Analysis of Temporal Mammogram Pairs to Detect and Characterise Mass Lesions

Acknowledgements

The work presented in this thesis was done at the Radiology Department of the University Medical Centre St. Radboud Nijmegen. The project was funded by a grant from the Dutch Cancer Society (KUN 2001-2380) and sponsored by fundraising activities of cycling club 'Bergh in het zadel'.

The printing of this thesis was kindly supported by the Dutch Cancer Society.

Cover Image: 'Encircled Breasts' © Deborah Lee Soltesz, 2006
Printing: Grafimedia, RUG, Groningen

ISBN-10: 90-9020550-0
ISBN-13: 978-90-9020550-2

Analysis of Temporal Mammogram Pairs to Detect and Characterise Mass Lesions

een wetenschappelijke proeve op het gebied van de Medische Wetenschappen

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de Rector Magnificus prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op dinsdag 6 juni 2006
des ochtends om 10.30 uur precies

door

Sheila Timp
geboren op 29 maart 1973
te 's Gravenhage

Promotor:	Prof. dr. C.C.A.M. Gielen
Copromotor:	Dr. ir. N. Karssemeijer
Manuscriptcommissie:	Prof. dr. J.G. Blickman
	Prof. dr. H.J. Kappen
	Prof. dr. ir. M.A. Viergever (Universiteit Utrecht)

Contents

1 Introduction **1**

1.1 Breast Cancer Epidemiology and Risk Factors 1
1.2 Normal Structure and Function of the Breast 2
1.3 Breast Tumours 3
1.4 Breast Cancer Screening 6
1.5 Imaging Modalities 8
1.6 Computer Aided Detection and Diagnosis 11
1.7 Multi View CAD 14
1.8 Temporal Changes in Breast Tissue 17
1.9 Nomenclature 18
1.10 Overview of this Thesis 21

2 Single View Computer Aided Diagnosis **23**

2.1 Pre-processing 23
2.2 Pixel Level Mass Detection Algorithm 26
2.3 Region Segmentation and Feature Calculation 27
2.4 Classifier Training and Testing 35
2.5 Performance Evaluation 35

3 Mass Segmentation based on Dynamic Programming **39**

3.1 Introduction 39
3.2 Segmentation Methods 43
3.3 Experiments to Evaluate Segmentation Methods 52
3.4 Results 54
3.5 Discussion 60

4 Temporal Changes in Masses **63**

4.1 Dataset 63
4.2 Temporal Change Analysis. 65

4.3 Results 67
4.4 Discussion 68

5 Registration to find Corresponding Masses in Temporal Images 71
5.1 Introduction 71
5.2 Registration Procedure 73
5.3 Experiments to Evaluate Regional Registration 80
5.4 Results 82
5.5 Discussion 89

6 Interval Change Analysis for the Detection of Masses. 93
6.1 Introduction 93
6.2 Single View and Temporal CAD programme 96
6.3 Mass Detection Experiments 102
6.4 Results 103
6.5 Discussion 110

7 Interval Change Analysis for the Characterisation of Masses 113
7.1 Introduction 113
7.2 Single View and Temporal CAD Programme 115
7.3 Mass Characterisation Experiments & Results 122
7.4 Discussion 127

8 Effect of Temporal CAD on Radiologists' Performance 131
8.1 Introduction 131
8.2 Description CAD Programme and Observer Experiment 133
8.3 Results Reading with CAD and Double Reading 138
8.4 Discussion 139

Bibliography 143

Summary 153

Samenvatting 157

Dankwoord 161

Curriculum Vitae 163

Chapter 1

Introduction

This preparatory chapter provides some background material and literature required for this thesis. For further reading we suggest one of the following books: Vainio & Bianchini (2002); Homer (1997); Underwood (1992); Friedrich & Sickles (2000). This chapter is organised as follows. Section 1.1 to 1.5 give general information about breast cancer, about screening programmes to detect breast cancer, and about modalities that are used to image the breast. In Section 1.6 we describe the use of computer aided detection and diagnosis (CAD) systems and review some important studies that evaluate potential benefits of using CAD. At the moment multi view CAD systems are being developed that include information from multiple views. Section 1.7 summarises recent advances in this field. In this thesis we focus on the design of a multi view CAD system that incorporates information about temporal changes that take place between two consecutive screening rounds. Section 1.8 shortly discusses the objective for this approach. Section 1.9 clarifies definitions and nomenclature used in this thesis. Finally, in Section 1.10, we present an overview of this thesis.

1.1 Breast Cancer Epidemiology and Risk Factors

Incidence Breast cancer is a very common disease. It is the most common cancer for females and the second most common cancer for males and females combined. In the year 2000 it accounted for 22% of all new cancers in women. In the western world, this percentage is even 27% and about 1 in 10 to 12 women will have to face breast cancer. In most European countries the aged standardised mortality rates for breast cancer range from 15 to 30 for every 100,000 women making breast cancer the most important cause of cancer-related mortality for women (Levi *et al.* 2004). The average age of women when they are diagnosed with breast cancer is 64 years. One third of all women diagnosed with

breast cancer is younger than 50 years. Breast cancer can also develop in men, although this is rare. Male breast cancer accounts for about 1% of all breast cancer cases.

In The Netherlands the breast cancer incidence for women is 140 per 100,000. There are about 12,000 new cases of breast cancer annually and about 3500 women die of the disease yearly (Dutch Cancer Registry 2003).

Risk Factors for Breast Cancer Although it is not possible to say what exactly causes breast cancer, some factors may increase or change the risk of developing breast cancer. These include, in order of importance, female sex, age, having a family history of breast cancer, and having a previous diagnosis of breast cancer or ductal carcinoma in situ. Other factors that slightly increase the risk of developing breast cancer are the following: a long interval between menarche and menopause, obesity, not having children or having a first child after 35 years of age, not breastfeeding, taking combined Hormone Replacement Therapy (HRT) after menopause (especially when taken for 5 years or longer), putting on a lot of weight in adulthood, drinking alcohol (more than 2 standard drinks a day), taking oral contraceptives (this appears to increase the risk only during the period of taking the pill) and having previously been diagnosed with lobular carcinoma in situ or atypical hyperplasia.

1.2 Normal Structure and Function of the Breast

Figure 1.1 shows the most important anatomical structures of the breast. The breast consists of two components. The first component is concerned with milk production and is known as the epithelial component. The second component consists of fat and connective tissue. This component supports and protects the structure of the breast.

The epithelial component of the breast consists of a tree-like branching pattern of milk ducts that come together at the nipple. The leaves of this tree are formed by the lobules which are the secretory units of the breast. Each lobule consists of a number of acini connecting to an intra-lobular duct. The acini are composed of two types of cells: epithelial and myo-epithelial. The epithelial cells secrete a variety of glyco-proteins and during lactation they also produce milk. The myo-epithelial cells are capable of contracting during breastfeeding. Each intra-lobular duct connects with an extra-lobular duct, and this together with the lobule, is called the terminal ductal lobular unit.

The extra-lobular ducts within the same area link together to form sub-segmental ducts, which in turn form segmental ducts. These ducts drain milk from different segments or lobes of the breast. In total, the breast consists of 15-20 lobes, which are roughly pyramidal in shape with the apex directed towards the nipple.

The non-epithelial component of the breast consists mainly of fatty tissue. There are no muscles in the actual breast, but there are a series of muscles behind and underneath

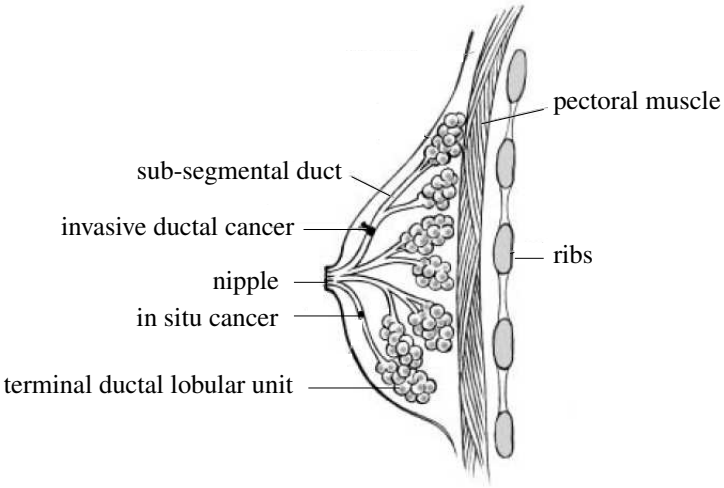


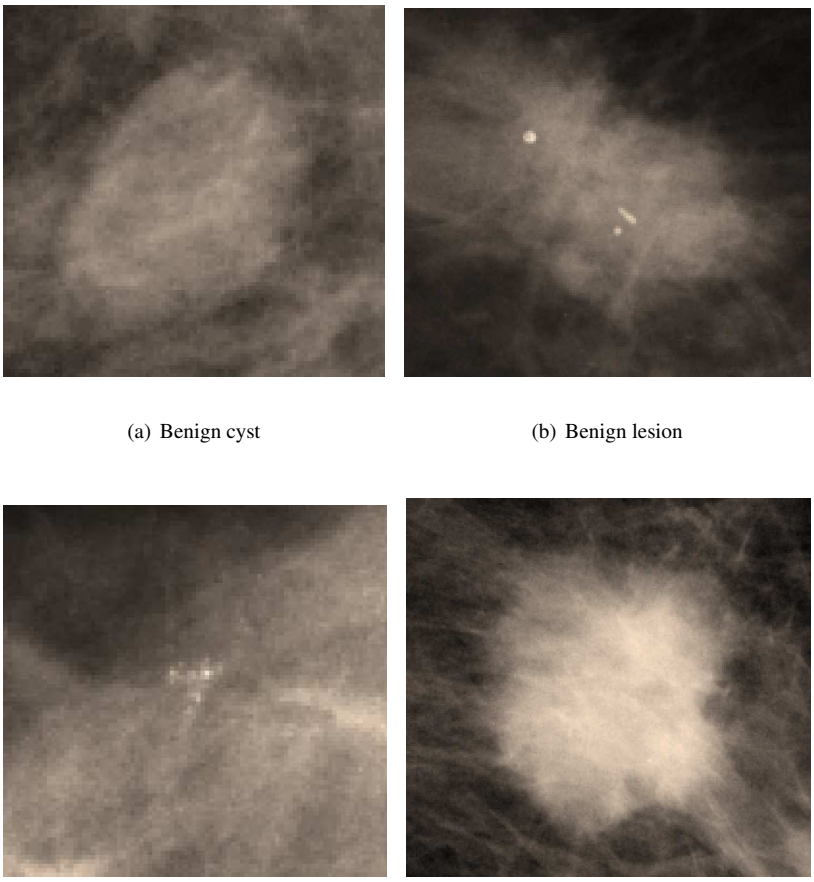
Figure 1.1: *Anatomy and structure of the breast.*

the breasts. These muscles work together with a ligament called Cooper ligament to support the weight of the breasts.

1.3 Breast Tumours

We distinguish three types of breast tumours: benign breast tumours, in situ cancer and invasive cancer. Figure 1.2 shows an example of each category.

