 7 measurements were obtained, and in addition three measurements of the bran. Six varieties were on average identified 88.7% correctly using 15 features.

Five varieties of wheat were studied by Myers and Edsall (1989). Kernels were analyzed in two positions: 1) placed with the crease down and 2) the side view of the kernels. Morphological features and Fourier magnitudes were used for identification. On the basis of 22 features the identification was 90.8% correct.

A laser scanning system was constructed by Thomson and Pomeranz (1991) for obtaining an elevation profile image and an intensity profile image of wheat kernels. Fourteen features were calculated from this 3-D information and used for identification. Two varieties in two test sets showed an identification rate of 92% and 94%, respectively.

Many good results have been achieved, but satisfactory identification of varieties seems to require many features from different sides of the kernels (*i.e.*, transverse section, side view and height profile). Furthermore, the works are characterized by considering many different approaches to achieve higher recognitions. With regard to the classification part, *i.e.*, the discriminant analysis, the learning algorithm approach by Brogan and Edison (1974) might turn out to be of considerable value in future work. Similarly, it could be suggested to introduce a reject class, so that the atypical seeds are rejected, and higher classification rates of the remaining seeds are obtained.

Objectives

The purpose of the present investigation is to evaluate the capability of digital image analysis techniques for identification of weed seed species. The evaluation will be seen in the light of future development of an automatic implementation. Specific objectives are:

- Description of the species by features from different analyses.
- Adjustment and improvement of image analysis for the weed seed application.
- Evaluation of shape and texture descriptors

using discriminant analysis.
- Improvement of identification with respect to flexibility and robustness.

2. Equipment

The first important condition of a successful analysis is a high quality of the images. For image acquisition of weed seeds a microscope was connected to a black and white CCD camera (Charge Coupled Device). A short introduction to video microscopy will provide an impression of which factors affect the quality of the image, and how the image may be improved. After description of the imaging device used in present investigation, problems arising from the image acquisition procedure are discussed.

The video camera

The two types of video camera, the vidicon and the CCD camera, will be outlined. The general reference for this description is Inoue (1986).

The vidicon family of camera tubes is incased in a glass envelope about 2 cm in diameter and 10-15 cm in length. At the front end, the optical image is focused on the 'target' which consists of three layers. An electron beam scans the target in a raster of single pixels. The electron beam, generated by the electron gun, sweeps the target, and charges up its back surface, a photoconductive layer, with electrons. In the dark the photoconductive layer is an insulator, but when light strikes the layer the resistance decreases nearly proportionally to intensity of illumination. This phenomenon gives rise to the video signal. The primary difference among tubes is the composition of the target material giving variations in sensitivity, noise and resolving power.

In the CCD camera the raster of picture elements is built into a semiconductive chip. Each light sensing element is a silicon photodiode, electrically isolated from its neighbours in a grid structure. The sensor element consists of a transparent electrode positioned above the silicon semiconducting material. In this material a charge packet is produced by illumination during the time of exposure. The readout of the charge packet is started by turning off the electrode above the packet and turning on an adjoining electrode. The packet of electrons immediately move to the adjoining electrode which is located in a light protected area (the 'isolation transfer gate'). By the same procedure the packets of electrons representing each pixel are moved out of the isolation transfer gate and enter an amplifier which measures the amount of photoelectrons produced by successive photodiodes and produces the output video signal.

For both camera types common desirable features are high resolution to capture fine details, a good contrast and a high sensitivity to illumination. However, a more basic property is the optimal utilization of the usable light range of the camera, *i.e.*, the best camera response to the brightness of the scene. One of the important features to control this is the so-called gamma of the video imaging device. The gamma is defined as the exponent of the function that relates the signal output and the degree of illumination. Some cameras provide built-in gamma compensation circuits to obtain an effective gamma of 1.0. But also other circuits may be incorporated to adjust the usable light range of the camera. An automatic gain control (AGC) regulates amplification of the video camera in relation to brightness of the scene, and an 'auto black' circuit adjusts the signal black to the darkest value in the picture (regardless of its absolute value). This means that the curve describing the relation between input illumination and output signal is modified to be a straight line by the gamma compensation circuit, the slope of the line is regulated by the AGC, and the level of the curve is controlled by the auto black function.

One or more unwanted properties may occur in video cameras. Blooming is a phenomenon where a high pixel signal amplifies the signal of neighbour pixels. In this way highly illuminated regions in the image are spread to larger regions. Another error is called burn, which

particularly occurs in vidicons, where exposure to excessively bright light causes a burned in image in the camera tube. If the output signal varies when the sensor is uniformly illuminated, it is called shading, and geometrical errors appear when circles get elliptic or egg shaped. Finally, permanent discrete defects are called blemishes. This might be light spots, missing pixels, reduced sensitivity in certain regions etc..

The microscope

The other important component of the equipment is the light microscope which is connected to the video camera. The optical image plane from the lens system is positioned on the target or the CCD chip in the cameras. The quality of the microscope image is mainly determined by the following properties: Resolution, depth of focus, working distance, degree of magnification and correction of lens aberrations which is explained in more detail in many microscope textbooks, *e.g.*, Pluta (1988).

The resolving power of the lens is determined by the relation between lens diameter and focal length. For a point in focus on the optical axis the angle (α) between the optical axis and the most divergent light ray, and the refractive index (n) define the numerical aperture (NA) as a measure of the resolving power:

$$NA = n \cdot \sin(\alpha)$$

For air n=1.0. NA is related to the other properties mentioned, except lens aberrations. For example, increasing the working distance for a given lens diameter the angle α will decrease. Similarly, a lower magnification is caused by a smaller lens curvature and consequently a larger focal length which will result in a decrease in NA.

A diaphragm in front of the lens may regulate the angular aperture. The reason for this way of choosing a smaller resolving power is that NA and the depth of focus are inversely related. A smaller aperture will increase the depth of focus, but decrease the resolution.

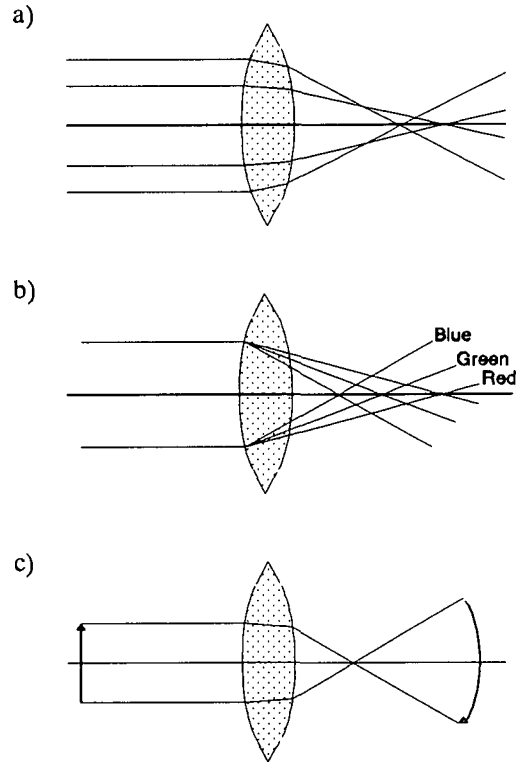


Figure 2.1 Three types of lens aberrations: a) Spherical aberration, b) chromatic aberration and c) field curvature.

The upper limit of numerical aperture for dry objectives is 0.95 corresponding to the angle α = 70°. An ordinary 40X dry objective will usually have a numerical aperture of 0.65-0.95 depending on the working distance.

The last important quality factor is the degree of correction for lens aberrations. Microscope objectives are divided into different types corresponding to their correction. The achromats are objectives with the simplest degree of correction. Three kinds of geometric aberrations are in achromats corrected for one colour. This is spherical aberration, which occurs when the center rays and the periphery rays focus at different points on the optical axis (Figure 2.1a), and coma and astigmatism, which concerns off axis points and oblique incoming rays, respectively. The chromatic aberration which is

caused by variation of the refractive index with wavelength (Figure 2.1b) is, in achromats, corrected for two colours. The best correction is in the apochromats with chromatic correction for three colours and the three geometric corrections for two colours.

The remaining correction is for the field curvature which cause the image points in focus on the optical axis to be farther away than the off-axial image points (Figure 2.1c). This makes a curved image where either the center is sharp and the periphery is blurred or vice versa. Microscope objectives with the prefix 'plan' are corrected for this aberration.

These corrections are adjusted to a normal optical path in the microscope. To preserve this when connecting microscope and video camera a video adapter is needed. The video adapter contains a simple convex lens which in some way replaces the lens in the eye of an observer.

A practical example

Evaluations of sources of errors in imaging devices are very rare. However, a practical example of the actual occurrence of errors has been described by Tappan *et al.* (1987) using two CCD cameras and one vidicon camera. Nine different errors were observed. 1) Reflection of light from a wet surface increased object size. 2) The outer rows and columns of pixels in the image edge contained pixels with no relation to the scene. This was more noticeable with the vidicon than with the CCD cameras. 3) The CCD camera lens showed shading effect at large apertures. 4) Camera warm-up for the vidicon caused a change in grey levels towards the brighter level during the first hour. 5) The area of a dark object increased in the presence of another similar object. This was explained by the auto black function. 6) The pixel aspect ratio was slightly different from 1.0. 7) For both camera types the grey level changed by change in position of the object. 8) The grey scale range was reduced due to incorrect connection of the device (parallel connection). 9) The size of an object was affected by different apertures.

Equipment used

A stereomicroscope of the type Olympus SZ-Tr was used in this investigation. This is a bin-objective binocular microscope with a photo-tubus. It contains a zoom control ring with a magnification range of 0.7 to 4.0 for the objective, and a numerical aperture range of 0.04 to 0.08. The optical path for each objective has a 6 degree angle difference to the vertical. The eyepiece (ocular) was of type FK3.3x and the video adapter is MTV-3. There is only the simplest type of correction for lens aberrations, and no diaphragms on the entire system.

The video camera was a Philips LDH 0660/10, a black and white CCD camera with AGC, but no auto black and gamma regulation.

As illuminator a Schott KL 1500 was used with two optical fibre bundles to illuminate the seeds from two directions.

The computer was a Compaq Deskpro 386/20 with a PCVISION plus Frame Grabber.

The image of size 512x512 pixels and 256 grey levels was shown on a separate monitor.

Image acquisition

The images of the weed seeds were used for analysis of both shape and texture. Therefore, certain conditions should be met to achieve an acceptable image. The edge of the seed should appear sharp in the image, the image of the seed should be large and the surface structure should be sharp too. The small depth of focus with large magnifications made a compromise necessary, so that the magnification of the seed images limited within a 250x250 pixel window was regarded as satisfactory. All seeds within the same species were acquired at the same degree of magnification. However, the variation between species was too large to keep the degree of magnification unaltered for all species. Therefore, a special millimetre scale, divided into one hundredth and one tenth of a millimetre, was used for calculating the actual size of a seed.

To avoid chromatic aberrations green filters were used to make light almost monochro-

matic, and remaining geometric aberrations (including optical distortion) were reduced by centring the image. However, gloss occurred on some seed surfaces, because illumination was not made diffuse.

Unfortunately, background colour and illumination intensity could not be held uniform when a satisfactory segmentation result should be achieved. It was not possible to find a common background colour for both the light and dark coloured seeds. A white background colour had a blurring effect on the seed surface which was unacceptable for seeds with high surface structure. Therefore, the black background was used whenever possible. Also, the lowest possible light intensity was used to reduce overlightning, *i.e.,* the cutting off grey levels above 255. The consequence of the variation in illumination intensity was that seed colour measurements, like average grey level, could not be used for identification.

One solution to the uniform background and illumination problem was suggested in a separate study by Petersen and Krutz (1991). The use of a colour camera in combination with a fast segmentation technique seemed sufficient.

A minor error was noticed during the investigation: An artificial decrease in brightness (shadow) occurred after a sharp transition from light to dark regions in the scanning direction and close to the left image edge. The width of the shadow was 1-2 pixels and the decrease in grey level could reach values about 20. The transition from dark to light caused a corresponding increase in brightness. Images with light background showed a dark shadow inside the seed and a bright shadow outside. In images with dark background this was reversed. However, it was concluded, that this error had no or negligible effect on size and shape analyses, and probably only a small effect on the texture analyses.

World coordinates and image coordinates

When the image has been captured, it is stored in a 512x512 pixel array. Because the pixels were not quadratic, a difference between world coordinates and image coordinates appeared. The ratio of pixel height to pixel length is called the aspect ratio. An aspect ratio of 0.68 was measured which means that x pixels with unit sides in the world coordinate system are mapped into $x/0.68$ pixels in the image coordinate system (Figure 2.2). This, of course, has certain implications on the algorithms used for the image analyses. For area determinations the number of pixels per millimetre was obtained in both the horizontal and vertical direction, but for length measurements a standard aspect ratio was used for the corrections. To get a corrected image on the image monitor the height of the image was adjusted on the control panel.

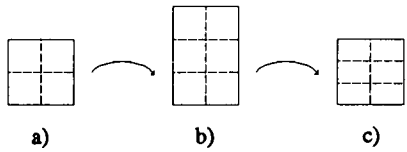


Figure 2.2 Non-quadratic pixels causing discrepancy between the world coordinate system (a) and the image coordinate system (b). The image is shown on the adjusted monitor (c).

3. Seed anatomy

The flowering process results in a fruit with one or many seeds. The seed is the fertilized and ripened ovule which in turn is enclosed in the fruit, which is the ripened ovary. Different types of fruit exist. The dry fruits are often classified as dehiscent or indehiscent. In the latter class only one seed is enclosed in the fruit, so opening is not necessary - the entire fruit functions as a single seed. The indehiscent fruits are classified into different types. Among these are 1) the achene, characterized by a fruit wall tightly fitting around the seed, 2) the nut, in which the fruit wall is a hard and bony, 3) the caryopsis, in which the seed coat is firmly adnated to the fruit wall, 4) the samara, similar to the achene, but winged, and 5) the cypsela, in which the fruit is developed from an inferior ovary and surrounded with the receptacle (Compositae).

An anatomical study of seed coats is important to give an impression of the formation of the many shapes and structures appearing on seeds. For this reason the most relevant anatomical concepts related to the development of angiosperm seeds will be introduced, and selected results from an anatomical and morphological investigation on the ontogenesis of the seed coat of *Stellaria media* will be presented.

The following description is mainly based on Fahn (1987), Boesewinkel and Bouman (1984), and Bhatnagar and Johri (1972).

Ovule development

The mature seed consists of the seed coat on the outer side and inside of the embryo together with some endosperm or perisperm, which functions as the nutrient tissue during germination. The early seed, the ovule, is attached to the placenta by a stalk, called funiculus. The ovule consists of a nucellus surrounded by one or two integuments. The integuments develop into the seed coat. At the nucellar apex a small opening is left by the integuments. This opening is called the micropyle.

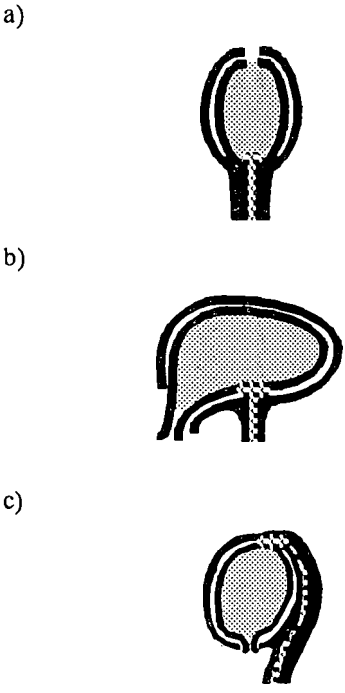


Figure 3.1 Three main types of ovules. a) The orthotropous, b) the campylotropous, and c) the anatropous.

Different types of ovules exist (Figure 3.1). They originate from the ontogeny of the ovule, and characterize the shape of the seed. The main types of ovules are a) the orthotropous or the atropous, in which the ovule apex is straight in line with funiculus, b) the anatropous, in which the ovule apex is rotated backwards, and c) the campylotropous, a median type which may be developed from an initial orthotropous or anatropous ovule (Bocquet, 1959). Different marks may be distinguished on the seed coat: 1) The micropyle, 2) hilum, the scar left by funiculus, and 3) the raphe, a long ridge formed by the fusion of the funiculus with the integuments on the anatropous ovules.

A survey of the different types of local outgrowths of the seed, or seed appendages, is

presented in Kapil *et al.* (1980). The first type mentioned is the fleshy outgrowths, called arils. This type includes the caruncle, a small, disc-like appendage attached to the micropylar region, and the strophiole, an outgrowth limited to the raphal region. The term, elaiosome, is a general ecological word for all fleshy and edible outgrowths. Seeds with elaiosomes are often dispersed by ants because of the food value of the appendage. Another type of appendage is wings which may surround the periphery or be restricted to minor parts of the seed coat. The last type is the hairy seeds which are divided into three groups: 1) Dispersed hairs in woolly seeds, 2) one- or two-sided tufts, and 3) a crown or ring of hairs.

Systematics of seeds

The classic anatomical grouping of seeds according to the location of the mechanical cell layer of the seed coat is testal seeds, in which the mechanical layer is placed in the outer integument, and tegmic seeds, in which the layer is placed in the inner integument. This is further divided into exo-, meso-, and endotestal seeds, and exo-, meso-, and endotegmic seeds depending on wether the mechanical cell layer originates from the outer, middle, or inner cell layer of the integuments, respectively.

Another classification approach is related to the surface structure of the seed coats as it is observed by SEM (scanning electron microscopy). The surface characters are grouped into four categories (Barthlott, 1981):

1) The cellular arrangement, in which the pattern of the surface cells may be characteristic for the taxa.

2) The primary sculpture, in which the cell shape, particularly the curvature of the outer cell wall, has some important aspects:

a) The outline of cells (elongated or isodiametric).

b) The anticlinal cell walls (cell boundaries) which may be straight, curved or lobed.

c) The relief of cell boundary which may be channelled or raised with or without special structures.

d) The curvature of the outer periclinal wall (*i.e.*, cell walls parallel to the surface), which may be flat, concave or convex with or without unicellular appendages (trichomes).

3) The secondary sculpture, in which the single cell wall may show certain characteristics. The structural categories are:

a) Cuticular sculptures which may be patterns of high diversity.

b) Secondary wall thickenings occurring in patterns on the inner side of the periclinal walls.

c) Other structures as micro-papillae.

4) Tertiary sculpture consisting of epicuticular secretions, such as waxes and other lipophilic substances. This is not very common on seed surfaces.

An example

Some of the structures mentioned will be illustrated by the development of ovules of *Stellaria media*. Ovaries of different sizes were treated for SEM investigation. From the morphological appearance of the seed coat, the ovule development was divided into five categories: 1) Ovules with raised anticlinal walls (Figure 3.2) 2) transition to smooth surface, where the rise of anticlinal walls disappear, the periclinal wall becomes convex, and the contour of the single surface cell becomes lobed (Figure 3.3) 3) smooth surface (Figure 3.4) 4) transition to mature surface (Figure 3.5), and 5) mature seed (Figure 3.6 and Figure 3.7).

From the initial orthotropous ovule a campylotropous shape soon developed. The bending process to the campylotropous shape continued over the entire growth period until the micropyle and hilum were close against each other as shown in Figure 3.7. The bending process was a result of cell enlargement of the surface cells close to the micropyle. This resulted in an asymmetric seed where the dorsal cells were flatter on the micropylar side than on the hilar side, and the two 'shoulders' were of unequal size and shape.

The primary sculpture in the mature seed was characterized by lobed cell boundaries with a

channelled relief and special pearl-like structures while the outer periclinal cell walls were concave. The secondary sculpture was limited to some light wart-like appearances on the cell surface.

Other scanning electron micrographs of seeds from the same plant family are shown in Figure 3.8 - 3.12. In general, they are of campylotropous shape with a concave, lobed cell form.

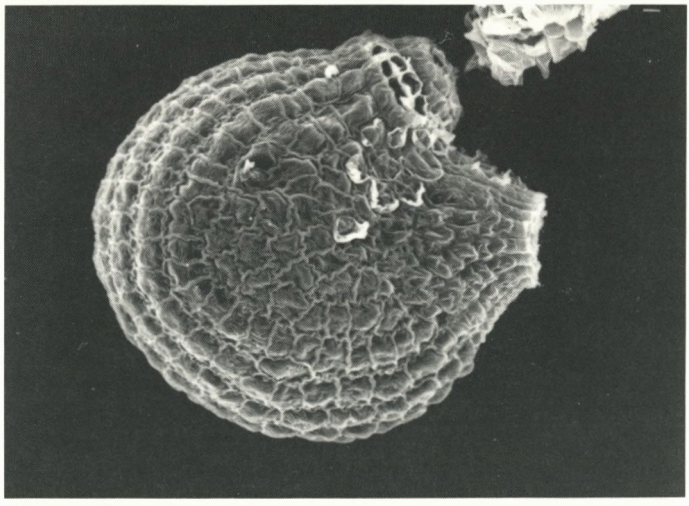


Figure 3.2 Ovule of Stellaria media with raised anticlinal cell walls

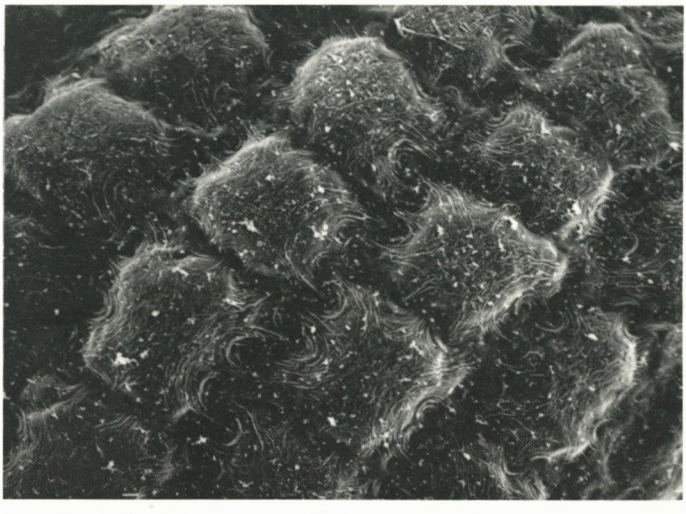


Figure 3.3 Transition to smooth seed coat surface

Figure 3.4 Ovule of S.media with smooth surface

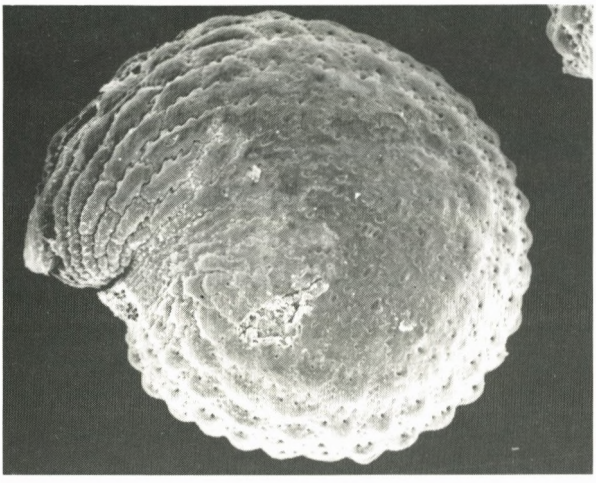


Figure 3.5 Transition to mature seed coat surface

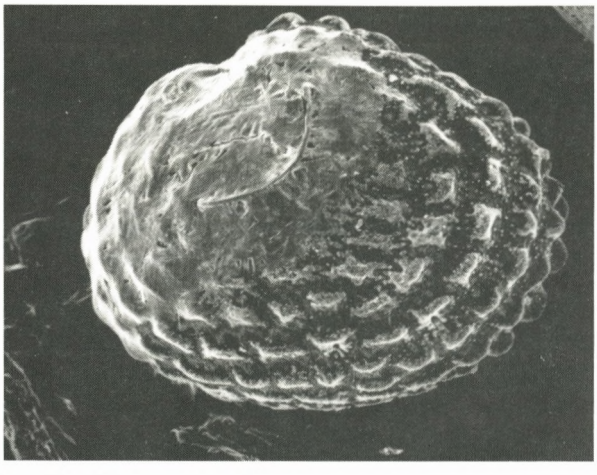


Figure 3.6 Early mature seed of S.media

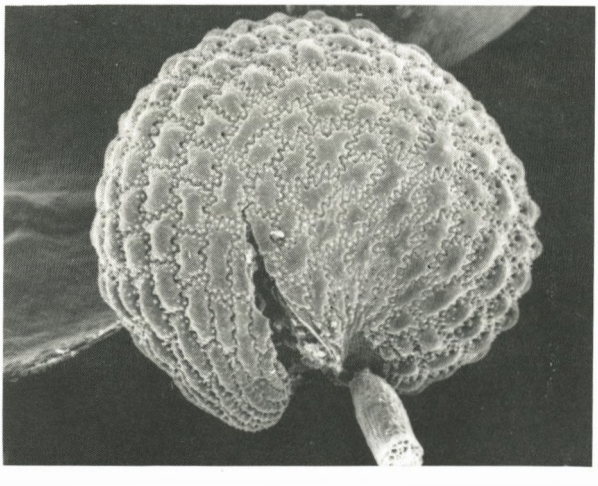


Figure 3.7 Late mature seed of S. media

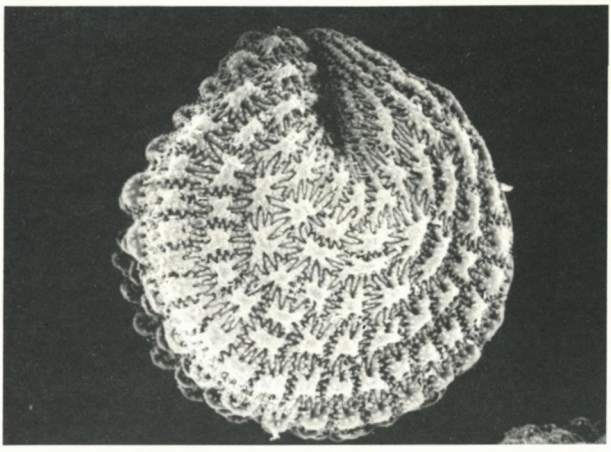


Figure 3.8 Mature seed of Stellaria gramina

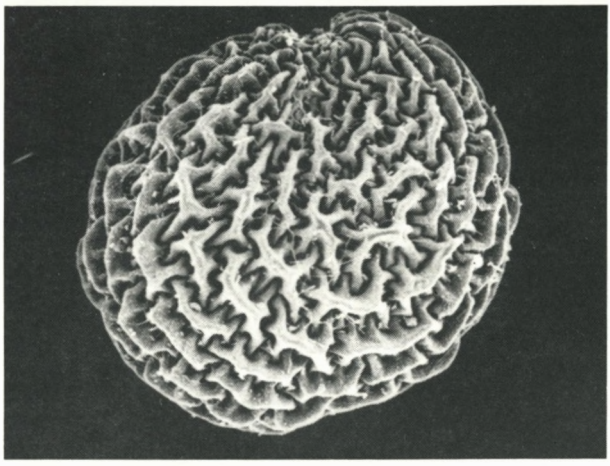


Figure 3.9 Mature seed of Silene vulgaris

Figure 3.10 Mature seed of Silene noctiflora

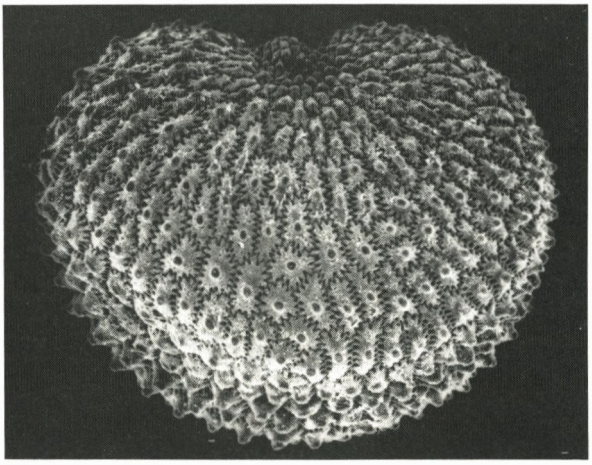


Figure 3.11 Mature seed of Melandrium album

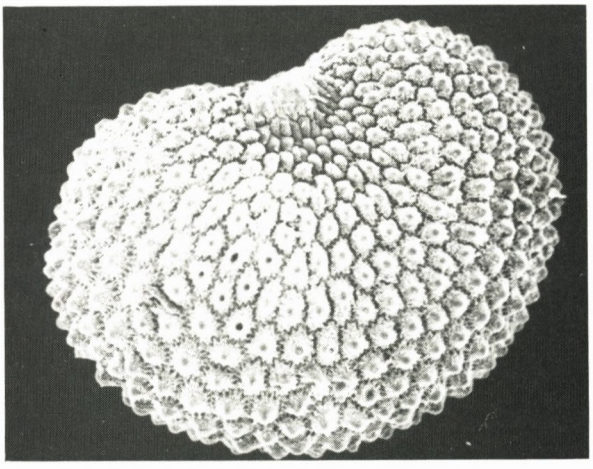
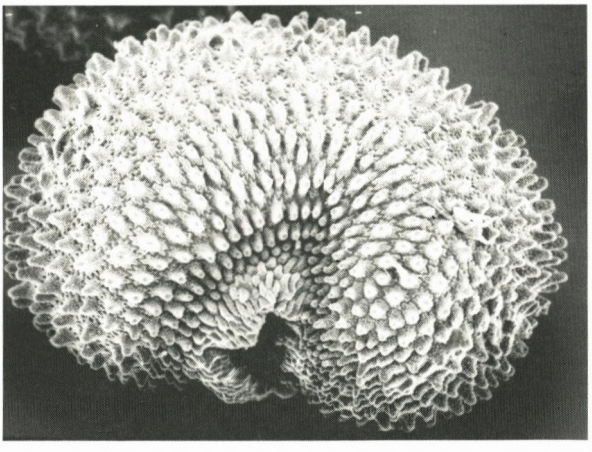


Figure 3.12 Mature seed of Melandrium rubrum



4. Description and analysis of images of the selected weed seeds

When images of seeds are analyzed, a number of features, which constitute a special kind of description of the seed, will be calculated. For comparison of these features with the actual appearance of the seed, pictures of representatives of each species will be presented. After this visual presentation of the variation between species, the variation within species and the technical sources of variation will be briefly evaluated.

When an image of a seed is acquired the following processing and analyzing is dependent on the quality of the image. Details which have never been captured by the camera, will never be described by any analysis. It is equally important that segmentation, as the processing step between image acquiring and image analysis, preserves the fine details of the seeds. Therefore, to evaluate the performance of the different analyses the image acquisition and segmentation should be properly matched to produce an image of acceptable quality. Various analyses of shape and texture are performed on the segmented image. These analyses will be described in this chapter while the next step - classification - will be described in the following chapter.

Selection of the weed seeds

A weed may loosely be defined as a plant out of place, and about 200 Danish plant species are regarded as weeds. Among these, some are of larger economical importance than others depending on their presence in the fields. An investigation of the frequency of the weed seeds in Danish arable soils was carried out by Jensen (1969). Seeds in soil samples taken from 57 cereal and root crop fields at a depth of 0-20 cm were determined by four different methods (washing of soil samples, sowing in greenhouse, sowing outdoor autumn, and sowing outdoor spring). The species were placed in a group according to the highest number of seeds determined by one of the four methods:

Group I: > 1000 living seeds per m^2.
Chenopodium album, Gnaphalium uliginosum, Juncus bufonius, Plantago major, Poa annua, Sagina procumbens, Spergula arvensis, Stellaria media.

Group II: 200-999 living seeds per m^2.
Aphanes microcarpa, Arabidopsis thaliana, Arenaria serpyllifolia, Capsella bursapastoris, Chrysanthemum segetum, Myosotis arvensis, Polygonum aviculare, P. convolvulus, P. persicaria, Scleranthus annuus, Trifolium repens, Veronica arvensis, V. persica, Viola spp. (V. arvensis + V. tricolor).

Group III: 50-199 living seeds per m^2.
Anagallis arvensis, Aphanes arvensis, Cerastium caespitosum, Erophila verna, Hordeum vulgare, Lamium amplexicaule, Matricaria maritima, M. matricarioides, Medicago lupulina, Mentha arvensis, Myosotis stricta, Poa trivialis, Polygonum lapathifolium, Rorippa islandica, Rumex acetosella, Scirpus setaceus, Senecio vulgaris, Sonchus asper, Trifolium spp. (T. campestre + T. dubium), Urtica urens, Veronica polita, V. serpyllifolia.

Group IV: < 50 living seeds per m^2.
Agropyrum repens, Cirsium arvense, Galeopsis spp. (G. bifida + G. tetrahit), Lamium purpureum, Ranunculus repens, Sinapis arvensis, Sonchus arvensis, Stachy palustris, Taraxacum spp., Tussilago farfara, Vicia spp. (V. angustifolia + V. sativa).

However, the criteria for selection of species in the present investigation was not entirely based on frequency of occurrence. First of all, the selection was limited to the dicotyledonous plants. The criteria for selection among these were as follows:

1) Various types of seeds should be represented.

1) Some species should be closely related botanically and have a uniform appearance.

Table 4.1: List of selected species with families in botanical order, genera and species in alphabetic order.

Family	Genus, species	Danish name
Urticaceae	*Urtica urens*	*Liden Nælde*
Polygonaceae	*Polygonum convolvus*	*Snerle Pileurt*
	Polygonum lapathifolium	*Bleg Pileurt*
	Rumex acetosa	*Almindelig Syre*
	Rumex crispus	*Kruset Skræppe*
	Rumex obtusifolius	*Butbladet Skræppe*
	Rumex thyrsiflorus	*Dusk-syre*
Caryophyllaceae	*Arenaria serpyllifolia*	*Markarve*
	Melandrium album	*Aften-pragtstjerne*
	Melandrium rubrum	*Dag-pragtstjerne*
	Silene noctiflora	*Nat-limurt*
	Silene vulgaris	*Blæresmælde*
	Stellaria gramina	*Græsbladet Fladstjerne*
	Stellaria media	*Fuglegræs*
Chenopodiaceae	*Chenopodium album*	*Hvidmelet Gåsefod*
Ranunculacaea	*Ranunculus repens*	*Lav Ranunkel*
Papaveraceae	*Papaver rhoeas*	*Korn-valmue*
Cruciferae	*Brassica campestris*	*Agerkål*
	Capsella bursa-pastoris	*Hyrdetaske*
Geraniaceae	*Geranium dissectum*	*Kløftet Storkenæb*
Euphorbiaceae	*Euphorbia exigua*	*Liden Vortemælk*
	Euphorbia helioscopia	*Skærm-vortemælk*
	Euphorbia peplus	*Gaffel-vortemælk*
Violaceae	*Viola arvensis*	*Ager-stedmoderblomst*
Boraginaceae	*Myosotis arvensis*	*Mark-forglemmigej*
Labiatae	*Lamium amplexicaule*	*Liden Tvetand*
Solanaceae	*Solanum nigrum*	*Sort Natskygge*
Scrophulariaceae	*Veronica arvensis*	*Markærenpris*
	Veronica persica	*Storkronet Ærenpris*
Plantaginaceae	*Plantago major*	*Glat Vejbred*
Compositae	*Cirsium arvense*	*Agertidsel*
	Chrysanthemum segetum	*Gul Okseøje*
	Matricaria chamomilla	*Vellugtende Kamille*
	Matricaria inodora	*Lugtløs Kamille*
	Matricaria matricarioides	*Skive-kamille*
	Sinapis arvensis	*Agersennep*
	Sonchus arvensis	*Agersvinemælk*
	Sonchus asper	*Ru Svinemælk*
	Sonchus oleraceus	*Almindelig Svinemælk*
	Taraxacum vulgare	*Fandens Mælkebøtte*

3) Frequency of occurrence in the fields was also considered.