Benign Diseases Benign tumours of the breast comprise fibro-adenoma, duct papilloma, adenoma and connective tissue tumours. The most common benign breast tumour is the fibro-adenoma. This tumour is a combined product of both connective tissue and epithelial cells. Most benign masses are circumscribed due to the absence of infiltration. Figure 1.2(a) shows a characteristic example of a benign mass. The shape is oval and the border is sharply delineated. On the other hand, benign masses may also present suspicious, as shown in Figure 1.2(b). Mammographically we cannot distinguish this benign mass from a malignant lesion.



(a) Benign cyst (b) Benign lesion

(c) DCIS (d) Infiltrative ductal cancer

Figure 1.2: *Appearance of breast lesions. Figure 1.2(a) shows a characteristic example of a benign mass. Figure 1.2(b) shows a benign mass which presents as a malignant lesion. Figure 1.2(c) shows an example of ductal carcinoma in situ. The last figure shows an infiltrative malignant cancer with characteristic ill-defined and spiculated borders.*

Non invasive Breast Cancer Non invasive—in situ—cancer consist of malignant cells that replace the normal epithelial cells lining the ducts or lobules. These malignant cells are still confined to the basement membrane and have not yet invaded the breast stroma or lymphatics. There are two non invasive forms of breast cancer: ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS).

- Ductal carcinoma in situ (DCIS) is a malignancy of the epithelial cells lining the lactiferous ducts—usually the terminal ducts—without penetration of the ductal basement membrane. The prognosis of untreated DCIS is not precisely known, as most patients are treated with mastectomy. One estimates that about one third to one half of the untreated patients eventually will develop invasive cancer, usually in the same quadrant of the breast as the first lesion. Mammographically DCIS is often characterised by the presence of micro-calcifications. When there is extensive fibrosis, DCIS may also present as a palpable mass.

- In lobular carcinoma in situ (LCIS) we find that the lobules are expanded by a uniform population of small yet atypical cells. Usually this process obliterates the lumen of acini. These atypical cells do not penetrate through the walls of the lobules. LCIS rarely gives rise to mammographic abnormalities. It is often found in biopsies that have been done for other reasons such as removal of benign lesions. LCIS is a risk factor for developing breast cancer in either breast. The majority of patients are therefore managed by careful follow up.

Invasive Breast Cancer Invasive breast cancer, also known as infiltrating cancer, occurs when malignant cells have spread beyond the ducts or lobules to other parts of the breast or body. Invasive cancers vary in size from less than 10 mm in diameter to over 80 mm, but are usually 20-30 mm at presentation.

Ductal carcinoma accounts for about 80% of all invasive breast cancer cases. These tumours are believed to arise from epithelial cells of the terminal ductal lobular unit. It is thought that ductal carcinoma may start as either DCIS or arise de nova. Less common types of breast cancer include lobular carcinoma, medullary carcinoma, tubular carcinoma, mucinous carcinoma, cribriform carcinoma and papillary carcinoma.

Breast cancers can infiltrate locally to the skin and the muscle, or metastasise to more distant sites via lymphatics or the bloodstream. The most common spread via lymphatics is to the axillary lymph nodes. Metastasis via the blood stream most frequently involves the lung and the liver, but adrenals and brains are also common sites for metastasis. When a woman has invasive breast cancer the prognosis depends among others on the histological grade and behavioural characteristics of the tumour, and the presence of metastatic spread . Considering histology we can grade tumours for their degree of differentiation. Well differentiated tumours often have a better prognosis than tumours that are poorly differentiated. Behavioural characteristics that influence the prognosis are the growth

rate and the receptor status of a tumour. Tumours with lower cell growth rates generally behave better. The presence of oestrogen receptors indicates that the tumour cells have a higher degree of functional differentiation resulting in a better prognosis. Tumour spread is also associated with a worse prognosis than when there is no evidence of metastasis. Although these factors may predict how individual cancers will behave, this has not lead to an improvement of patient survival. Screening on the other hand might improve survival rates due to earlier detection of breast cancer. The next section gives an overview of breast cancer screening programmes and the effect on breast cancer mortality rates.

1.4 Breast Cancer Screening

The aim of breast cancer screening is early detection of breast cancers while keeping the number of false positive detections at a minimum. The earlier most breast cancers are detected, the better the prognosis and treatment options for the patient. A higher recall rate, i.e. the percentage of mammographically screened women that is recalled for further assessment, generally improves the detection rate. This however will also lead to an increase in the number of false positive detections resulting in unnecessary examinations and additional costs. Most countries have recall rates between 3% and 5%.

An important trial to the effect of screening was done between 1977 and 1984 in Sweden (Tabár *et al.* 1985). This trial concerned 162,981 women aged 40 and older who were living in the counties of Kopparberg or Ostergötland. The women were divided at random into two groups. Each woman in the study group was offered screening every 2 or 3 years depending on age. Women in the control group were not offered screening. Results obtained after seven years of follow up showed a 31% reduction in breast cancer mortality and a 25% reduction in the rate of advanced breast cancers for the group invited to screening. These findings confirmed the results of an earlier trial by Shapiro *et al.* (1982). Many countries initiated national screening programmes for breast cancer after the results of the Swedish two counties trial were published in Tabár *et al.* (1985). Finland and Sweden started their programmes in 1986, the United Kingdom in 1988, and the Netherlands in 1989.

Different trials have been done to determine whether these screening programmes were achieving their goals. The eight most important trials are the following: Chu *et al.* (1988), Alexander *et al.* (1999), Bjurstam *et al.* (1997), Frisell *et al.* (1997), Tabár *et al.* (1995), Miller *et al.* (1992a), Miller *et al.* (1992b), Andersson *et al.* (1988), and Andersson & Janzon (1997). Most of these trials show a significant reduction in breast cancer mortality, especially for women aged 50–70 years. These results have been used to guide screening programmes world wide. Recently a pair of Danish investigators, Gotzsche and Olsen, criticised the quality of a majority of these trials (Gotzsche & Olsen 2000; Olsen & Gotzsche 2001). They found randomisation imbalances and inconsistencies

in six of the trials. The only trials they considered good were the Canadian trial (Miller *et al.* 1992a; Miller *et al.* 1992b) and the initial trial of the Malmo report (Andersson *et al.* 1988). These two studies show no benefit from screening mammography. Therefore they concluded that mammography is ineffective in reducing breast cancer mortality. Various authors reacted and stated that randomisation was not a major problem. Furthermore they pointed out that it is difficult to develop and implement a perfect trial. It seems thus acceptable to include the data from the six criticised trials. These all show a significant reduction in breast cancer mortality (Jackson 2002).

Although the arguments of Gotzsche and Olsen may not be of substantial importance, the possible benefits of screening must be weighed against the risks, such as psychological trauma of receiving a false positive result, and costs. Furthermore the efficacy of screening mammography, especially for women in the age group from 40-49 years, remains controversial. In the Netherlands, the United Kingdom, Sweden, and Finland women from 50 to 70/75 are invited every 2 or 3 years for screening. The American Cancer Society (ACS) recommends annual mammography for all women beginning at age 40.

Screening in the Netherlands The Dutch Breast Cancer Screening Programme started in 1989 and reached its full population capacity in 1997. In the Netherlands the screening programme offers all women between 50 and 70 years a biennial screen examination, resulting in 750,000 invited women each year. All women receive a personal letter with a fixed appointment that can be changed on request. Non attenders receive a reminder about 2 to 3 months later. At the first screening examination two mammographic views—medio lateral oblique and cranio caudal—are obtained. At subsequent examinations only medio lateral oblique views are obtained unless additional views are necessary. Films are developed immediately at the screening unit. A radiographer judges each film on technical quality and decides whether additional views are necessary. Afterwards two radiologists independently read all films in batches. Consensus between the two readers is required for a referral.

Some studies have been done to evaluate the effectiveness of the Dutch Breast Cancer Screening Programme (Otto *et al.* 2003; Fracheboud *et al.* 1998; Otten *et al.* 2005). Otto *et al.* (2003) assessed the effect of screening on breast cancer mortality rates, taking into account the phased implementation of the screening programme. For this purpose they used population statistics from 27,948 women aged 55–74 who died of breast cancer between 1980 and 1999. They found that breast cancer mortality rates started to fall between 1991 and 1996. This decrease became significant in 1997 and remained so in subsequent years. Their analysis shows that the point at which breast cancer mortality rates changed into a downward trend coincided with the start of the screening. This means that the programme already prevented death from advanced disease in the first years after implementation of the programme. Fracheboud *et al.* (1998) studied the fol-

lowing outcomes of the Dutch screening programme between 1990 and 1995: attendance rate, detection performance and compliance. In these years the attendance rate was on average 78% with little differences between screening rounds and age groups. Of 1,000 initially (and two years thereafter) screened women, 13.4 (6.6) women were referred for further investigation, 9.7 (4.4) got a biopsy and 6.4 (3.4) turned out to have breast cancer. The positive predictive value of screening and biopsy were 47% (51%) and 66% (78%) respectively.

A characteristic feature of the Dutch screening programme is the low referral rate (1.05% of all screened women), which in most other programmes is at least twice as high. In a recent study Otten *et al.* (2005) estimated the effect of a change in recall rate on the detection of breast cancer. For that purpose they used a set of 495 screen negative mammograms, 250 from control subjects and 245 from women who subsequently developed breast cancer. Fifteen radiologists with a specialisation in breast cancer screening read all mammograms. They annotated all suspicious regions and gave each region a rating. These ratings were used to measure the effect of different recall rates on the detection of cancers and on the number of false positive detections. Results show that lowering the threshold for recall, especially for recall rates between 1%-4%, leads to an improvement in breast cancer detection rates at an acceptable false positive rate. By further increasing the recall rate they found that cancer detection levels off with a disproportionate increase in the number of false positive detections.

1.5 Imaging Modalities

At the moment the modality of choice for breast cancer screening is mammography. For additional examinations, or when mammography is not sufficient, other modalities might be used. These include ultrasonography (US) and contrast enhanced magnetic resonance imaging (MRI). In this section we shall give an overview of the different modalities.

Mammography

Mammography is an X-ray technique developed specifically for the breast. It is based on the differential absorption of X-rays between the various tissue components of the breast such as fat, connective tissue, tumour tissue and calcifications. Mammography is used both as a clinical tool to examine symptomatic patients and for screening. Requirements for mammography are high contrast, high spatial resolution, and minimal radiation exposure. High contrast is needed because differences in density between normal and pathologic structures of the breast are small. The detection of micro-calcifications requires both high contrast as well as a high spatial resolution. Minimal radiation exposure is essential as in screening programmes women frequently undergo mammography, often

annually or bi-annually.

Mammographically we can recognise breast cancer by the presence of a focal mass lesion or micro-calcifications. Below we describe both characteristics. Less frequent signs of malignancy are architectural distortions and asymmetric breast tissue.