As mentioned earlier the last criterion was not of very high priority. Certain species were selected due to the second criterion without being of high economical importance, and a single species with high frequency of occurrence was rejected because of segmentation difficulties (see later). Finally, the selection of species was restricted to the species available.

A collection of weed seeds was delivered by the Danish Botanical Garden. From this collection species were selected for this investigation. These are listed in Table 4.1 in botanical order from the phylogenetical more primitive to the more advanced.

It is often so, that closely related species share certain common seed characters. In Figure 4.1 the fruits belonging to the same genus, *Rumex*, are presented. They are 3-angled achenes with a smooth, shining, dark brown surface. The fruits are borne invested with calyx-wings which were removed in this study to define a standard condition. Figure 4.2 shows six species belonging to three genera of *Caryophyllaceae*. They all have a highly structured surface originating from the out-bulging of the seed coat cells. Figure 4.3 shows seeds of the *Euphorbia* genus. They are variously pitted, have a visible raphe and an elaiosome (caruncle). Seeds used from this genus were all placed with the raphe side up and contain an intact elaiosome. The genus, *Matricaria*, from the *Compositae*, are disc-flowers with the flowers born on a common receptacle. A calyx-tube adnate completely to the ovary, and the fruit (cypsela) is 3-5-ribbed with a pappus (formed from the calyx) as a short crown appearing on *M. inodora* and *M. matricaria* (Figure 4.4). Two other genera of the *Compositae* family are shown in Figure 4.5. This is fruits of the *Sonchus*, which are 10-20-ribbed, and the pappus consists of a soft white bristle usually falling away. The pappus on the *Taraxacum* fruit is supported by a small beak, thus giving a different scar when the bristle is removed. The other species of weed seeds included in this investigation are grouped in big seeds (Figure 4.6), small seeds (Figure 4.7), round and oval seeds (Figure 4.8) and the last group of species as mixed (Figure 4.9).

The variation of the appearance of the seeds within species is illustrated in Figure 4.10 by six different seeds of *Silene vulgaris*. This variation is characteristic for seeds of the *Caryophyllaceae* family, and when selecting seeds for the automatic identification the most extreme seeds in this family were avoided.

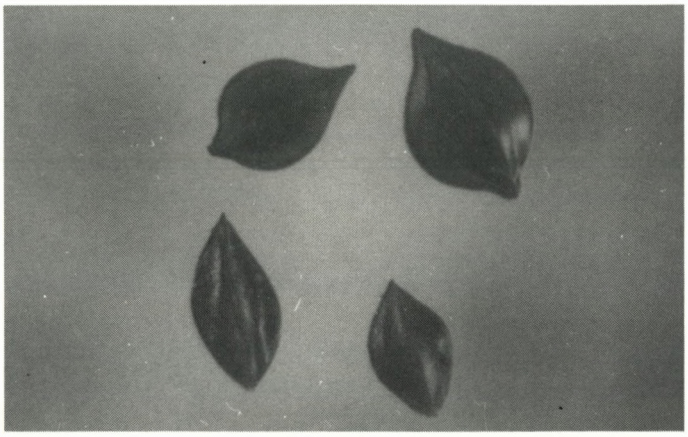


Figure 4.1 Image of four seeds of Rumex: R.obtusifolius (upper left), R.crispus (upper right), R.acetosa (lower left) and R.thyrsiflorus (lower right).

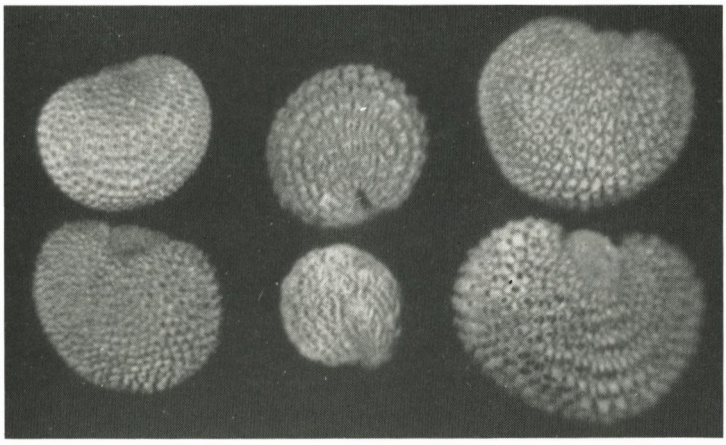


Figure 4.2 Image of six seeds of Caryophyllaceae: Melandrium album (upper left), Melandrium rubrum (lower left), Stellaria media (upper middle), Stellaria gramina (lower middle), Silene noctiflora (upper right), and Silene vulgaris (lower right).

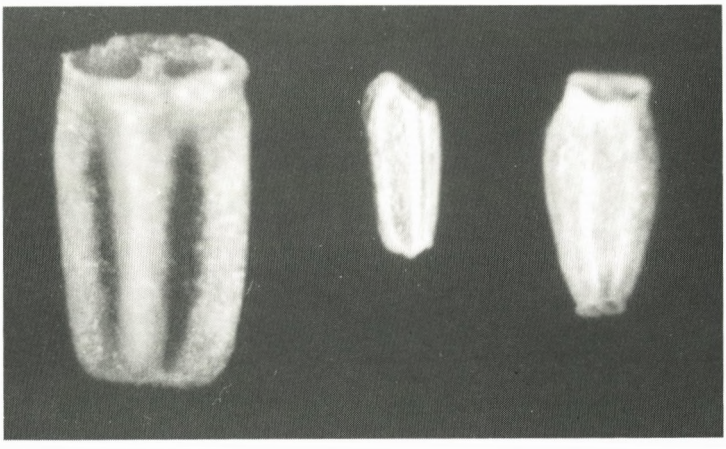


Figure 4.3 Image of three seeds of Euphorbia: Euphorbia helioscopia (left), Euphorbia peplus (middle) and Euphorbia exigua (right).

Figure 4.4 Image of three seeds of Matricaria: M.inodora (left), M. chamomilla (middle) and M.matricarioides (right).

Figure 4.5 Image of fruits of Sonchus and Taraxacum from left to right: Taraxacum vulgare, Sonchus oleraceus, Sonchus arvensis and Sonchus asper.

Figure 4.6 Image of big seeds: Polygonum convolvolus (upper left), Polygonum lapathifolium (upper middle), Ranunculus repens (upper right), Chrysanthemum segetum (lower left) and Cirsium arvense (lower right).

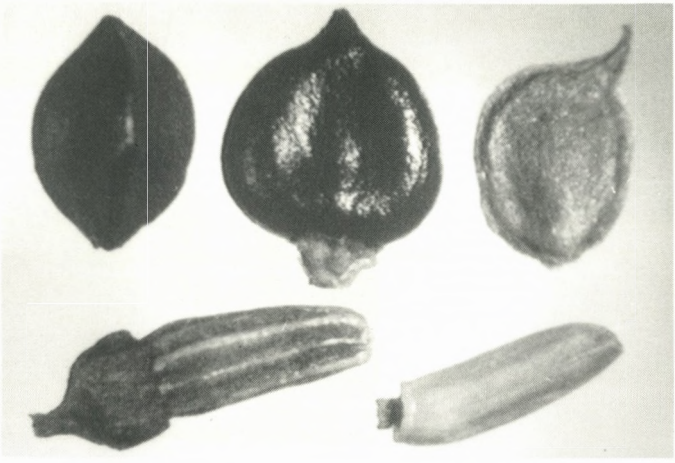


Figure 4.7 Image of small seeds: Papaver rhoeas (upper left), Arenaria serpyllifolia (upper right), Capsella bursa-pastoris (lower left) and Veronica arvensis (lower right).

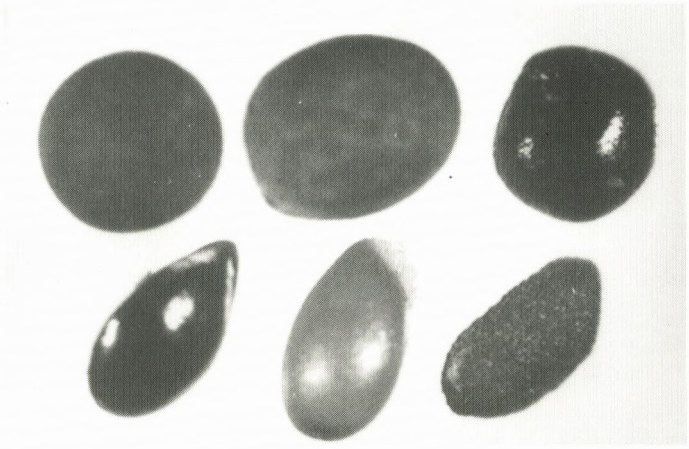


Figure 4.8 Image of round and oval seeds: Sinapis arvensis (upper left), Brassica campestris (upper middle), Chenopodium album (upper right), Myosotis arvensis (lower left), Viola arvensis (lower middle) and Plantago major (lower right).

Figure 4.9 Image of seeds of Geranium dissectum (upper left), Veronica persica (upper middle), Solanum nigrum (upper right), Urtica urens (lower left) and Lamium amplexicaule (lower right).

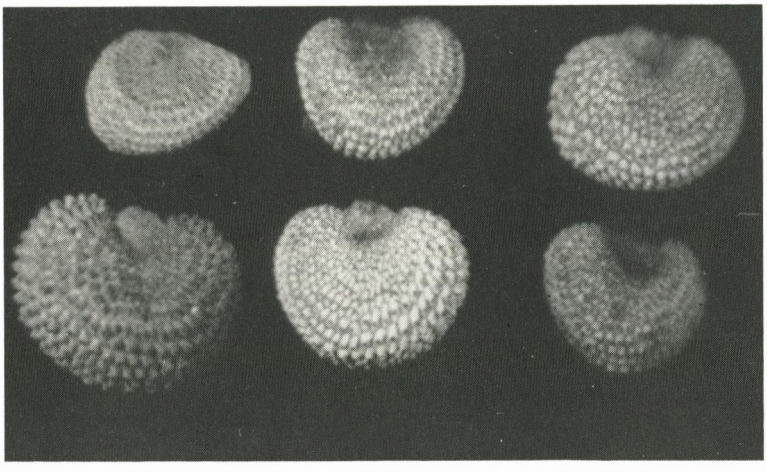


Figure 4.10 Image of six seeds of Silene vulgare

Technical error sources

Various possible errors in a computer vision system are described in general terms in chapter 2. Now, a few simple measurements will illustrate the magnitude of the following technical sources of variation: 1) Focusing, 2) optical magnitude, 3) illumination intensity (overlightning and AGC), 4) gloss, and 5) orientation of the seed. The features used are later defined in connection with the description of the image analyses. Other kinds of error are also present, such as different optical aberrations, shading, dust etc. but these were difficult to measure separately and/or considered of minor importance.

The influence of the human focus control was estimated by measuring the contrast of the seed surface. Ten images of the same scene (a *Silene vulgaris* seed) were captured with renewed focusing, and the mean and standard deviation of the average greytone, the contrast, extracted from the grey level cooccurrence matrices (GLCM) defined in section 4.7, and the run percentage, expressing the relative number of adjacent pixels with the same grey level, were calculated as shown in Table 4.2. The relative deviations appear to be relatively small (0.4 - 5 percent of the mean), indicating a fairly uniform focusing control.

Depth of focus has a significant influence on the surface structure of the image, and by increasing the magnification the depth of focus will decrease, resulting in an increasingly blurred image. Therefore, the change in the degree of magnification is another factor affecting the contrast of the image. A series of images was captured of the same seed, *Silene vulgaris*, at different degrees of magnification. The results are presented in Table 4.3 showing about 0.15 increase in contrast units per one thousand pixel decrease in area. Within the magnification degree normally used the contrast may vary about 2 units or 15 percent for this seed. However, in the present study all images of seeds belonging to the same species were acquired at the same magnification, but all future test seeds will be influenced by this error.

Table 4.2 Deviation arising from focusing.

	Mean	Std. dev.
Avg. greytone	197.6	0.93
Contrast	9.74	0.51
Run Percentage	0.81	0.0034

Table 4.3 Effect of magnification on contrast and run percentage

Area (in pixels)	Contrast	Run Percentage
42877	9.64	0.82
33179	11.69	0.83
24694	13.72	0.85
18291	12.43	0.84
14413	14.46	0.86
12330	14.83	0.86

Computer vision systems are in general very sensitive to illumination. Shadows in different degrees appear depending on the light intensity and the number, the position and the angle of the lamps etc. Therefore, it is very important to have illumination in a well-defined state. Unfortunately, this was not possible in this study, because the segmentation procedure required stronger illumination of some seeds than of others. This caused, of course, variation in grey levels, but some special problems arise from differences in illumination. If light intensity was too high the grey level of the pixels was cut off (above grey level 255). This will cause a reduction in image contrast. On the other hand, low intensities may not enlighten the fine details, thus causing a decrease in contrast (blurred image). These two effects may explain the optimum for contrast measurements in a window size of 120x120 pixels in a series of images of a *S. vulgaris* seed at different light intensities. The contrast values are shown in Table 4.4 together

Table 4.4 Effect of different illumination intensities on contrast, run percentage and number of pixels values equal to 255. Seed was Silene vulgaris, and measurements were based on window size of 120x120.

Avg. Greytone	Con- trast	Run Per- centage	Pixels =255
164.2	8.54	0.83	0
187.8	10.08	0.84	67
197.6	10.98	0.84	417
222.2	11.15	0.70	3450
232.1	9.70	0.59	5681
246.7	5.15	0.33	10268

with the average greytone of the seed surface as a measure of light intensity and run percentage, reflecting the reciprocal spot size. The last column shows the number of pixels of value equal to 255. The run percentage is not affected by changes in intensity until the effect from overlightning is above a certain level. Below this level, the increase in illumination intensity seems to cause increased grey level differences (contrast), but with approximately the same spot sizes (run percentage).

The different background colours had a great influence on the image of the seed. The effect appeared to be related to the AGC and maybe some stray light (*i.e.*, reflected light from the objective on the specimen). To evaluate this effect a new series of images of a *S. vulgaris* seed were acquired using light and dark backgrounds at different illumination intensities.

The results are shown in Table 4.5. The influence of a light background is to decrease the gain thus causing a considerable decrease in average greytone and contrast, when illumination intensity is medium to high, while for low intensities, *i.e.*, from illumination intensity "1" to "2" in Table 4.5, it showed an increasing effect. At the high intensity level the effect from overlightning becomes visible on the contrast values, when using a dark background.

Gloss appears on seeds with smooth surfaces. Variation depends on the orientation of the

Table 4.5 Effect of light and dark background on average greytone and contrast. Scale of illumination: 1-low to 5-high.

Illumi- nation	Back- ground	Avg. Greytone	Con- trast
1	light	106.4	2.84
1	dark	103,2	2.39
2	light	137.2	5.82
2	dark	127.5	6.34
3	light	159.7	4.31
3	dark	158.5	9.40
4	light	168.3	3.26
4	dark	192.8	12.24
5	light	174.6	2.79
5	dark	229.8	10.53

seed and the number of lamps. The *Rumex crispus* seeds, for example, show most gloss when orientated with the principal axis across the illumination direction and almost no gloss when orientation is parallel to illumination. An increase in the number of lamps will increase the gloss, but decrease the variation due to orientation. In Table 4.6 the effect of orientation of a *R. crispus* seed is presented for average greytone, contrast and run percentage. The effect is highest for average greytone and contrast, while run percentage only changes slightly.

Table 4.6 Effect of gloss on R. crispus seed. Position with low gloss is parallel to illumination direction.

Position	Avg. Greytone	Contrast	Run Percentage
parallel	119.1	2.74	0.74
cross	148.9	5.68	0.74

Table 4.7 Mean and variance components for four Silene vulgaris seeds measured on each side in three orientations.

	Area (in mm^2)	Compactness	Run Percentage
Mean	1.45	1.74	0.79
Var(seeds)	0.0078	0.046	0.0010
Var(side)	0.0001	0.011	0.00007
Var(orient.)	0.0002	0.014	0.00007

Finally, the effect of orientation was estimated for a *S. vulgaris* seed with structured and non-glossy surface. Images of four different seeds, two sides and three different orientations were analyzed. The variance components were calculated for area in mm^2, compactness (reflecting shape) and run percentage (Table 4.7). The variation caused by orientation was very low for area and run percentage and relatively high for compactness. Furthermore, the variation for side and orientation is approximately equal.

In general, the texture analyses were shown to be sensitive to the effect of the technical error sources. The most important errors were variation in illumination intensities and the AGC operation with different backgrounds. The consequences are that the average greytone measurements are considered useless in this study, and that all seeds as far as possible are captured using the dark background. The effect of a different degree of magnification is also important for future analysis of unknown test seeds.

Some species in the basis collection were not included in the investigation because of technical limitations. The seeds of the species *Cerastium fontanum, C. semidecandrum, C. glomeratum and Juncus bufonius* were considered too small to be properly represented and seeds of *Spergula arvensis* were not possible to segment with the dark background, and segmentation with light background blurred the surface structure unacceptably.

4.1 Segmentation

Segmentation is a process where the image is divided into regions according to certain criteria. An important special case is the separation into background and object. When an image of satisfactory quality is captured, a segmentation often follows. The result of this segmentation is the basis for all the following image analyses, and, therefore, the precision of the segmentation is very important. Numerous segmentation methods exist, and in the following they are briefly introduced. After this, special attention will be paid to the very popular thresholding methods.

Segmentation methods

Two basic concepts are included in the definition of segmentation:1) Connectedness (the neighbouring regions of an image are connected), and 2) homogeneity (each region is homogeneous with respect to one or more integrated definitions). Based on these two concepts Borisenko *et al.* (1987) expressed the general form of segmentation by four conditions: 1) All regions of the image are divided and no regions are overlapping, 2) all regions are connected, 3) the chosen conditions for homogeneity are met (or true) for all regions, and 4) all connected regions can not be further merged with regard to the homogeneity conditions.

The first group of segmentation methods, described by Borisenko *et al.* (1987), consisted of the clustering methods. Thresholding is an example of a clustering method.

The second group of segmentation methods is the region growing methods. Here, the starting points ('kernels' or 'atoms') may be selected arbitrary or as zero variance segments. The growing of these points may be controlled by different features.

The third group is the splitting methods. Two major algorithms are found in this group: The split-and-merge (Horowitz and Pavlidis, 1976) and the split-and-link (Pietikanen *et al.*, 1982) algorithm. As indicated by the names the splitting methods also include merging procedures.

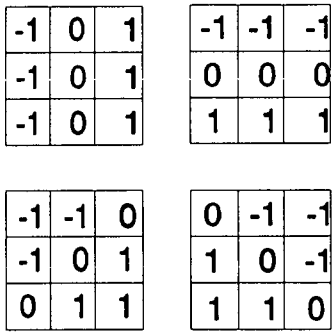


Figure 4.1.1 Simple edge masks for the four possible directions.

Purely splitting is just a special case in the algorithms.

The last group of segmentation methods in Borisenko *et al.* (1987) is the edge detection group. The edge is the transition between regions where homogeneity breaks down. Two different approaches exists: 1) Edge fragment extraction where the edge is detected using local analysis with edge masks (Figure 4.1.1), and 2) edge tracking where adjacent edge points are found sequentially from some starting point.

Finally, a combination of the different methods may solve the actual segmentation problem. An example of such a hybrid technique is given by Yakimovsky (1976) where edge analysis is combined with region growing.

Thresholding

Weszka (1978) described a threshold operator in the general form as a function

$$T(x,y,N(x,y),g(x,y))$$

where g(x,y) is the grey level of the point (x,y) and N(x,y) is some local property of the same point. Then, three different threshold types were defined:

1) If T only depends on g(x,y), the threshold is of the global type.

2) If T depends on both g(x,y) and N(x,y), the threshold is of local type.

3) If T depends on the values x,y (*i.e.*, position in image) as well as g(x,y) and N(x,y), the threshold is the dynamic type.

The global thresholds are often based on the grey level histogram. If an object is known to be darker (or lighter) than the background, and also is known to occupy some percentage of the entire picture, the grey level that maps this pixel percentage into the object is chosen from the histogram as the threshold. If the size of the object is unknown, the threshold may be chosen from the valley between two modes on the histogram.

The grey level histogram may not always be clear bimodal, and for this reason some methods try to improve the shape of the histogram. One method is to weight the greytone of each pixel using a difference operator at that pixel. Pixels on the border, *i.e.*, those with high difference values, are weighted low, and others are weighted high. Another possibility is to make histograms of only the pixels with high gradient values, *i.e.*, pixels close to the boundary. Finally, the values from a difference operator on each pixel may be used more directly. Here, Watanabe (1974) introduced a method where the difference value for each greytone was calculated. The grey level with the highest proportion of high-difference values was chosen as the global threshold.

In local thresholds methods the value of a center pixel is compared to some neighbours in a certain distance. The allocation of the center pixel may depend on the difference between center and neighbours, and different allocation rules may be dependent on the grey level of the center. Other methods chose the threshold from two-dimensional plots of grey level versus gradient value, or use different histograms of edge points for high and low grey levels respectively.

In dynamic threshold selection by Chow and Kaneko (1972) local statistics were calculated in overlapping windows of the image. From a local histogram analysis a threshold was determined for all windows satisfying a bimodal test. Then, thresholds were obtained for every point in the

image using linear interpolation. In this way the threshold of a point depended on the proximity to the boundary points first obtained.

Mardia and Hainsworth (1988) described a comparative study of eight different thresholding methods. All the methods were iterative where the initial threshold was the mean grey level for the two group situation. In the successive segmentations the threshold was based on the parameters of the populations estimated by the previous segmentation. This proceeded until convergence was achieved. Different approaches were used in construction of the thresholds. Among these were Bayesian methods, spatial methods and use of post-smoothing.

Finally, a more complex form of thresholding was described by Kohler (1981). This method was based on multiple thresholds using both pixel similarity and pixel difference information.

Object smoothing

Although the original object has a smooth surface, segmentation methods often produce segments with non-smooth boundaries. Therefore, a post-processing smoothing may be valuable. A general smoothing method is called morphological image processing. This type of processing modifies the spatial form or structure of objects within an image (Pratt, 1991). The smoothing of the object boundary is performed through a series of so-called erosion - dilation cycles. An erosion operation performs a shrinking of the objects, and a dilation will make them grow, and both processes are limited to a single ring of boundary pixels.

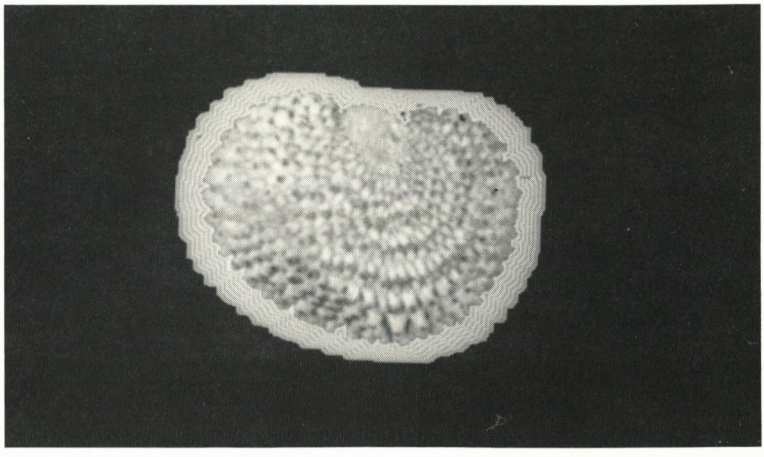
If only the boundary needs smoothing the shrink/expand methods described by Niblack (1985) may be sufficient. The function is similar to erosion/dilation, but no filters are used. The shrinking process removes the outer pixel layer, whereby small irregularities disappear. The following expanding process adds a new pixel layer to the boundary. If the operation starts with expanding followed by shrinking small concavities on the boundary disappear. Figure 4.1.2a shows several expanding processes on a weed seed, and in Figure 4.1.2b the resulting image after the same number of shrinking operations.

The segmentation algorithm

The image segmentation algorithm used in this study was of the global threshold type, where the threshold was a constant determined from measurements of the background at normal illumination.

Two conditions should be met: 1) The seed is approximately centered in the image, and 2) the seed does not exceed a window size of one fourth of the total image. This was chosen to get a reasonably sharp image of the seed sur-

Figure 4.1.2a Expanding the boundary



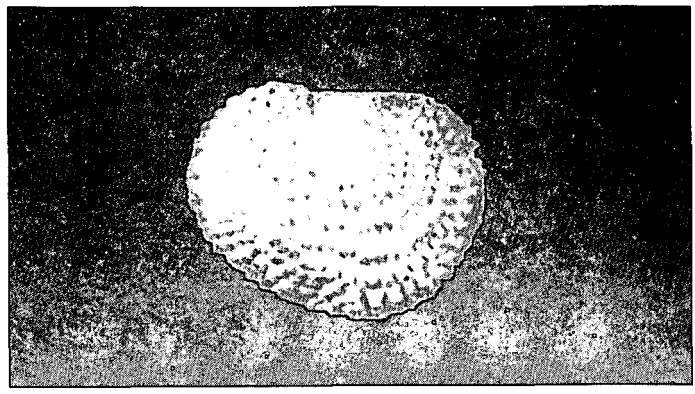


Figure 4.1.2b The result of shrinking the previously expanded boundary

face as earlier described. This means that only the center window had to be considered. Furthermore, it was not found possible to segment both light and dark seeds using the same background colour. Therefore, two procedures were developed, one for the black background and one for the light background. In the following description a black background is regarded.

The algorithm starts in the upper left corner of the window, and moves to the right until a value bigger than the threshold is met. During this movement each pixel is set to zero indicating the background segment. When the value above the threshold is met, this movement is repeated for the next (lower) row of pixels and so on, until the bottom row is reached. Then, the procedure is repeated from the upper right corner with movement in the opposite direction. Finally the procedure goes in the top-down and in the bottom-up directions.

In the next step the algorithm searches from the center of the window in horizontal direction until the first background value (zero) is met. A border tracking routine (see later) then traces the boundary between background and values above zero. All the pixel values on the boundary below the threshold are set to zero, and the tracing is repeated until no value below the threshold is met on the entire boundary. This ensures that all boundary bendings are filled with zeroes. Alternatively, the segmentation could be implemented as a region filling routine which, started from the background, should fill the background segment with zeroes. However, the algorithm described allows pixels in the middle of the seed surface to be below the threshold, but the disadvantage is an increased computation time compared to direct operations on the LUTs.

The border tracking routine

This algorithm is fundamental for detection of the boundary. As already mentioned is was used in the segmentation, but it is also used in the post-smoothing procedure.

The boundary is regarded as the outer ring of pixels in consecutive order of an object. The algorithm starts from the center of the segmentation window and goes to the right until a pixel of zero value is met. Then, it proceeds tracking the border in a clockwise direction. The basic principle in the algorithm is that a right background neighbour (value zero) cause a downward movement, and a left background neighbour cause an upward movement. This vertical movement stops when the center pixel is a background pixel or no vertical neighbour pixel belongs to the background. This initiates a horizontal movement along the same pixel row. In the border tracking program a direction up or down is found after detecting the first background pixel, and then move to this pixel (here named the shift pixel). This pixel is not

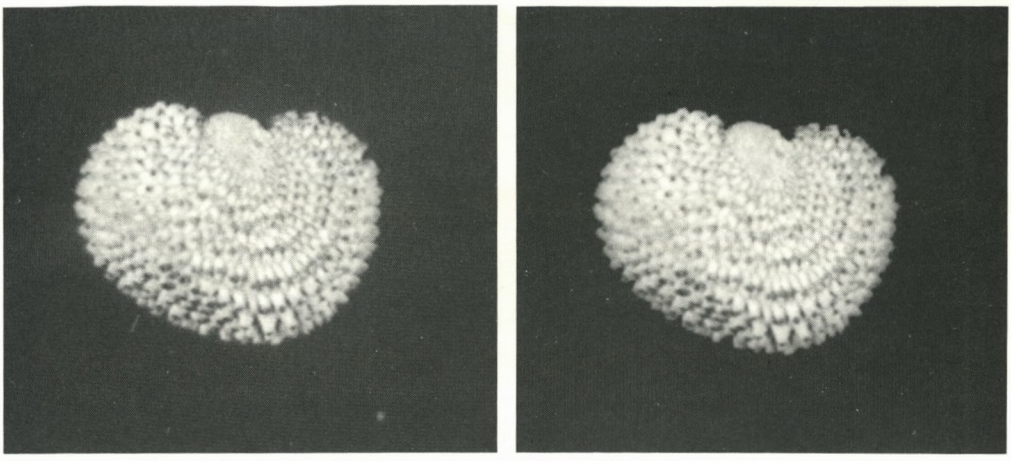


Figure 4.1.3 Image of seed of S.vulgaris before (left) and after (right) segmentation

yet saved as a boundary point, but is used to decide which new action to take. There are four cases depending on the value of the shift pixel and the previously decided up/down movement: zero/down, zero/up, nonzero/down, nonzero/up. In each case the routine now moves in a horizontal direction and saves the boundary pixels (except the shift pixel), or goes back and saves the shift pixel and proceeds with a new up/down decision (straight vertical movement). (If the shift pixel has value zero, the boundary pixels saved during the horizontal movement are above, in the zero/down case, or below, in the zero/up case, the actual position.) The horizontal movement proceeds until a certain condition is met, and the routine is repeated with a new up/down decision. The structure of the movement decisions is shown in appendix A (algorithm 1).

The algorithm was constructed to create a boundary of eight-connected neighbouring points, when the following conditions were met:

1. The nonzero shift pixel should not be saved as a boundary point unless it is part of a vertical boundary line.

2. When the horizontal movement stops, the actual pixel is not saved if it is nonzero and placed in the same row or column as the previous boundary pixel, *i.e.*, the nonzero/up and nonzero/down case stopped by a lower or upper nonzero pixel, respectively.

Finally, it is possible to use backtracking, when the end of a line of a single pixel width is reached. Otherwise, the boundary will contain the same points two times.

A minor weakness was discovered in this algorithm. It is not possible to shift from a right movement to an upward movement along a single pixel line, because the up/down decision rule will force the movement downward. Only the first pixel of this line will be registered. However, this was not corrected because single pixel lines were unwanted in these segments.

Post-smoothing

The border tracking routine may use 'backtracking', when the algorithm has reached the end of a pixel line (*i.e.*, with a length above one pixel) of a single pixel width. However, these lines are considered as artifacts, and besides they cause difficulties in the following shape analysis programs. Therefore, these lines were cut off after the segmentation as a gentle smoothing operation.

An example of an image before and after segmentation is shown in Figure 4.1.3.

4.2 Simple measurements

After extraction of the object segment, the next step is often to represent the boundary in a suitable way. Therefore, different representation schemes are introduced, and the algorithm for the selected boundary representation is described. Based on the boundary information some simple measurements are calculated for the weed seed images. These measurements are of morphological nature, and may therefore be of taxonomical value. However, they will provide an overview of the data set used for classification.

Contour representation

The straightforward way of representing a curve in a digitized image is in the x,y coordinates of the pixels. An alternative scheme is the so-called Freeman chain code presented by Freeman (1961, 1974). The principle is to describe the curve by a sequence of angles using a standard distance. For the grid of pixels it is clear that for a given curve point the next point is located in one of eight possible directions. In the Freeman chain code these directions are labeled from zero to seven in counterclockwise directions as shown in Figure 4.2.1. In a curve represented by the successive directions only the first point has to be registered in an absolute sense.

Different techniques of manipulating chain coded curves are described by Freeman (1961, 1974). For example, expansion is performed by replacing each digit (or link) by a set of identical digits, where the number of digits in the set depends on the expansion ratio, and rotation is performed by adding a number to each digit. Thus, adding '1' to each digit will rotate the curve 45 degree. However, more interesting for image analysis is the determination of features like length and area. Thus, the length is

$$L = n_e + n_o\sqrt{2}$$

3	2	1
4		0
5	6	7

Figure 4.2.1 The Freeman chain code.

where n_e and n_o are the numbers of even and odd valued links, respectively. For area determination the integration with respect to the x axis is used. Denoting the chain links a_i, the change in the y coordinate between links is denoted $a_{iy} = y_i - y_{i-1}$ and similarly for a_{ix}, the area encircled in a clockwise sense is then

$$S = \sum_{i=1}^{n} a_{ix}(y_{i-1} + \frac{1}{2}a_{iy})$$

A closely related representation of the contour is termed shape numbers, which are used to measure the similarities between shapes (Briebiesca and Guzman, 1980). For this purpose the derivative of a chain code is defined as follows. The chain code is first simplified to a four-connected representation, and then the concave corners are assigned the value '1', the straight links the value '2', and the convex corners the value '3'. The procedure for obtaining the shape numbers is

1. Extracting a region limited by a curve boundary, *e.g.*, a silhouette of some object.
2. Overlay a grid of arbitrary cell size.
3. A new region is formed with all the cells that fall above 50% inside the region (silhuette).
4. The chain code of the new region is obtained. From this all possible derivatives of that chain (depending on the starting point) are constructed.
5. For normalization the derivative which is the minimum number (*e.g.*, reading the chain of digits as a single number - the shape number) is selected.

For further normalization the orientation of the grid has to coincide with the major axis of the grid. Also, to get a unique shape number

the order of the shape number, *i.e.*, the number of digits in the shape number, has to be specified.

When comparing two shapes by shape numbers the degree of similarity is the wanted measure. Briebiesca and Guzman (1980) proposed to use the largest order of the shape numbers, where they are identical, as similarity measure. And the distance between shapes was defined to be the inverse of their degree of similarity.

Polygonal approximation is a contour representation, where the boundary points are reduced to some extent and replaced by straight line segments. It is here the goal to represent the boundary with as few line segments as possible, but also with the smallest possible error (distance between line and boundary). There are two different approaches to the construction of the polygon (Teh and Chin, 1989). One is to fit straight lines along the boundary, where the lengths depend on some error threshold. The second approach concentrates upon a dominant point search. The dominant points will be the corners of the approximated polygon, and they are located where larger changes in curvature are registered. However, curvature, defined as the rate of change of slope as a function of arc length, has the problem for the digital curve that slope angles only differ by multiples of 45 degrees. To solve this problem a region of support is introduced, where line segments of a length > 1 will smooth the angles. The dominant point detection, therefore, is dependent on 1) the measure of curvature used, and 2) the size of region of support. But difficulties arise from the selection of the size of this region. If it is too large dominant points of fine features will get lost, and if too small also non-dominant points will be determined as dominant points.

Teh and Chin (1989) presented a review of five earlier algorithms for detecting dominant points and proposed a new algorithm where the region of support varies from point to point dependent on local properties. A comparison of the algorithms showed that detection of dominant points relies heavily on the region of support and only to a minor degree on the curvature estimation method.

Finally, the centroidal profile should be mentioned as a possible way to represent the contour (Freeman, 1978). This profile is a normalized plot of the distance from the boundary to the centroid (*i.e.,* the center) of the silhouette, as a function of distance along the boundary. To remove the dependency of scale the profile values should be divided by the maximum value. Furthermore, it was suggested that a fixed number of observations and selection of the maximum value as the initial value could provide some uniformity in the profiles.

Simple shape description

The simple measurements used as shape descriptors are numerous. In a review Levine (1985) presented a small collection. Among these were:

1. Angle regularity (closed polygon). If the number of vertices is n, the number of boundary points is m, and θ_k the interior angle of the k'th boundary point, then

$$A_1 = \frac{1}{n}[(\theta_1 - \theta_m) + \sum_{k=1}^{m-1} (\theta_{k+1} - \theta_k)]$$

2. Compactness or circularity.

$$A_2 = \frac{Perimeter^2}{4 \times \pi \times Area}$$

This has the minimum value 1 for a circle.

3. Side regularity. If the number of sides in the polygon is n, the average side length is L, and the length of the k'th side is l_k, then side regularity is

$$A_3 = \frac{[\sum_{k=1}^{n} (l_k - L)^2]^{\frac{1}{2}}}{2 \times L \times (n - 2)}$$

This is proportional to the coefficient of variation of length.

4. Average bending energy normalized for

size. The curvature, R(k), is the change in tangent direction divided by the distance between the k'th and the k-1'st point. The energy is

$$A_4 = 1 - \frac{4\pi^2}{P\sum_{k=0}^{m-1}|R(k)|^2}$$

where P means perimeter.

5. Elongation or eccentricity.

$$A_5 = \frac{|D - W|}{D}$$

where D and W are the lengths of the major and minor axes, respectively.

Algorithms for area, center of gravity and perimeter

The area of the weed seed is presented in mm^2. The algorithm first calculated the number of pixels, and later the calibration with measurements of the standard mm scale in horizontal and vertical direction was performed.

An array of boundary points was input for this algorithm. An integration method was used for calculating the area and the center of gravity (see appendix A, algorithm 2). The integration can be done with respect to x or y. When following the contour some values were counted negative and some positive depending on the direction. This direction was determined by the position of the previous neighbour and the position of the next neighbour in the boundary.

These positions could be named like the chain code (Figure 4.2.1). Thus, the previous/next pixel combination of code value 4/0 should update the sum by positive values, whereas the opposite 0/4 code contributed with negative values. The combination 0/1 updated with a positive value and a negative value (one less than the positive), and the combination 2/6 made no updating. Similarly, all possible combinations were considered.

The perimeter length was calculated straightforwardly by setting horizontal pixel length equal one and vertical length equal to aspect ratio (r). Diagonal length was then $\sqrt{1+r^2}$. The sum was divided by the horizontal number of pixels per mm. Finally, the compactness was calculated as shown in the previous section.

The contour representation

A representation of the contour analog to the centroidal profile was used for obtaining other simple features. This representation will in the following be termed the time series representation, and it was constructed from the algorithm description of Dubois and Glanz (1986). The general principle of this algorithm was to approximate the boundary by an ordered sequence of N angularly equispaced radius vectors from the centroid and the boundary. The vector lenghts were used as the observations as shown in Figures 4.2.2 - 4.2.7 for different weed seeds. In this algorithm it was possible to obtain more than N observations when bendings on the boundary caused the algorithm to move in the opposite direction (counterclockwise). However, a single modification of the Dubois and Glanz algorithm was introduced in this investigation. A constant value for N was used. For N = 200 most seeds had a sampling distance of about 2-3 pixels on the average. When small radii occurred, the sampling distance between consecutive points on the boundary could fall below the pixel length. Therefore, sampling of the next point was omitted until the equiangular boundary interval contained the next boundary pixel coordinate. No interpolation between pixel coordinates was used. This explains the shorter time series in Figure 4.2.6.

Algorithm used for some simple measurements

The time series representation of contour was used for the calculation of some characters, but in order to depress noise effects the time series was smoothed by a mean filter with distance 3. If r_t denotes the radius length at time t the

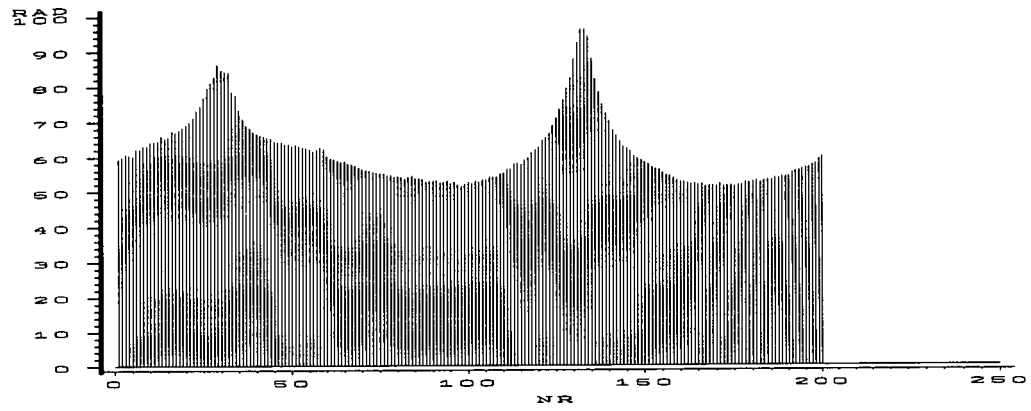


Figure 4.2.2 Time series representation of a seed of Rumex crispus

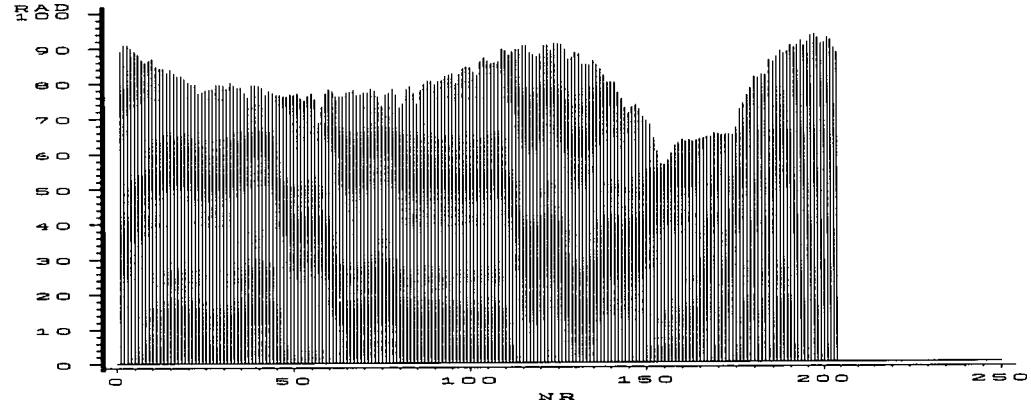


Figure 4.2.3 Time series representation of a seed of Silene vulgaris

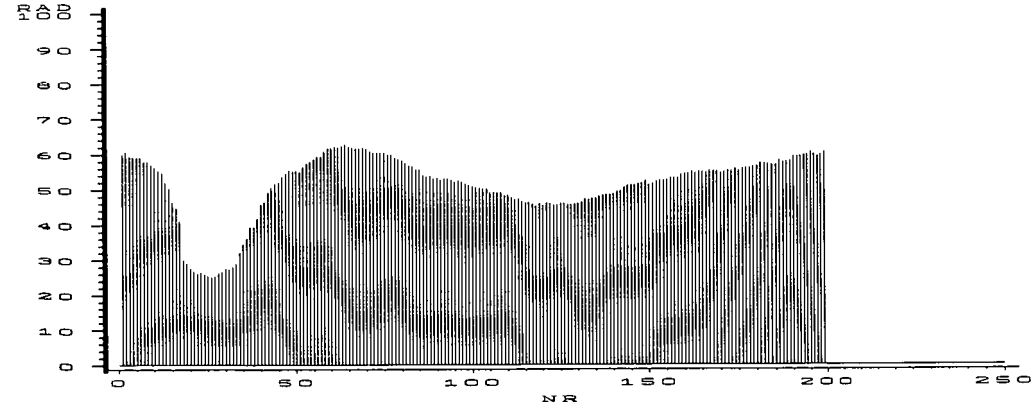


Figure 4.2.4 Time series representation of a seed of Papaver rhoeas

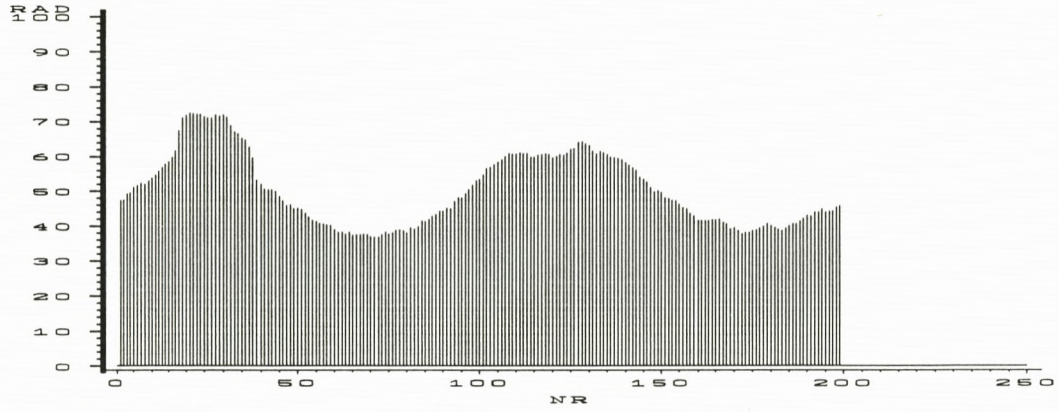


Figure 4.2.5 Time series representation of a seed of Euphorbia exigua

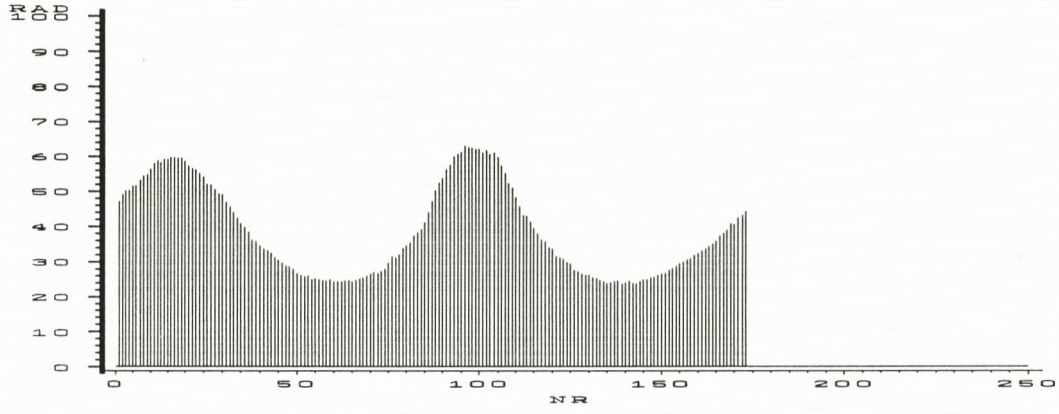


Figure 4.2.6 Time series representation of a seed of Matricia chamomilla

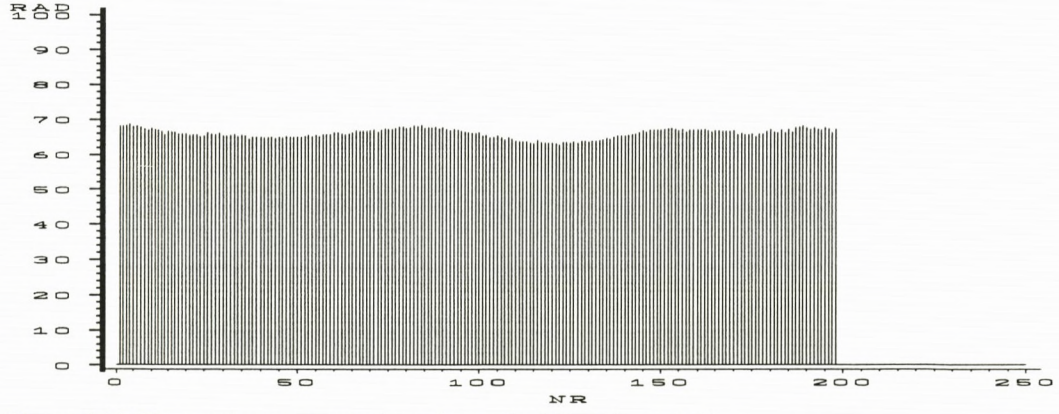


Figure 4.2.7 Time series representation of a seed of Sinapis arvensis

Table 4.2.1 Average and coefficient of variation (C.V.) of the area (in mm^2) and the compactness of each species (25 seeds per species).

	Average		Coefficient of Variation	
Species	**Area**	**Compact-ness**	**Area**	**Compact-ness**
Urtica urens	1.46	1.32	0.152	0.042
Polygonum convolvus	4.65	1.30	0.130	0.036
Polygonum lapathifolium	5.06	1.27	0.126	0.032
Rumex acetosa	1.83	1.46	0.108	0.035
Rumex crispus	2.51	1.40	0.106	0.042
Rumex obtusifolius	2.05	1.36	0.131	0.037
Rumex thyrsiflorus	1.20	1.31	0.124	0.049
Arenaria serpyllifolia	0.17	1.32	0.130	0.025
Melandrium album	1.22	1.42	0.113	0.044
Melandrium rubrum	1.33	1.90	0.101	0.084
Silene noctiflora	1.29	1.43	0.089	0.040
Silene vulgaris	1.38	1.67	0.123	0.109
Stellaria gramina	0.59	1.36	0.095	0.036
Stellaria media	0.75	1.42	0.087	0.035
Chenopodium album	1.31	1.17	0.152	0.013
Ranunculus repens	4.23	1.54	0.130	0.056
Papaver rhoeas	0.35	1.37	0.128	0.035
Brassica campestris	1.79	1.16	0.159	0.020
Capselle bursa-pastoris	0.47	1.37	0.062	0.038
Geranium dissectum	2.48	1.31	0.064	0.030
Euphorbia exigua	1.00	1.55	0.074	0.040
Euphorbia helioscopia	3.12	1.41	0.041	0.044
Euphorbia peplus	1.15	1.37	0.031	0.028
Viola arvensis	1.14	1.34	0.095	0.040
Myosotis arvensis	1.06	1.29	0.136	0.063
Lamium amplexicaule	1.38	1.78	0.105	0.059
Solanum nigrum	2.26	1.28	0.131	0.032
Veronica arvensis	0.44	1.28	0.175	0.029
Veronica persica	1.15	1.45	0.213	0.056
Plantago major	0.88	1.36	0.122	0.050
Cirsium arvense	2.94	1.98	0.096	0.105
Chrysanthemum segetum	2.75	1.75	0.176	0.068
Matricaria chamomilla	0.34	1.64	0.104	0.047
Matricaria inodora	1.28	1.67	0.245	0.122
Matricaria matricarioides	0.49	1.83	0.132	0.045
Sinapis arvensis	1.58	1.14	0.137	0.009
Sonchus arvensis	2.59	1.87	0.166	0.088
Sonchus asper	1.85	1.94	0.091	0.082
Sonchus oleraceus	1.87	2.26	0.082	0.094
Taraxacum vulgare	2.43	3.46	0.116	0.127

Table 4.2.2 Average of eccentricity, sharpness, depth, maxima, high maxima, medium maxima and coefficient of variation (C.V.) (25 seeds per species).

Species	Eccen-tricity	Sharp-ness	Depth	Maxi-ma	High max.	Medium max.	C.V.
Urtica urens	0.42	0.43	9.2	9.6	0.11	0.28	0.15
Polygonum convolvus	0.37	0.38	7.3	11.3	0.11	0.18	0.13
Polygonum lapathifolium	0.26	0.30	2.3	23.7	0.05	0.08	0.06
Rumex acetosa	0.54	0.57	20.4	5.9	0.37	0.10	0.24
Rumex crispus	0.42	0.40	7.9	12.7	0.15	0.06	0.13
Rumex obtusifolius	0.39	0.36	5.6	15.8	0.13	0.06	0.12
Rumex thyrsiflorus	0.41	0.42	9.6	10.0	0.14	0.16	0.15
Arenaria serpyllifolia	0.36	0.45	3.2	23.3	0.08	0.27	0.08
Melandrium album	0.29	0.46	2.6	27.2	0.11	0.25	0.09
Melandrium rubrum	0.26	0.71	2.6	32.9	0.29	0.34	0.07
Silene noctiflora	0.29	0.45	2.5	27.4	0.10	0.24	0.08
Silene vulgaris	0.32	0.67	3.0	28.4	0.26	0.32	0.09
Stellaria gramina	0.21	0.45	2.8	23.7	0.06	0.30	0.06
Stellaria media	0.19	0.61	2.8	25.9	0.19	0.38	0.04
Chenopodium album	0.16	0.20	1.6	25.5	0.01	0.04	0.04
Ranunculus repens	0.45	0.41	4.9	18.6	0.10	0.14	0.13
Papaver rhoeas	0.54	0.47	7.7	14.2	0.10	0.20	0.16
Brassica campestris	0.19	0.26	2.9	19.6	0	0.10	0.06
Capsella bursa-pastoris	0.48	0.57	16.8	6.6	0.29	0.20	0.21
Geranium dissectum	0.31	0.43	6.0	14.0	0.07	0.32	0.11
Euphorbia exigua	0.53	0.81	11.1	11.2	0.44	0.19	0.22
Euphorbia helioscopia	0.31	0.59	3.8	22.2	0.20	0.32	0.11
Euphorbia peplus	0.47	0.49	11.5	10.1	0.21	0.16	0.19
Viola arvensis	0.48	0.50	15.1	7.2	0.24	0.17	0.21
Myosotis arvensis	0.40	0.43	11.4	8.6	0.13	0.23	0.15
Lamium amplexicaule	0.66	0.70	16.3	8.7	0.31	0.28	0.32
Solanum nigrum	0.35	0.36	4.6	15.6	0.04	0.19	0.11
Veronica arvensis	0.39	0.43	10.4	9.1	0.10	0.26	0.16
Veronica persica	0.41	0.81	6.7	17.3	0.38	0.27	0.15
Plantago major	0.46	0.57	13.5	8.8	0.28	0.18	0.19
Cirsium arvense	0.72	1.21	32.2	5.5	0.59	0.09	0.43
Chrysanthemum segetum	0.66	0.89	21.3	7.6	0.44	0.17	0.36
Matricaria chamomilla	0.63	0.93	21.6	7.0	0.49	0.22	0.33
Matricaria inodora	0.60	0.71	14.3	9.7	0.30	0.24	0.32
Matricaria matricarioides	0.68	0.83	21.1	7.0	0.40	0.26	0.38
Sinapis arvensis	0.10	0.20	1.0	32.8	0	0.02	0.03
Sonchus arvensis	0.68	0.80	21.3	6.9	0.42	0.16	0.39
Sonchus asper	0.67	0.81	26.4	6.4	0.47	0.11	0.37
Sonchus oleraceus	0.70	0.96	18.7	9.0	0.47	0.22	0.40
Taraxacum vulgare	0.79	1.84	22.7	9.3	0.66	0.15	0.51

Table 4.2.3 Coefficient of variation of eccentricity, sharpness, depth, maxima, high maxima, medium maxima and coefficient of variation (C.V.) (25 seeds per species).

Species	Eccen-tricity	Sharp-ness	Depth	Maxi-ma	High max.	Medium max.	C.V.
Urtica urens	0.09	0.21	0.39	0.23	0.59	0.38	0.13
Polygonum convolvus	0.09	0.23	0.46	0.29	0.86	0.57	0.14
Polygonum lapathifolium	0.13	0.13	0.29	0.20	0.38	0.52	0.16
Rumex acetosa	0.05	0.19	0.29	0.24	0.38	1.08	0.10
Rumex crispus	0.09	0.18	0.44	0.31	0.45	1.24	0.17
Rumex obtusifolius	0.09	0.16	0.28	0.23	0.35	1.66	0.11
Rumex thyrsiflorus	0.11	0.18	0.42	0.28	0.62	0.98	0.19
Arenaria serpyllifolia	0.13	0.11	0.23	0.16	0.58	0.40	0.14
Melandrium album	0.11	0.13	0.23	0.13	0.62	0.29	0.12
Melandrium rubrum	0.11	0.20	0.16	0.09	0.40	0.28	0.14
Silene noctiflora	0.11	0.17	0.20	0.11	0.57	0.39	0.14
Silene vulgaris	0.16	0.17	0.16	0.12	0.43	0.30	0.17
Stellaria gramina	0.20	0.12	0.23	0.15	0.65	0.39	0.25
Stellaria media	0.16	0.12	0.12	0.07	0.47	0.28	0.17
Chenopodium album	0.22	0.13	0.31	0.18	1.88	0.60	0.21
Ranunculus repens	0.09	0.22	0.28	0.20	0.55	0.71	0.16
Papaver rhoeas	0.09	0.29	0.43	0.25	0.53	0.62	0.16
Brassica campestris	0.30	0.22	0.54	0.36	.	1.09	0.39
Capsella bursa-pastoris	0.04	0.21	0.31	0.28	0.53	0.76	0.06
Geranium dissectum	0.14	0.17	0.37	0.22	1.25	0.33	0.15
Euphorbia exigua	0.03	0.16	0.19	0.20	0.25	0.49	0.06
Euphorbia helioscopia	0.11	0.15	0.19	0.13	0.36	0.22	0.12
Euphorbia peplus	0.04	0.18	0.39	0.27	0.36	0.78	0.05
Viola arvensis	0.06	0.23	0.31	0.28	0.37	0.68	0.07
Myosotis arvensis	0.19	0.40	0.66	0.26	1.06	0.64	0.28
Lamium amplexicaule	0.05	0.17	0.22	0.26	0.29	0.56	0.08
Solanum nigrum	0.19	0.17	0.31	0.22	0.81	0.54	0.21
Veronica arvensis	0.14	0.25	0.43	0.27	1.34	0.56	0.19
Veronica persica	0.17	0.22	0.24	0.12	0.32	0.41	0.23
Plantago major	0.13	0.21	0.44	0.38	0.49	0.73	0.20
Cirsium arvense	0.04	0.26	0.41	0.36	0.31	1.18	0.08
Chrysanthemum segetum	0.06	0.33	0.38	0.30	0.40	0.85	0.12
Matricaria chamomilla	0.05	0.17	0.30	0.27	0.30	0.64	0.09
Matricaria inodora	0.13	0.20	0.26	0.19	0.38	0.51	0.21
Matricaria matricarioides	0.04	0.15	0.28	0.23	0.36	0.73	0.07
Sinapis arvensis	0.36	0.14	0.61	0.27	.	1.31	0.44
Sonchus arvensis	0.06	0.22	0.26	0.27	0.32	0.96	0.12
Sonchus asper	0.06	0.26	0.45	0.35	0.39	0.98	0.10
Sonchus oleraceus	0.02	0.25	0.26	0.28	0.32	0.64	0.04
Taraxacum vulgare	0.03	0.23	0.31	0.36	0.19	0.74	0.08

smoothed value is

$$R_t = (r_{t-1}+r_t+r_{t+1})/3$$

Then the following measurements were calculated:

1. Eccentricity is calculated as the relative difference between the global maximum (A) and the global minimum (B)

$$\frac{|A - B|}{A}$$

2. Sharpness. The average slope, which is the difference between a local minimum and the subsequent local maximum divided by the distance.
3. Depth. The average difference between a local minimum and the next local maximum.
4. Maxima. The number of local maxima in the time series.
5. High maxima. The number of local maxima with a slope above 0.9 divided by maxima (*i.e.*, relative to the total number of local maxima).
6. Medium maxima. The number of local maxima with a slope between 0.5 and 0.9 divided by maxima.
7. Coefficient of variation (C.V.).

Results

The measurements of area and compactness are presented in Table 4.2.1 as an average of 25 seeds per species together with the corresponding coefficients of variation within species. The shape features based on the time series representation are shown in Table 4.2.2 and the corresponding coefficients of variation within species are shown in Table 4.2.3.

The average size of the seeds used in this study varied between 5.06 mm^2 for *P. lapathifolium* and 0.17 mm^2 for *A. serpyllifolia*. The coefficients of variation of size within species were between 0.031 for *E. peplus* and 0.245 for *M. inodora*. The relatively large range in area and low values of coefficient of variation indicate, that size will be an important feature for recognition of the seeds.

The five shape features 1) compactness, 2) eccentricity, 3) sharpness, 4) high maxima, and 5) C.V. showed extreme values for *Sinapis arvensis* and *Taraxacum vulgare*, which were the most circular and the most linear/oblong shape among the selected species, respectively. The range in compactness was between 1.14 and 3.46, which was relatively lower than for size, but the coefficient of variation within species was very small, and ,therefore, this feature looks promising for recognition. The eccentricities were between 0.096 and 0.79 with a reasonably small deviation within species, which should be fairly good for recognition. Sharpness was in a narrow range (0.18 - 1.84), and showed only medium values of the relative deviation. Depth showed a high range (0.87 - 32.19), but unfortunately also very high deviations. The distribution of maxima were in general not very promising, mainly because of the high coefficients of variation. Finally, the C.V. ranged from 0.027 to 0.508 with relative small deviations, and this makes it a good candidate for recognition.

The values of the most promising features were describing the macroscopic shape, meaning that the minor structures on the contour were not very well expressed. This is not satisfactory, because the information in these structures is regarded essential for seed identification.

4.3 Moment invariants

The moment invariants is a standard analysis of shape where some features invariant for rotation, size and translation are calculated. In general, a moment is the mean value of a power of a variate, and moment ratios are often used to characterize the shape of a distribution, *e.g.*, skewness and kurtosis. However, the moment invariants do not have the same simple shape interpretation as skewness and kurtosis.

Theory

The definition of the two-dimensional $(p + q)$'th order moment is

$$m_{pq} = \sum_i \sum_j i^p j^q I(i,j) \quad p,q=0,1,2,...$$

where i,j are the image coordinates and $I(i,j)$ is the grey level value. For shape analysis $I(i,j)$ is replaced by the binary representation of the image, *i.e.*, the full silhouette. From this the area (A) is given by m_{00}, and the center of gravity (c_i, c_j) is calculated by

$$c_i = \frac{m_{10}}{m_{00}}, \quad c_j = \frac{m_{01}}{m_{00}}$$

Normalization for translation is then achieved by centering the coordinates (central moments):

$$\mu_{pq} = \sum_i \sum_j (i-c_i)^p (j-c_j)^q \quad p,q=0,1,2,...$$

Normalization for size is achieved when dividing by (Hu, 1961, 1962)

$$m_{00}^{\frac{p+q}{2}+1}$$

Hu (1961, 1962) proposed seven moment invariants for rotation and translation using second and third order central moments

$$M_1 = \mu_{20}+\mu_{02}$$
$$M_2 = (\mu_{20}-\mu_{02})^2+4\mu_{11}^2$$
$$M_3 = (\mu_{30}-3\mu_{12})^2+(3\mu_{21}-\mu_{03})^2$$
$$M_4 = (\mu_{30}+\mu_{12})^2+(\mu_{21}+\mu_{03})^2$$
$$M_5 = (\mu_{30}-3\mu_{12})(\mu_{30}+\mu_{12})[(\mu_{30}+\mu_{12})^2 - 3(\mu_{12}+\mu_{03})^2] + (3\mu_{21}-\mu_{03})(\mu_{21}+\mu_{03})[3(\mu_{30}+\mu_{12})^2 - (\mu_{21}+\mu_{03})^2]$$
$$M_6 = (\mu_{20}-\mu_{02}[(\mu_{30}+\mu_{12})^2-(\mu_{21}+\mu_{03})^2] + 4\mu_{11}(\mu_{30}+\mu_{12})(\mu_{21}+\mu_{03})$$
$$M_7 = (3\mu_{21}-\mu_{03})(\mu_{30}+\mu_{12})[(\mu_{30}+\mu_{12})^2 - 3(\mu_{21}+\mu_{03})^2] - (\mu_{30}-3\mu_{12})(\mu_{21}+\mu_{03})[3(\mu_{30}+\mu_{12})^2 - (\mu_{21}+\mu_{03})^2]$$

For scale normalization the central moments should be replaced by the normalized central moments, *i.e.*, $\eta_{pg} = \mu_{pq}/m_{00}^{1/2(p+q)+1}$, in the equations above.

However, this scale normalization is not appropriate when using the boundary instead of the silhouette or the grey level image as input. Dudani *et al.* (1977) used invariant moments derived from both the silhouette and the boundary for aircraft identification. They proposed the size normalization to be based on the radius defined as

$$r = \sqrt{\mu_{20}+\mu_{02}}$$

This radius is directly proportional to the size of the image. The new invariant moments were then

$$M_1' = (\mu_{20}+\mu_{02})^{\frac{1}{2}} \times B = r \times B$$
$$M_2' = M_2/r^4$$
$$M_3' = M_3/r^6$$
$$M_4' = M_4/r^6 M_5' = M_5/r^{12}$$
$$M_6' = M_6/r^8$$
$$M_7' = M_7/r^{12}$$

where B is the distance along the optical axis. The advantage of using moments derived from the boundary is that they contain more information about the high frequency portions of the object than those derived from the silhouette. On the other hand, the low frequency portions are better represented by moments derived from the silhouette (Dudani *et al.*, 1977).

Table 4.3.1 Moment invariants derived from boundaries of eight images of a Stellaria media seed at different sizes.

Area	M_1'	M_2'	M_3'	M_4'	M_5'	M_6'	M_7'
29050	-5.01	-1.73	-11.36	-13.65	-26.27	-15.03	-26.83
24388	-4.94	-1.78	-11.19	-13.46	-27.69	-17.36	-26.10
18430	-4.88	-1.74	-11.22	-13.39	-28.24	-16.60	-26.03
14868	-4.82	-1.74	-10.90	-12.05	-23.97	-13.47	-24.95
12027	-4.76	-1.73	-11.02	-13.61	-25.96	-14.63	-28.14
9257	-4.70	-1.77	-10.56	-12.42	-25.60	-16.56	-24.17
7724	-4.64	-1.72	-10.91	-14.17	-32.05	-17.35	-27.03
6773	-4.57	-1.77	-11.24	-14.41	-27.27	-15.88	-28.36

1977).

Although theoretically invariant, the moments have been shown to vary when applied to images of the same scene in different magnifications and rotations. Wong and Hall (1978) used invariant moments to match radar to optical images using a cross-correlation technique. The calculation of the moment invariants for a set of grey level images with different size and rotation showed standard deviations up to 7.5 % of the mean (in logaritmic scale). Significant variations of the invariant moments due to scale and rotation were also found by Hsia (1981).

This matching by the moment invariants was improved by Goshtasby (1985). The normalized invariant moments from two circular windows were used to determine the similarity (cross-correlation) between the windows. A two stage matching technique was developed where first the zero-order moment was used to determine the likely positions in the image for a match with a template. Then, higher order moments were used to determine the best match among the likely ones. The limiting of the search area in the first stage resulted in a speed-up by factors 13.1 and 11.5 in two experiments with satellite images.

Working with matching of grey level images it is important to consider differences in contrast (k) between images ($f(i,j)$). For the relation between two images characterized by

$$f_1(i,j) = kf_2(i'j')$$

the normalized central moments are not invariant. Maitra (1979) proposed six new functions which were invariant under scale, translation, rotation as well as contrast change.

Algorithm for moment invariants

To make a fast analysis only the boundary was used as input to the program. A new boundary was constructed by sampling by a distance of one horizontal pixel length. Then the number of sampling points is equal to the perimeter length when measured in the horizontal pixel length. The seven invariant moments proposed by Hu(1961, 1962) were calculated using the size normalization from Dudani *et al.* (1977), except for the first moment which was divided by $perimeter^2$ for size normalization. Finally, the natural logarithm were taken for narrowing the range of values.

Test for size invariance

A small series of images of the same seed captured at different degrees of magnification were used to test for size invariance. The seed used were *Stellaria media*, and the results are presented in Table 4.3.1. The highest range of values is found for M_5', and in general the last three moment invariants showed greater vari-

ation than the first. Although M_1' showed relatively high stability there was a clear trend of falling values from the big to the small seed size. No such trend was present in the other moment invariants, indicating that the size normalization for M_1'is not entirely adequate.

4.4 Time series analysis

The time series representation of the contour was used in an earlier section for calculation of simple measurements. Now, a linear modeling of the time series will be used to obtain new features to describe the series.

Theory

Basically a time series is a set of data $\{y_t \mid t = 1,...,n\}$ in which t indicates the time for the observation y_t (Diggle, 1990). A time series is called stationary if the probability structure (*i.e.*, the joint distribution) of the variables $\{Y_t\}$ is unaffected by a shift in the time origin. Thus, a first order stationary process is characterized by having no trend, *i.e.*, $E[Y_t]$ is constant, and a second order stationary process shows no trend and the autocovariance function of $\{Y_t\}$ depends only on the differences between observation times. Often the second order stationarity is used instead of the strict stationary assumption.

An important class of linear stationary models is called autoregressive-moving average (ARMA) processes. This is a mixture of two kinds of models where the autoregressive process is defined in terms of its predecessors

$$Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-i} + Z_t$$

where Z_t is white noise, *i.e.*, independent random variables with mean zero and common variance. The process is abbreviated by AR(p), where p is the order of the process.