- Mass lesion. Most breast tumours, benign as well as malignant ones, present as a focal mass lesion. A task of radiologists therefore is to discriminate between benign and malignant lesions. When a radiologist considers a lesion suspicious for containing a malignancy the woman will be referred for additional examinations. The most important sign of malignancy is the presence of spiculation. This is a stellate pattern of lines directed towards the centre of a lesion. The border of a mass may also give information about the potential malignancy of a lesion. Benign masses are often characterised by sharp, circumscribed borders. Malignant masses on the other hand frequently have ill-defined or spiculated borders. The sharpness of the border however can not be used as solitary criterion for malignancy as some malignant masses, for example medullary carcinoma, colloid carcinoma and intracystic carcinoma, have circumscribed borders as well. Moreover benign masses may have poorly defined margins, for instance due to overlapping breast tissue or fibrosis. When a lesion is probably benign or when multiple similar masses are found in the breast the patient is often placed in a follow up protocol. Otherwise further examination is necessary to determine the nature of the mass.

- Micro-calcifications. Another sign of malignancy is the presence of micro-calcifications. Micro-calcifications develop in microscopically small cavities inside the lobuli or ducti. Micro-calcifications inside the lobular unit are often due to benign conditions such as adenosis or fibro-adenoma. Micro-calcifications of ductal origin are more suspicious and may be the first sign of breast cancer. Intra-ductal micro-calcifications can be diagnosed as benign or malignant by analysing the shape of the cluster and the shape of the individual micro-calcifications. Studies show that irregular, pleomorphic shapes have a higher probability of being associated with malignant disease than those with round shapes and uniform size.

Missed Cancers A problem of screening programmes is the large percentage of missed cancers. Studies show that during screening radiologists fail to detect 15–25% of breast cancers that are visible in retrospect (Goergen *et al.* 1997; Bird *et al.* 1992). Moreover, when minimal signs are taken into account, estimates of missed cases increase to 50% (Timp *et al.* 2002a). The most important causes of these false negative screening examinations are errors of perception. Eye-tracker studies have classified these errors into three main categories:

1. Search errors. In these cases the radiologist overlooked the abnormality. Eye-tracker experiments show that foveal sight never reached the lesion.

2. Detection errors. The lesion has been seen but the visual dwell time was shorter than a certain threshold, for instance one second.

3. Interpretation errors. These lesions are consciously evaluated but acted on inappropriately.

Without considering recorded eye movements, one may define search and detection errors as those that occur when a radiologist does not report the presence of a visible lesion and interpretation errors as those that occur when the lesion is reported but not considered actionable. Recent studies indicate that the majority of errors are due to misinterpretation and that inefficient search only makes a minor contribution to the error rate (Karssemeijer *et al.* 2003; Manning *et al.* 2004).

Digital Mammography Although most radiologists are still more comfortable with the use of screen film combinations, disadvantages are obvious. Once an image is printed, it can no longer be manipulated, and any information available in the digital data but not captured on the printed image will be lost. Furthermore screen film combinations have important limitations in detecting subtle soft tissue lesions, especially in the presence of dense glandular tissue (Lewin *et al.* 2001).

To overcome these limitations full field digital mammography (FFDM) has been introduced. FFDM offers several advantages over film mammography: easier access to images, use of CAD, improved means of transmission, retrieval, and storage of images, and the use of a lower average dose of radiation without a compromise in diagnostic accuracy. In a recent study Pisano *et al.* (2005) compared the diagnostic accuracy of digital and film mammography. In this study a total of 49,528 asymptomatic women presenting for screening underwent both digital and film mammography. Breast cancer status was ascertained by a breast biopsy or a follow-up mammogram. This study showed that the overall diagnostic accuracy of digital and film mammography was similar, digital mammography however turned out to be more accurate in women under the age of 50 years, women with radiographically dense breasts, and pre-menopausal and peri-menopausal women. The major disadvantage of adopting digital mammography is its cost: at the moment digital systems cost about 1.5 to 4 times as much as film systems. On the other hand the fact that a CAD system can easily be incorporated and the possibility of retrieving archived images will reduce costs as well. A cost-effectiveness analysis is needed to weight the additional costs against the advantages of FFDM and the gain in diagnostic accuracy.

Ultrasonography

The role of ultrasonography (US) in breast imaging is a subject of ongoing discussion. US is generally accepted as the method of choice for the differentiation between a simple

cyst and a solid mass. US also plays a role in guiding intervention procedures such as needle aspiration, core needle biopsy, and pre-biopsy needle localisation. Another usage of US is the detection and staging of lymph nodes (Rahbar *et al.* 1999). Studies performed to evaluate US as a screening modality failed to establish its efficiency and it has been concluded that US should not be used as a screening tool. On the other hand US may play a role when it is used as an adjunct to mammography. Zonderland *et al.* (1999) reported an improvement in detection accuracy of 7.4% when US was used as an adjunct to mammography to analyse lesions from one of the following categories: circumscribed lesions that could be cysts, mammographically visible lesions, or palpable lesions that were not visible on the mammogram.

Magnetic Resonance Imaging

High-resolution contrast enhanced MRI of the breast has recently emerged as a sensitive instrument for the detection of breast cancer. MRI proved useful in screening younger women with dense breasts who are at a special high risk of developing breast cancer, e.g. having a strong family history or hereditary risk of breast cancer (Stoutjesdijk *et al.* 2001). MRI can also be used as an adjunct to mammography for selected patients. Finally MRI of the breast has the potential to be a powerful aid in pre-surgical planning (multi-focal cancer detection).

MRI however has a significant false positive rate, is not readily available in all areas, and is more expensive than mammography or ultrasonography. Other limitations are the fact that MRI requires contrast injection and that it can cause problems with claustrophobia. At the moment MRI therefore remains limited to specific problem solving situations and patients at high risk for cancer.

1.6 Computer Aided Detection and Diagnosis

In recent years a major effort has been made to develop computer aided detection and diagnosis (CAD) programmes to assist radiologists with the detection and characterisation of breast lesions. Computer aided *detection* systems identify and mark suspicious regions to bring them to the attention of a radiologist. These systems prevent that a radiologist fails to consciously see an abnormality and thus minimise search and perception errors. Computer aided *diagnosis* systems on the other hand aim at minimising interpretation errors.

CAD systems can be used for the detection and characterisation of mass lesions—including architectural distortions and asymmetry—and for the detection and characterisation of micro-calcifications. In the sequel we shall restrict ourselves to systems for mass lesions. Most of these CAD systems follow a two step procedure. The first step

concerns the detection of suspicious locations inside the breast area. In the second step a segmentation algorithm determines a contour at the most suspicious locations. For each segmented region several features are calculated to discriminate between benign lesions, malignant lesions, and false positive detections.

Currently commercial systems are only available for computer aided *detection*. These CAD systems are intended to be used after the radiologist has completed an evaluation of the images without CAD prompts and has made an initial decision whether recall is required. If a radiologist identifies an abnormal area of concern on a mammogram during initial reading and that area does not get marked by CAD, the radiologist is still advised to interpret the mammogram as positive and to recall the patient for further work-up. CAD is proposed as an adjunct to mammography to decrease search and detection errors. The radiologist, not CAD, determines if a clinically significant abnormality exists and decides whether further diagnostic evaluation is warranted. The hope is that these CAD systems will improve the sensitivity of mammography without substantially increasing mammography recall rates. In the next section we discuss the effectiveness of systems for computer aided detection in clinical practice. In the future, CAD systems for computer aided *diagnosis* will also become available to help radiologists with the diagnostic process.

1.6.1 Effectiveness of Computer Aided Detection Systems

Two types of studies have been done to evaluate the effectiveness of using computer aided detection systems in clinical practice: prospective and retrospective studies.

Retrospective Studies These studies retrospectively evaluate the effect of CAD on the detection of initially missed cancers (Warren Burhenne *et al.* 2000; Karssemeijer *et al.* 2003; Brem *et al.* 2003; Birdwell *et al.* 2001). Warren Burhenne *et al.* (2000) conducted a large retrospective study to potential benefits of CAD on mammographically missed cancers. For this study they collected more than 1000 screening mammograms that led to the detection of biopsy-proven cancer. For about half of the cases (427) they also obtained the prior mammograms for retrospective review. At retrospective review, 67% (286 of 427) of the breast cancers was visible on the prior mammograms. A panel of radiologists performed a blinded review to evaluate prior mammograms with a visible lesion. A CAD systems also analysed these prior mammograms. The recall rates of 14 radiologists were measured with and without using a CAD system. Without CAD the radiologists had a false-negative rate of 21%. CAD prompting could have potentially helped to reduce this false-negative rate by 77% without an increase in the recall rate. Results of this study indicate a potential for CAD to help the breast radiology community with detecting breast cancers.

Karssemeijer *et al.* (2003) estimated the potential contribution of CAD by measuring the performance of a CAD system in identifying lesions initially missed at screening. For this purpose they used screening mammograms of 500 cases, consisting of the mammograms at time of referral and all previous screening examinations. A CAD programme analysed the most recent prior mammograms and assigned each suspicious region a malignancy score. Ten experienced radiologists also indicated suspicious regions on these mammograms and rated each finding. The scores were combined in a way to simulate the following three reading modes: *single reading*, *double reading* and *reading with CAD*. For *single reading* the scores from the individual radiologists were used. To simulate double reading the scores of two radiologists were combined for each finding. For *reading with CAD* the score assigned to each finding by the radiologist was combined with the CAD score at the area of the finding. True positive findings of the CAD system that the radiologists had overlooked were ignored. Results show that the sensitivity of the radiologists increased by 7.0% for *reading with CAD* and by 10.5% for *double reading* compared to *single reading*. This study shows the potential benefit of CAD for the detection of breast cancer. Brem *et al.* (2003) also studied the performance of radiologists and CAD on missed cancers. For this purpose they used a dataset consisting of 177 missed cancers and 200 normal cases. Three radiologists independently read each mammogram. The CAD system also marked suspicious regions on each mammogram. Then they estimated the number of additional tumours that would have been detected when a radiologist was used as a second reader and when the CAD system was used in addition to the radiologist. With double reading 123 extra malignancies would have been detected, with CAD 80. This study shows that both double reading and CAD improve the detection of cancers. Another study to the effect of CAD on missed cancers has been done by Birdwell *et al.* (2001). They analysed the characteristics of 115 missed cancers and studied the potential utility of CAD. From these 115 missed cancers, 35 were calcifications and 80 were mass lesions. CAD correctly marked 30 of 35 missed calcifications and 58 of 80 missed masses. The mean number of marks of the CAD system was 4.3 for each four view mammogram, of which one third marked the missed cancers.

Although most of these studies report a positive effect of CAD on the detection of cancers, it is difficult to measure the effect that false positive CAD marks have on screening outcomes. The majority of these detections will indicate areas that a radiologist will choose to dismiss because no abnormal appearing characteristics are present. A radiologist however will need extra time to evaluate each CAD mark and some marks will also appear suspicious to the radiologist. This may increase referral rates which results in additional examinations and extra costs. To investigate the effect of CAD in clinical practice prospective studies may be more informative.