The moving average model is a linear filter applied to $\{Z_t\}$:

$$Y_t = Z_t + \sum_{j=1}^{q} \beta_j Z_{t-j}$$

This is abbreviated MA(q) for a process of order q. Thus, the mixed ARMA(p,q) model is defined by the equation

$$Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-i} + Z_t + \sum_{j=1}^{q} \beta_j Z_{t-j}$$

For nonstationary time series the appearence of trend or seasonal variation will dominate all other features. To remove the nonstationary elements, the differencing technique is often used. If a seasonal fluctuation appears with a period s, the time series should be differenced by subtracting the observations from time t-s,

$$D_s y_t = y_t - y_{t-s}$$

When the differencing technique is integrated into the ARMA process, the process is called autoregressive integrated moving average (ARIMA), which is an extension to non-stationary processes.

A valuable tool to characterize and identify the ARIMA process is the correlogram of the time series. This is a graph of the sample autocorrelation coefficient r_k versus the time lag k, each r_k defined as

$$r_k = g_k / g_0$$

where

$$g_k = \sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y}) / n$$

For white noise r_k is approximately distributed as N(0,1/n), and, therefore, the limits $\pm 2/\sqrt{n}$ are used to assess r_k for significant departure from zero.

The partial autocorrelation is the extra increase in the autocorrelation by extending the

model order of the time series. The partial correlogram is constructed from partial autocorrelations versus the time lag.

The correlogram of an autoregressive process is characterized by a smooth curve (like damped sine waves or exponentials), whereas the partial correlogram shows a cut-off after some lag p above which no significant dependency between variables appear. The lag p determines the highest relevant model order of an autoregressive process. For moving average processes the charateristics of these correlograms are opposite.

The periodogram may be used alternatively to the correlogram for detecting a seasonal fluctuation pattern of the time series. A Fourier transform (see next section) is used to regard the time series in the frequency domain. In this domain the frequency of the p'th harmonic (sine) wave is

$$\omega_p = 2\pi p/n$$

and the corresponding amplitude is denoted R_p. Then the periodogram is a plot of

$$I(\omega_p) = nR_p^2/4\pi$$

against the frequency (or alternatively the period). A high amplitude at a low frequency indicates seasonal variation in the time domain.

Previous work

Dubois and Glanz (1986) constructed the time series representation using the equiangular sampling method of radius vectors already described. The number of vectors used was 64, and the classification rates from different model orders showed that more complicated shapes required a higher order for accurate representation. However, shapes were successfully classified for model orders lower than the optimum (*i.e.*, giving highest classification results), and for this reason a determination of model order was regarded as unnecessary.

Kartikeyan and Sarkar (1989) represent the contour by equal distance points, and the distances from the centroid to the contour points yield the time series. It was mentioned that a sampling interval as small as the pixel dimension will cause high errors due to limited spatial resolution. Sampling intervals above 2-3 units were recommended. They removed trend and periodic components detected by spectral techniques. Then the order of the AR model was obtained by determining the highest lag for which the partial autocorrelation coefficient was significant. This model was then reduced by selecting a subset of the variables.

The linear AR model was compared with the non-linear model:

$$Y_t = \sum_{i \in S} \beta_i Y_{t-i} + \sum_{(u,v) \in G} g_{uv} Y_{t-u} Y_{t-v} + e_t$$

where the significant linear term S was identified from the partial correlations, and from the residual of the linear series the set G was estimated (see for details Kartikeyan and Sarkar, 1989).

The shape of an aircraft was used as a prototype, and minor changes were imposed to make five test shapes. Only the feature vectors from the nonlinear model showed significant differences when compared with the vector of the prototype. It was therefore concluded, that linear models might be insufficient for shapes with differentiating features at detailed levels. In such situations overfitted AR models might also result in poor recognition performance.

A bivariate model was presented by Das *et al.* (1990). If the coordinates of the contour are termed $\mathbf{x}_i$, $i = 1,...,n$, and the process mean vector is $\boldsymbol{\alpha}$, the centering is

$$\mathbf{y}_i = \mathbf{x}_i - \boldsymbol{\alpha}$$

and the bivariate circular autoregressive model was

$$\mathbf{y}_i = \sum_{j=1}^{p} \mathbf{A}_j \mathbf{y}_{i-j} + \sqrt{\mathbf{B}}\mathbf{v}_i$$

with $\mathbf{A}_i$ as the 2×2 coefficient matrices, **B** as the covariance matrix, and $\sqrt{\mathbf{B}}\,\mathbf{v}_i$ as the mean zero white noise.

From this model a set of classification features was developed. The feature vectors were twice as large as their 1-D counterparts. The classification results were compared to a equiangular sampling method and a equal-curve-length sampling method using AR models of orders one to four. The results showed a superiority of the bivariate model even at significantly lower model orders.

Eom and Park (1990) represented the contour by a centriodal profile, where points were sampled with equal distance. The model fitted was the circular autoregressive, and a maximum likelihood method was suggested as classification method. The results were compared with the results of Dubois and Glanz (1986). It was claimed that the poorer results from the latter work could be explained by sampling bias. The method of Dubois and Glanz will undersample peaks far from the centroid, and oversample where the contour is close to the centroid.

However, it can be expected that this problem will increase by decreasing the number of sampling points, and that a high number of sampling points may reduce this bias.

Algorithm for time series analysis

For estimation of parameters the time series is regarded circular with the period N, so that $y_t = y_{t+N}$. The model used was the autoregressive type cited in Dubois and Glanz (1986).

$$y_t = \alpha + \sum_{i=1}^{p} \theta_i y_{t-i} + \sqrt{\beta}\, w_t$$

where

y_t = *current radius length*
α = *constant to be estimated*
θ_i = *coefficients to be estimated*
β = *residual variance to be estimated*
$\sqrt{\beta}\, w_t$ = *error term*

The parameters were estimated by the method of least squares. Using the notation

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ . \\ . \\ y_N \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} y_{1-1} & . & . & y_{1-p} & 1 \\ . & . & . & . & . \\ . & . & . & . & . \\ y_{N-1} & . & . & y_{N-p} & 1 \end{bmatrix}$$

the solution is

$$\begin{bmatrix} \hat{\theta}_1 \\ . \\ . \\ \hat{\theta}_p \\ \hat{\alpha} \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

and the residual variance is estimated by

$$\hat{\beta} = \frac{1}{N}\sum_{t=1}^{N} \left(y_t - \hat{\alpha} - \sum_{i=1}^{p} \hat{\theta}_i y_{t-i} \right)^2$$

The parameter α is proportional to the mean radius vector length, and to normalize it for size it was divided by $\sqrt{\beta}$.

Experimental results

A small experiment was performed to get an idea of how to describe the seeds satisfactorily by AR models. Different types of seeds exist, some with a complicated (structured) surface and some with a smooth surface. A seed of each type was selected, and the time series obtained was plotted. These plots are shown in Figure 4.4.1 for the *Silene vulgaris* seed, the complicated seed shape, and in Figure 4.4.4 for the *Rumex obtusifolius* seed, the smooth shape. The time series show no trend because they are circular, but a dominating frequency corresponding to a half period was visible in the periodograms in Figure 4.4.7 and 4.4.10, respectively. To remove this the time series were differenced by this period, and the resulting series are presented in Figure 4.4.2 and 4.4.5. This was repeated, but now for one third of the

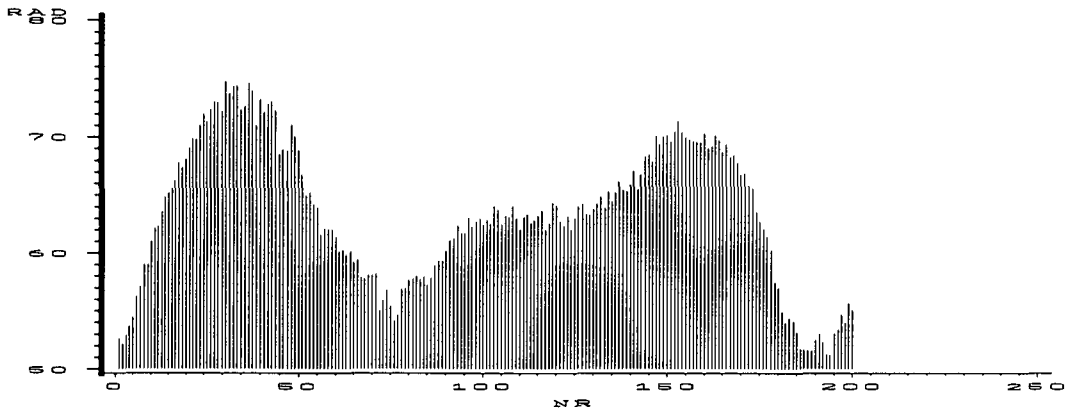


Figure 4.4.1 Radius vectors of a Silene vulgaris seed

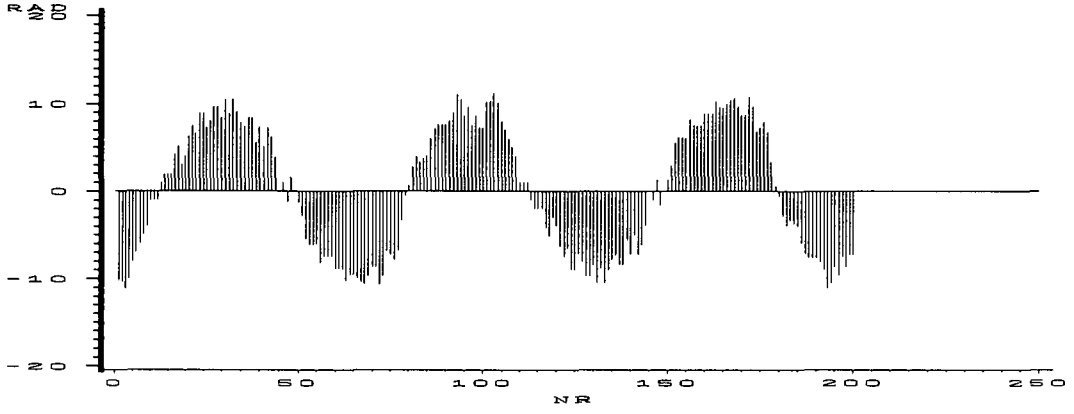


Figure 4.4.2 Differencing with lag equal to period/2 of time series in Figure 4.4.1.

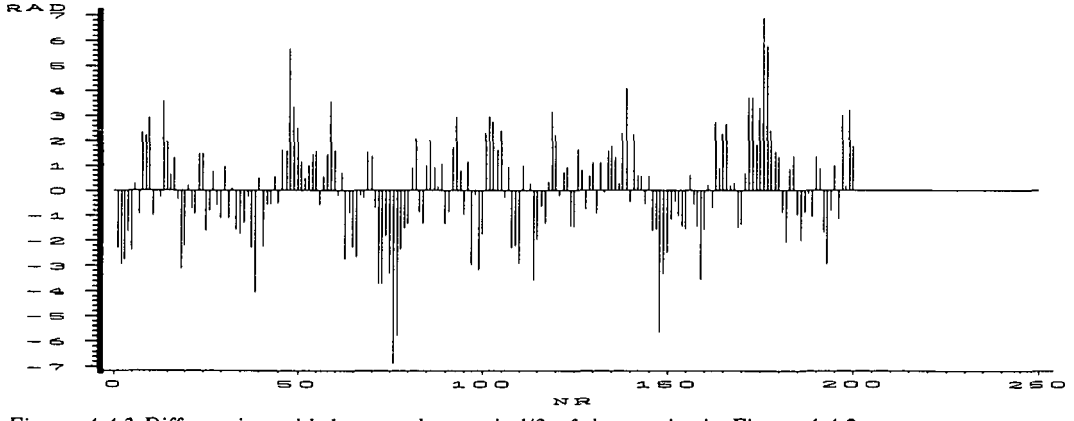


Figure 4.4.3 Differencing with lag equal to period/3 of time series in Figure 4.4.2.

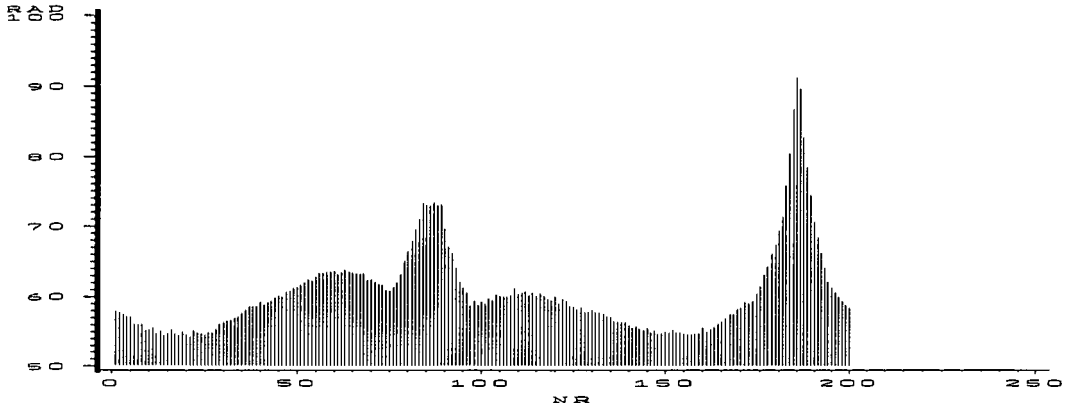


Figure 4.4.4 Radius vectors of a Rumex obtusifolius seed.

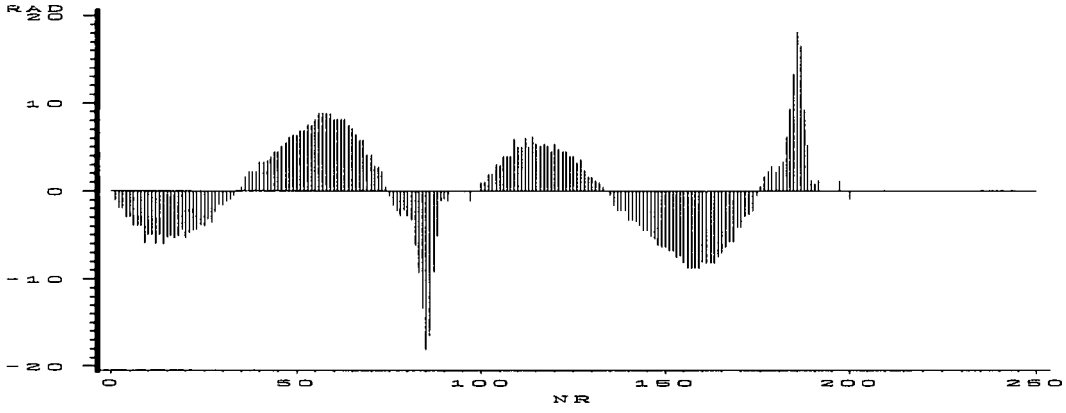


Figure 4.4.5 Differencing with lag equal to period/2 of time series in Figure 4.4.4

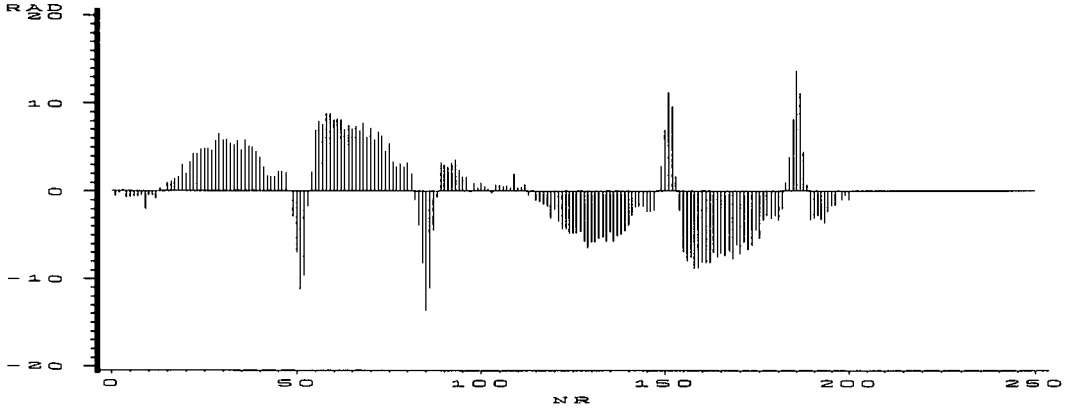


Figure 4.4.6 Differencing with lag equal to period/3 of time series in Figure 4.4.5.

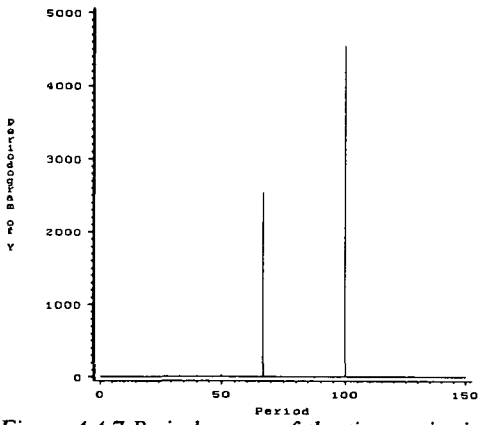


Figure 4.4.7 Periodogram of the time series in Figure 4.4.1.

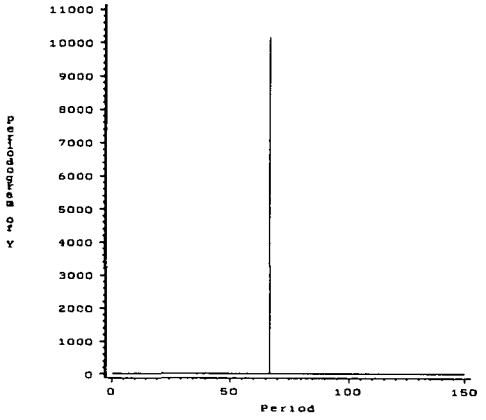


Figure 4.4.8 Periodogram of the time series in Figure 4.4.2.

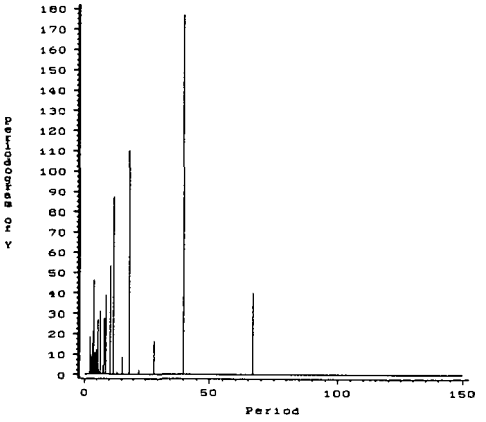


Figure 4.4.9 Periodogram of the time series in Figure 4.4.3.

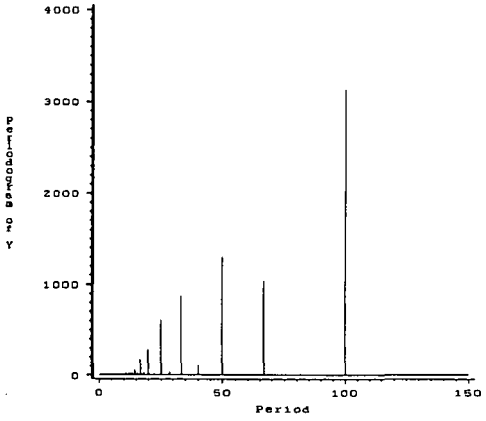


Figure 4.4.10 Periodogram of the time series in Figure 4.4.4.

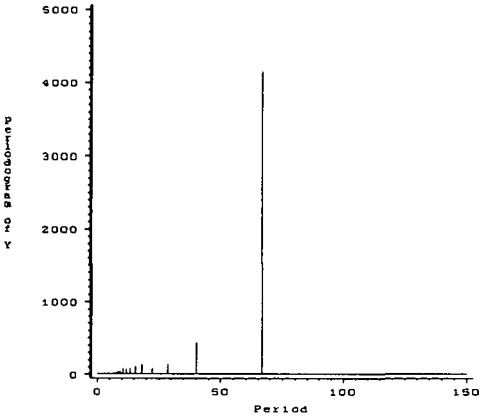


Figure 4.4.11 Periodogram of the time series in Figure 4.4.6.

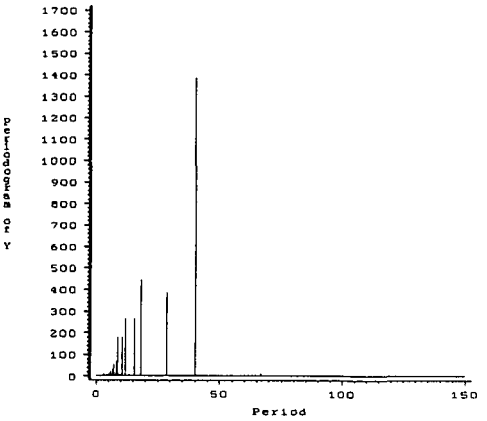


Figure 4.4.12 Periodogram of the time series in Figure 4.4.6.

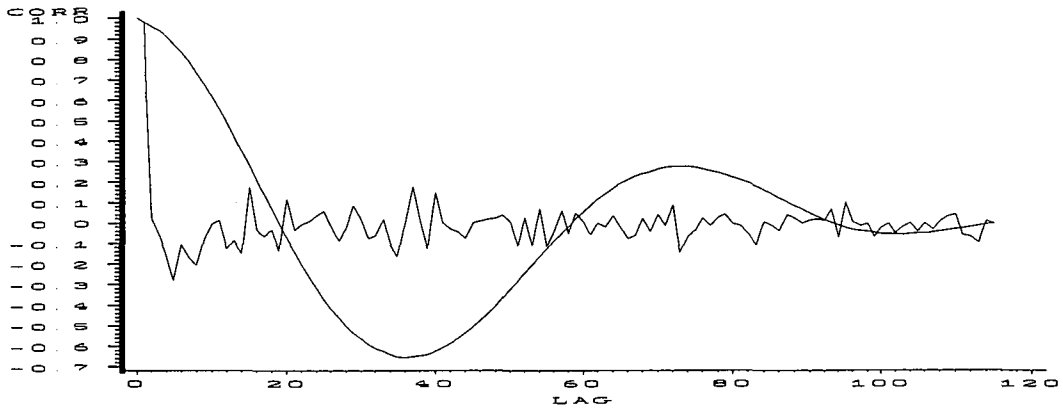


Figure 4.4.13 Autocorrelation and partial autocorrelation for the time series in Figure 4.4.1.

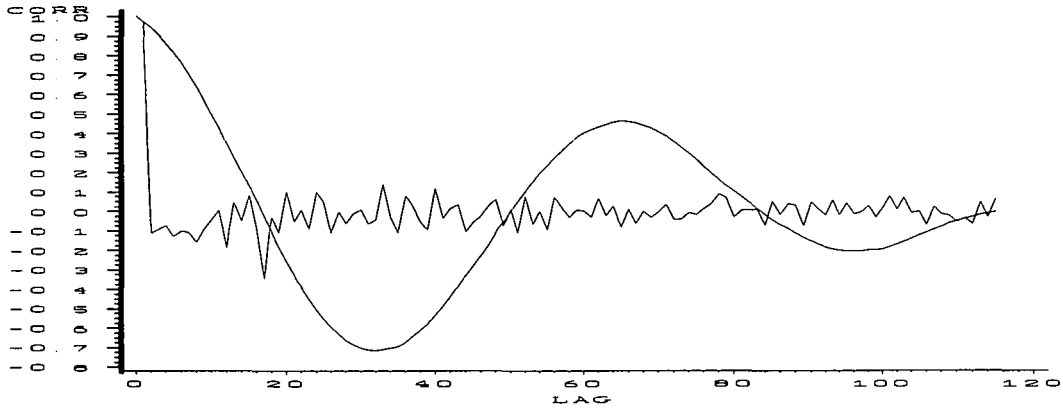


Figure 4.4.14 Autocorrelation and partial autocorrelation of the time series in Figure 4.4.2.

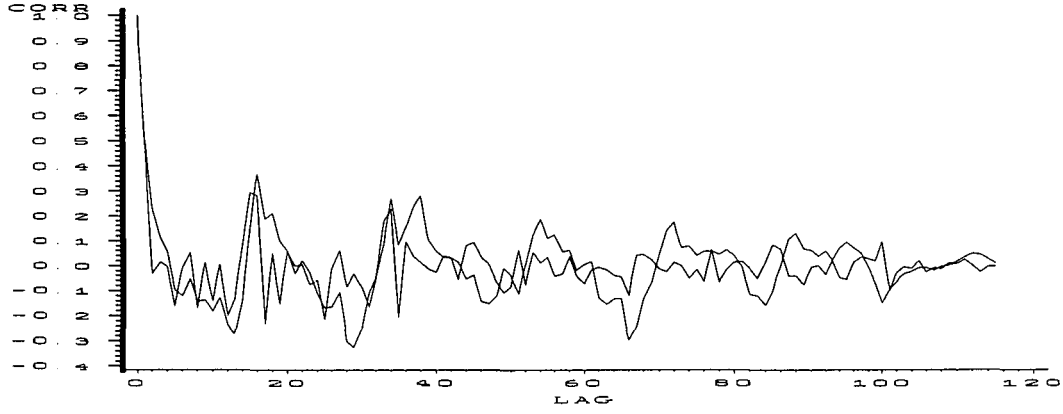


Figure 4.4.15 Autocorrelation and partial autocorrelation of the time series in Figure 4.4.3.

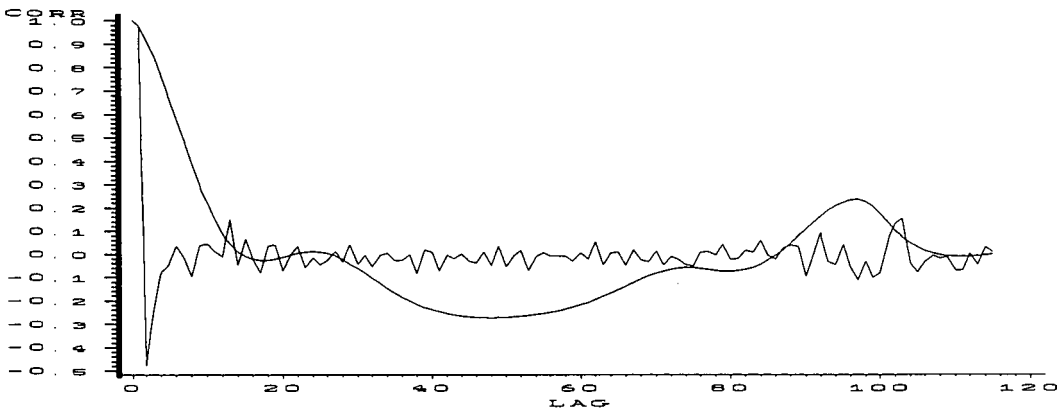


Figure 4.4.16 Autocorrelation and partial autocorrelation of the time series in Figure 4.4.4.

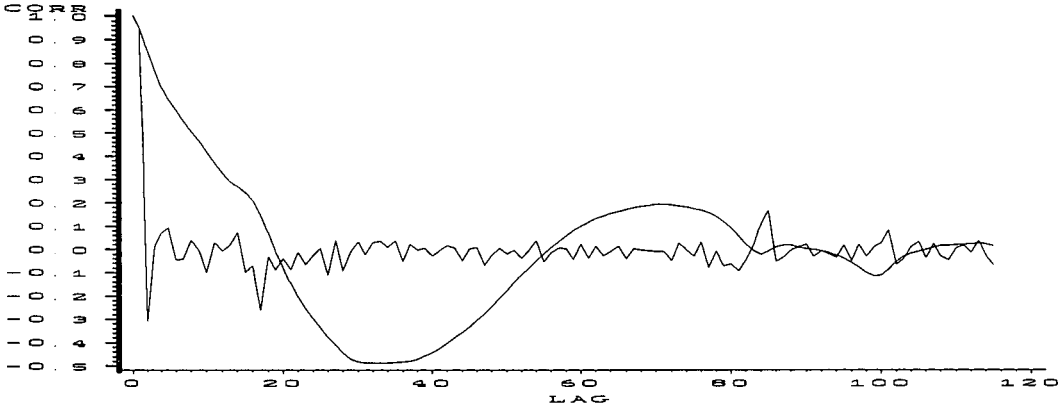


Figure 4.4.17 Autocorrelation and partial autocorrelation of the time series in Figure 4.4.5

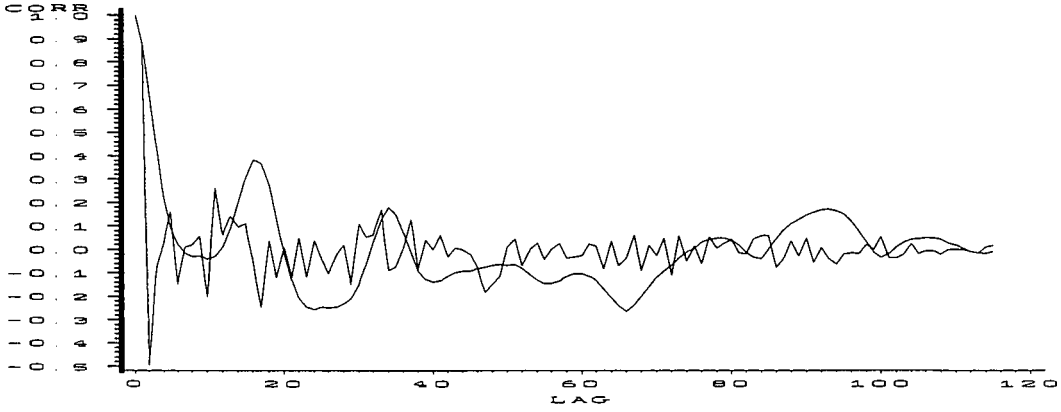


Figure 4.4.17 Autocorrelation and partial autocorrelation of the time series in Figure 4.4.6

Table 4.4.1 Classification rates from AR(p) at increasing model orders p, based on the original time series alone.

Model order	Caryophyllaceae	Rumex	All species
1	50.8	66.3	57.0
2	60.8	62.5	61.5
3	60.8	58.8	60.0
4	65.0	51.3	59.5
5	65.8	47.5	58.5
6	69.2	50.0	61.5
7	73.3	53.8	65.5
8	75.8	48.8	65.0
9	70.8	47.5	61.5

total period (Figure 4.4.3 and 4.4.6). The resulting series are clearly different for the two seeds. Apparently, the high frequency components in the new time series are more pronounced for the *S.vulgaris* seed, thus providing more information on the detailed level of the shape.

The corresponding autocorrelation and partial autocorrelation functions are shown in Figure 4.4.13-18. The autocorrelation function is for all series the slowly damped curve, which indicates that the series might be well modeled by an autoregressive process. For determination of the model order the partial autocorrelations are interesting. Values outside the limits ± 0.14 are significantly different from zero, and for the original (undifferenced) series such values occur for time lag 1, 5 and 8 for *S.vulgaris*, but only for lag 1 and 2 for *R.obtusifolius*.

From this two different approaches were possible for modeling the seeds. One approach was to fit the series to a model of a certain (high) order, and in the second approach differencing was included in combination with a low order model. For testing the two methods a special series of images consisting of six species of *Caryophyllaceae (M.album, M.rubrum, S. noctiflora, S.vulgaris, S.gramina* and *S.media)* which was acquired with focus on the contour, plus four species of *Rumex (R.acetosa, R.crispus, R. obtusifolius* and *R.thyrsiflorus)* with

Table 4.4.2 Classification rates from AR(2) parameters of original time series alone and combined with parameters from differenced time series

Time series	*Caryophyllaceae*	*Rumex*	All species
Original	60.8	62.5	61.5
1. differencing	62.5	60.0	61.5
2. differencing	72.5	65.0	69.5

normal focus. Twenty seeds per species were used. The classification rates (as percentage correct identification) shown in Table 4.4.1 and 4.4.2 were obtained from linear discriminant analysis using the cross-validation technique.

In Table 4.4.1 the classification rates were obtained for model orders one to nine. For seeds of *Carryophyllaceae* the rates increased up to model order eight, while the rates for seeds of *Rumex*, in general, decreased for increasing orders, thus reflecting a heavily overfitted model. The classification rate for all ten species was highest at model order seven. This approach of using high model orders only favoured the complicated shapes, and, therefore, is not satisfactory as a general method. The second approach was to use a model of order two, and combined with the differencing technique to provide information about the detailed level of the shapes. This means that the classification rates from the original time series were based on three parameters, *i.e.*, θ_1, θ_2, and $\alpha/\sqrt{\beta}$, and after the first differencing these parameters were combined with the two obtained from fitting an AR(2) model to the differenced series (where the constant α is zero). After the second differencing two parameters more were added, thus giving seven parameters in total. The results presented in Table 4.4.2 show that this method increased the rates of the species of *Caryophylaceae* without decreasing the rates of the *Rumex* species. In fact, the classification rate increased slightly for the *Rumex* species too, and the overall rate were above the best overall rate of the previous approach in Table 4.4.1.

The conclusion is, obviously, to recommend the second approach, and this will consequently be the method used in the seed recognition work presented later.

4.5 The Fourier transform and template matching

The use of the Fourier transform is another common way to obtain information about the shapes. From this analysis the information is available as spectral information, *i.e.*, frequencies and amplitudes of the waves approximating the contour.

Theory

In general, a Fourier transform is an approximation with trigonometric functions (sine and cosine) to an arbitrary function. The mathematical expression is dependent on the function to be approximated. If the function is periodic it will be expanded as a Fourier series, otherwise as a Fourier integral. An illustrative introduction to this is provided by Eriksson *et al.* (1970-), and the transform of discrete functions is presented in Gonzalez and Wintz (1987).

For a periodic, continuous or piecewise continuous function $f(t)$ with the period 2π, *i.e.*, $f(a) = f(a + 2\pi)$, the Fourier series is

$$f(t) = \mu + \sum_{n=1}^{\infty} (a_n \cos(nt) + b_n \sin(nt))$$

This expression converges to f for all points of continuity (for points of discontinuity it converges to the mean of the function limits from right and from left). The coefficients a_n and b_n are termed the Fourier coefficients. This expression may also be regarded as a sum of harmonic waves with the amplitude given by the Fourier coefficients. It may be noted, that the frequency of the wave increases with n, and it is the amplitude that determines the weight of the single wave in the expression.

To show the weights of the harmonic components of a periodic function an amplitude spectrum may be constructed. For calculation of the amplitude, the magnitude of the Fourier coefficients is used, *i.e.*, $A_n = \sqrt{a_n^2 + b_n^2}$. This leads to the polar expression of the Fourier series,

$$f(t) = \mu + \sum_{n=1}^{\infty} A_n \cos(nt - \alpha_n)$$

where (A_n, α_n) are the polar coordinates of (a_n, b_n). Each harmonic wave is now represented by a single trigonometric function (cos), the n'th harmonic amplitude (A_n) and the phase angle (α_n). The amplitude spectrum shows how fast the Fourier series converges to f. If the amplitudes are relatively large for high frequencies the series converges slowly, which, for example, appears for functions with sharp edges.

It is often useful to write the Fourier series using complex notation:

$$f(t) = \sum_{-\infty}^{\infty} c_n e^{int}, \quad i = \sqrt{-1}$$

where $f(t)$ may be complex, *i.e.*, $f(t) = x(t) + iy(t)$. It may be noted that the summation now includes negative values of n. For a real function the Fourier coefficients are related as

$$c_0 = \mu$$
$$c_n = \frac{1}{2}(a_n - ib_n)$$
$$c_{-n} = \frac{1}{2}(a_n + ib_n)$$

It is possible to remove the limitation of the periodic interval of 2π, so that the period is of arbitrary length, $2l$. This gives

$$t = \frac{t_{(2\pi)} \times \pi}{l}$$

$$f(t) = \sum_{-\infty}^{\infty} c_n e^{int\pi/l}$$

The intervals between the frequencies are π/l, meaning that an increase in period provides closer frequencies. If the period increase indefinitely (*e.i.*, the non-periodic case) the frequencies becomes continuous.

The corresponding expression for the Fourier coefficients is

$$c_n = \frac{1}{2l}\int_{-l}^{l} f(t)e^{int\pi/l}dt$$

If the function f is non-periodic, it is expanded as a Fourier integral. This may be imagined as a limiting case where the period is arbitrarily large. However, for the analysis of seed images the discrete case is more relevant.

When f is a discrete function of t assuming the values 0,1,2,...,N-1, it is expanded as a discrete Fourier series

$$f(t) = \sum_{u=0}^{N-1} F(u)e^{i2\pi ut/N}$$

with the inverse process

$$F(u) = \frac{1}{N}\sum_{t=0}^{N-1} f(t)e^{-i2\pi ut/N}$$

for u = 0,1,2,...,N-1. However, it may be recalled from the Nyquist theorem that the highest harmonic frequency is half the sampling frequency, in this case N/2. Therefore, in the following the interval for u is changed to [-N/2+1, N/2] by translating the last N/2 coefficients to the negative axis for u. This is allowable due to the periodic properties of the Fourier transform.

The discrete Fourier transform uses N^2 multiplications to calculate all values of $F(u)$. Thus, the number of multiplications increases with the square of the number of values sampled. Therefore, algorithms have been developed to reduce the number of operations. They are referred to as fast Fourier transforms (FFT), and take a particularly simple form when the number of sampling values is a power of two.

Previous work

Many different approaches have been tried to extract spectral information in image analysis. Brill (1968) and Zahn and Roskies (1972) represented the contour of a closed curve by a function which relates the angular bendings to the distances from a starting point. The angular bending was normalized, so that it became zero for a circle, and the length was normalized to the interval $[0;2\pi]$, *i.e.*, size normalization. This normalized function was a piecewise, continuous and periodic function which may be expanded as a Fourier series as earlier expressed. For recognition purposes normalized descriptors were obtained in the form of amplitudes and invariant functions of the phase angles, thus becoming independent of the starting point.

Bennett and MacDonald (1975) used the Fourier transform to study the quantization noise from six different contour tracing algorithms. A discrete curvature function of the angular bendings at the sampling points was constructed from each algorithm. Considerable variation among the amplitude spectra of the different curvature functions was found. However, the one dimensional angle versus length representation of the contour in both approaches has difficulties in reconstructing a closed boundary curve from the Fourier descriptors.

Granlund (1972), Richard and Hemami (1974) and Wallace and Wintz (1980) represented the boundary curve in the complex plane, so that the curve was described as $f(l) = x(l) + iy(l)$, where l is the length from some starting point, and $x(l)$ and $y(l)$ are the coordinate values at that position. When using sampling points for boundary representation the spectral information should be obtained by the discrete Fourier transform. They proposed different normalization procedures for Fourier descriptors invariant for size, rotation, translation and

starting point (Richard and Hemami (1974) normalized only for size). The advantage of this boundary representation is the possibility to reconstruct it from the inverse transform, and in the present study the normalization method of Wallace and Wintz (1980) is adopted.

Persoon and Fu (1977) proposed a model of line patterns where the contour was described by two components: 1) A skeleton, and 2) a thickness. For objects which can be modeled in this way, the relation between the Fourier descriptor for the contour and the skeleton was presented. This relation was shown to be used for obtaining the skeleton of an object.

Kuhl and Giardina (1982) presented a procedure for obtaining the Fourier coefficients of a contour which was represented as a Freeman chain code. But, also, a representation similar to the complex periodic function, $x(l) + iy(l)$, was regarded. The difference was that $x(l)$ and $y(l)$ were expanded separately as Fourier series, and these two expressions were combined in matrix form. For this representation each harmonic described an ellipse, and an illustration of an elliptic approximation to a contour was shown. The normalized descriptors were based on the elliptic properties, and called the elliptic Fourier features. Adopting this approach Lin and Hwang (1987) presented new shape invariants using matrix trace operators.

The normalized Fourier descriptors

The normalization procedure was based on the approach described in Wallace and Wintz (1980) and Gonzalez and Wintz (1987) where the contour was represented as uniformly sampled points in the complex plane. A modification of this normalization method applied to weed seeds was earlier presented in Petersen (1991).

Regard a discrete, periodic function which has a discrete Fourier transform with the frequencies of the coefficients in the range from -N/2 + 1 to N/2. The basis for the normalization method is the orientation and starting point operations in this transform system. That is, if the contour is rotated in the spatial domain by the angle θ, the equivalent expression in the frequency domain is the multiplication

$$F(u) \times e^{i\theta}$$

for all u.

If the starting point is shifted T radians, this corresponds to a multiplication of each $F(u)$:

$$F(u) \times e^{iuT}$$

This means that shifting the starting point and rotating the contour will change the Fourier coefficients by changing the phase angle (the additional term in the exponent of e). This is used to define a normalized orientation through a defined value of the phase angles of the most important Fourier coefficients. As a rule $F(1)$ will always be the coefficient of largest magnitude when the contour is traced in counterclockwise direction, and the contour does not cross itself (Wallace and Wintz, 1980). Now, the normalization is performed by setting the phase angles of the two Fourier coefficients of largest magnitude equal to zero. If the largest is $F(1)$, as noted above, and the second largest is $F(k)$, this normalization will in general not provide a unique solution. A theorem, proved by Wallace and Wintz (1980), states that zero phase angles for $F(1)$ and $F(k)$ is satisfied by $|k-1|$ combinations of starting point and orientation. Therefore, only for $k = 2$ the starting point and orientation is unique for zero phase angle.

The phase angle of $F(k)$ is calculated as

$$\alpha_k = \arctan\left(\frac{Im(F(k))}{Re(F(k))}\right)$$

regarding the interval $[0;2\pi]$. If the normalized phase angle of $F(k)$ is called α_k', then

$$\begin{cases} \alpha_1' = \alpha_1 + \theta + T = 0 \\ \alpha_k' = \alpha_k + \theta + kT = 0 \end{cases}$$

$$\rightarrow \begin{cases} \theta = -(T + \alpha_1) \\ T = -(\alpha_k - \alpha_1)/(k - 1) \end{cases}$$

This leads to the first (ambiguous) solution:

$$F^{(1)}(u) = F(u) \times e^{i(\theta + uT)}$$
$$= F(u) \times e^{i[(u - k)\alpha_1 + (1 - u)\alpha_k]/(k - 1)}$$

The next $|k\text{-}1|\text{-}1$ possible solutions are obtained by rotating $\theta = 2\pi/(k\text{-}1)$ radians and shifting the starting point by $T = 2\pi/(k\text{-}1)$ in the opposite direction to the rotation. This is performed by successively multiplying the previous solution:

$$F^{(2)}(u) = F^{(1)}(u) \times e^{i(u - 1)2\pi/(k-1)}$$

and so on. The number of possible solutions is by this method reduced to $|k\text{-}1|$. Wallace and Wintz (1980) achieved the best solution by maximizing the function

$$\sum_{u=1}^{N-1} Re[F(u)]|Re[F(u)]|$$

among the $|k\text{-}1|$ possible solutions. During the investigation of weed seeds another criterion function was proposed (Petersen, 1991), so that the unique solution was achieved by maximizing

$$\sum_{t=0}^{N-1} x(t)|x(t)|$$

where $x(t)$ is the spatial x-coordinate. Usually, the possible solutions will orientate the seed, in such a way that a fitted ellipse has the major axis horizontally. This makes two directions of the contour possible. Maximizing of the criterion function above will favour the orientation where the acute or jagged end is to the right. If the seed is approximately symmetric around the major axis (*e.g.*, the *Rumex* species) there will be one orientation after normalization. For asymmetric seeds there will be two possibilities depending on which side of the seed turns up or down. This corresponds to the two mirror images around the x-axis.

In few cases the $|k\text{-}1|$ solutions were above two, and for an uneven number of solutions the system was improved by choosing the solution with the smallest angle between the major axis and the horizontal. For higher and even numbers the two solutions giving the smallest angle were selected, and the final one was the one with maximal criterion function.

Normalization for size were obtained by dividing all Fourier coefficients with the magnitude of *F(1)* and multiplying by a constant. Finally, the position was normalized by setting $F(0) = 0+i0$.

Template matching

For the purpose of template matching a more geometric interpretation will be valuable. When the transform pair, described in the previous section, is used (*i.e.*, uniform sampling in counterclockwise direction and the frequency variable u in the range from -N/2+1 to N/2) *F(0)* determines the position of the center of gravity, and *F(1)* is the fundamental harmonic, which describes a circle of a size equal to the contour area. If all the transformed values of a contour, *F(u)*, are set to zero except *F(1)*, the inverse transform gives a circle centered at origo. If both *F(-1)* and *F(1)* keep values different from zero, an ellipse will be formed from the inverse transform. This may be used to fit an ellipse to an object. If values of *F(u)* at higher frequencies are preserved, the ellipse will be distorted, so that the reconstructed shape will converge to the original shape for increasing number of frequencies preserved. This is illustrated in Figure 4.5.1, where different intervals of *F(u)* are set to zero using a seed of *Silene vulgaris*. It may be noted that setting high frequency Fourier coefficients to zero, a smoothed version of the original shape is obtained by reconstruction.

Template matching has the simple purpose to fit a test contour to a template (*i.e.*, a prototype contour). For making the best fit the contours should be normalized to the same size, position and orientation.

The templates in present investigation were

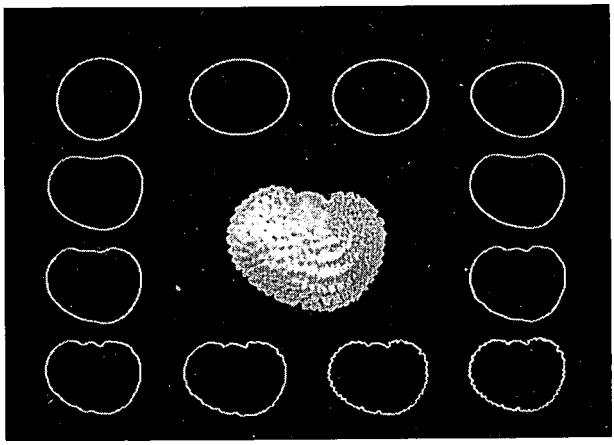


Figure 4.5.1 Reconstruction of boundary of a Silene vulgaris seed using different subsets of Fourier coefficients, F(u). Intervals for the integer u, rowwise, from upper to lower: [0,1], [-1,1], [-1,2], [-2,2], [-6,6], [-8,8], [-13,13], [-16,16], [-24,24], [-32,32], [-64,64], [-255,256] (= all).

constructed by normalizing five seeds in each species followed by averaging in the spatial domain. The averaging produced a smoothed template contour as shown in appendix B. To reduce the mismatch within the species a standard smoothing procedure was used where only frequencies in the range from -16 to 16 were included, *i.e.*, $F(u)$ at other frequencies were set to zero. This smoothing procedure was applied to both test and template contours.

From the boundary 512 points were obtained by uniform sampling in counterclockwise direction, and the Fourier transform was calculated using a FFT algorithm (Press *et al.*, 1987). The size of the normalized seed was standardized by dividing all $F(u)$ by $|F(1)|$ and multiplying by the constant 65. This resulted in an area of about 19700 pixels for the reconstructed silhouette.

The normalized template and test seed boundaries were the basis for calculating the mismatch. First each boundary was checked for lacking pixels, *i.e.*, holes in the boundary. In this case new boundary pixels were constructed by linear interpolation. Then non-overlapping boundary segments were extracted, and a filling routine was performed to count the pixels in the non-overlapping areas as shown in Figure 4.5.2. The control structure of this filling routine is described in algorithm 3 in Appendix A. More detailed description on algorithms for

Table 4.5.1 The classification rates for matching using two orientation methods of the seed contours.

	Normalization	Closest fit
Caryophyllaceae	63	68
Rumex	78	79
All species	69	72

contour filling can be found elsewere (Pavlidis, 1990). Thus, the mismatch is

$$(T \setminus S) \cup (S \setminus T)$$

where T is the template area and S is the area of the test seed. The procedure was repeated for the mirror image around the x-axis of the template contour, and the smallest of the two mismatch values was kept.

For fish species recognition by Strachan and Nesvadba (1990) a similar measure of mismatch was used, except for normalizing the measure by the combined area, T U S. Two faster algorithms for similarity/mismatch were presented by Sze and Yang (1981). They were based on counts of overlapping boundary pixels, thus determining the number of non-matching pixels or the ratio between the matching and non-matching pixels. However, these algorithms are

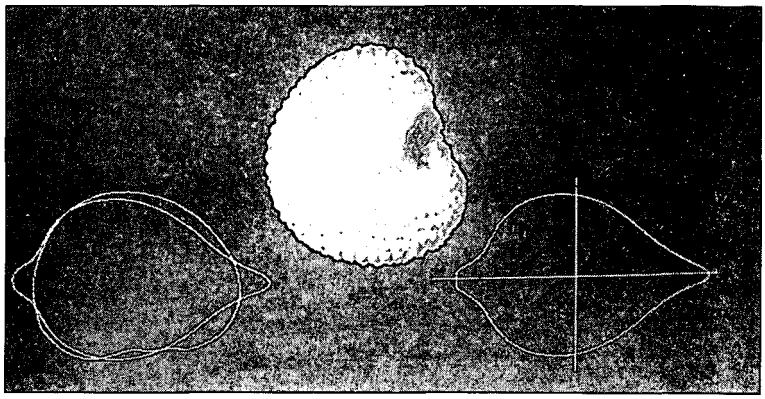


Figure 4.5.2a Matching shape of a Silene vulgaris seed to a template. Lower right: Boundary of Rumex crispus seed (template). Lower left: Matching the two boundaries.

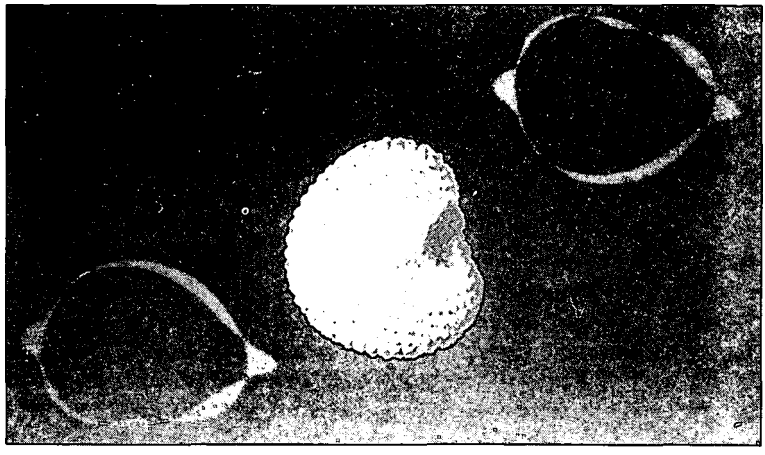


Figure 4.5.2b Determination of mismatch between test and template shape in Figure 4. 5.2a. Lower left: Grey area is the mismatch area. Upper right: Same situation using the mirror image of the test seed boundary.

not as precise for mismatch, because the distance between boundaries is not considered.

In Petersen (1991) a linear discriminant analysis was performed after matching test seeds of 10 species against all 10 templates. Six species of *Caryophyllaceae (M.album, M.rubrum, S.noctiflora, S.vulgaris, S.gramina* and *S.media)* and four species of *Rumex (R.crispus, R.obtusifolius, R.thyrsiflorus* and *R.acetosa)* were used with 20 seeds of each species. The mismatch values were obtained from the normalized seeds, but also from changing the orientation of test seed to obtain the closest possible fit to the template. The classification rates of both matching methods are summarized in Table 4.5.1, showing the smallest difference for the *Rumex* species. However, the difference in classification rates was based on relatively small differences in mismatch values (see Petersen, 1991).

In the present investigation only the mismatch values between test seeds and template within the same species are presented in Table 4.5.2. The magnitude of the mean and standard deviation provides an impression of the difficulties of matching seeds of different species. The matching of a test seed to all possible templates is regarded as a very time consuming process, but the matching may be used after a classification using other analyses to verify the result or to discriminate among a few possibilities.

The amplitude discrimination

Regard again a boundary curve in the complex plane described as $f(t) = x(t) + iy(t)$. The amplitudes of the discrete Fourier transform are calculated as

$$A_u = |F(u)| = \sqrt{Re[F(u)]^2 + Im[F(u)]^2}$$

which are always invariant to rotation, starting point and position. For the purpose of discriminating seed species using the amplitude information, 512 amplitudes were calculated from the normalized Fourier coefficients. The results may be plotted in amplitude spectra (scaled by taking the logarithm of the amplitude and adding two) as shown in Figure 4.5.3-4.5.9 for seeds of a few selected species. For classification purposes the amplitude information should be reduced heavily. Thus, when using linear discriminant analysis the number of features should not exceed one third of the group size, to get stable discriminant functions (Seber, 1984).

Different methods for data reduction are possible. In Petersen (1991) the amplitude spectra of seeds of each species were studied, and the contours were reconstructed by setting successively higher frequency components to zero. This showed which frequency bands were important for reconstruction of certain important shape features. This was the basis for grouping the amplitudes into 7-8 frequency bands. In general, the smooth seed shapes were discriminated only from low frequency information, and the more jagged contours could exploit amplitude information of higher frequencies. In Figure 4.5.3-4.5.9 the spectra are arranged in descending order of contour jaggedness, and it is characteristic that all spectra have the highest amplitudes at low frequencies, but for some species in certain bands also significant local maxima appear. For example, the spectra of *Stellaria media* shows a local maximum at frequency 30 corresponding to 30 swellings of equal size around a circular shape. This number corresponds approximately to the number of outbulging surface cells in the periphery of *S.media* seeds. For comparison the seeds of *Silene vulgaris* (Figure 4.5.4) have a much higher number of out-bulging cells in the periphery, which is reflected in relatively high amplitudes above frequency 50. From these considerations it seems reasonable to construct frequency bands of increasingly larger range. The following frequency bands are proposed:

1) $u = -3, 3$
2) $u = -4, 4$
3) $u = -5, 5$
4) $u = \pm[6;14]$
5) $u = \pm[15;24]$
6) $u = \pm[25;44]$
7) $u = \pm[45;64]$
8) $u = \pm[65;128]$

Within these bands the amplitudes are summed and normalized, so that the total is equal to one.

Another method of reducing the amount of data is to construct the canonical variates (discriminant coordinates in Seber, 1984). The principle behind this method is summarized in the following.

Let $\mathbf{x}_{ij}$ be the j'th observation in group i represented as a vector of dimension d. Then the multivariate analog of the within and the between sum of squares is defined as (using the usual notation):

$$\boldsymbol{W} = \sum_i \sum_j (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{i.})(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{i.})'$$

and

$$\boldsymbol{B} = \sum_i \sum_j (\bar{\boldsymbol{x}}_{i.} - \bar{\boldsymbol{x}}_{..})(\bar{\boldsymbol{x}}_{i.} - \bar{\boldsymbol{x}}_{..})'$$

respectively. It is possible to transform data by choosing a vector $\mathbf{c}'$, so that the observation vector is projected to a one-dimensional space, *i.e.*, $z_{ij} = \mathbf{c}'\mathbf{x}_{ij}$. To achieve maximum separation of the groups in one direction the ratio of the one dimensional sum of squares has to be maximized. That is maximization of

$$\frac{\boldsymbol{c}'\boldsymbol{B}\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{W}\boldsymbol{c}} = F_c$$

Table 4.5.2. Mismatch values from matching normalized test seeds to template within species

Species	Mean	Std Dev
Urtica urens	*1206*	*459*
Polygonum convolvus	*1395*	*378*
Polygonum lathifolium	*1216*	*461*
Rumex acetosa	*1588*	*410*
Rumex crispus	*1608*	*570*
Rumex obtusifolius	*1585*	*465*
Rumex thyrsiflorus	*1579*	*690*
Arenaria serpyllifolia	*1460*	*405*
Melandrium album	*1304*	*353*
Melandrium rubrum	*1953*	*368*
Silene noctiflora	*1530*	*460*
Silene vulgaris	*1938*	*755*
Stellaria gramina	*1629*	*330*
Stellaria media	*1420*	*328*
Chenopodium album	*1239*	*420*
Ranunculus repens	*2601*	*1403*
Papaver rhoeas	*2876*	*680*
Brassica campestris	*1438*	*424*
Capsella bursa-pastoris	*1354*	*442*
Geranium dissectum	*1211*	*373*
Euphorbia exigua	*1376*	*395*
Euphorbia helioscopia	*1177*	*237*
Euphorbia peplus	*1000*	*244*
Viola arvensis	*1419*	*559*
Myosotis arvensis	*1931*	*841*
Lamium amplexicaule	*2672*	*424*
Solanum nigrum	*2153*	*602*
Veronica arvensis	*1591*	*571*
Veronica persica	*2011*	*673*
Plantago major	*2113*	*625*
Cirsium arvense	*2115*	*582*
Chrysanthemum segetum	*2030*	*976*
Matricaria chamomilla	*1800*	*647*
Matricaria inodora	*2918*	*1316*
Matricaria matricarioides	*1703*	*626*
Sinapis arvensis	*879*	*246*
Sonchus arvensis	*2009*	*787*
Sonchus asper	*2308*	*1088*
Sonchus oleraceus	*2483*	*1033*
Taraxacum vulgare	*2498*	*741*

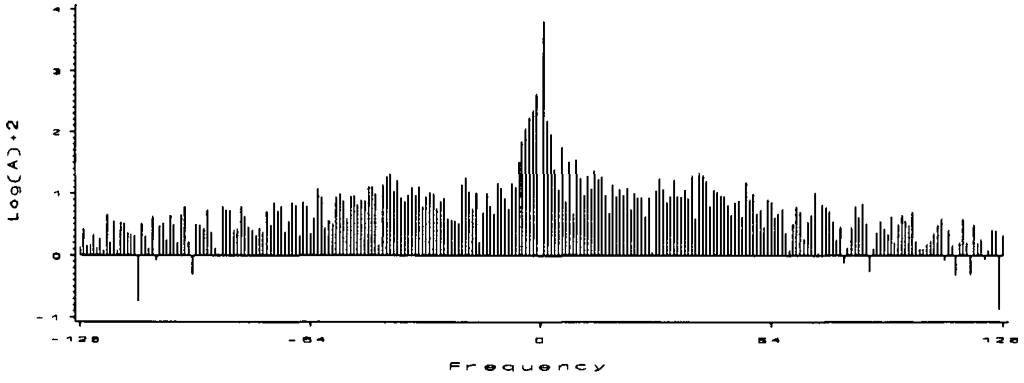


Figure 4.5.3 Amplitude spectrum of seed of Melandrium rubrum

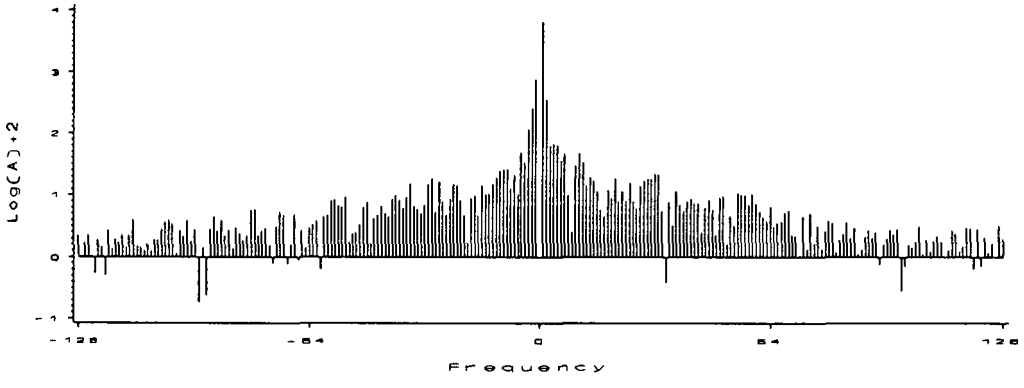


Figure 4.5.4 Amplitude spectrum of seed of Silene vulgaris

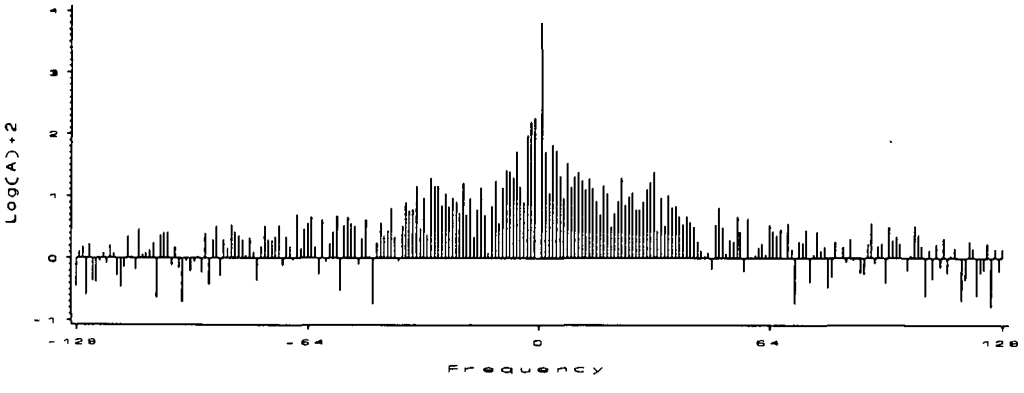


Figure 4.5.5 Amplitude spectrum of seed of Stellaria media

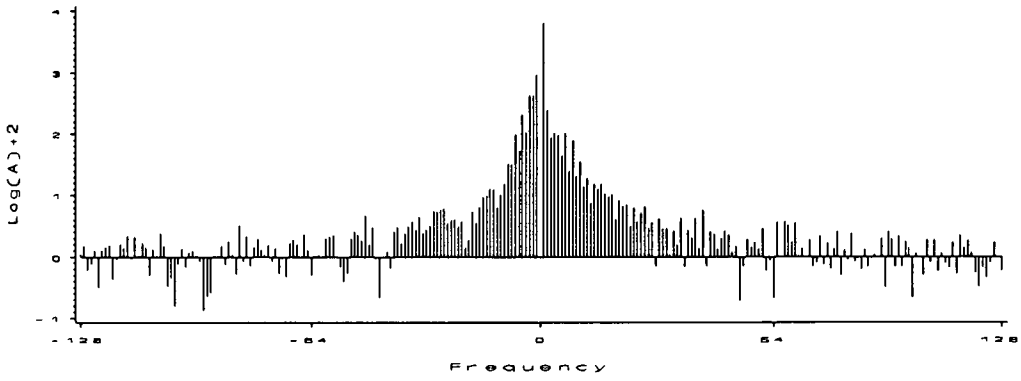


Figure 4.5.8 Amplitude spectrum of seed of Rumex obtusifolius

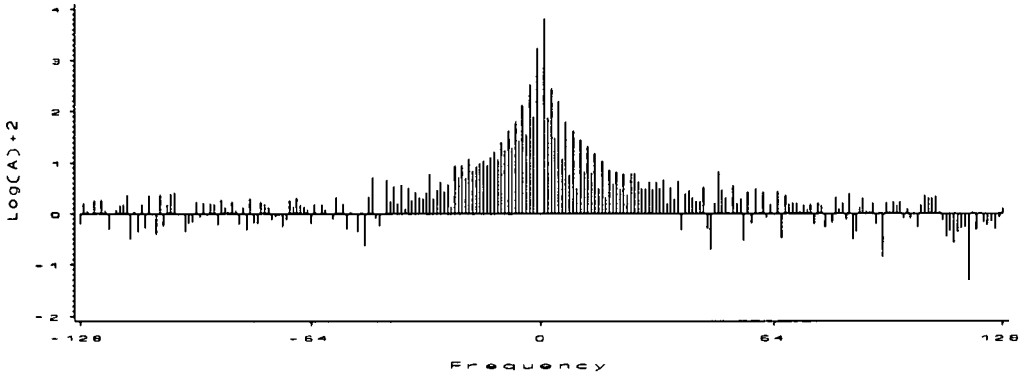


Figure 4.5.9 Amplitude spectrum of seed of Rumex acetosa

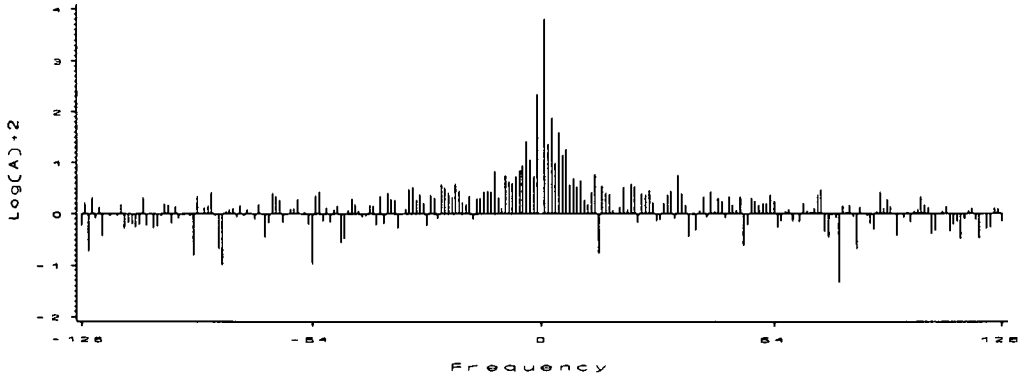


Figure 4.5.10 Amplitude spectrum of seed of Sinapis arvensis

Table 4.5.3 Classification rates using amplitudes as features in canonical discriminant analysis. The transformation matrix was constructed from data in the training set, and applied on data in the test set.

Training seeds	1 - 10		16 - 25		1-5 & 21-25	
Classification set	training	test	training	test	training	test
Caryophyllaceae	96.7	65.6	98.3	70.0	91.7	72.2
Rumex	90.0	65.0	85.0	76.7	80.0	76.7
All species	94.0	65.3	93.0	72.7	87.0	74.0

Training seeds	1 - 15		11 - 25		6 - 20		1 - 25
Classification set	training	test	training	test	training	test	training
Caryophyllaceae	92.2	75.0	90.0	75.0	88.9	75.0	86.7
Rumex	78.3	67.5	81.7	72.5	81.7	77.5	79.0
All species	86.7	72.0	86.7	74.0	86.0	76.0	83.6

This may be repeated for *d-1* other orthogonal directions. The solution has been shown to occur for **c** equal to the eigenvector of the largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$, and the other orthogonal directions with maximal group separation occur for **c** equal to the other *d-1* eigenvectors. Thus, to obtain the canonical variates a transformation matrix $\mathbf{C}' = (\mathbf{c}_1,...,\mathbf{c}_k)$ is constructed from the k first eigenvectors ($k <= d$), so that $\mathbf{z}_{ij} = \mathbf{C}\mathbf{x}_{ij}$.

From the amplitude spectra a new observation vector was constructed by adding the amplitudes at each positive frequency to the amplitudes at the corresponding negative frequency. Moreover, these values were added two and two, so that the number of observation was reduced to one fourth. The first 30 values were included in the observation vector, thus representing amplitude information in the frequency range from -60 to 60 excluding zero. A data set of six species of *Caryophyllaceae (M.album, M.rubrum, S.noctiflora, S.vulgaris, S.gramina* and *S.media)* and four species of *Rumex (R.crispus, R.obtusifolius, R.thyrsiflorus* and *R.acetosa)* was used with 25 seeds per species. From these a training set and a test set were selected. First the training set consisted of 10 seeds per species selected by the seed labels: 1-10, 16-25, and 21-25 & 1-5. Another series was constructed by switching test and training sets, and finally the training set consisted of all seeds. From the training set the 8 first canonical coefficients, *i.e.*, the $\mathbf{C}'$ matrix with $k = 8$, was obtained, and applied to the data in the test set. A linear discriminant analysis with crossvalidation was used to classify seeds in each data set separately.

The results are presented in Table 4.5.3 as classification rates. This shows that high rates appear for small training sets, while the test sets had much lower rates, thus reflecting a difference in the distribution of features in the feature space between test and training sets. The problem when using canonical variates for data reduction is, in general, that the transformation matrix for the training set is not the optimal transformation matrix for the test set. This is also illustrated by the results in Table 4.5.3, and the method will not be used in the present investigation. A higher number of observations might reduce the difference between test and training seeds, but the level of

recognition is not expected to be very high as indicated by the classification rate of all seeds in the training set.

4.6 Other shape / texture descriptors

Two other shape analyses were applied to the weed seeds. One of these, the fractal dimension analysis, was also implemented as texture analysis. These analyses have not changed considerably during the project, and will only be briefly introduced.

Fractal dimensions

The fractal dimension corresponds closely to the visual perception of roughness. This may be illustrated by regarding a curve with many irregularities and of a certain length L. For conventional differential methods the result of measuring the curve using a yardstick of length e converge to L as e goes to zero. If the curve irregularities are of a type where new curly details show up for all kinds of resolutions, the length will converge to infinity as e goes to zero. The difference between the two approaches is explained by extending the usual topological dimensions from 1, 2 and 3 to the fractal dimensions, which also cover real values between the integers. For the last type of curve the fractal dimension D is above one, and for some constant F the equation

$$L(e) = F \times e^{1-D} \qquad 1 \le D < 2$$

describes, that length is dependent on e and converges to infinity for $D > 1$ as e goes to zero. Otherwise, for $D = 1$ (a straight line) the length equals the true length F.

For surface measurements a similar equation is used to describe the area dependency on the yardstick:

$$A(e) = F \times e^{2-D} \qquad 2 \le D < 3$$

In shape analysis the digital boundary values have a fixed length for the chosen resolution, and cannot be genuine mathematical fractals. However, to approximate the fractal method new boundaries at lower resolutions have to be created artificially, by successively smoothing the former boundary. If the original boundary was smooth only small changes occur when constructing the next boundaries.