Prospective Studies There are two types of studies that prospectively evaluate the effect of CAD: sequential reading studies and studies based on historical controls. In sequential reading studies radiologists first read each mammogram without CAD followed by a review of the CAD prompted findings. Freer & Ulissey (2001) did a large study with 12,860 screening mammograms. All mammograms were first interpreted without CAD, immediately followed by a re-evaluation of areas marked by the CAD system. The effect of CAD was measured on recall rate, positive predictive value for biopsy, cancer detection rate and tumour stage at detection. Freer & Ulissey (2001) found an increase in recall rate (from 6.5% to 7.7%), no change in positive predictive value, a 19.5% increase in the number of cancers detected and an increase in the proportion of early stage (0 and I) malignancies from 73% to 78%. Helvie *et al.* (2004) performed a study with 13 radiologists to evaluate the additional effect of using CAD on a dataset consisting of mammograms from 2,389 patients. A CAD programme for the detection of masses and micro-calcifications processed each image and indicated all suspicious regions. Each radiologist read a part of the cases and assessed mammograms first without CAD and then with CAD. The detection performance of CAD and the radiologists was identical (91%), corresponding with detecting 10 out of 11 cancers. The detection performance of the radiologists increased from 91% to 100% when using CAD. A severe limitation of this study is the small number of cases and consequently the small number of breast cancers. The cancer that was detected by the CAD system but not by the radiologists was an area of micro-calcifications identified as ductal carcinoma in situ.

Studies based on historical controls compare the screening performance before and after the introduction of a CAD system. Gur *et al.* (2004) assessed changes in mammography recall and cancer detection rates after the introduction of a computer-aided detection system into a clinical radiology practice. In total they used the outcomes of 24 radiologists who interpreted 115,571 screening mammograms: 59,139 with CAD and 56,432 without CAD. They found that the introduction of computer-aided detection was not associated with statistically significant changes in recall rate and breast cancer detection rates. It should be noticed however, that the 95% confidence intervals obtained in this study (-11% to 19%) allow for a wide range of detection rate changes. Recently Cupples *et al.* (2005) evaluated the performance of radiologists before and after the introduction of a CAD system. They found that screening with CAD increased the detection rate by 17.7%, primarily due to increased detection of invasive cancers $\leq$ 1cm.

1.7 Multi View CAD

Most current CAD systems separately analyse each mammographic view to detect and characterise abnormalities. Radiologists on the other hand generally combine information from multiple mammographic views. Besides images of the left and right breast they

often have views from previous screening rounds and views from different projections. When a radiologist discovers a suspicious region in one view, he or she will try to find a corresponding region in the other views. Views from different projections, typically cranio caudal (CC) and medio lateral oblique (MLO) views, allow for a better characterisation of each detected region than the use of a single view. Prior views are useful to study changes in the appearance of a region over time. Contra-lateral views provide a reference to the appearance of different tissues in the breast and help to determine the relative suspiciousness of a region. By combining information from all views radiologists estimate the suspiciousness of each region and decide whether further investigation is required. Studies report a positive effect on either recall rate or an improvement in mass detection performance when using multiple views in mammography screening compared to single-view mammography, cf. (Wald *et al.* 1995; Sickles *et al.* 1986; Thurfjell *et al.* 2000; Callaway *et al.* 1997).

Given the positive effect of multi view systems on radiologists' performance we expect that fusion of information from different views will improve CAD systems as well. There have been some studies to the effect of using multiple views in CAD programmes. These studies combine information from medio lateral oblique and cranio caudal views (Good *et al.* 1999; Paquerault *et al.* 2002), from left and right views (Yin *et al.* 1991; Lau & Bischof 1991; Bovis *et al.* 2000), or from previous and current views (Vujovic *et al.* 1995; Kok-Wiles *et al.* 1998; Hadjiiski *et al.* 2001b). The next two paragraphs summarise work that has been done to combine views from either different projections or from left and right breasts. Section 1.8 discusses the use of views obtained at different time moments.

Different Projections of the Same Breast. The most obvious multi view approach is the combination of information from different projections of the same breast. Common projections are cranio caudal (CC) and medio lateral oblique (MLO) views. Radiologists use both views to determine the suspiciousness of a lesion and whether to refer a woman for further examination. Most CAD programmes only work on single view images and then combine evidence from both views in the following way. First they assign all detected regions from both projections to the same case. Then the cancer detection rate and the false positive rate are determined per case. So a tumour (and also a false positive) is counted as detected when it is found on either view.

Few studies combine evidence from MLO and CC views in a more intelligent way, for instance by linking similar structures in both breasts. Good *et al.* (1999) developed a method to match corresponding regions in CC and MLO views. They first determined all possible region pairs consisting of one region from the MLO view and one region from the CC view. Each pair was identified as either a true mass pair or a false mass pair. A true mass pair consists of two regions which are both projections of the same mass lesion. A false mass pair is a pair of regions in which either one is a false positive

detection or—in case multiple tumours are present in the same breast—in which both regions indicate different mass lesions. For each pair multi view features were calculated. A Bayesian network classified all pairs exclusively on these features. Results show that multi view information was useful to discriminate between true and false mass pairs. Paquerault *et al.* (2002) used a similar technique, but instead of using only multi view features they used a fusion scheme to combine the classifier score from the multi view features with the single view detection score. They found that the fusion information from the two view detection scheme improved the lesion detectability and reduced the number of false positives compared to the one view scheme. Van Engeland *et al.* (2002, 2006) also worked on the combination of information from MLO and CC views. In 2002 they developed a matching algorithm that used feature probability distributions to link suspicious regions in CC and MLO views. Results from this study show that the combination of feature vectors from both views slightly improved the mass detection performance. Recently they presented a new matching algorithm that correctly linked all true positive detections in 82% of the cases. They however found that the gain in detection performance was rather low (Van Engeland *et al.* 2006).

Left and Right Views. Some studies have been done to evaluate the use of information from left and right views. In general the left and right breasts of a woman are more or less symmetric. An asymmetric appearance can be suspicious, depending on the underlying cause. A common cause of an asymmetric appearance is the presence of a visible mass lesion in one view. According to the BI-RADS system, that is used to guide breast cancer diagnostics, the word asymmetry should be reserved for cases where the left and right breast have an asymmetric appearance *without* the presence of a clearly visible mass lesion (D'Orsi & Kopans 1997). The BI-RADS system discriminates between asymmetric breast tissue and the presence of a focal asymmetric density. Asymmetric breast tissue is defined relative to the contra lateral breast as a greater volume of breast tissue, greater density of breast tissue, or more prominent ducts. Asymmetric breast tissue is present on 3% of all mammograms and is nearly always benign (Piccoli *et al.* 1999). A focal asymmetric density is visible as an asymmetry of tissue density, but completely lacking the conspicuity of a true mass. A focal asymmetric density is suspicious as it may represent a mass lesion with ill-defined or obscured borders.

Some CAD programmes have been developed to determine the degree of asymmetry between right and left breasts. These programmes often aim to find all kinds of asymmetries, in particular asymmetry due to the presence of a mass lesion in one view, as this is suspicious for the presence of a malignancy. The conventional approach is as follows. First left and right images are registered, for instance by matching the breast boundaries of each image. This results in a mapping between the two images and makes it possible to compare features from corresponding locations in left and right breasts. The methods differ in their choice of image features used for measuring local asymmetry. Yin *et al.*

(1991) and Karssemeijer & Te Brake (1998) used brightness; Lau & Bischof (1991) used brightness and texture. Karssemeijer & Te Brake (1998) only found a small benefit when using asymmetry as an additional feature in their detection scheme. Instead of comparing all locations between left and right breasts, radiologists often compare anatomically similar regions in both breasts. Miller & Astley (1993) used this approach in a preliminary study and compared corresponding non fat regions in left and right breasts. For each region they calculated shape- and grey-level characteristics. They measured the degree of asymmetry as the difference between feature values of corresponding regions. On a small set they found that these asymmetry measures were useful for the detection of masses. To our knowledge, however, no further studies have been published since then that confirm the usefulness of asymmetry for automated detection of breast lesions.

1.8 Temporal Changes in Breast Tissue

The goal of this thesis is to design a CAD system that captures useful temporal information and to investigate the possible benefits of this approach. One of the reasons for temporal changes in the breast is the growth or development of a lesion. Besides changes due to developing lesions other factors also influence breast tissue at a given time and may thus change the radiographic appearance over time. These factors include ageing, involution, hormonal interactions, and lifestyle indicators such as diet and exercise (Heine & Malhotra 2002). Therefore when comparing previous and current mammograms, radiologists should take into account normal changes that occur in breast tissue.

At the moment most radiologists compare current mammograms with previous ones to improve the detection of tumours and to reduce the number of false positive interpretations. Several studies confirm the usefulness of this approach. In a recent study Roelofs *et al.* (2006) retrospectively determined the influence of comparing current mammograms with priors on breast cancer detection. Twelve experienced radiologists each read 160 mammograms, once with and once without using prior mammograms. Results obtained in this study show that the number of false positive detections was reduced with on average 44% when priors were used while maintaining the same sensitivity level. According to Callaway *et al.* (1997) the presence of previous mammograms significantly reduces the number of additional examinations and ultrasound examinations. Bassett *et al.* (1994) reviewed 1432 randomly selected screening mammography examinations and evaluated the effect of having priors and found that a comparison with previous examinations has a positive impact on clinical management and cancer detection in a limited number of cases. White *et al.* (1994) found that previous mammograms are judged valuable in interpreting current studies in 89% of the cases. Some studies also investigate the additional time and cost involved in obtaining previous mammograms. When prior mammograms have to be obtained from other facilities this may result in substantial

labour and cost (Bassett *et al.* 1994). In the future the use of FFDM and PACS systems will reduce these costs considerably. This may lead to a more positive cost-effectiveness analysis. Furthermore, when prior mammograms are easily available, CAD systems that measure temporal changes can be implemented more easily as well.

Considering the positive effect of prior views for radiologists we expect that CAD systems may improve as well when temporal information is used. The use of temporal information may improve the detection and classification performance of a CAD system for the following reasons. First, comparing the current mammogram with mammograms from previous screening rounds may bring to attention subtle signs of malignancy such as a small mass or new or increasing calcifications (White *et al.* 1994). These changes be overlooked if the previous mammogram is not available for comparison. Radiologists often use this technique to detect developing abnormalities. CAD programmes can also implement this technique to increase the number of true positive detections. Second, suspicious regions on the current view can be evaluated more precisely when this region is compared with the corresponding region on the previous view. For example, if a mass is detected on the current view, a radiologist or CAD system can use the previous view to determine whether this lesion is new or already existed. If the mass was already visible on the prior, the size and the contrast of both lesions can be compared to estimate the malignancy of a lesion. A third advantage of using prior mammograms for CAD systems is that additional clues can be found to remove false positive detections. Many false positive detections are caused by mammographic structures that are present on both current and prior mammograms. These structures will have a similar appearance on both mammograms. Examples are crossing vessels and benign lymph nodes. Analysis of temporal changes can be used to measure the similarity between the region on prior and current views. When both regions are similar, it is likely that the region represents a false positive detection or a slowly growing benign mass.

1.9 Nomenclature

In this section we clarify some nomenclature that we use in this thesis.