The method used in this study was presented by Peleg *et al.* (1984). The original boundary was obtained and represented in the time series representation, $u_e(t)$ say. For this series $e = 0$, and new series were constructed for $e = 1,2,...$, defined as

$$u_e(t) = \max \left\{ u_{e-1}(t) + 1, \max_{|s-t| \le 1} u_{e-1}(s) \right\}$$

This means that either the value at the previous resolution is added by one, or the neighbour value at the previous resolution is taken, depending on which one is the largest. This is done similarly in the opposite direction, where the new constructed series are termed $b_e(t)$, and $u_0(t) = b_0(t)$. Thus, the series are defined as

$$b_e(t) = \min \left\{ b_{e-1}(t) - 1, \min_{|s-t| \le 1} b_{e-1}(s) \right\}$$

The area measured by subtracting the two series of the same e value is

$$A(e) = \sum_t (u_e(t) - b_e(t))$$

Then the length corresponding to the increase in area between e and $e-1$ on average of the upper and lower layer is

$$L(e) = \frac{(A(e) - A(e-1))}{2}$$

When plotting L against e on a log-log scale a straight line should appear for fractal surfaces. Then the 'fractal signature' is calculated as the slope of the straight line, $S(e)$, fitting the three

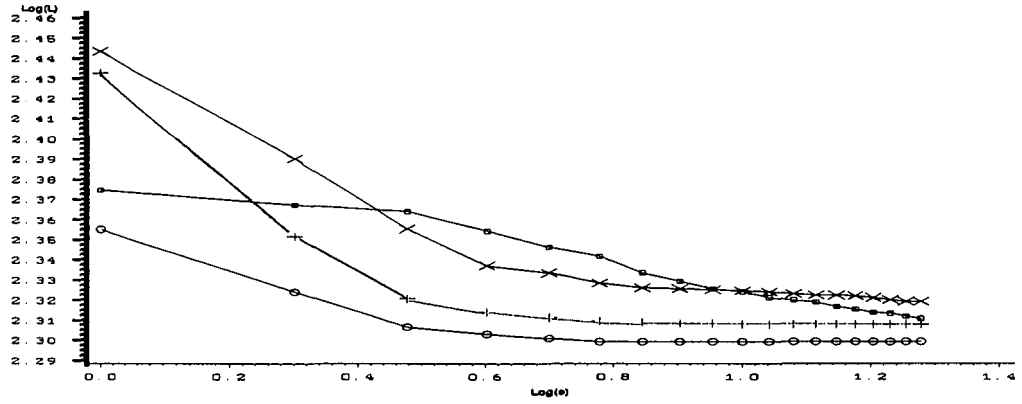


Figure 4.6.1 Plot of length versus e (extension) in a log-log scale. Symbols are (+) for seed of Melandrium rubrum, (×) for seed of Silene vulgaris, (○) for seed of S.media and (□) for seed of Rumex crispus.

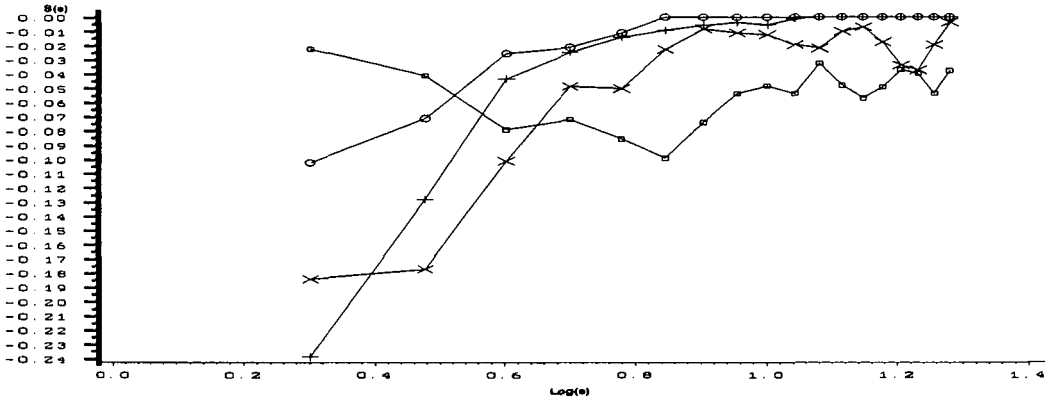


Figure 4.6.2 Plot of fractal signature, S(e), versus log(e) (extension). Symbols are (+) for seed of Melandrium rubrum, (×) for seed of S. vulgaris, (○) for seed of S.media and (□) for seed of Rumex crispus.

points (log(*e-1*), log(L(*e-1*))), (log(*e*), log(L(*e*))) and (log(*e+1*), log(L(*e+1*))). This local slope should be constant for fractal surfaces and equal to the exponent to *e*.

In practice it showed up that many seed shapes were smoothed relatively fast, to such an extent, that all values in the time series were equal, and, therefore, the fractal signature goes to zero (Figure 4.6.1 and 4.6.2). The rough seeds with many small swellings on the contour, and with round 'macro' shape (as *Melandrium rubrum*) converged very fast to zero, whereas the smooth contour with two high local maxima (as *Rumex crispus*) converged very slowly, although it started at a level closer to zero. Intermediate shape types with a decreasing number of swellings and a more complicated macro shape than *M.rubrum* are shown in

Figure 4.6.1 and 4.6.2. (*Silene vulgaris* and *Stellaria media*) too.

It is the fractal signatures that are used as features in the present investigation of shapes. For the extension to surfaces in texture analysis the grey levels were replacing the time series and measurements of volume and area were calculated similarly to area and length above.

Rapid transformation

Rapid transformation was introduced by Reitboeck and Brody (1969), and applied to shape analysis by Ma *et al.* (1986). Basically, the transform works by adding and subtracting values in an observation vector of size $N = 2^M$, M is an integer.

The observation vector consists of values from the input variables which in the first step are divided into two groups numbered from 0 to N/2 - 1 and from N/2 to N - 1. Thus, the first transform vector is constructed from the equations

$$X_i^{(1)} = X_i + X_{i+N/2}, \qquad i = 0,1,\ldots,N/2-1.$$
$$X_{i+N/2}^{(1)} = |X_i - X_{i+N/2}|, \qquad i = 0,1,\ldots,N/2-1.$$

Then each of the two groups in the first transformation vector is divided into two subgroups, and so on until no more division into subgroups is possible (when subgroup size equals one observation). The general expression is

$$X_{m+2ns}^{(R)} = | X_{m+2ns}^{(R-1)} + X_{m+(2n+1)s}^{(R-1)} | \qquad \prod_{m=0}^{s-1} \prod_{n=0}^{t-1}$$
$$X_{m+(2n+1)s}^{(R)} = | X_{m+2ns}^{(R-1)} - X_{m+(2n+1)s}^{(R-1)} |$$

where R denotes the number of transformations from one to M, t is the number of subgroups in the Rth transformation, and s is the size of the subgroup. For shape analysis the observation vector consists of values representing the contour.

To use this transformation system as a shape descriptor the original variables have to be invariant to position, rotation, size and starting point. This is obtained by applying a distance transform. The underlying principle is to calculate the center of gravity and the length of the radii to the boundary. The transform system itself was shown (Reitboeck and Brody, 1969, and Ma *et al.*,1986) to be invariant to cyclical shifts, which makes the transform of the time series representation invariant to rotation and to shift in starting point. Finally, scaling invariance is achieved by dividing all radii lengths by the mean radius length.

The total procedure for rapid transformation is: 1) Obtain all the boundary points and the center of gravity, 2) sample 128 boundary points with equal distance, 3) calculate the distances of the sampled points to the center, 4) divide the radii length by the mean radius length to obtain the normalized observation vector, and 5) make the rapid transform on the normalized observation vector as described above.

The features obtained from this transform was the sum of the transform values in eight intervals of i:

1) [1;15]
2) [16;31]
3) [32;47]
4) [48;63]
5) [64;79]
6) [80;95]
7) [96;111]
8) [112;127]

Only 15 transform values contributed to the first feature, because the value for $i = 0$ is the sum of all normalized values in the observation vector, which is equal to one.

The transform is not simple to interpret, so a number of artificial time series were constructed from sampled sine waves. The transformation showed that for a number of sine waves which equals a power of two, high frequency information was represented with high values in the first transform values (*i.e.*, low i values). Decreasing frequencies moved this maximum in transform values toward $X_{N/4}$ (for two sine waves). For an equal number of sine waves (not a power of two) the global maximum was also placed at $X_{N/4}$, and for an uneqal number

of sine waves (1,3,5...) the global maximum was at the center (*i.e.*, $X_{N/2}$) with relatively high transform values before and after. On both sides of the global maximum a special pattern of local maxima and minima were observed.

4.7 Texture matrices

Texture analysis concerns the grey levels in the digitized image. The grey levels reflect the shape and the colour of the entire surface. In images of weed seeds, such as *Silene noctiflora*, the out-bulging surface cells are visible as swellings on the contour, but also as grey levels where high levels represents elevated parts and low levels represents valleys. However, this pattern is disturbed by colouration of the seed, for example, when dark pigmentation spots appear on top of the swellings. This pigmentation is only registered in the surface grey levels, and not by the boundary positions of the contour.

Most analyses of shape have an analogous analysis of texture. Only a shift from the one-dimensional representation with radii lengths corresponding to increasing angles to grey levels in the two-dimensional space is needed to go from the shape to the texture representation. Thus, the Fourier transform, moment invariants, the time series and the fractional dimensions are also available in the two-dimensional space. In the present study the methods of texture matrices were applied to the weed seed images using a program constructed by Olsen (1988) who performed an extensive investigation of texture in ultrasound liver images for diagnostic purposes.

The Grey Level Histogram (GLH)

The frequency of occurrence of the grey levels in an image may be presented in a simple histogram. The usual way to characterize a distribution is by estimating the features mean, variance, skewness and kurtosis. Regarding the values of the histogram the formulas are:

$$\mu = \sum_i w_i i$$

$$s^2 = \sum_i w_i i^2 - \mu^2$$

$$\beta_1 = (\sum_i w_i i^3 - 3\mu\sum_i w_i i^2 + 2\mu^3)/s^3$$

$$\beta_2 = \frac{\sum_i w_i i^4 - 4\mu\sum_i w_i i^3 + 6\mu^2\sum_i w_i i^2 - 3\mu^4}{s^4} - 3$$

where i denotes the grey levels and w_i is the frequency of occurrence at that grey level.

The Grey Level Cooccurrence Matrix (GLCM)

The Grey Level Cooccurrence Matrix was introduced by Haralick *et al.* (1973), and has since been commonly used in analysis of texture. This method is also based on the frequency of occurrence of the grey levels as in the GLH analysis, but it is extented to a 2-D (discrete) distribution. This means that the frequency of occurrence is registered for all combinations of two variables. This distribution is, as for the GLH, also characterized by certain features.

The two variables in the GLCM analysis are 1) the grey level of a center pixel within some neighbourhood in the image, and 2) the grey level of a neighbour pixel in a chosen direction and distance from the center pixel. The distribution may be illustrated in a spatial histogram or equivalently in a matrix. It should be noted that these distributions are symmetric as each neighbour pixel also becomes a center pixel and *vice versa*.

In the present investigation the GLCM is constructed by considering neighbour pixels in the four directions (north, south, east, west) and with distance one. This means that each pixel contributes four counts to the two dimensional histogram. Eventually, the dimension of the matrix / 2-D histogram is reduced from 256x256 to 32x32 by averaging, and normalized to sum to one.

Before looking at the features estimated from these matrices, it would be valuable to have a visual impression of the matrices themselves. Figure 4.7.1 - 4.7.3 show the 2-D distribution of grey levels from three images of *Silene vulgaris, Rumex crispus and Euphorbia exigua*,respectively. In general, the highest frequency of

Figure 4.7.1 Plot of GLCM from image of Silene vulgaris.

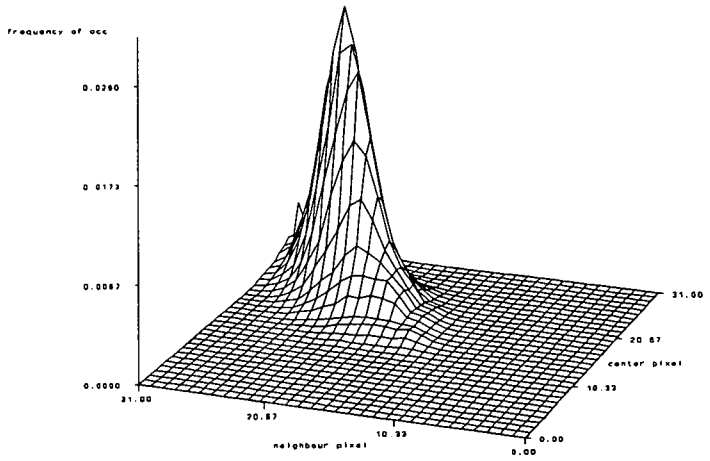


Figure 4.7.2 Plot of GLCM from image of Rumex crispus.

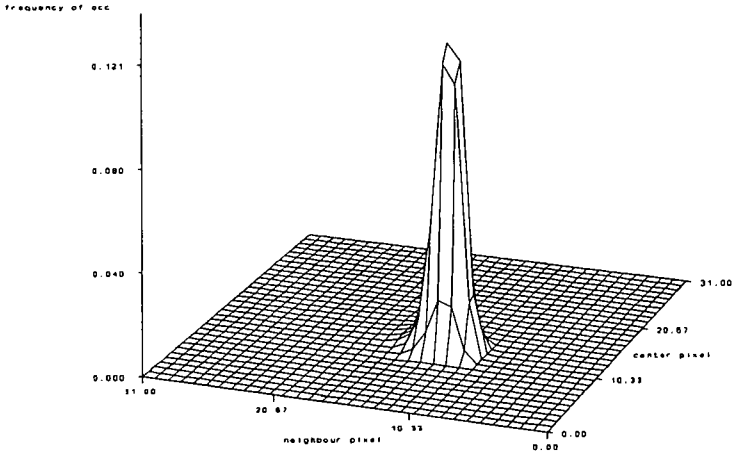


Figure 4.7.3 Plot of GLCM from image of Euphorbia exigua

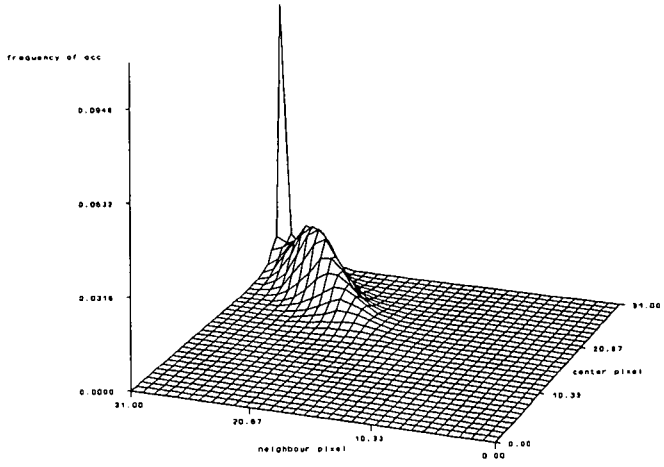


Table 4.7.1 Definition and interpretation of GLCM features proposed by Haralic et al. (1973). Notation: p(i,j) is frequency of occurrence from row i (center pixel grey level) and column j (neighbour pixel grey level), $p_x(i)$ is sum of row i, N_g is largest grey level value, μ is the mean of p(i,j), μ_x, μ_y, σ_x and σ_y are the means and standard deviations of p_x and p_y whereas p_{x+y} and p_{x-y} is the sum of the right- and left diagonals, respectively. A matrix size of 32x32 is assumed.

---

1. Angular second moment:

$$f_1 = \sum_i \sum_j (p(i,j))^2$$

Gives high values for high frequencies of occurrence in the matrix. This may arise from a very homogeneous textured surface.

2. Contrast:

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)^2 \right\}_{|i-j|=n}$$

Gives high values if frequencies of occurrence are concentrated in the two corners (1,32) and (32,1), thus indicating high grey level differences between center pixel and neighbours.

3. Correlation:

$$f_3 = \frac{\sum_i \sum_j (ij)^2 p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Values close to the diagonal (1,1) to (32,32) contribute relatively more to correlation than values further away.

4. Sum of Squares:

$$f_4 = \sum_i \sum_j (i - \mu)^2 p(i,j)$$

Gives high values if grey levels in the image are very different. That is, if frequencies of occurrence are concentrated equally in the opposite corners of the matrix.

5. Inverse Different Moment:

$$f_5 = \sum_i \sum_j \frac{p(i,j)}{1 + (i-j)^2}$$

Gives high values if frequencies of occurrence are concentrated around the diagonal (1,1) to (32,32) in the matrix. This occurs in images with very smooth transitions in grey levels.

6. Sum Average:

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i)$$

Gives high values for frequencies of occurrence concentrated in the (32,32) corner of the matrix.

7. Sum Variance:

$$f_7 = \sum_{i=2}^{2N_g} (i - f_6)^2 p_{x+y}(i)$$

Gives high values for frequencies of occurrence equally concentrated in the (1,1) and (32,32) corner.

8. Sum Entropy:

$$f_8 = \sum_{i=2}^{2N_g} p_{x+y}(i) \log(p_{x+y}(i))$$

Gives high values if frequencies of occurrence are equal in (right-) diagonal sums of the matrix.

9. Entropy:

$$f_9 = - \sum_i \sum_j p(i,j) \log(p(i,j))$$

Gives high values for frequencies of occurrence equally distributed over the matrix. This comes from a very complex image with both contrast, smooth transitions and many grey levels present.

10. Difference Variance: f_{10} = ***variance of*** $p_{x-y}(i)$

Gives high values if frequencies of occurrence are equally concentrated in the (1,32) and (32,1) corners.

11. Difference Entropy: $f_{11} = -\sum_{i=0}^{N_g-1} p_{x-y}(i)\log(p_{x-y}(i))$

Gives high values if frequencies of occurrence are equal in (left-) diagonal sums of matrix.

12. Information Measures of Correlation: $f_{12} = \dfrac{f_9 + \sum_i \sum_j p(i,j)\log(p_x(i)p_y(j))}{\max\{-\sum_i p_x(i)\log(p_x(i)),\ -\sum_j p_y(j)\log(p_y(j))\}}$

This is a very complex measure without a simple interpretation.

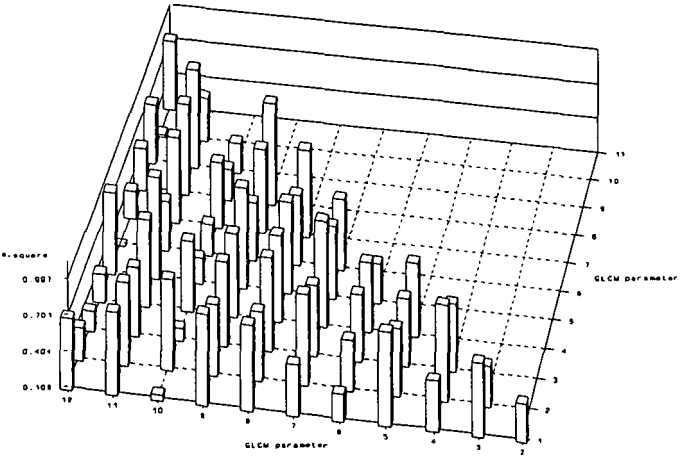


Figure 4.7.4 Overwiev of squared correlations for GLCM parameters. Parameter number follows the enumeration in Table 4.7.1.

occurrence occurs close to the diagonal, showing a high dependency between the center pixel and its neighbours even for the seed of *S.vulgaris* (Figure 4.7.1), which has a relatively high variation in the surface pattern. In Figure 4.7.3 a distinct peak appears at the highest grey level. This comes from overlightning of the light elaiosome (caruncle), and the effect is that small differences in colour between neighbours disappear. Finally, a dark smooth seed surface (without gloss) is represented in Figure 4.7.2.

Fourteen features were proposed by Haralick *et al.* (1973) to be estimated from the GLCM. Twelve of these were used in this study and they are defined and explained in Table 4.7.1. Among these features high positive and negative correlations occur. This is illustrated in Figure 4.7.4 where the squared correlation is shown as pillars. This correlation was based on seed images of ten species with 25 seeds per species. The highest negative correlations appeared for *Difference Entropy / Inverse Different Moment* (-0.983) and for *Entropy / Inverse Different Moment* (-0.979), and the highest positive correlations were found between *Sum of Squares / Sum Variance* (0.999), *Entropy /* Sum Entropy (0.988) and *Entropy / Difference Entropy* (0.976).

The Generalized Cooccurrence Matrix (GCM)

Davis *et al.* (1979) proposed a texture analysis where the matrix registered local features. This could be local edges or extrema occurring in the object surface. In general, this analysis is performed by edge operators followed by thresholding of the value obtained.

The procedure for constructing the GCM was to filter the image in four directions using the egde masks shown in Figure 4.1.1 (page 25). The highest response was evaluated against the sum of the three others. If the ratio was above a threshold of 0.60 the edge direction of the highest response was marked in a new image matrix at the same pixel position, otherwise the value is zero.

When a vertical edge was registered, the two horizontal neighbours were checked for edge information, and likewise for other edge directions. Thus, the matrix was constructed by counting combinations of 4 edge directions between the edge pixel and the two diametric neighbours. If both neighbours were egde pixels then the center edge pixel contributed two counts in the 4x4 matrix. Thus, when a pixel and its diametric neighbours had the same edge value, *i.e.*, corresponding to the diagonal in the GCM, then the original image would have shown a gradual shift in grey level in the same direction (parallel edges). If not, the image would have shown two edges crossing each other.

The features calculated from this matrix were Angular Second Moment, Contrast and Correlation. ASM expresses high variation among neighbour edge directions, Contrast is high for crossing edges and correlation is high for parallel edges in the image.

A similar matrix was constructed for registration of local extrema. For each pixel was registered whether it is the minimum or maximum in a 3x3 neighbourhood. Then the number of identified extrema was counted within a 5x5 neighbourhood (for each pixel). This score was grouped into three levels: 1) low for values 0 - 2, 2) medium for values 3 - 5, and 3) high for values above 5. The total, *i.e.*, the number of low, medium and high scores obtained, was registered in a 3x3 matrix for the cooccurrence of the 3 levels of local minima and maxima. When calculating ASM from this matrix high values are obtained for high frequencies of occurrence. Correlation is high when the number of localminima and maxima are equal within the neighbourhoods, and contrast is high when the difference between maxima and minima is high.

The Grey Level Run Length Matrix (GLRLM)

The last type of texture matrices in this study was the Grey Level Run Length Matrix presented by Galloway (1975). The number of adjacent pixels in the same direction with the same grey level was registered as the run length. Only the horizontal and the vertical directions were used in this study, and the order of the matrix was reduced to 32 grey level intervals by adding the counts of run length within the same interval. For practical reasons run lengths above 10 was divided into two or more, so that 10 became the maximum run length.

Figures 4.7.5-4.7.7 show the GLRLM of the same images as used in Figures 4.7.1-4.7.3. It might be noted that the GLRLM of the seeds of *Silene vulgaris* shows, compared to *Rumex crispus*, small run length and more grey levels. The GLRLM of seed of *Euphorbia exigua* has some resemblance to the one of *S.vulgaris* except for the small peak at maximum run length and grey level arising from the elaiosome.

Five features were proposed by Galloway (1975) to be calculated from GLRLM as presented in Table 4.7.2.

4.8 Processing time

Estimation of the processing time of the image analyses is an important issue for evaluation of the actual possibilities for system implementation. However, the processing time is dependent on the available hardware and software. For

Figure 4.7.5 Plot of GLRLM from image of Silene vulgaris.

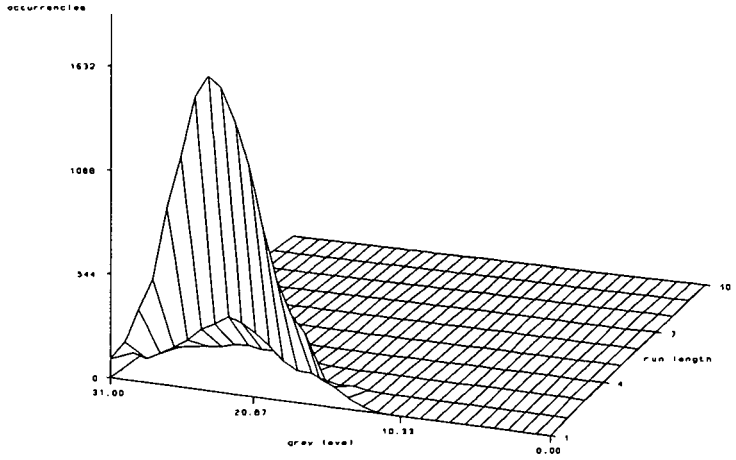


Figure 4.7.6 Plot of GLRLM from image of Rumex crispus

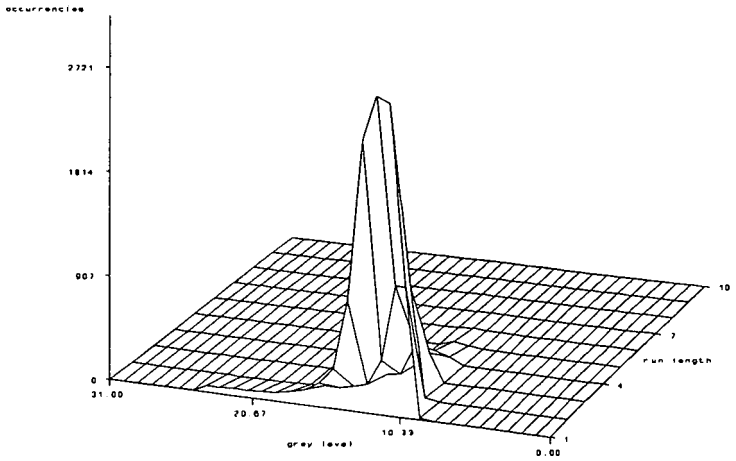


Figure 4.7.7 Plot of GLRLM from image of Euphorbia exigua

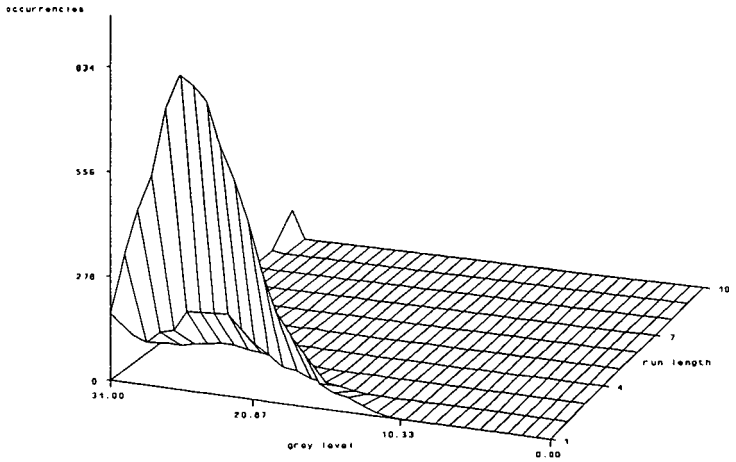


Table 4.7.2 Definition and interpretation of GLRLM features (Galloway, 1975). Notation: p(i,j) is element in row i (grey level) and column j (run length), N_g is highest number of grey levels, N_r is highest run length and N^2 is total number of pixels in the image.

1. Short Run Emphasis:	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)/j^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}$ Gives high values for short run lengths (complex texture).
2. Long Run Emphasis:	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} j^2 p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}$ Gives high values for long run lengths (homogeneous texture).
3. Grey Level Distribution:	$\frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_r} p(i,j))^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}$ Gives high values if frequencies of occurrence of run lengths are distributed over very few grey levels.
4. Run Length Distribution:	$\frac{\sum_{i=1}^{N_r}(\sum_{j=1}^{N_g} p(i,j))^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}$ Gives high values if frequencies of occurrence are distributed over very few run lengths.
5. Run Percentage:	$\frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)}{N^2}$ Inverse to average run length. Low for homogeneous texture.

example, faster algorithms, bigger machines orspecialized hardware for image analysis may be available. A few examples of specialized chips developed for image analysis may illustrate some of the possibilities.

Andersson (1985) has developed a chip which can compute the moments of an intensity array (image). From the calculation of zero to second order moments the area, center of gravity of the object and angle to major axis for the approximated ellipse is found. Furthermore, combination of moments to form invariants are useful to recognize the objects as earlier explained. Using the VLSI moment generator chip the zero to second order moments of a grey scale image of size 256x256 pixels were calculated in real time (60 frames per second). Sugai *et al.* (1987) described the architectural

Table 4.8.1 Estimates of processing time in the research system

Process description	time in seconds		average time/feature
Segmentation of window (256x256)	9.83		
Contour tracking	0.27		
Subtotal		*10.10*	
Area, center of gravity and perimeter	0.94		0.47
Other simple measurements, 7 features	0.16		0.023
Moments invariants (incl. sampling), 7 features	1.48		0.21
Time series features (differencing method)	2.37		0.34
Normalized FFT (incl. sampling), 512 complex points	6.48		0.81
Template matching	11.48		
Fractal dimension (shape), 18 features	1.92		0.11
Rapid transformation	0.44		0.055
Subtotal (shape analyses)		*25.27*	
Transfer image values (256x256 pixels) to texture analysis	0.54		
Histogram features (GLH)	0.88		0.22
GLCM features	9.01		0.75
GCM features (edge and extrema)	11.81		1.97
GLRLM features	1.54		0.308
Fractal dimension (texture), 18 features	43.00		2.39
Subtotal (texture analyses)		*66.78*	
Total		**102.15**	0.78

features of Toshiba's single-chip VLSI processor for image processing (T9506). It is capable of performing FFT (fast Fourier transform), spatial filtering, affine transform (of the type y = ax + b) and histograms in real time. Typical performance for 4096-point complex FFT was 2.0 ms, for spatial filtering (3x3 on 512x512 pixels) 262 ms, for a histogram 400 ns/pixel, and for an affine transform 400 ns/-pixel.

In Table 4.8.1 the time estimates are related to the relatively slow research system used in this study. A very time consuming task is the segmentation process, and it should be noted that a colour vision system as described in Petersen and Krutz (1992) may be needed for segmentation. This colour system was PC based and the segmentation was relatively fast (about 1-2 seconds). The average time/feature is calculated by dividing time in seconds by the number of features which were produced by the analysis. This value may not be the relevant value to know if only a part of the features from an analysis is needed. In most analyses there is general processing work to perform whether one or more features is needed. Furthermore, the processing time for the eight Fourier transform features is slightly underestimated, because calculations of the amplitudes are performed in a separate programming system, and, therefore, are not included in the presented estimates. The measurements presented in Table 4.8.1 shows that shape analysis, in general, is much faster than texture analysis. However, the GLH and GLRLM analysis are both very fast texture analyses.

5. Image classification

There are two different ways of classifying objects. One way is to find relations among the objects with the purpose of grouping them. For example, the phylogenetic relationship among plant species is expressed in the plant features which are used in taxanomic classification into genera, families etc. Statistical methods covering this kind of classification are called clustering, and the general principle is to group the observation vectors into clusters of a certain similarity. The second way of classification is to assign objects into defined groups. The statistical method for this classification is called discriminant analysis, and this is the usual kind of classification which follows image analysis for recognition purposes.

Another statistical method which may be applied to the same data set as discriminant analysis is the one-way multivariate analysis of variance. However, this method should be used if the problem is to find differences among the groups by testing for differences among the group mean vectors. Discriminant analysis is concentrating on assigning the individual observations to the respective groups, and this will be the discussed in more detail in the following sections. In this connection two main issues are important for the seed recognition project: 1) Evaluation of the discriminating power of each image analysis earlier described, and 2) improving the recognition by combining features from different analyses.

5.1 Classification results of separate analyses.

Theory

The task of discriminant analysis is to find the decision rule which assigns an object described by a number of d features to one of several groups G_i (i = 1,...,n) in a population. The special case where the feature vector is multinormally distributed will be introduced based on the presentation of Seber (1984).

The simplest case is discrimination by one feature (*e.g.*, size of object) and two groups. If we know the probability density function of this feature for each group, $f_1(x)$ and $f_2(x)$ say, the object should be assigned to the group with the highest probability density. That is, assign to group G_1 if

$$f_1(x) > f_2(x)$$

This is called the likelihood ratio method.

This method may be improved if we know that a proportion π_1 of the total population belongs to G_1 and the remaining π_2 belongs to G_2. In this case we assign to G_1 if

$$\pi_1 f_1(x) > \pi_2 f_2(x)$$

which is the Bayesian classifier.

If we assume that x is normally distributed in each group as $N(\mu_i,\sigma_i^2)$ then

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

if further $\sigma_1 = \sigma_2 = \sigma$ for the two groups then

$$f_1(x)/f_2(x) = e^{-\frac{(x-\mu_1)^2-(x-\mu_2)^2}{2\sigma^2}}$$

Setting this expression equal to one (or π_2/π_1) gives the threshold for group separation.

The corresponding expression for a multivariate normal distribution of feature vectors $\mathbf{x}_i$ with dispersion matrices $\Sigma_1 = \Sigma_2 = \Sigma$ is

$$f_1(x)/f_2(x) =$$

$$\exp[(\mu_1-\mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)]$$

In the univariate case a threshold is used for separation of groups, in the bivariate case a

line, and in the multivariate case it is the so-called hyperplanes which separate groups in the multidimensional feature space. The hyperplane for separating two groups is defined by setting the discriminant functions equal to $\log(\pi_2/\pi_1)$:

$$(\mu_1-\mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) = \log(\pi_2/\pi_1)$$

In general the distribution of the features is not known, but assumed, and the parameters are estimated from a population, often called the training set. For discrimination three special cases are considered of practical importance:
1) The resubstitution method: In this method the parameters of the discriminant functions are estimated from the same population which is classified into groups. The numbers of incorrectly classified observations m_i of the n_i observations in group G_i define the apparent error rates as

$$e_{i,app} = m_i/n_i$$
$$e_{app} = \pi_1 e_{1,app} + \pi_2 e_{2,app}$$

2) The cross-validation method (jackknifing): This method estimates the discriminant functions (*i.e.*, hyperplanes) from the sample data minus one observation. This omitted observation is then classified as the unknown observation. This is repeated until all observations are classified in this way. The corresponding error rate is

$$e_{i,c} = a_i/n_i$$
$$e_c = \pi_1 e_{1,c} + \pi_2 e_{2,c}$$

where a_i is the number of misclassified from G_i.
3) Training and test samples: This method uses a separate population (training data) for construction of the discriminant functions, and another population for testing the classification results. From the assumed normality of the distribution the error rate may be estimated by calculating the 'area' of the region where the density function is overlapped by a density function from another group. For the two group problem the region is estimated by $\hat{R}_1 = \{x: f_1(x|\hat{\theta}_1)/f_2(x|\hat{\theta}_2) > \pi_2/\pi_1\}$ where $\hat{\theta}_i$ is the estimated parameters of the probability density function. The so-called plug-in estimate for misclassification for group *1* is

$$\hat{e}_{1,plug} = \int_{\hat{R}_2} f_1(x|\hat{\theta}_1)\, dx$$

The separation of groups in the feature space depends on how well the parameters of the distribution functions are estimated. If this estimation is based on very few observations the group separation will become correspondingly uncertain. Therefore, as a rule of thumb, the number of observations in each group should exceed three times the number of estimated parameters for the group.

Another way of stabilizing the classification is by rejecting outliers. When working with biological objects as weed seeds a large variation appears due to the natural variation, damage of the seeds, seeds covered by foreign material etc. Therefore, it will be of high practical value to reject objects that do not fit very well into any group. There are two criteria for such outliers:
1) A low density estimate which shows that the observed feature values are rare for the group the observation has been assigned to.
2) A low posterior probability that the observation belongs to the assigned group. Using the notation above the posterior probability of x belonging to group *i* is

$$p(i|x) = \frac{\pi_i f_i(x)}{\sum_j \pi_j f_j(x)}$$

where the summation is over all groups.

If groups are close in the feature space, the group density near the dividing hyperplane will be relatively high, and at the same time the probability for belonging to that group is low.

Therefore, both a threshold for density and for probability should be considered when introducing a rejection mechanism in the classification procedure.

Results

Classification was performed using 40 species and 25 seeds per species, and all results were obtained by the cross-validation method. In Table 5.1.1 - 5.1.4 the classification rates of each analysis are presented. It should be emphazised that they are based on different numbers of features. The GLCM analysis has 11 features, Simple measurements 9, Fractal dimension (texture and shape), Fourier transform and Rapid transform have each 8, Autoregresion and Moments invariants have 7, GCM has 6, GLRLM has 5, and GLH has only 3 features.

In future practical applications it might be relevant to classify seeds into groups other than the species. This could be botanical groups such as genus or family, but also an unrelated group of species might be sufficient as identification. In the present study the family relations are illustrated by classifying the seeds into families as well as species.

The first important result of the shape analysis in Tables 5.1.1 - 5.1.2 is the ranking of the analyses due to the discriminatory power: Simple measurements - Fourier transformation - Rapid transformation - Autoregression - Fractal dimension - Moments invariants.

This shows that the so-called simple measurements are relatively powerful, but it should be recalled that it contains both size and shape information. If size was removed from the features in the simple measurements the classification rate (species) would decrease to 57.5 percent. On the other hand, if size was combined with the autoregressive parameters this analysis would increase classification rate from 41.7 to 63.3 percent. Thus, shape analysis is considerably improved if size information is included. Features of the two transformations performed well in this study, but the moment invariants were remarkable poor. Although the moments invariants showed no variation when rotating and enlarging a seed contour artificially, the analysis seemed to be very sensitive for biological and technical variation. Even the pattern of misclassifications differed considerably from the other analyses, and, therefore, the moment invariants are not considered in the following description.

Misclassifications were not restricted to botanical related groups although the classifications into family groups increased the rate considarably. A general problem for shape analysis is that many species have seeds of an approximate elliptic form. This causes apparently high confusions among the following species for classification into family groups:

Polygonum convolvus, Rumex acetosa, Rumex thyrsiflorus, Viola arvensis, Myosotis arvensis, Lamium amplexicaule, Veronica arvensis, Plantago major, Matricaria chamomilla, Sonchus arvensis.

Also seeds of smooth and round shapes provided high misclassifications among families:

Chenopodium album, Brassica campestris, Sinapis arvensis.

Misclassifications of certain species were mainly restricted to species of the same genus. High confusion appeared between *Rumex crispus* and *Rumex obtusifolius* but not between the other *Rumex* species. In *Euphorbia* confusions appeared mainly among *E. exigua* and *E. peplus*, in *Matricaria* high confusions were found between *M. chamomilla* and *M. matricarioides* and between *M. matricarioides* and *M. inodora*, and in *Sonchus* all three species could be highly confused. For classification into family groups a characteristic increase in classification rate was found for species of *Caryophyllaceae* and *Compositae*. Full recognition was achieved for five species in the simple measurements analysis and for one species in Rapid transform. This suggests that combining shape with size provides a more unique description of the seed. However, one species was fully recognized from shape alone in the rapid transform analysis, but this type of seed shape (heart shaped) was outstanding in this study. Two single species were in general poorly recognized, *Myosotis arvensis* and *Plantago major*, both with a smooth and approximate elliptic shape.

The ranking of the texture analyses after discriminatory power is: Fractal dimension - GLCM - GLRLM - GCM - GLH. The relative

Table 5.1.1 Cross-validation classification rates in percent. Shape analyses of 25 seeds per species.

From (fam.)	species	Simple measurem. species	Simple measurem. family	Autoregression species	Autoregression family	Moment invariants species	Moment invariants family
(Urt.)	*Urtica urens*	76	76	68	68	44	44
(Pol)	*Polygonum convolvus*	92	96	4	20	24	44
	Polygonum lapathifolium	92	92	76	76	4	16
	Rumex acetosa	92	92	44	48	4	20
	Rumex crispus	76	100	64	88	20	28
	Rumex obtusifolius	76	92	40	80	0	20
	Rumex thyrsiflorus	48	48	8	28	32	36
(Car.)	*Arenaria serpyllifolia*	100	100	72	76	36	44
	Melandrium album	64	92	56	88	8	32
	Melandrium rubrum	80	100	64	100	64	76
	Silene noctiflora	68	100	52	84	16	52
	Silene vulgaris	64	100	16	88	36	64
	Stellaria gramina	88	100	52	72	28	52
	Stellaria media	96	100	80	96	20	52
(Che.)	*Chenopodium album*	80	80	72	72	8	8
(Ran.)	*Ranunculus repens*	100	100	64	64	72	72
(Pap.)	*Papaver rhoeas*	100	100	52	52	36	36
(Cru.)	*Brassica campestris*	80	80	40	40	28	28
	Capsella bursa-pastoris	88	88	60	60	16	16
(Ger.)	*Geranium dissectum*	84	84	76	76	0	0
(Eup.)	*Euphorbia exigua*	92	100	48	72	16	20
	Euphorbia helioscopia	100	100	60	60	8	16
	Euphorbia peplus	76	76	36	60	8	12
(Vio.)	*Viola arvensis*	72	72	40	40	0	0
(Bor.)	*Myosotis arvensis*	52	52	8	8	20	20
(Lab.)	*Lamium amplexicaule*	88	88	32	32	60	60
(Sol.)	*Solanum nigrum*	64	64	36	36	8	8
(Scr.)	*Veronica arvensis*	80	80	24	24	4	12
	Veronica persica	80	80	48	48	12	28
(Pla.)	*Plantago major*	28	28	8	8	20	20
(Com.)	*Cirsium arvense*	64	100	8	72	32	64
	Chrysanthemum segetum	52	92	24	60	8	52
	Matricaria chamomilla	76	96	16	20	76	84
	Matricaria inodora	52	72	36	68	4	32
	Matricaria matricarioides	88	100	12	72	40	64
	Sinapis arvensis	76	76	72	72	60	60
	Sonchus arvensis	40	100	12	44	0	52
	Sonchus asper	72	88	16	64	56	80
	Sonchus oleraceus	84	100	20	60	28	64
	Taraxacum vulgare	100	100	52	64	92	100
Total classification rate		77.0	87.1	41.7	59.0	26.2	39.7

Table 5.1.2 Cross-validation classification rates in percent. Shape analyses of 25 seeds per species.

From (Fam.)	species	Fractal dimension species	Fractal dimension family	Fourier transform species	Fourier transform family	Rapid transform species	Rapid transform family
(Urt.)	*Urtica urens*	16	16	76	76	84	84
(Pol.)	*Polygonum convolvus*	32	60	52	92	72	92
	Polygonum lapathifolium	72	96	92	92	100	100
	Rumex acetosa	20	28	96	96	96	100
	Rumex crispus	56	84	68	100	68	96
	Rumex obtusifolius	48	76	72	100	48	100
	Rumex thyrsiflorus	12	68	48	92	68	96
(Car.)	*Arenaria serpyllifolia*	52	64	96	100	64	100
	Melandrium album	32	84	60	88	36	92
	Melandrium rubrum	80	100	72	100	52	96
	Silene noctiflora	48	72	64	100	52	96
	Silene vulgaris	56	100	44	96	36	92
	Stellaria gramina	12	32	80	88	60	84
	Stellaria media	36	44	84	100	56	80
(Che.)	*Chenopodium album*	60	60	80	80	80	80
(Ran.)	*Ranunculus repens*	4	4	96	96	88	88
(Pap.)	*Papaver rhoeas*	40	40	96	96	92	92
(Cru.)	*Brassica campestris*	64	64	68	68	56	56
	Capsella bursa-pastoris	20	20	60	60	76	76
(Ger.)	*Geranium dissectum*	44	44	84	84	76	76
(Eup.)	*Euphorbia exigua*	36	36	88	96	84	96
	Euphorbia helioscopia	48	48	72	72	72	72
	Euphorbia peplus	24	28	96	96	68	76
(Vio.)	*Viola arvensis*	32	32	76	76	56	56
(Bor.)	*Myosotis arvensis*	4	4	36	36	48	48
(Lab.)	*Lamium amplexicaule*	20	20	68	68	96	96
(Sol.)	*Solanum nigrum*	32	32	76	76	68	68
(Scr.)	*Veronica arvensis*	12	12	24	24	56	56
	Veronica persica	56	56	92	92	40	40
(Pla.)	*Plantago major*	12	12	44	44	32	32
(Com.)	*Cirsium arvense*	28	56	76	100	72	100
	Chrysanthemum segetum	20	68	60	80	56	100
	Matricaria chamomilla	20	32	20	64	76	96
	Matricaria inodora	12	44	56	68	76	88
	Matricaria matricarioides	44	76	36	92	44	92
	Sinapis arvensis	96	96	84	84	84	84
	Sonchus arvensis	16	76	36	96	48	100
	Sonchus asper	44	88	52	92	44	92
	Sonchus oleraceus	52	92	56	88	52	92
	Taraxacum vulgare	80	100	88	88	96	100
Total classification rate		37.3	54.1	68.1	83.4	65.7	84.0

Table 5.1.3 Cross-validation classification rates in percent. Texture analyses of 25 seeds per species.

From (Fam.)	species	GLH species	GLH family	GLCM species	GLCM family	GCM species	GCM family
(Urt.)	*Urtica urens*	36	36	68	68	76	76
(Pol.)	*Polygonum convolvus*	8	24	44	68	20	52
	Polygonum lapathifolium	68	80	80	84	44	60
	Rumex acetosa	68	80	76	100	80	100
	Rumex crispus	60	100	92	92	68	72
	Rumex obtusifolius	40	92	56	100	48	100
	Rumex thyrsiflorus	28	64	48	96	20	96
(Car.)	*Arenaria serpyllifolia*	52	56	36	36	20	20
	Melandrium album	20	40	40	92	40	88
	Melandrium rubrum	36	56	56	80	52	88
	Silene noctiflora	32	56	56	96	40	96
	Silene vulgaris	12	36	52	64	28	68
	Stellaria gramina	28	40	84	88	76	84
	Stellaria media	48	48	80	88	88	92
(Che.)	*Chenopodium album*	64	64	84	84	52	52
(Ran.)	*Ranunculus repens*	24	24	60	60	36	36
(Pap.)	*Papaver rhoeas*	20	20	36	36	40	40
(Cru.)	*Brassica campestris*	68	68	92	92	40	60
	Capsella bursa-pastoris	20	32	60	72	44	72
(Ger.)	*Geranium dissectum*	56	56	44	44	40	40
(Eup.)	*Euphorbia exigua*	16	24	48	52	20	52
	Euphorbia helioscopia	0	16	76	76	36	80
	Euphorbia peplus	0	20	100	100	72	72
(Vio.)	*Viola arvensis*	52	52	84	84	64	64
(Bor.)	*Myosotis arvensis*	44	44	64	64	60	60
(Lab.)	*Lamium amplexicaule*	24	24	36	36	44	44
(Sol.)	*Solanum nigrum*	28	28	12	12	56	56
(Scr.)	*Veronica arvensis*	44	72	68	100	76	84
	Veronica persica	52	96	56	92	48	48
(Pla.)	*Plantago major*	68	68	68	68	44	44
(Com.)	*Cirsium arvense*	4	20	20	40	20	48
	Chrysanthemum segetum	4	24	28	52	16	72
	Matricaria chamomilla	8	36	28	68	52	84
	Matricaria inodora	4	32	36	64	36	48
	Matricaria matricarioides	4	36	32	68	32	68
	Sinapis arvensis	36	36	32	40	16	52
	Sonchus arvensis	44	44	68	68	60	72
	Sonchus asper	20	48	44	64	4	28
	Sonchus oleraceus	24	32	60	68	36	64
	Taraxacum vulgare	4	36	20	48	88	88
Total classification rate		31.7	46.5	55.6	70.1	45.8	65.5

Table 5.1.4 Cross-validation classification rates in percent. Texture analyses of 25 seeds per species.

From (Fam.)	species	GLRLM species	GLRLM family	Fractal dimension species	Fractal dimension family
(Urt.)	*Urtica urens*	72	72	68	68
(Pol.)	*Polygonum convolvus*	12	28	28	40
	Polygonum lapathifolium	12	32	64	64
	Rumex acetosa	68	100	88	100
	Rumex crispus	76	76	88	92
	Rumex obtusifolius	8	100	52	100
	Rumex thyrsiflorus	56	96	56	96
(Car.)	*Arenaria serpyllifolia*	28	28	64	64
	Melandrium album	28	64	84	100
	Melandrium rubrum	20	76	60	100
	Silene noctiflora	44	88	80	96
	Silene vulgaris	64	92	44	100
	Stellaria gramina	68	84	84	88
	Stellaria media	72	80	76	88
(Che.)	*Chenopodium album*	56	56	80	80
(Ran.)	*Ranunculus repens*	32	32	40	40
(Pap.)	*Papaver rhoeas*	24	24	40	40
(Cru.)	*Brassica campestris*	52	76	84	84
	Capsella bursa-pastoris	76	76	36	36
(Ger.)	*Geranium dissectum*	60	60	92	92
(Eup.)	*Euphorbia exigua*	24	32	52	52
	Euphorbia helioscopia	84	84	44	44
	Euphorbia peplus	88	88	92	92
(Vio.)	*Viola arvensis*	48	48	92	92
(Bor.)	*Myosotis arvensis*	36	36	40	40
(Lab.)	*Lamium amplexicaule*	28	28	88	88
(Sol.)	*Solanum nigrum*	20	20	60	60
(Scr.)	*Veronica arvensis*	68	100	72	80
	Veronica persica	68	88	36	48
(Pla.)	*Plantago major*	84	84	84	84
(Com.)	*Cirsium arvense*	36	80	76	88
	Chrysanthemum segetum	24	68	48	68
	Matricaria chamomilla	88	96	68	80
	Matricaria inodora	16	44	44	64
	Matricaria matricarioides	28	68	36	56
	Sinapis arvensis	20	28	56	76
	Sonchus arvensis	48	76	28	44
	Sonchus asper	32	32	52	64
	Sonchus oleraceus	72	84	56	56
	Taraxacum vulgare	28	68	20	24
Total classification rate		46.7	64.8	61.3	71.7

low performance of GLRLM, GCM and GLH may partly be explained by the low number of features in these analyses. The general impression is that texture analysis in general has lower classification rates than shape analysis although they are based on many more observations. However, it should not be concluded that texture analysis in general is weaker than shape analysis. The texture analyses were suffering from a lower number of features and the lack of features reflecting seed colour or average greytone, which were removed from the analyses because the illumination intensity could not be standardized using the equipment described earlier. And in a separate investigation (Petersen and Krutz, 1992) it was shown that colour information was important for weed seed discrimination. However, it is possible to compare the two analyses of fractal dimension of which the one applied to texture increased the classification rate 24 percent.

The pattern of misclassifications in the texture analyses shows in general high confusion among species of smooth surface and among species of structured surface with little confusion between the two types.

When only using texture analysis it is, of course, possible to detect confusion between seeds of very different shapes. The high confusion between the round *Chenopodium album* and the heart shaped *Polygonum lapathifolium* may illustrate that. However, some confusions between smooth and structured species were detected, such as between *Lamium amplexicaule* and *Taraxacum vulgare* and between *Sinapis arvensis* and *Arenaria serpyllifolia*. In the first case gloss and colour spots probably influenced the confusion, and in the second case a loss of contrast of the structured seed (*e.i., Arenaria serpyllifolia*) due to use of light background is suspected to cause the confusion. Finally, high confusions occurred between species with big and small seeds, such as between *Polygonum convolvus* and *Capsella bursa-pastoris* and between *Ranunculus repens* and *Papaver rhoeas*.

This pattern of misclassification confirms the initial suggestion that some kind of combined analysis of size, shape and texture will be relevant. Furthermore, the pattern of misclassification is different from one shape analysis to another and from one texture analysis to another. Therefore, a better classification rate might be achieved by constructing a combined shape analysis or combined texture analysis.

5.2 Combination of image analyses

When combining analyses it is important to reduce the number of features, so that poor discriminators are detected and eliminated. After this reduction of features the remaining may be combined to form a more powerful analysis without increasing the number of features considerably. Another method for combination of analyses is to use a hierarchical scheme where one classification follows another. In this scheme the succeeding classification should take advantage of the results obtained from the preceeding classification. Both approaches will be illustrated by the weed seed data.

Theory

A general strategy for variable reduction is to collect variables which are merely suspected to be of importance to provide the basis for discrimination among the groups in the population. During the investigation it is then the intention to eliminate those variables which are found to be redundant or irrelevant.

The stepwise techniques in discriminant analysis parallel the regression methods for variable selection. Within this concept of evaluation variables which contribute additional information to group separation there are three approaches: Forward selection, backward elimination, and stepwise selection. In forward selection the variables are assessed by a chosen criterion, and the best variable is selected. The remaining variables are then added one at a time based on the optimization of the criterion. This continues until a stopping rule is met, where none of the remaining variables contri-

bute significantly to the group separation. In backward elimination the examinations starts with all variables, and the principle is to examine whether a variable supplies additional information independently of the remaining variables. The variable of the lowest significance due to the criterion is eliminated in each step. The forward selection and backward elimination are combined in the stepwise selection starting with the forward selection. For each variable selected backward elimination is applied to the selected subset of variables. Thus it is possible to eliminate selected variables which in combination with others turn out to be redundant. The procedure terminates when none of the selected variables can be excluded, and no further variables can be included.

A well-known disadvantage of the stepwise technique is that the selected subset is not compared with the original variable set, and that there is no guarantee that the selected subset will be optimal with respect to the chosen criterion when the optimization of the criterion is performed in each step. In addition, in forward selection and backward elimination there is no evaluation of the selected variable at later steps in the procedure.

Different criteria have been proposed for the selection procedure. In McKay and Campbell (1982) the use of canonical variable (as previously described) is illustrated. In short, the canonical variates are for grouped data what the principal components are for ungrouped data. In most cases only the first canonical variable should be considered, but if other variables are important a certain linear combination of coefficients for all the canonical variates might be considered to produce a score. The variable with lowest canonical coefficient is eliminated, and the canonical variate coefficients are recalculated after each deletion.

The principle from analysis of covariance is another that may also be used for selection of variables. If the variable x_{p+1} is tested for increase of group separation provided by the variables $x_1,...,x_p$, the new variable is treated as the response and the older variables as covariates. A significant increase in group separation is now tested as a difference in group mean values (*i.e.*, parallel regression lines when plotted) by a partial F statistic.

Finally, the often used method called Wilks lambda should be mentioned. In a one-way multivariate analysis of variance with g groups and n_i observations in G_i the within groups matrix is

$$\boldsymbol{E} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_{i.})'$$

and the between groups matrix

$$\boldsymbol{H} = \sum_{i=1}^{g} n_i (\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})'$$

Then Wilks lambda for p variables in the feature vector is

$$\Lambda(1,...,p) = \frac{|\boldsymbol{E}(1,...,p)|}{|\boldsymbol{E}(1,...,p) + \boldsymbol{H}(1,...,p)|}$$

and the multiplicative increment

$$\Lambda(p+1) = \frac{\Lambda(1,...,p+1)}{\Lambda(1,...,p)}$$

is obtained by adding the variable x_{p+1} to the preceeding p variables. This is also called the partial Λ-statistic. The significance of change in group separation can be tested by the F-statistic

$$F = \frac{n-g-p}{g-1} \frac{1-\Lambda(p+1)}{\Lambda(p+1)}$$

As shown Wilks lambda takes into consideration both differences among groups and cohesiveness within groups.

A special problem occurs for the F tests because they assume that the variable added to the subset is an arbitrary variable rather than the one which maximizes F (Rencher and Larson, 1980). Actually, the F-statistic does not

Table 5.2.1 Selected features from a stepwise selection procedure and the corresponding classification rates (cross-validation). Forty species and 25 seeds per species were used.

Group I *(Shape)*	**Group II** *(Shape)*	**Group III** *(Texture)*
C.v.	Rapid transform, 3'rd feat.	Short Run Emphasis
Size	Size	Fractal dim., 4'th feat.
Fractal dim., 1'st feat.	Fourier transform, 6'th feat.	Skewness
Eccentricity	Fourier transform, 8'th feat.	Fractal dim., 1'st feat.
Compactness	Rapid transform, 5'th feat.	Grey Level Distribution
Autoregression (θ_2)	Rapid transform, 2'nd feat.	Run Length Distribution
Moment invariants (M_1')	Fourier transform, 4'th feat.	Fractal dim., 2'nd feat.
Autoregression (θ_1)	Rapid transform, 4'th feat.	Variance
Classification rates (%):		
81.9	84.6	86.9

have an F distribution when it is maximized in each step, and is therefore biased under the selection. A consequence of this choosing maximal F values is that too many variables are likely to be included in the subset. To compensate for this effect a more conservative significance level might be used by dividing α by the number of remaining variables for possible inclusion.

Before entering further examinations a variable is tested in a tolerance test as one minus the squared multiple correlation between that variable and all variables already selected. If the new variable is a linear combination (or almost) it is excluded because inaccuracies may occur from computing the inverse of almost non-singular matrices.

The hierarchical method

One strategy for combining analyses is the hierarchical approach. This strategy was proposed in Petersen (1991) for combining results of two different classification procedures, *i.e.*, assigning observations due to smallest mismatch and applying the discriminant analysis allocation rules, respectively. In this system each subsequent classification procedure should include information about the results of the preceeding classifications. The method is simply to transform the results of the first classification into prior probabilities for the following classification. The main advantage of this procedure is that it introduces flexibility. Thus, it makes possible first to choose a small and fast analysis and later extend it with a new analysis without including the features of the already performed analysis again. This strategy could include the two rejection conditions (*i.e.*, the density estimate and the probability estimate) to obtain a more stable process and higher results of the accepted observations (seeds).

In the present investigation the full analysis is divided into three steps corresponding to combination of three analyses. In each step a cross-validation method is used and initially the prior probabilities are equal in all steps. However, in steps II and III the posterior probabilities are multiplied with the corrected prior probabilities (calculated from the preceeding classification results) to obtain the corrected classification result (*i.e.*, a corrected posterior probability). When including rejection conditions in this approach, it becomes slightly more complicated. After the first two classification steps there are

the following possibilities:
After STEP I:
- accepted seeds
- rejected seeds

After STEP II:
- seeds accepted in step I and II
- seeds accepted in step I
- seeds accepted in step II
- seeds rejected in step I and II

This means that the actual prior probability used in step II and III depends on which of the above groups the seed belongs to.

Finally, the corrected prior probabilities are calculated from the distribution of seeds classified into species. For example, if group A after the first classification contains observations from groups A, B, and C in the ratio 14:4:2, then prior probabilities for the next classification are 0.7, 0.2 and 0.1, respectively, for all observations classified to group A in the first step.

This system has a conservative effect. If the first classification process assigns an observation from group B to group A, the next classification requires a relatively high probability to replace it, because the observations normally receive a high prior probability for belonging to the same group. If the first and second classifications are identical (*i.e.*, use the same features), no replacements to other groups will occur unless the prior probability for belonging to the first assigned group is below equal prior probabilities. Therefore, to make many changes in group assignment the second classification should be based on features which are independent of the features in the first classification.

Results

It is now the intention to try different approaches of combining image analyses to obtain a broader view of the existing possibilities. First, features related to shape or texture could be grouped separately in order to construct relatively simple and fast analyses. If wished, these analyses could be combined to a more complex analysis. Alternatively, the best combination of all existing features could be collected in one large analysis.

The results presented in section 5.1 may suggest that two 'natural' groups of shape analysis could be constructed: 1) features mainly calculated from a time series representation, and 2) features from the two transforms. To each of the two groups the size feature was added, and the features of moment invariants and the compactness were added to the first group. A third group was constructed of all the texture features. In Table 5.2.1 the groups are called group I, II, and III, respectively.

Among the textural features it might be attempted to remove all features of one or more analyses without losing significantly in discriminatory power. This would contribute considerably to a faster combined analysis. Actually, it was found that an initial removal of all GLCM and GCM features only decreased classification rates 0.7 percent compared to the best obtained. Therefore, a combined analysis of features selected from the remaining texture analyses was constructed (group III in Table 5.2.1).

In Table 5.2.2 the results of the hierarchical combination are shown for the rejection thresholds of: 1) a density estimate equal to 10% of the estimated group mean density, and 2) the posterior probability of belonging to the allocated group equal to 0.5. In step I a relative high classification rate was obtained among the accepted seeds, but also a very high percentage of rejected seeds. When combining this classification with the next in step II only a slight increase in classification rate from 88.1 to 89.4 occured, but the amount of accepted seeds increased considerably. In the third step this trend continued so that the classification rate of accepted seeds reached 91.3 percent and only 3 percent of the seeds were rejected. During this procedure the rate of the individual species generally increased, but occasionally it also decreased. Such decrease was mainly caused by acceptance and misclassification of previously rejected seeds, but also it occurred that correctly classified seeds in the first step became misclassified in the next step. Thus, Table 5.2.2

Table 5.2.2 Hierarchical classification in three steps with rejection. See text for details.

From (Fam.)	species	STEP I. species	STEP I. family	STEP II. species	STEP II. family	STEP III. species	STEP III. family
(Urt.)	*Urtica urens*	95	95	96	96	96	96
(Pol.)	*Polygonum convolvus*	100	100	100	100	92	100
	Polygonum lapathifolium	100	100	100	100	100	100
	Rumex acetosa	100	100	100	100	96	100
	Rumex crispus	78	100	79	100	96	100
	Rumex obtusifolius	83	100	83	100	88	100
	Rumex thyrsiflorus	86	86	96	96	88	100
(Car.)	*Arenaria serpyllifolia*	100	100	100	100	96	96
	Melandrium album	77	100	88	100	83	100
	Melandrium rubrum	82	100	76	100	79	100
	Silene noctiflora	65	100	63	100	80	100
	Silene vulgaris	76	100	70	100	65	96
	Stellaria gramina	95	100	100	100	96	100
	Stellaria media	80	100	96	100	88	100
(Che.)	*Chenopodium album*	100	100	100	100	100	100
(Ran.)	*Ranunculus repens*	100	100	100	100	96	96
(Pap.)	*Papaver rhoeas*	100	100	100	100	96	96
(Cru.)	*Brassica campestris*	95	95	86	86	80	80
	Capsella bursa-pastoris	96	96	96	96	100	100
(Ger.)	*Geranium dissectum*	100	100	100	100	100	100
(Eup.)	*Euphorbia exigua*	100	100	96	100	100	100
	Euphorbia helioscopia	100	100	100	100	92	92
	Euphorbia peplus	83	100	100	100	100	100
(Vio.)	*Viola arvensis*	100	100	90	90	88	88
(Bor.)	*Myosotis arvensis*	93	93	75	75	83	83
(Lab.)	*Lamium amplexicaule*	100	100	100	100	100	100
(Sol.)	*Solanum nigrum*	85	85	96	96	88	88
(Scr.)	*Veronica arvensis*	89	89	87	87	88	92
	Veronica persica	88	88	95	95	91	91
(Pla.)	*Plantago major*	43	43	35	35	88	88
(Com.)	*Cirsium arvense*	65	100	83	100	92	100
	Chrysanthemum segetum	58	100	91	100	83	100
	Matricaria chamomilla	90	100	88	100	88	100
	Matricaria inodora	77	85	88	88	75	90
	Matricaria matricarioides	87	100	88	100	96	100
	Sinapis arvensis	91	91	87	87	92	92
	Sonchus arvensis	57	100	76	100	100	100
	Sonchus asper	95	100	86	100	96	100
	Sonchus oleraceus	88	100	82	100	100	100
	Taraxacum vulgare	100	100	100	100	96	100
Total classification rate		88.1		89.4		91.3	96.5
% rejected		24.3		11.1		3.0	

shows that the step III classification (texture analysis) has improved the classification rates of *Plantago major, Silene noctiflora* and the three *Sonchus* species significantly, but also caused a decrease for several other species.

If the hierarchical method is found suitable for the practical implementation, it might be valuable to obtain more knowledge concerning the effect of the rejection thresholds. In Table 5.2.3 the results of increasing the group probability threshold from 0.5 to 0.7 respective increasing the density estimate threshold from 10% to 50% of the group mean density are presented. Increasing the rejection thresholds caused in both cases an increase in total classification rate and in percentage rejected seeds. However, the increase in group probability threshold was more favourable causing a slightly higher classification rate and a lower increase in percentage rejected seeds.

Finally the 'best' possible combination of features from the total set should be found. From a series of stepwise selections the first 31 features were found as listed in Table 5.2.4. This number of features causes the problem of stability in estimating the discriminant functions. A relatively simple way of evaluating this stability, and hence selecting the optimal number of features is to plot cross-validation classification rates versus feature numbers as shown in Figure 5.2.1 (no rejection used). It appeared that the maximum classification rate of 97.7 percent was achieved when using 25 features. This was a remarkably high number of features compared to the group size in the data set. The corresponding resubstitution classification rates are also presented in this figure, and as expected they showed higher values.

Table 5.2.3 Hierarchical classification at different rejection conditions.

Rejecting thresholds			
Probability	0.5	0.7	0.5
Density	0.1	0.1	0.5
Rates in %	91.3	93.6	93.3
Rejected seeds in %	3.0	6.1	15.9

In Table 4.8.1 the estimated average processing time/feature was presented. These estimates were accumulated for the features selected in Table 5.2.4 and Figure 5.2.1. The accumulated values are shown in Figure 5.2.2, but it should be emphasized, that they under-

Table 5.2.4 Ranking of the best 31 features from a stepwise selection procedure

1) C.v.
2) Size
3) Short Run Emphasis (GLRLM)
4) Contrast (GLCM)
5) Skewness (GLH)
6) Compactness
7) Run Length Distribution (GLRLM)
8) Fract. dim. texture, 1'st feat.
9) Fourier transform, 8'th feat.
10) Fract. dim. texture, 3'rd feat.
11) Grey Level Distribution (GLRLM)
12) Fract. dim. shape, 1'st feat.
13) Correlation (GLCM)
14) Fourier transform, 6'th feat.
15) Fract. dim. texture, 2'nd feat.
16) Fract. dim. texture, 6'th feat.
17) Contrast, local extrema (GCM)
18) Rapid transform, 4'th feat.
19) Eccentricity
20) Fourier transform, 4'th feat.
21) Rapid transform, 2'nd feat.
22) Rapid transform, 3'rd feat.
23) Correlation, local extrema (GCM)
24) Moment invariants, 1'st feat.
25) Difference Variance (GLCM)
26) Angular Second Moment (GLCM)
27) Sum Entropy (GLCM)
28) Long Run Emphasis (GLRLM)
29) Difference Entropy (GLCM)
30) Inverse Different Moment (GLCM)
31) Run Percentage (GLRLM)

estimate the real processing time of the image analyses in the research system. In many analyses there is a general processing part, which is constant whether one or many features are needed. It should also be considered that image acquisition, segmentation and classification were not included in the calculation of these processing times.

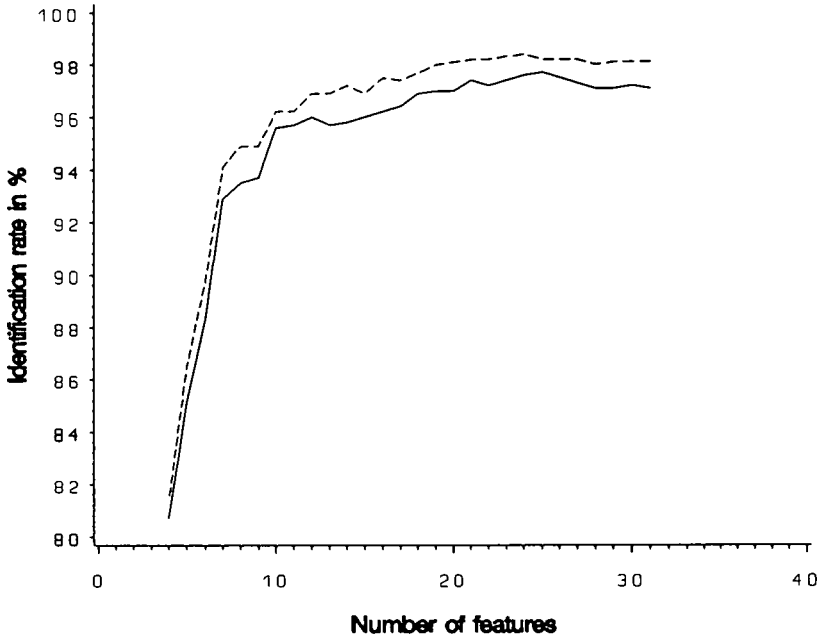


Figure 5.2.1 Classification rates from cross-validation (solid line) and resubstitution (dashed line) versus number of features. Forty species and 25 seeds per species used. The features are listed in Table 5.2.4.

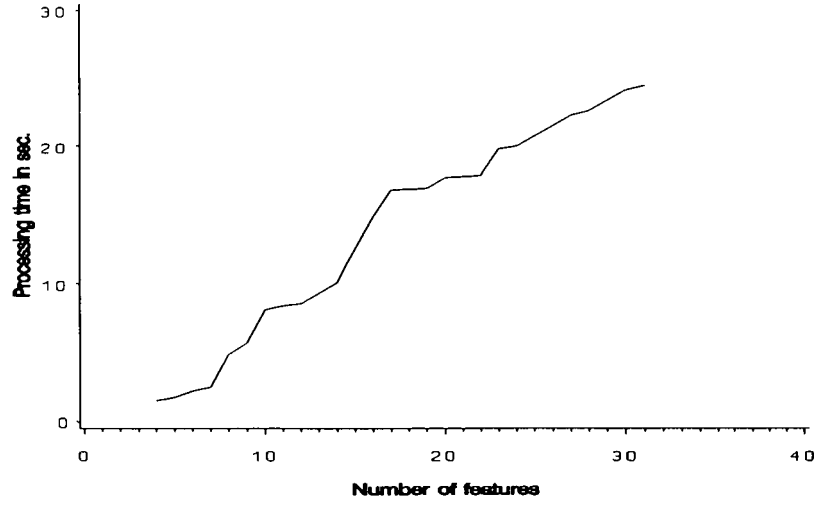


Figure 5.2.2 Accumulated average processing time of the features in Figure 5.2.1 and Table 5.2.4.