Case A *case* includes all available mammograms of one woman. Figure 1.3 shows a case that includes the mammograms from three consecutive mammographic exams. A *mammogram* includes all images obtained at the same time. Mammograms often contain two or four views. A two view mammogram usually consists of the right and left MLO view, whereas a four view mammogram consists of left and right MLO and CC views. There are two types of mammography exams: screening and clinical. Screening mammography is an x-ray examination of the breasts in an asymptomatic woman (that is the woman has no complaints or symptoms of breast cancer). When a radiologist sees

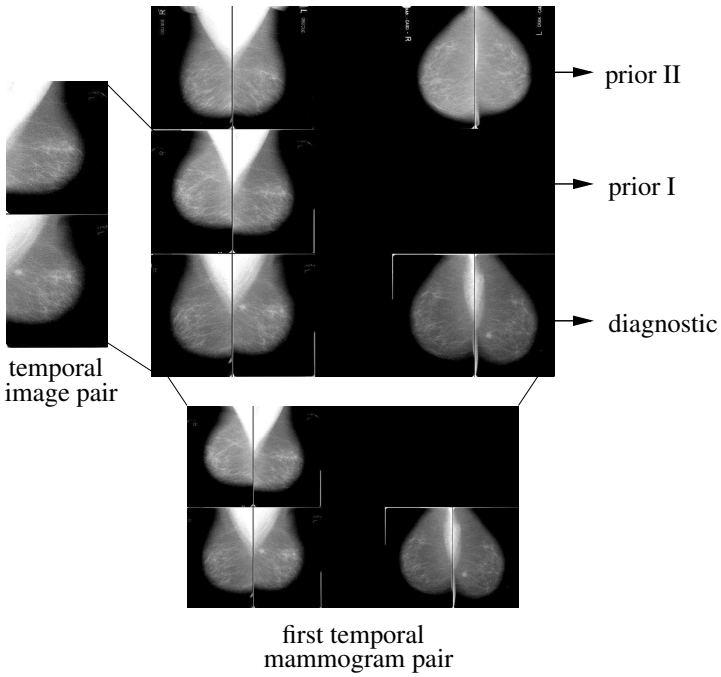


Figure 1.3: *Example of three consecutive mammographic exams of the same woman. Mammograms are displayed in chronological order. The bottom row represents the diagnostic mammogram, this is either a referral or a clinical mammogram. A malignant lesion is present in the left-MLO image of the diagnostic mammogram and its corresponding prior mammogram. The mammograms from two consecutive screening rounds form a temporal mammogram pair. This case provides two temporal mammogram pairs. The bottom and middle rows show the first mammogram pair, in which the diagnostic mammogram represents the current mammogram. This mammogram pair consists of two temporal image pairs (left and right MLO current-prior) and two single views (left and right CC). The top and middle rows form the second mammogram pair, in which the mammogram prior to diagnosis represents the current mammogram. This mammogram pair contains two temporal image pairs (left and right MLO current-prior).*

an abnormality on a mammogram he will refer the woman for further examination. This last screening mammogram is therefore called a referral or recall mammogram. Clinical mammography on the other hand is an x-ray examination of the breast in a woman who either has a breast complaint (for instance a breast lump found during self-exam) or has had an abnormality found during screening mammography. Cancers that are detected between two screening rounds are called interval cancers. When a tumour is present we call the most recent mammogram the diagnostic mammogram. For screen detected cancers this is the referral mammogram; for interval cancers the clinical mammogram.

Mammograms from previous screening rounds are called prior or previous mammograms. When the mammograms from more than one prior screening round are available, we sometimes number them. Prior I indicates the most recent prior mammogram, prior II the second most recent prior mammogram and so on.

An expert radiologist re-examined all mammograms in our database and indicated possibly benign and malignant lesions. All malignant lesions were confirmed by biopsy. All benign lesions were either proven by biopsy or by additional assessment such as ultrasound or follow-up. Other findings were assumed to contain no pathology and were classified as false positive detections.

Temporal Pairs To determine temporal changes we often use mammograms from two consecutive screening rounds. These form a temporal mammogram pair. The case in Figure 1.3 contains two temporal mammogram pairs. In each temporal pair we call the most recent mammogram the *current* mammogram and the mammogram from one screening round earlier the *prior* or *previous* mammogram. The first temporal mammogram pairs consists of a diagnostic mammogram and the mammogram one screening round prior to diagnosis, the prior I mammogram. In this temporal pair we call the diagnostic mammogram the current mammogram and the prior I mammogram the prior or previous mammogram. The second mammogram pair consists of the prior I and the prior II mammogram. In this pair we call the prior I mammogram the current mammogram and the prior II mammogram the prior one.

Lesions A breast *lesion* is a lump or mass that is either felt by palpation or has been detected by mammography. Mammographically we distinguish three types of mass lesions: focal mass lesions, architectural distortions and asymmetry. Histologically lesions can be classified as benign or malignant. A *region* or *finding* is a segmented area inside the breast that has been detected by a radiologist or a CAD system. This region can contain a benign lesion, a malignant lesion, or normal breast tissue. In the last case we call this region a false positive detection.

1.10 Overview of this Thesis

This thesis is organised as follows. Chapter 2 describes our general CAD programme. This programme detects suspicious regions inside the breast and assigns each region a measure representing the likelihood that the region contains a mass lesion, the so-called *mass likelihood*. The next step in the CAD programme is the segmentation of suspicious regions. In Chapter 3 we develop a segmentation algorithm based on dynamic programming and compare the efficiency of this algorithm with other segmentation methods from literature.

In Chapter 4 we analyse the temporal behaviour of mass lesions. First we determine for each lesion whether it is visible on previous views or not. When a lesion is not visible on the prior view we determine a location on the prior view that corresponds with the location of the lesion on the current view. We then determine features for prior and current regions and study which features change during time and which features stay constant. We also investigate differences in temporal behaviour between lesions that are visible on the prior view and lesions that are not.

Chapter 5 to 7 present a CAD programme that includes temporal information. As a first step each region on the current view is linked to a corresponding location on the prior view. Chapter 5 describes a regional registration method to accomplish this. The next steps of the temporal CAD programme are segmentation of prior regions and calculation of temporal features. Temporal features aim to measure changes and similarities between a region on the current view and the corresponding region on the prior view. We use two kinds of temporal features: difference features and similarity features. Difference features calculate the (relative) change between feature values of the current region and feature values of the prior region. Similarity features measure whether both regions are comparable in appearance. Chapter 6 and 7 evaluate the effect of temporal features on the performance of a CAD system for the detection and characterisation of mass lesions.

The last chapter investigates the potential contribution of a temporal CAD system in clinical practice to help radiologist with the task of mass characterisation. For this purpose we compare the following three reading modes: *single reading*, *independent reading with CAD* and *independent double reading*.

Chapter 2

Single View Computer Aided Diagnosis [1]

In this chapter we explain our single view computer aided diagnosis (CAD) system. Figure 2.1 gives an overview of the whole method. We start with applying some pre-processing algorithms to each mammographic image: segmentation of the breast boundary and the pectoral muscle, peripheral enhancement and pectoral fading. Then we apply a pixel level mass detection algorithm that calculates several features at each location in the breast area. A neural network classifier combines these features into a single score, the so-called *mass likelihood*, which indicates whether the location is suspicious for containing a mass or not. After that we select the most suspicious locations for further processing. This includes segmentation of the image at the selected locations and calculation of features for each segmented region. Finally a second classifier combines these features into a *malignancy score* that represents the likelihood that the region is malignant. We use this score to evaluate the effectiveness of our CAD programme.

2.1 Pre-processing

The single view CAD programme contains three pre-processing steps. These are illustrated in Figure 2.2. In the first, the image is segmented into breast area and background region. For this purpose we use an algorithm developed by Karssemeijer (1998). This algorithm applies a global thresholding technique to segment the breast tissue from the background. Then the location of the pectoral edge is determined. As the acquisition of mammograms is a standard procedure, we can indicate a region of interest (ROI) where

[1] A part of this Chapter has been published in Varela *et al.* (2006)

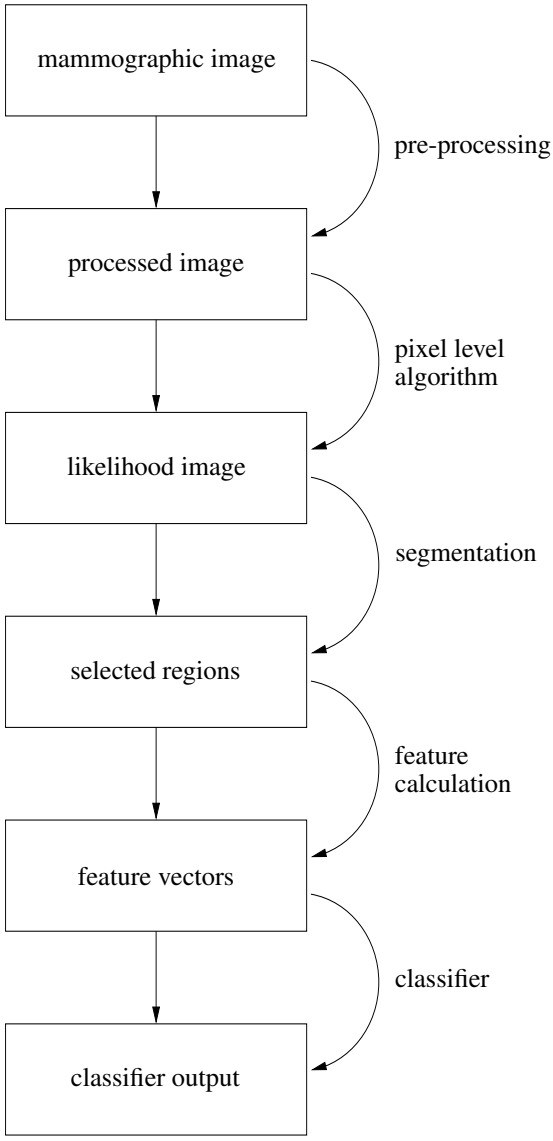


Figure 2.1: *Overview of single view CAD method*

the pectoral edge is probably located. Inside this ROI we calculate the gradient magnitude and direction ϕ by applying the 3x3 Sobel operator. We transform this gradient image to Hough space using the following line parametrisation:

$$\rho = m \sin(\phi) + n \cos(\phi).$$

The range of the parameter ϕ is constrained by the measured gradient direction $\phi_{m,n}$ at location (m, n) by:

$$|\phi_{m,n} - \phi| < \delta\, \phi.$$

After having processed all points in the ROI we discretise the Hough space resulting in a set of boxes, the so-called Hough accumulators. Each line increments a count (initialised at zero) in the corresponding Hough accumulator with weight factor w, where w is based on the gradient magnitude. After considering all pixels inside the ROI, we select the Hough accumulator with the highest value. The selected peak in Hough space is back-projected in the image space. The resulting straight line represents the pectoral edge which segments the pectoral region from the rest of the breast area.