6. Discussion and conclusions

This chapter discusses some difficulties related to a practical implementation based on the experiences obtained in the seed project. Finally, the conclusions are drawn to assess the the results of the image analyses applied to microscope images of weed seeds.

6.1 Future directions

Image analysis has shown to be a highly reliable method for automatic recognition of weed seeds. However, for practical implementations of such a system some problems remain to be solved of which the major ones are:

1) Transportation of seeds from a sample to the video camera.
2) Automatic focusing to ensure a sharp image of the seeds.
3) Automatic magnification to a defined state.
4) Standardization of illumination intensity.
5) Requirements to speed and reliability.

Westerlind (1984) presented a system for transportation of bigger seeds (*i.e.*, grains) one by one from a sample to a video camera. This system had several containers of grain samples which in turn were emptied into a vibrating unit. The vibration caused a separation and upward movement of the single grains to a rotating plate with a v-shaped track. The grain was delivered to this track, and by rotation moved it to the video camera. A computer starts the image analysis and decides whether the seed is to be accepted or rejected. During this processing time the seed continues to rotate until a point is reached where it has to be removed from the plate. The seed is then removed by jet-air into one box if accepted and into another if rejected.

This system is now placed at the Danish Seed Testing Station, and it seems to work well for cereal grains, but certain changes will be necessary for weed seeds due to the smaller size with larger variation and the large colour variation.

The first difficulty arising from the seed variation is the requirement of different illumination intensities in order to obtain a satisfactory image. Colour cameras might be the solution, but this has not yet been verified. However, colour cameras are recommended for obtaining a uniform and fast segmentation procedure of both light and dark seeds. For colour segmentation it should be noticed that a serial acquisition of the three colour channel images usually will require a short movement stop (*e.g.*, of the rotating plate) for each seed passing by. Use of high-speed cameras might overcome this disadvantage.

Furthermore, in order to obtain a satisfactory size of the seed image, it might be necessary to adjust the optical magnification when shifting from a big to a small seed and *vice versa*. This implies several difficulties. First, texture analysis is sensitive to degree of magnification, which means that, if the magnification is changed, all seeds should be magnified to a certain standard area if texture is included in the analysis. Such changes in degree of magnification will require time and consequently a prolonged movement stop of the seeds. Secondly, changes in degree of magnification above a certain degree provide blurring of the image, and, therefore, both magnification and focusing should be controlled by the computer. Alternatively, the seeds could initially be sorted mechanically according to size, and then go to different cameras. This will split up the sample in subsamples, and increase the complexity of the system.

The speed problem is dependent on many factors such as hardware performance, software optimization, number of analyses needed, number of groups (*i.e.*, species) included and reliability demands. If speed is critical, dividing the process into an image acquisition and an image processing section might be considered. In the first section the image is captured and stored on a disk. Later the segmentation, analysis and classification are performed by one or more computers. This approach looses the possibility of dividing the real seeds into an accepted and

rejected group, of which the latter goes to human identification. However, this might not be a problem, if the image of the seed itself is satisfactory for human identification.

An alternative to a full automatic system might be a semi-automatic one. A system where humans place the seeds under the microscope and capture the image should be straightforward to implement. However, control of the magnification to a standard area is still regarded as important. Such a system could be used for employment of untrained staff or be part of training courses in seed identification.

6.2 Conclusion

Automatic recognition of weed seeds is an important issue for seed testing and agricultural weed research. Therefore, fundamental investigations with image analysis for seed description were initiated. Seeds of forty species of weeds were selected for evaluation of the image analyses. Microscope images with one seed per image were captured by a black and white CCD-camera. The images were segmented using a general thresholding method followed by a gentle smoothing of the seed boundary. The investigation concentrated on evaluation of different image analyses applied to the weed seeds.

Certain limitations in the vision system were considered important:

1) It was not possible to segment dark and light seeds in black and white images using the same background colour.

2) Use of a uniform illumination intensity was not possible in order to obtain a satisfactory image quality.

3) The texture analyses are sensitive to the degree of magnification.

The first mentioned limitation may be solved by use of colour cameras, but in the present investigation the background colour was shifted manually. Unfortunately, when using the light background colour a more blurred image was produced. The consequence of the second limitation was a removal of all features which reflected seed colour from the textural characterization of the seeds. Eventually, the third limitation was overcome by setting the degree of magnification constant for all species belonging to the same species in this study.

During the project a few methods of shape analysis were improved and adjusted to the weed seed problem. The general difference to many other applications is that the minor structures created by the out-bulging of the seed coat cells or seed appendices were important for the identification. Consequently, these minor structures should be represented by the features of the analysis. This was the background for the special modification of the autoregressive method and the Fourier transform. When applied to 40 species, 25 seeds per species, the performance of the various shape analyses ranged from 26.2% to 77.0% in classification rate (the cross-validation method). The performance of the textural analyses ranged from 31.7% to 61.3%. Ranking the analyses by discriminatory power leads to the following list:

1) Simple measurements
2) Fourier transform
3) Rapid transform
4) Fractal dimensions (texture)
5) Grey Level Cooccurrence Matrix (GLCM)
6) Grey Level Rnu Length Matrix (GLRLM)
7) Generalized Cooccurrence Matrix (GCM)
8) Autoregression
9) Fractal dimensions (shape)
10) Grey Level Histogram (GLH)
11) Moment invariants

However, it should be mentioned that these analyses are based on a different number of features, which means that single features of poorly performing analyses might be better than single features of well performing analyses.

When classifying the seeds to family groups an increment in identification rate of 15.3% on average of all analyses was obtained. However, certain species showed confusions mainly within the family group, but others showed high confusions with species of foreign families.

Estimates of processing time showed that there was considarable variation among all analyses, but, in general, shape analyses were faster than texture. On average a shape feature

only used about 16% of the processing time of a texture feature. In this study the average processing time per feature (excluding segmentation) was 0.78 seconds.

In addition, it was concluded that a combination of features describing both size, shape and texture would be favourable for the recognition. A hierarchical classification approach including two rejecting conditions was studied. This method should provide an increase in classification rate by combining features and rejecting poorly classified seeds. Furthermore, the hierarchical combination could introduce flexibility in the identification process. However, the classification rate was only moderate, which may partially be explained by the conservative effect in reassignment of misclassified seeds. The best result obtained was a 93.6% recognition using a rejecting probability threshold of 0.7 and a rejecting density threshold of 0.1 times the average group density.

Another approach was to combine features in a single discriminant analysis. For this purpose the 31 'best' combined features were selected by a series of stepwise selection procedures. The cross-validation classification rates showed a maximum of 97.7% when including 25 features in the analysis. The distribution of these features was 1 size feature, 10 shape features and 14 texture features.

The classification results obtained in this study seem encouraging for continued research in and construction of an automatic system for weed seed identification.

References

Andersson, R.L. 1985. Real-Time Gray-Scale Video Processing Using a Moment-Generating Chip. *IEEE Journal of Robotics and Automation* 1(2), 79-85.

Barthlott, W. 1981. Epidermal and seed surface characters of plants: systematic applicability and some evolutionary aspects. *Nordic Journal of Botany* 1, 345-355.

Bennett, J.R. and J.S. Mac Donald. 1975. On the Measurement of Curvature in a Quantized Environment. *IEEE Transactions on Computers* 24(8), 803-820.

Berlage, A.G., T.M.Cooper and J.F.Aristazabal. 1988. Machine Vision Identification of Diploid and Tetraploid Ryegrass Seed. *Transaction of the ASAE* 31(1), 24-27.

Bhatnagar, S.P. and B.M. Johri. 1972. Development of Angiosperm Seeds. *In:* Kozlowski, T.T. (ed.) Seed Biology. Academic Press, New York, London. Vol. I. pp. 77-150.

Bocquet, G. 1959. The Campylotropous Ovule. *Phytomorphology* 10, 222-227.

Boesewinkel, F.D. and F. Bouman. 1984. The Seed: Structure. *In:* Johri, B.M. (ed.) Embryology of Angiosperms. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, pp. 567-610.

Borisenko, V.I., A.A.Zlatopol'skii and I.B. Muchnik (1987). Image Segmentation (state-of-the-art survey). *Automation and Remote Control* 48, 837-879.

Bribiesca, E. and A. Guzman. 1980. How to Describe Pure Form and How to Measure Differences in Shapes Using Shape Numbers. *Pattern Recognition* 12, 101-112.

Brill, E.L. 1968. Character Recognition via Fourier Descriptors. Presented at WESCON, Session 25, Qualitative Pattern Recognition Through Image Shaping, Los Angeles, Calif., August.

Brogan, W.L. and A.R. Edison. 1974. Automatic Classification of Grains via Pattern Recognition Techniques. *Pattern Recognition* 6, 97-103.

Chen, C., Y.P.Chiang and Y.Pomeranz. 1989. Image Analysis and Characterization of Cereal Grains with a Laser Range Finder and Camera Contour Extractor. *Cereal Chem.* 66(6), 466-470.

Chow, C.K. and T. Kaneko. 1972. Boundary Detection of Radiographic Images by a Threshold Method. *In* 'Frontiers of Pattern Recognition', ed. S. Watanabe, Academic Press, New York, London, 61-82.

Davis, L.S., S.A. Johns and J.K. Aggarwal. 1979. Texture Analysis Using Generalized Co-Occurrence Matrices. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 1(3), 251-259.

Das, M., M.J. Paulik and N.K. Loh. 1990. A Bivariate Autoregressive Modelling Technique for Analysis and Classification of Planar Shapes. *IEEE Transactions on Pattern Recognition and Machine Intelligence* 12(1), 97-102.

Diggle, P.J. 1990. Time Series (A Biostatistical Introduction). *Oxford Statistical Science Series, Clarendon Press, Oxford.*

Draper, S.R. and P.D.Keefe. 1989. Machine vision for the characterization and identification of cultivars. *Plant Varieties and Seeds* 2, 53-62.

Draper, S.R. and A.J.Travis. 1984. Preliminary observations with a computer based system for analysis of the shape of seeds and vegetative structures. *J. natn. Inst. agric. Bot.* 16, 387-395.

Dubois, S.R. and F.H. Glanz. 1986. An Autoregressive Model Approach to Two-Dimensional Shape Classification. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 8(1), 55-66.

Dudani, S.A., K.J. Breeding and R.B. McGhee. 1977. Aircraft Identification by Moment Invariants. *IEEE Transactions on Computers* 26(1), 39-46.

Eom, K.-B. and J. Park. 1990. Recognition of Shapes by Statistical Modeling of Centroidal Profile. *Proceedings 10th International Conference on Pattern Recognition, 16-21*

June, Atlantic City, New Jersey, USA, Vol. 1, 860-864.

Eriksson, T., T. Lagerwall and O. Beckman. 1970. Fysik 1 (Mekanik Värmeläre). *Almqvist & Wiksell, Stockholm.*

Fahn, A. 1987. Plant Anatomy. *Pergamom Press, Oxford, New York, Beijing, Frankfurt, Sao Paulo, Sydney, Tokyo, Toronto.*

Freeman, H. 1961. On the Encoding of Arbitrary Geometric Configurations. *IRE Transactions on Electronic Computers* 10(2), 260-268.

Freeman, H. 1974. Computer Processing of Line-Drawing Images. *Computing Surveys* 6(1), 57-97.

Freeman, H. 1978. Shape Description via the Use of Critical Points. *Pattern Recognition* 10, 159-166.

Galloway, M.M. 1975. Texture Analysis Using Gray Level Run Lengths. *Computer Graphics and Image Processing* 4, 172-179.

Gonzalez, R.C. and P. Wintz. 1987. Digital Image Processing. Addison-Wesley Publishing Company, London, Amsterdam, Don Mills (Ontario), Sydney, Tokyo.

Goshtasby, A. 1985. Template Matching in Rotated Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 7(3), 338-344.

Granlund, G.H. 1972. Fourier Preprocessing For Hand Print Character Recognition. *IEEE Transaction on Computers* 21, 195-201.

Haralick, R.M., K. Shanmugam and I. Dinstein. 1973. Textural Features for Image Classification. *IEEE Transaction on Systems, Man, and Cybernetics* 3(6), 610-621.

Horowitz, S.L. and T. Pavlidis. 1976. Picture Segmentation by a Tree Traversal Algorithm. *Journal of the Association for Computing Machinery* 23(2), 368-388.

Hsia, T.C. 1981. A Note on Invariant Moments in Image Processing. *IEEE Transactions on Systems, Man, and Cybernetics* 11(12), 831-834.

Hu, M.-K. 1961. Pattern Recognition by Moments Invariants. *Proceedings of the IRE* 49(9), 1428.

Hu, M-K. 1962. Visual Pattern Recognition by Moment Invariants. *IRE Transactions on Information Theory*, 8(2), 179-187.

Inoué, S. 1986. Video Microscopy. *Plenum Press, New York and London.*

Jensen, H.A. 1969. Content of Buried Seeds in Arable Soil in Denmark and its Relation to the Weed Population. *Dansk Botanisk Arkiv* 27(2),1-56.

Jensen, H.A. 1989. The Seed Testing Board during 100 years. *In* Report from the Danish State Seed Testing Station for the 118. working year from the 1st July 1988 to 30th June 1989. (Danish with English summary). pp. 73-78.

Kapil, R.N., J.Bor and F.Bouman. 1980. Seed appendages in Angiosperms. I. Introduction. *Bot. Jahrb. Syst.* 101(4), 555-573.

Kartikeyan, B. and A. Sarkar. 1989. Shape Description by Time Series. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 11(9), 977-984.

Keefe, P.D. 1990. Observations concerning shape variation in wheat grains. *Seed Sci. & Technol.* 18, 629-640.

Keefe, P.D. and S.R.Draper. 1986. The measurement of new characters for cultivar identification in wheat using machine vision. *Seed Sci. & Technol.* 14, 715-724.

Keefe, P.D. and S.R.Draper. 1988. An auto mated machine vision system for the morphometry of new cultivars and plant genebank accessions. *Plant Varieties and Seeds* 1, 1-11.

Klitgård, K. 1989. The Seed Testing Board during 100 years. *In* Report from the Danish State Seed Testing Station for the 118. working year from the 1st July 1988 to 30th June 1989. (Danish with english summary). pp. 153-172.

Kohler, R. 1981. A Segmentation System Based on Thresholding. *Computer Graphics and Image Processing* 15, 319-338.

Kuhl, F.P. and C.R. Giardina. 1982. Elliptic Fourier Features of a Closed Contour. *Computer Graphics and Image Processing* 18, 236-258.

Levine, M.D. 1985. Vision in Man and Machi-

ne. *McGraw-Hill Book Company.*

Lin, C.-S. and C.-L. Hwang. 1987. New Forms of Shape Invariants from Elliptic Fourier Descriptors. *Pattern Recognition* 20(5), 535-545.

Ma, J., C.-K. Wu and X.-R. Lu. 1986. A Fast Shape Descriptor. *Computer Vision, Graphics, and Image Processing* 34, 282-291.

Maitra, S. 1979. Moment Invariants. *Proceedings of the IEEE* 67(4), 697-699.

Mardia, K.V. and T.J. Hainsworth. 1988. A Spatial Thresholding Method for Image Segmentation. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 10(6), 919-927.

McKay, R.J. and N.A. Campbell. 1982. Variable selection techniques in discriminant analysis. I. Description. *British Journal of Mathematical and Statistical Psychology*, 35, 1-29.

Myers, D.G. and K.J.Edsall. 1989. The applica tion of image processing techniques to the identification of Australian wheat varieties. *Plant Varieties and Seeds* 2, 109-116.

Neuman, M., H.D.Sapirstein, E.Shwedyk and W.Busuk. 1987. Discrimination of Wheat Class and Variety by Digital Image Analysis of Whole Grain Samples. *Journal of Cereal Science* 6, 125-132.

Neuman, M., H.D.Sapirstein, E.Shwedyk and W.Busuk. 1989a. Wheat Grain Colour Analysis by Digital Image Processing I. Methodology. *Journal of Cereal Science* 10, 175-182.

Neuman, M., H.D.Sapirstein, E.Shwedyk and W.Busuk. 1989b. Wheat Grain Colour Analysis by Digital Image Processing II. Wheat Class Discrimination. *Journal of Cereal Science* 10, 183-188.

Niblack, W. 1985. An Introduction to Digital Image Processing. *Strandberg Publishing Company, Birkerød, Denmark.*

Olsen, K.J. 1988. Texture Analysis of Ultra sound Images of Livers. *IMSOR, The Technical University of Denmark, Ph.D. Thesis nr. 51.*

Pavlidis, T. 1990. Algorithms for Graphics and Image Processing. *Computer Science Press.*

Peleg, S., J. Naor, R. Hartley and D. Avnir. 1984. Multiple Resolution Texture Analysis and Classification. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 6(4), 518-523.

Persoon, E. and K.-S. Fu. 1977. Shape Discrimination Using Fourier Descriptors. *IEEE Transactions on Systems, Man, and Cybernetics* 7(3), 170-179.

Petersen, P.E.H. 1991. Shape Analysis of Weed Seeds Using the Fourier Transform. *Proceedings of the 7th Scandinavian Conference on Image Analysis, Ålborg, Denmark, August 13-16, Vol I,* 56-63.

Petersen, P.E.H. and G.W. Krutz. 1991. Automatic Identification of Weed Seeds by Color Machine Vision. *(submitted to Seed Science and Technology).*

Pietikäinen, M., A. Rosenfeld and I. Walter. 1982. Split-and-Link Algorithms for Image Segmentation. *Pattern Recognition* 15(4), 287-298.

Pluta, M. 1988. Advanced Light Microscopy. Volume 1. Principles and Basic Properties. *Elsevier & PWN-Polish Scientific Publishers.*

Pratt, W.K. 1991. Digital Image Processing. *John Wiley & Sons, Inc. New York, Chichester, Brisbane, Toronto, Singapore.*

Press, W.H., B.P. Flannery, A. Teukolsky and W.T. Vetterling. 1987. Numerical Recipes (The Art of Scientific Computing). *Cambridge University Press, Cambridge.*

Price, T.V. and C.F. Osborne. 1990. Computer Imaging and Its Application to Some Problems in Agriculture and Plant Science. *Critical Reviews in Plant Sciences* 9(3), 235-266.

Reitboeck, H. and T.P. Brody. 1969. A Transformation with Invariance under Cyclic Permutation for Application in Pattern Recognition. *Information and Control* 15, 130-154.

Rencher, A.C. and S.F. Larson. 1980. Bias in Wilks' Λ in Stepwise Discriminant Analysis. *Technometrics,* 22(3), 349-356.

Richard, C.W. and H. Hemami. 1974. Identification of Three-Dimensional Objects

Using Fourier Descriptors of the Boundary Curve. *IEEE Transaction on Systems, Man, and Cybernetics* 4(4), 371-378.

Roberts, H.A. 1981. Seed Banks in Soils. *Advances in Applied Biology* 6, 1-55.

Sapirstein, H.D., M.Neuman, E.H.Wright, E.Shwedyk and W.Bushuk. 1987. An Instrumental System for Cereal Grain Classification using Digital Image Analysis. *Journal of Cereal Science* 6, 3-14.

Seber, G.A.F. 1984. Multivariate Observations. *John Wiley & Sons, New York.*

Segerlind, L.J. and B. Weinberg. 1972. Grain Kernel Identification by Profile Analysis. *ASAE paper* 72-314.

Strachan, N.J.C. and P. Nesvadba. 1990. Fish Species Recognition by Shape Analysis of Images. *Pattern Recognition* 23(5), 539-544.

Sugai, M., A. Kanuma, K. Suzuki and M. Kubo. 1987. VLSI Processor for Image Processing. *Proceedings of the IEEE* 75(9), 1160-1165.

Symons, S.J. and R.G.Fulcher. 1988a. Determination of Variation in Oat Kernel Morphology by Digital Image Analysis. *Journal of Cereal Science* 7, 219-228.

Symons, S.J. and R.G.Fulcher. 1988b. Determination of Wheat Kernel Morphological Variation by Digital Image Analysis:I. Variation in Eastern Canadian Milling Quality Wheats. *Journal of Cereal Science* 8, 211-218.

Symons, S.J. and R.G.Fulcher. 1988c. Determination of Wheat Kernel Morphological Variation by Digital Image Analysis:II. Variation in Cultivars of Soft White Winter Wheats. *Journal of Cereal Science* 8, 219-229.

Sze, T.W. and Y.H. Yang. 1981. A Simple Contour Matching Algorithm. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 3(6), 676-678.

Tappan, J.H., M.E. Wright and F.E. Sistler. 1987. Error Sources in a Digital Image Analysis System. *Computers and Electronics in Agriculture* 2, 109-118.

Teh, C. and R.T.Chin. 1989. On the Detection of Dominant Points on Digital Curves. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 11(8), 859-872.

Thomson, W.H. and Y. Pomeranz. 1991. Classification of Wheat Kernels Using Three-Dimensional Image Analysis. *Cereal Chem.* 68(4), 357-361.

Travis, A.J. and S.R.Draper. 1985. A computer based system for the recognition of seed shape. *Seed Sci. & Technol.* 13, 813-820.

Wallace, T.P. and P.A. Wintz. 1980. An Efficient Three-Dimensional Aircraft Recognition Algorithm Using Normalized Fourier Descriptors. *Computer Graphics and Image Processing* 13, 99-126.

Watanabe, S. and the CYBEST Group. 1974. An Automated Apparatus for Cancer Prescreening: CYBEST. *Computer Graphics and Image Processing*, 3, 350-358.

Westerlind, E. 1984. Datoriserad sortering som hjälpmedel vid antalsbestämning i utsäde av stråsäd. *Medd. Statens Utsädekon.*, 59, 83-98.

Weszka, J.S. 1978. A Survey of Threshold Selection Techniques. *Computer Graphics and Image Processing* 7, 259-265.

Wong, R.Y. and E.L. Hall. 1978. Scene Matching with Invariant Moments. *Computer Graphics and Image Processing* 8, 16-24.

Yakimovsky, Y. 1976. Boundary and Object Detection in Real World Images. *Journal of the Association for Computing Machinery* 23(4), 599-618.

Zahn, C.T. and R.Z. Roskies. 1972. Fourier Descriptors for Plane Closed Curves. *IEEE Transactions on Computers* 21(3), 269-281.

Zayas, I., F.S. Lai, and Y.Pomeranz. 1986. Discrimination Between Wheat Classes and Varieties by Image Analysis. *Cereal Chem.* 63(1), 52-56.

Zayas, I., Y.Pomeranz and F.S.Lai. 1985. Discrimination Between Arthur and Arkan Wheats by Image Analysis. *Cereal Chem.* 62(6), 478-480.

Zayas, I., Y.Pomeranz and F.S.Lai. 1989. Discrimination of Wheat and Nonwheat Components in Grain Samples by Image

Analysis. *Cereal Chem.* 66(3), 233-237.

Appendix A

Algorithm 1: Border tracking

{*Notation:* (x,y) is the coordinate of the image, where y increase downward and x increase in right direction}

Find first boundary pixel (x,y) by moving horizontally from interior point to the rightmost object pixel;

```
repeat
    {decision rules for upward or downward movement}
    if (x+1, y) = 0 or
      ((x,y+1) = 0 and (x,y-1) > 0 and (x-1,y) >0 )   then
        go down
    else
        go up;

    {horizontal movement decisions}
    if down and (x,y) = 0 then
        repeat
            move to the left;
        until (x,y) > 0 or (x,y-1) = 0;
    else
        if down and (x,y) > 0 then
            repeat
                move to the right;
            until (x,y) = 0 or (x,y-1) > 0;
        else
            if up and (x,y) = 0 then
                repeat
                    move to the right;
                until (x,y) > 0 or (x,y+1) = 0;
            else
                if up and (x,y) > 0 then
                    repeat
                        move to the left;
                    until (x,y) = 0 or (x,y+1) > 0;

    if (x,y) = 0 then
        move to last boundary pixel;
until the initial boundary pixel is met;
```

Algorithm 2: Area and center

{*Notation:* The boundary array, B_i say, where $i = 1,...,n$ is used as input. A_y is the object area integrated in the y direction, and C_y denotes the y coordinate of the center of gravity}

```
(x,y) := first boundary point;
A_y := y; C_x := x/2; C_y := y/2;
repeat
    (x,y) := next boundary point;
    obtain Freeman code for previous and following boundary pixel;
```

{the procedure for obtaining the C_x value is analogous to the following}
decide to ADD and/or SUBTRACT the new y value;
if ADD **then**
begin
$C_y := (C_y * A_y + y^2/2)/(A_y + y)$;
$A_y := A_y + y$;
end;
if SUBTRACT **then**
begin
$C_y := (C_y * A_y - (y-1)^2/2)/(A_y -(y - 1))$;
$A_y := A_y -(y - 1)$;
end;
until last boundary point;

Algorithm 3: Fill and count

{*Notation:* x_i, y_i denotes the boundary coordinates placed in an array, and an index *i* ranging from 1 to maxindex.}
find smallest y value (Ymin)
repeat
{find rightmost *x* value}
for i := 1 **to** maxindex **do**
if $x_i > x_{first}$ **and** y_i = Ymin **then**
first := i;
IN := false;
repeat
index := 1;
{find last of consequtive points of same y value, and count pixels if not allready counted}
while $y_{first + index}$ = Ymin **do**
index := index + 1;
if $y_{first - 1}$ = Ymin + 1 **and** $y_{first + index}$ = Ymin + 1 **then**
LOCALMAX1 := true;
if $y_{first - 1}$ = Ymin - 1 **and** $y_{first + index}$ = Ymin - 1 **then**
LOCALMIN1 := true;
right := first + index - 1;
{find the nearest left point of same y value}
for i := 1 **to** maxindex **do**
if y_i = Ymin **and** $x_i < x_{right}$ **and** $x_i > x_{left}$ then
left := i;
index := 1;
{find last of consequtive points of same y value, and count pixels}
while $y_{left + index}$ = Ymin **do**
index := index + 1;
if $y_{left - 1}$ = Ymin + 1 **and** $y_{left + index}$ = Ymin + 1 **then**
LOCALMAX2 := true;
if $y_{left - 1}$ = Ymin - 1 **and** $y_{left + index}$ = Ymin - 1 **then**
LOCALMIN2 := true;
if not (LOCALMIN1 **or** LOCALMAX1) **or** IN **then**
begin

```
                {count pixels between right and left point}
                IN := true;
            end
        else
            first := right;

        if LOCALMAX2 or LOCALMIN2 and IN then
            begin
                {continue from the left point}
                first := left;
                IN := true;
            end
        else
            begin
                {jump over a gap}
                for i := 1 to maxindex do
                    if y_i = Ymin and x_i > x_first and x_i < x_left+index then
                        first := i;
                IN := false;
            end;
        if first not allready changed then
            {find new value of first}
            for i := 1 to maxindex do
                if x_i > x_left and y_i = Ymin then
                    first := i;
    until no new x_i for y_i = Ymin;
    {go one line down}
    Ymin := y_right + 1;
until no more points;
```

Appendix B

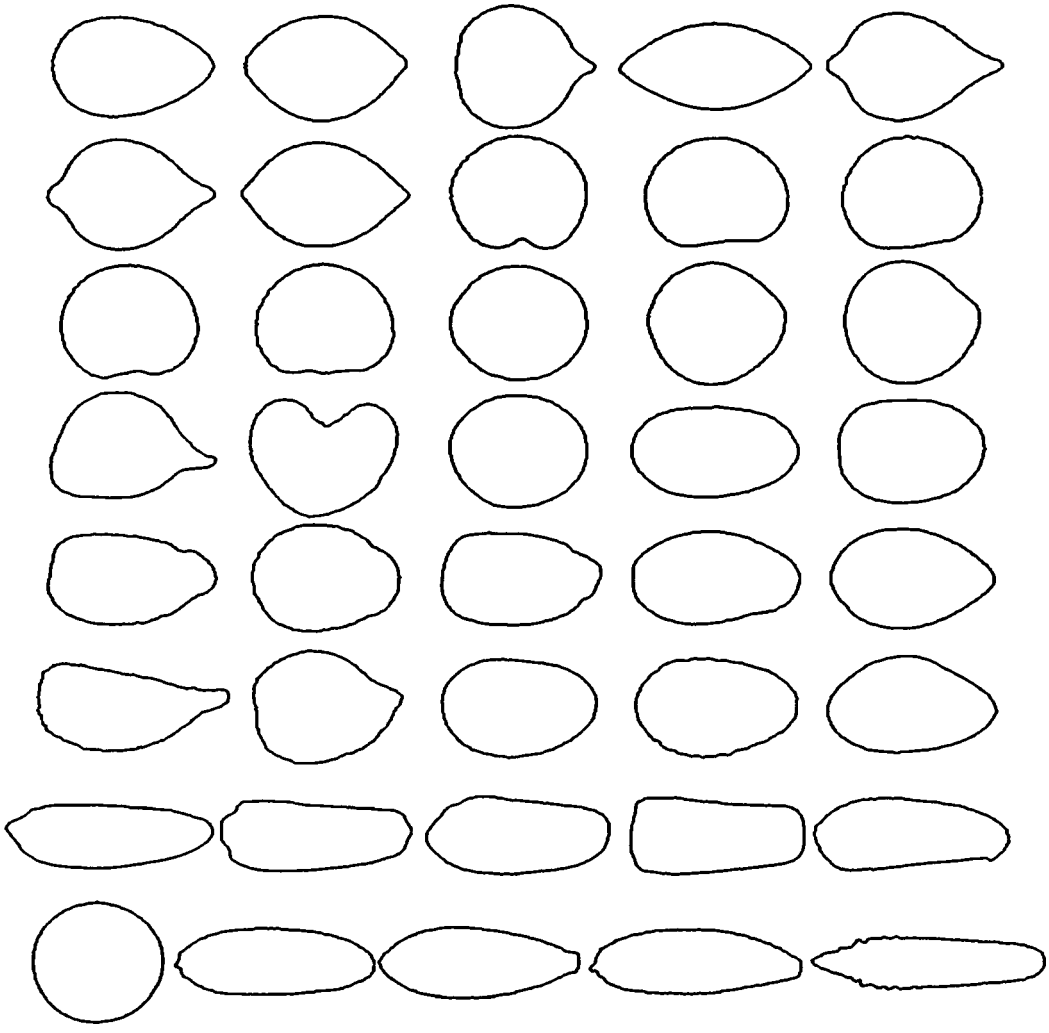


Figure B.1 Templates constructed by averaging 5 seeds within each species. The rowwise order corresponds to the order of species used in the tables.

Afdelinger under Statens Planteavlsforsøg

Direktionen

Direktionssekretariatet, Skovbrynet 18, 2800 Lyngby	45 93 09 99
Afdeling for Biometri og Informatik, Lottenborgvej 24, 2800 Lyngby	45 93 09 99

Landbrugscentret

Centerledelse, Fagligt Sekretariat, Forskningscenter Foulum, Postbox 23, 8830 Tjele	86 65 25 00
Afdeling for Grovfoder og Kartofler, Forskningscenter Foulum, Postbox 21, 8830 Tjele	86 65 25 00
Afdeling for Industriplanter og Frøavl, Ledreborg Allé 100, 4000 Roskilde	42 36 18 11
Afdeling for Sortsafprøvning, Teglværksvej 10, Tystofte, 4230 Skælskør	53 59 61 41
Afdeling for Kulturteknik, Flensborgvej 22, Jyndevad, 6360 Tinglev	74 64 83 16
Afdeling for Jordbiologi og -kemi, Lottenborgvej 24, 2800 Lyngby	45 93 09 99
Afdeling for Planteernæring og -fysiologi, Vejenvej 55, Askov, 6600 Vejen	75 36 02 77
Afdeling for Jordbrugsmeteorologi, Forskningscenter Foulum, Postbox 25, 8830 Tjele	86 65 25 00
Afdeling for Arealdata og Kortlægning, Enghavevej 2, 7100 Vejle	75 83 23 44
Borris Forsøgsstation, Vestergade 46, 6900 Skjern	97 36 62 33
Lundgård Forsøgsstation, Kongeåvej 90, 6600 Vejen	75 36 01 33
Rønhave Forsøgsstation, Hestehave 20, 6400 Sønderborg	74 42 38 97
Silstrup Forsøgsstation, Oddesundvej 65, 7700 Thisted	97 92 15 88
Tylstrup Forsøgsstation, Forsøgsvej 30, 9382 Tylstrup	98 26 13 99
Ødum Forsøgsstation, Amdrupvej 22, 8370 Hadsten	86 98 92 44
Laboratoriet for Biavl, Lyngby, Skovbrynet 18, 2800 Lyngby	45 93 09 99
Laboratoriet for Biavl, Roskilde, Ledreborg Allé 100, 4000 Roskilde	42 36 18 11

Havebrugscentret

Centerledelse, Fagligt Sekretariat, Kirstinebjergvej 10, 5792 Årslev	65 99 17 66
Afdeling for Grønsager, Kirstinebjergvej 6, 5792 Årslev	65 99 17 66
Afdeling for Blomsterdyrkning, Kirstinebjergvej 10, 5792 Årslev	65 99 17 66
Afdeling for Frugt og Bær, Kirstinebjergvej 12, 5792 Årslev	65 99 17 66
Afdeling for Planteskoleplanter, Kirstinebjergvej 10, 5792 Årslev	65 99 17 66
Laboratoriet for Forædling og Formering, Kirstinebjergvej 10, 5792 Årslev	65 99 17 66
Laboratoriet for Gartneriteknik, Kirstinebjergvej 10, 5792 Årslev	65 99 17 66
Laboratoriet for Levnedsmiddelforskning, Kirstinebjergvej 12, 5792 Årslev	65 99 17 66

Planteværnscentret

Centerledelse, Fagligt Sekretariat, Lottenborgvej 2, 2800 Lyngby	45 87 25 10
Afdeling for Plantepatologi, Lottenborgvej 2, 2800 Lyngby	45 87 25 10
Afdeling for Jordbrugszoologi, Lottenborgvej 2, 2800 Lyngby	45 87 25 10
Afdeling for Ukrudtsbekæmpelse, Flakkebjerg, 4200 Slagelse	53 58 63 00
Afdeling for Pesticidanalyser og Økotoksikologi, Flakkebjerg, 4200 Slagelse	53 58 63 00
Bioteknologigruppen, Lottenborgvej 2, 2800 Lyngby	45 87 25 10

Centrallaboratoriet

Centrallaboratoriet, Forskningscenter Foulum, Postbox 22, 8830 Tjele	86 65 25 00

Frederiksberg Bogtrykkeri a-s 920540