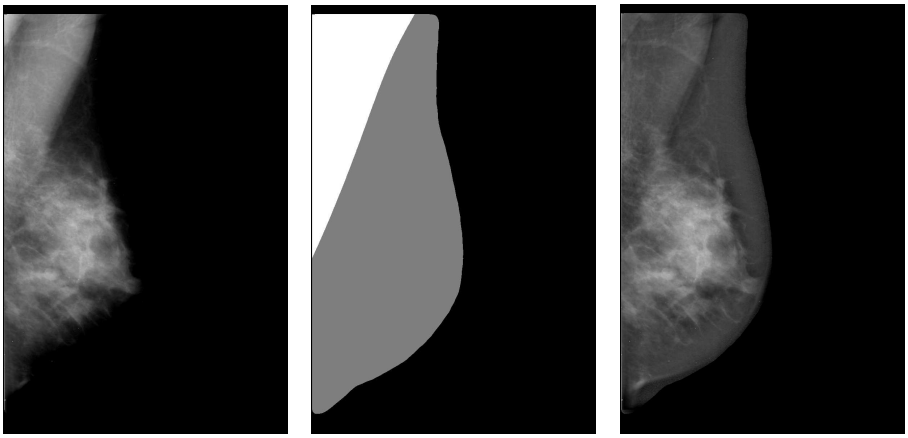


Figure 2.2: *Left the original image is shown. The middle figure shows segmentation of the image into breast area—including pectoral muscle (white)—and background tissue (black). The right figure shows the pre-processed image after pectoral fading and peripheral enhancement.*

In the next pre-processing step we adjust the grey values of the pectoral region to the grey values of rest of the breast area to make the boundary region more homogeneous. Without this adjustment problems can arise when calculating contrast measurements for tumours that are partly inside and partly outside the pectoral region. The algorithm first

calculates the mean grey value of all pixels inside the pectoral region with equal distance to the pectoral edge. Then the pixels inside the pectoral region with distance d to the pectoral edge are normalised as:

$$\tilde{y} = y + \overline{y}(0) - \overline{y}(d),$$

where $\tilde{y}$ is the normalised grey value, y the original grey value, $\overline{y}(d)$ the mean grey value of all pixels in the pectoral region with distance d to the pectoral edge and $\overline{y}(0)$ the mean grey value of all pixels that are exactly on the pectoral edge.

Finally, in the last pre-processing step, we apply a peripheral enhancement algorithm to the breast area to correct for differences in tissue thickness. This algorithm starts with calculating for each pixel the distance to the breast boundary. The maximum distance is denoted by d_{max}. Next we determine the mean grey value $g1$ and minimum grey value $g2$ of all pixels with a distance $d > \frac{2}{5}d_{max}$. We then define a threshold T as

$$T = \frac{1}{2}(g1 + g2)$$

and adjust all pixels inside the mammogram for which the smoothed grey value $y_s < T$ as:

$$y_p = \tilde{y} + (T - y_s),$$

where y_s is obtained by smoothing the original image with a Gaussian filter with a sigma of 5 mm. The grey value after pre-processing is given by y_p.

2.2 Pixel Level Mass Detection Algorithm

After pre-processing we apply a pixel level mass detection algorithm to all pixels in the breast area. This algorithm calculates at each location two features for the detection of stellate lesions and two features for the detection of focal masses. A neural network classifier combines these features into the so-called *mass likelihood*, which represents the likelihood that a mass is present at that location. Below we shortly describe the algorithm, for details see (Te Brake & Karssemeijer 1999) and (Karssemeijer & Te Brake 1996).

Features to Detect Stellate Lesions We use two features to detect spiculation, as this is a characteristic feature of malignant lesions. The spiculation features are based on the idea that stellate lesions show a pattern of lines directed towards the centre pixel of a lesion. To determine whether a spiculated lesion is present at a certain location (m, n) inside the image, we define a circular neighbourhood around (m, n). We estimate the line orientation at each location inside this neighbourhood using directional second order Gaussian derivatives. For spiculated lesions most pixels in this neighbourhood will have a line orientation towards the centre pixel (m, n). The first feature $f1$ is a

normalised measure of the fraction of pixels with a line orientation directed towards the centre pixel. We call this set of pixels F. For the second feature $f2$ we divide the circular neighbourhood into 24 angular sections. This feature measures to what extent the pixels in set F are uniformly distributed among all angular sections.

Features to Detect Focal Mass Lesions The approach for the detection of focal mass lesions is similar to the one used for the detection of spicules. We first define around each location (m, n) in the image a circular neighbourhood. Next we determine the gradient orientations at each location in this neighbourhood. When a focal mass lesion is present, pixels in this neighbourhood will have a gradient orientation towards the centre pixel (m, n). Otherwise the gradient directions will be random. We derive the following two features from the calculated gradient orientations. The first feature $g1$ is a normalised measure of the fraction of pixels with a gradient direction pointing towards the centre pixel. We call this set of pixels G. The second feature $g2$ indicates whether the pixels in set G are uniformly distributed among all angular sections.

Mass Likelihood A simple 3-layer feed-forward neural network trained on known abnormalities classifies each pixel using the above described features. The classifier output represents the likelihood that a mass is present at that location. Therefore we call this classifier output the *mass likelihood*. The corresponding image is called the *likelihood image*, see for example the middle row images in Figures 6.2 and 6.3. In the likelihood image each pixel is assigned a grey value corresponding with the *mass likelihood*, such that high grey values indicate suspicious locations and vice versa. We slightly smooth the likelihood image and then select locations with a high mass likelihood for further processing. When a selected location is closer than 1 cm from another selected location we remove the least suspicious selection as we expect that both selections belong to the same suspicious region. We apply the method at a high sensitivity level to ensure that most mass lesions are found. The average number of selected locations per image is ten.

2.3 Region Segmentation and Feature Calculation

The next step in the CAD programme concerns segmentation of the image at the selected locations and feature extraction. For segmentation we developed a new method based on dynamic programming. Chapter 3 describes this method in detail. After segmentation different features are calculated for each region: local area features, region features, and border features. Local area features only depend on the selected location. These features are thus independent of the segmentation. Region and border features on the other hand do depend on the contour. Region features represent characteristics of the segmented region; border features specifically aim at characterising the border of a region. We

use these features for either the detection of mass lesions or for discriminating between benign and malignant lesions.

2.3.1 Local Area Features

Table 2.1 summarises the local area features used in the single view CAD programme.

Local Spiculation Measures	
$f1$	presence of spiculation: concentration of spicules
$f2$	presence of spiculation: angular distribution of spicules
Local Mass Measures	
$g1$	presence of a focal mass: gradient concentration
$g2$	presence of a focal mass: angular gradient distribution
Mass Likelihood Measures	
l	mass likelihood, presence of a mass lesion
$l2$	mass likelihood relative to location 1 to 5
$l3$	mass likelihood relative to location 4 to 8
Location Features	
relx	relative x location
rely	relative y location
skindist	shortest distance to the skin line
pectdist	shortest distance to the pectoral edge

Table 2.1: *Description of the local area features used in our CAD programme. We calculate these features at the most suspicious locations in the breast area.*

Local Spiculation Measures For each selected location we determine the local spiculation measures $f1$ and $f2$, see Section 2.2. These features measure to what extent a stellate pattern is present.

Local Mass Measures For each selected location we determine the local mass measures $g1$ and $g2$, which represent the presence of a focal mass lesion, see Section 2.2.

Mass Likelihood Measures The first mass likelihood measure l—see Section 2.2—is the output from the neural network classifier and indicates whether a location is suspicious for containing a mass lesion. Other mass likelihood measures determine the

suspiciousness of a location relative to other locations in the breast area. When multiple, equally suspicious locations are present in the breast area, this might be due to properties of glandular tissue. A single suspicious location on the other hand is more likely to represent a real mass lesion. Therefore we first number all suspicious locations in the breast area in order of increasing mass likelihood, where the most suspicious location is called loc0, the second most suspicious location loc1, and so on. We determine two relative likelihood measures $l1$ and $l2$ by scaling the mass likelihood l with mass likelihood measures of other selected locations in the image. For $l2$ we scale the mass likelihood with the average mass likelihood of loc1 to loc4. For $l3$ we scale the mass likelihood with the average mass likelihood of loc5 to loc8.

Location Features Malignant lesions have a preference to develop in the upper outer quadrant of the breast (Caulkin *et al.* 1998). Therefore we include some features that indicate the location relative to the pectoral edge. For this purpose we define a new coordinate system which differs for medio lateral oblique (MLO) and cranio caudal (CC) views. In MLO views the fitted pectoral edge serves as the y-axis, in CC views the chest wall boundary of the image is taken as the y-axis. For both views the x-axis is the line perpendicular to the y-axis that has the longest distance from the y-axis to the breast boundary. The new coordinate system defines the relative x- and y-location of the centre of each region. To compensate for differences in breast size these coordinates are normalised with the effective radius of the breast $r = \sqrt{A/\pi}$, where A is the size of the segmented breast region.

2.3.2 Region Features

For each region we define the following three sections: the segmented region itself, the border region, and the surround region. The segmented region contains all pixels inside the contour. The border region forms a narrow band along the contour and contains all pixels with a distance of less than 1 mm to the contour, inside as well as outside the segmented region. Thus including the contour the width of this band is about 2.2 mm. The surround consists of all pixels outside the segmentation with distance of less than $0.6\,r$ from the contour, where r is the effective radius of the segmented region. The surround is about twice the size of the segmented region. Figure 2.3 shows a segmented region and its surround. Table 2.2 summarises the region features. In the sequel we use the following notation. We denote the set of pixels in the segmented region by I, the set of pixels on the contour by C, and the set of pixels in the surround by S. $N(X)$ is the number of pixels in set X. The mean grey level of the pixels in set X is denoted by $\overline{y}(X)$, the grey level standard deviation of the pixels in set X by $\sigma(y|X)$.

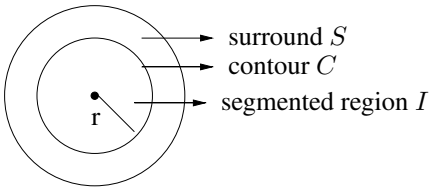


Figure 2.3: *A segmentation algorithm determines a contour for each region. The surround consists of all pixels within a distance of* $0.6\,r$ *from the contour, where* r *is the effective radius of the segmented region.*

	Spiculation Features
$\overline{f1}$	mean value of $f1$ inside segmented region I
$\overline{f2}$	mean value of $f2$ inside segmented region I
	Focal Mass Features
$\overline{g1}$	mean value of $g1$ inside segmented region I
$\overline{g2}$	mean value of $g2$ inside segmented region I
	Dense Tissue Features
D_B	fraction dense tissue in whole breast area
D_S	fraction dense tissue in surround S
$\overline{ll}(I)$	likelihood ratio between D and F of segmented region
$\overline{ll}(S)$	likelihood ratio between D and F in surround S
$lldiff$	ratio between $L(I)$ and $L(S)$
	Contrast Features
Int	mean grey value of segmented region I
$C1$	contrast difference between regions I and S
$C2$	normalised contrast difference
$C3$	contrast distance between regions I and S
$C4$	relative contrast difference
$C5$	relative contrast difference

Table 2.2: *Description of region features used for detection and classification.* I *denotes the segmented region,* S *the surround.* F *and* D *indicate the set of pixels in the fatty and dense parts of the breast.*

Gray Level Variance	
var1	grey level variance of region I
var2	difference in grey level variance between I and S
var3	grey level variance in I relative to fatty tissue
var4	grey level variance in I relative to dense tissue
Linear Texture	
$T1$	presence of linear texture in I
$T2$	presence of linear texture in I, normalised
$T3$	presence of linear texture in S
Border Features	
BC	continuity of the border of the segmented region
$\overline{FD}$	average first order directional derivative along border
$\overline{SD}$	average second order directional derivative along border
Other	
ID	iso-denseness of the segmented region
size	area of segmented region I
circularity	(perimeter)^2/size
pect	location in pectoral area or not
Wolfe	estimated Wolfe class
MC	number of micro-calcifications present in I

Table 2.2: *(cont.) Description of region features used for detection and classification. I denotes the segmented region, S the surround. F and D indicate the set of pixels in the fatty and dense parts of the breast respectively.*

Regional Spiculation Features For both local spiculation measures $f1$ and $f2$ we determine the average value inside the segmented region. We call these features $\overline{f1}$ and $\overline{f2}$.

Regional Mass Features For both local mass measures $g1$ and $g2$ we determine the average value inside the segmented region. We call these features $\overline{g1}$ and $\overline{g2}$.

Dense Tissue Features Dense tissue features provide information about the presence of dense tissue in the segmented region and its surround. We use a Gaussian mixture model to estimate the distribution of fatty and dense tissue in the breast area (Karssemei-

jer 1998). Based on this model we segment the breast into a fatty part and a dense part, indicated by F and D respectively. The first two dense features represent the fraction dense tissue in the whole breast B and in the surround S:

$$D_B = \frac{N_d(B)}{N_d(B) + N_f(B)}$$

and

$$D_S = \frac{N_d(S)}{N_d(S) + N_f(S)},$$

where $N_d(X)$ and $N_f(X)$ are the number of dense and fatty tissue pixels in the set X. For the dense and fatty parts of the breast we calculate the mean grey level and the grey level variance. Then we determine for each grey value y the log likelihood ratio between both tissue types:

$$ll(y) = \frac{(y - \overline{y}(F))^2}{\sigma^2(y|F)} - \frac{(y - \overline{y}(D))^2}{\sigma^2(y|D)} + \log(\sigma(y|F)) - \log(\sigma(y|D))$$

The third $\overline{ll}(I)$ and the fourth feature $\overline{ll}(S)$ give the mean value of the log likelihood ratio in the segmented region I and the surround S. The last feature is the ratio between $\overline{ll}(I)$ and $\overline{ll}(S)$:

$$lldiff = \overline{ll}(I)/\overline{ll}(S)$$

Intensity and Contrast Features The contrast of a region is a useful feature since tumour tissue absorbs more X-rays than fat and also slightly more than glandular tissue. The first contrast feature is the mean grey level of the segmented region,

$$\text{Int} = \overline{y}(I).$$

We define five distance measures to indicate differences in contrast between the segmented region and its surround. The first distance measure is the difference in intensity:

$$C1 = \overline{y}(I) - \overline{y}(S).$$

The second distance measure is the squared difference in intensity between the segmented region and its surround, divided by both standard deviations,

$$C2 = \frac{(\overline{y}(I) - \overline{y}(S))^2}{\sigma(y|I) + \sigma(y|S)}.$$

The third distance measure represents the distance between the grey level histogram of the segmented region and the surround area,

$$C3 = \sum_y |H(y|I) - H(y|S)|,$$

where $H(y|X)$ denotes the fraction of pixels in set X with intensity value y. To limit the number of entries, we divide the intensity range into 82 bins, each containing a range of 50 grey values. The fourth and fifth measure calculate the contrast difference relative to some dense tissue parameters,

$$C4 = \frac{\overline{y}(I) - \overline{y}(S)}{\overline{y}(D) - \overline{y}(F)}, \qquad C5 = \frac{\overline{y}(I) - \overline{y}(S)}{\sigma(y|F)}.$$

Variance Malignant masses often show little grey level variance compared to normal breast tissue. Therefore we define some features based on the grey level variance of the segmented region and its surround. The first feature is the grey level variance in the segmented region,

$$\text{var1} = \sigma^2(y|I).$$

The second feature is the ratio between the variance in the segmented region and its surround,

$$\text{var2} = \frac{\sigma^2(y|I)}{\sigma^2(y|S)}.$$

The third and fourth feature calculate the grey level variance of the segmented region relative to the variance measured in the fatty or dense parts of the breast,

$$\text{var3} = \frac{\sigma^2(y|I)}{\sigma^2(y|F)}, \qquad \text{var4} = \frac{\sigma^2(y|I)}{\sigma^2(y|D)}.$$

Linear Texture Normal breast tissue often has different texture characteristics than tumour tissue. Karssemeijer & Te Brake (1996) developed three texture measures to capture linear structures as these often indicate the presence of normal breast tissue. To determine these features we need the map of line orientations and magnitudes that we constructed for calculating our spiculation measures, see Section 2.2. This map contains for each location in the breast area a vector representing the line orientation and magnitude. We compute this map at two different scales using second order Gaussian derivatives with a sigma of 0.3 mm and a sigma of 0.6 mm. We then sum all vectors in the inside region using the double angle representation, resulting in a final *sum vector*. Next we calculate three different linear texture features. The first texture feature $T1$ is the magnitude of the *sum vector*. The second texture feature $T2$ is the magnitude of the *sum vector* divided by the sum of the magnitudes of all vectors in the segmented area. The third texture feature $T3$ is the magnitude of the *sum vector* in the surround area S.

Iso-denseness When dark areas are present inside a segmented region it is likely that the region is a normal structure. Tumours on the other hand are often dense compared to the surrounding tissue. Te Brake *et al.* (2000) developed a feature that measures the

"denseness" of a segmented region compared to the surround region. This feature first determines a threshold t that indicates the maximum of the 10% lowest grey values that are found inside the segmented region:

$$t = \arg\max_{k} \sum_{y=0}^{k} H(y|I) < 0.1.$$

The iso-denseness feature is the fraction of pixels in the surround area S with a value lower than the threshold t:

$$ID = \sum_{y=0}^{t} H(y|S).$$

A value close to one indicates a high likelihood for the presence of a tumour.

Morphological Features We include two morphological features. The first one is the size (area) of the segmented region. Studies show that malignant masses on average are larger than benign ones (Timp *et al.* 2005). The second morphological feature measures to what extent the segmented region is circularly shaped. We include this feature because benign masses often have a round or oval shape compared to a more irregular shape of malignant masses. We define circularity as

$$c = p^2/A,$$

where p is the perimeter and A the size of the region.

Pectoral Overlap This feature quantifies to what extent the segmented region is located inside or near the pectoral area.

Presence of Micro-calcifications. The presence of micro-calcifications at the location of a mass lesion is a sign of malignancy. Therefore we use a programme for the detection of micro-calcifications (ImageChecker, R2 Technology, Sunnyvale (CA)). As feature we use the number of calcifications found in the segmented region.

Wolfe Class Studies show that there is a relation between parenchymal patterns and the risk of developing breast cancer. Wolfe defined four types of parenchymal patterns, ranging from fatty to predominantly dense breasts. We use an automated programme from Karssemeijer (1998) to classify each breast as one of these parenchymal patterns.

Border Features Border features are especially useful to discriminate between benign and malignant lesions. Most benign lesion can be characterised as circumscribed or well-defined lesions. Margins of these lesions are sharply demarcated with an abrupt transition between the lesion and its surrounding tissue, which reflects the absence of infiltration. Malignant lesions on the other hand often have ill-defined or spiculated borders. Therefore we designed some quantitative measures that indicate to what extent the margin of a lesion is continuous and circumscribed. These features are described in detail in (Varela *et al.* 2005).

2.4 Classifier Training and Testing

The last step of the CAD programme involves training and testing a classifier. To this end we first select a subset of features appropriate for the task of the CAD system, which is either detecting masses or classifying masses as benign or malignant. A classifier trained on known abnormalities combines the selected features into a so-called *malignancy score*, which indicates the likelihood that a region is malignant. In this thesis we use different classifiers such as linear discriminant analysis, Support Vector Machines, k-Nearest Neighbour, and Neural Networks. For further reading the following books can be consulted: Duda *et al.* (2001); Bishop (1995); Ripley (1996); Fukunaga (1990). Training and testing of the classifier are implemented using a cross-validation or leave-one-out scheme where a part of the dataset is used for training and the other part for testing. In this way training and testing are done completely independent.

2.5 Performance Evaluation

For performance evaluation we use Receiver Operating Characteristic (ROC) and Free-response Receiver Operating Characteristic (FROC) methodology. FROC analysis is used when the CAD system aims at detecting masses; ROC analysis when the CAD system aims to characterise mass lesions as benign or malignant. Both methods are described below.

2.5.1 Classification of Masses

To evaluate the performance of the CAD system in classifying masses as benign or malignant we use Receiver Operating Characteristic (ROC) methodology. Figure 2.4 shows an ROC curve. ROC curves usually plot sensitivity—also called the *true positive fraction* or TPF—as a function of [1-specificity], called the *false positive fraction* or FPF. We can evaluate the performance of a CAD system case based, image based, and lesion based.

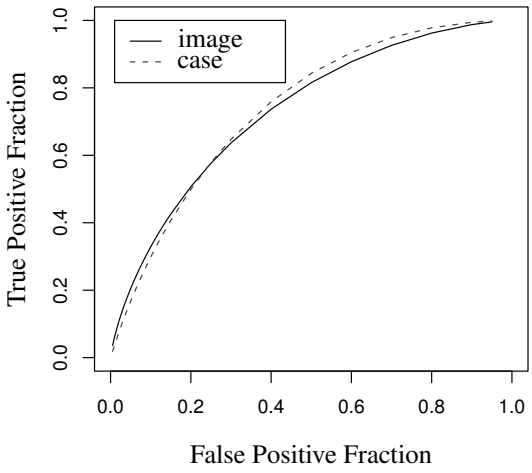


Figure 2.4: *ROC curve. The horizontal axis represents the true positive fraction (TPF), the vertical axis the false positive fraction (FPF).*

For image based evaluation we use the malignancy scores of single view images to determine the ROC curve. For case based evaluation we combine the malignancy scores from the CC and MLO view into a single *case based score*. The malignancy scores are often combined by taking the minimum, maximum, or the average of all malignancy scores. When the breast contains multiple lesions it is better to do a lesion based evaluation. In this evaluation we only combine the malignancy scores from both views when the lesions represent the same underlying mass lesion. For each type of evaluation we can use the area under the ROC curve—the A_z value—as a performance measure for the CAD system. A value close to one indicates high sensitivity and high specificity.

2.5.2 Detection of Masses

To evaluate the detection accuracy of CAD systems we use FROC methodology. Figure 2.5 shows an FROC curve. The horizontal axis indicates the average number of false positive detections per image, the vertical axis the fraction of correctly detected masses (sensitivity). We use a logarithmic scale for the x-axis to show the performance of the CAD system at a low number of false positive detections per image. For the FROC curve in Figure 2.5 we used a dataset consisting of 500 images from women that have been

referred during screening. This figure shows that almost all tumours are detected at a

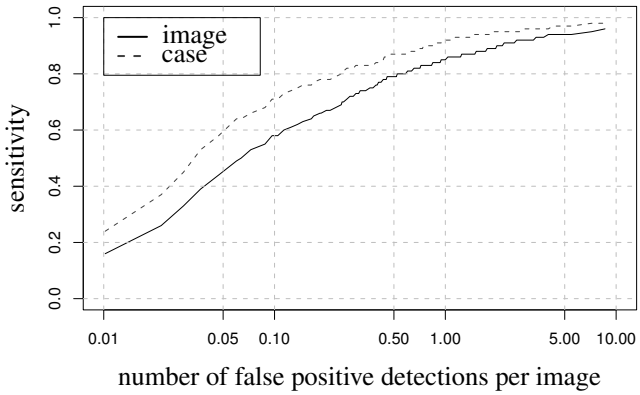


Figure 2.5: *FROC curve. The horizontal axis represents the number of false positive detections, the vertical axis the sensitivity.*

high false positive rate. Screening however asks for a referral rate of about 1-5%. The detection percentage at this referral rate is quite low, which illustrates the problem of current CAD systems. Figure 2.5 shows both image and case based curves. For image based analysis we consider a tumour as detected when the initial detection location is inside the ground truth. If multiple detections are found inside the same ground truth region they are considered as a single hit. We count detections outside the ground truth areas as false positives. For case based analysis we consider a tumour as detected when it is found on either the CC or the MLO view. To obtain some quantitative performance measure we can calculate the area under the whole FROC curve or under a part of the FROC curve, for instance from 0.1 to 1.0 false positive per image. We can use a logarithmic scale for the x-axis when determining the area under the FROC curve. The advantage of using a logarithmic scale is that greater weight is assigned to the part of the curve where the number of false positive detections is low, which corresponds with the operating point for normal screening situations.

Chapter 3

Mass Segmentation based on Dynamic Programming [1]

An important step in CAD programmes is the segmentation of mammographic lesions. After segmentation different features can be determined that depend on the contour. These include the region and border features that have been described in Section 2.3. Examples are size and contrast of a lesion, or the sharpness of the border. This chapter presents a robust and fast algorithm to accurately determine the contour of mammographic lesions. Furthermore we compare this method with two well known segmentation methods from literature: region growing and the discrete dynamic contour model. Section 3.2 explains each segmentation method. Then, in Section 3.3, we describe the experiments to evaluate the different segmentation methods. In the first experiment we evaluate the segmentation performance of each method by comparing the resulting contour with a manual outline of the lesion. In the second and third experiment we investigate the influence of the segmentation accuracy on the performance of a CAD system for the detection and characterisation of mass lesions. Section 3.4 presents the results.

3.1 Introduction

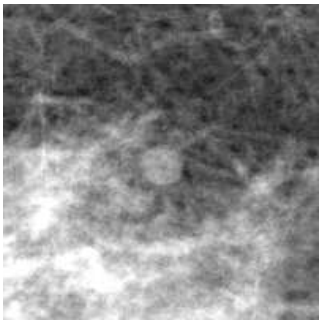
Globally segmentation methods fall into two main categories: region based and edge based. Methods of both categories have been applied to the segmentation of mammographic masses.

The first category assigns each pixel to a particular object or region. Examples are split-and-merge algorithms and region growing techniques. Region growing is one of the

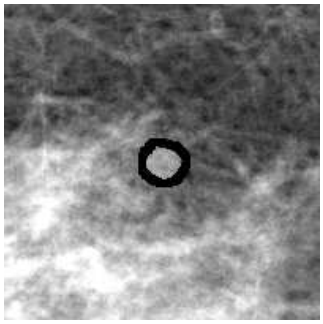
[1] The content of this chapter has been published previously in Timp *et al.* (2002b) and Timp & Karssemeijer (2004a).

most popular segmentation methods and many different approaches have been proposed. Kupinski & Giger (1998) developed two extended region growing techniques, one based on the radial gradient index and another based on simple probabilistic models. They tested these methods against a conventional region growing algorithm using a database of biopsy proven, malignant lesions and found that the new lesion segmentation algorithms more closely matched radiologists' outlines of these lesions. Guliato *et al.* (1998) proposed fuzzy region growing methods for segmenting breast masses and further classified the segmented masses as benign or malignant based on the transition information present around the segmented region. Petrick *et al.* (1999) applied object based region growing in combination with a density-weighted contrast enhancement filter to segment all significant structures within the breast. Region based segmentation algorithms have two main disadvantages. First, small and low-contrast structures have a tendency to grow into the background and become large regions even though the actual mass is quite small. An example is given in Figure 3.1(d). The region growing method fails to find the border of the mass and the resulting segmentation is too large. Second, structures containing internal gradients do not always grow to the correct border but can end up containing only a section of the true object.

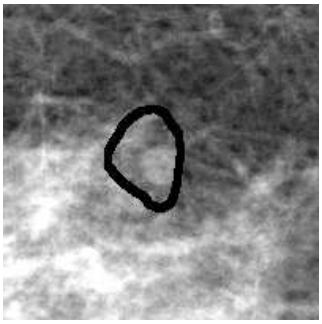
The second category are edge based algorithms. These algorithms aim at detecting the boundary of an object. Most algorithms first construct a so-called *edge image*. In the edge image each pixel is assigned a value according to the edge strength. Based on this image, pixels with strong edges are selected and linked to each other. In most cases the linked pixels will represent object boundaries. A disadvantage of the original edge based algorithms is that these do not guarantee a closed contour. To overcome this problem an active contour model (snake) was developed for contour detection. Dynamic contour models (snakes) have become en vogue with the snake model of McInerney & Terzopoulos (1996) and have since then been investigated and applied in various ways. The snake model builds a deformable contour consisting of connected spline segments and lets the contour approximate a desired form by minimising an energy function containing internal and external energy. The internal energy is the bending energy of the spline, the external energy is calculated by integrating image features, like the presence of lines and edges. Lobregt & Viergever (1995) developed a discrete version of the snake model (discrete contour model) and applied this model to medical images. The main drawback of these edge based models for the task of mammographic mass segmentation is that the algorithms heavily depend on being initialised with a contour that is close to the actual boundary. Otherwise the contour may stick to the first strong edge it finds. An example is shown is Figure 3.2(c). The initial estimate of the contour, shown in black, is too far from the mass boundary. As a result the model is not able to find the contour and instead is attracted to the pectoral muscle. Another known problem with deformable models is that the model may shrink owing to internal forces, when the edges are not strong enough or too far from the initial contour.



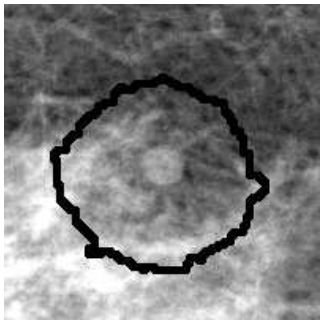
(a) Benign mass



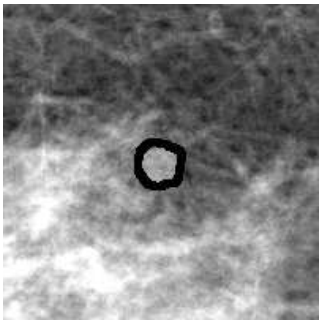
(b) Manual segmentation



(c) Discrete contour model

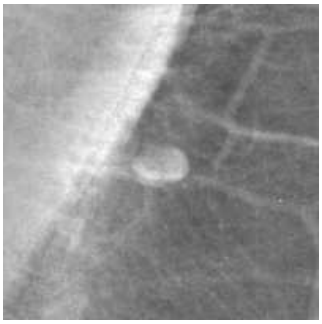


(d) Region growing

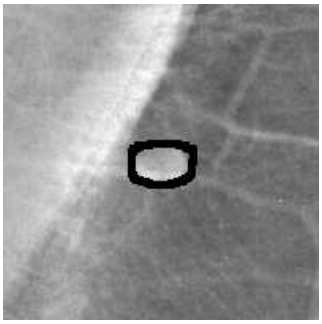


(e) Dynamic programming

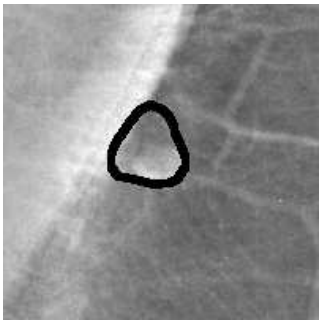
Figure 3.1: *Segmentation results for a benign mass.*



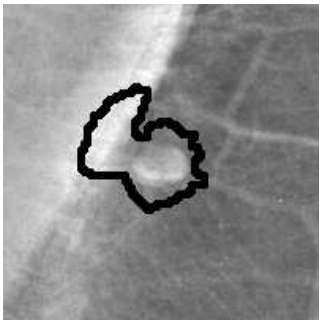
(a) Benign mass



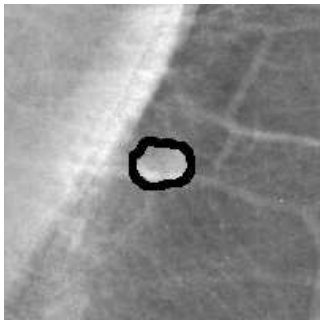
(b) Manual segmentation



(c) Discrete contour model



(d) Region growing



(e) Dynamic programming

Figure 3.2: *Segmentation results for a benign mass located near the pectoral muscle.*

There are a few studies that compare different segmentation methods (Te Brake *et al.* 1999; Timp *et al.* 2002b; Sahiner *et al.* 2001). In a previous study we compared three segmentation methods with manual segmentation (Timp *et al.* 2002b). In that study we did not evaluate the effect of the segmentation on the classification performance. Te Brake *et al.* (1999) compared the discrete contour model from Lobregt & Viergever (1995) with the region growing algorithms developed by Kupinski & Giger (1998) and evaluated the methods by comparing them to manual segmentation. Furthermore they studied the effect of the segmentation on the cancer detection performance. One of the region growing methods and the discrete contour model performed equally well in the segmentation task. In the detection experiment the discrete contour model had a higher performance in classifying each segmented region as normal or abnormal. Sahiner *et al.* (2001) compared a mass segmentation method based on an active contour model with manual segmentation and studied the effect of the segmentation on the classification accuracy. They found that the classification performance obtained with features extracted from a manually or an automatically segmented region were nearly identical.

In this study we develop a new segmentation method to overcome the problems of region growing and the discrete contour model. The new method uses both edge based information as well as a priori knowledge about the grey level distribution of an ROI (region of interest) around the mass. We select the best contour using an optimisation technique based on dynamic programming. To test the performance of this method, we compare our proposed method with region growing and the discrete contour model using an area overlap criterion. Furthermore we study the effect of the segmentation on the detection and characterisation of mammographic masses.

3.2 Segmentation Methods

In this section we describe the three segmentation methods used in this work. The first subsection describes the dynamic programming approach. In the second and third subsection we briefly review the region growing method and the discrete contour model.

3.2.1 Dynamic Programming

Dynamic programming is an optimisation technique that can be used to find the boundary of objects (Ballard & Brown 1982). For this purpose the boundary definition problem is first formulated as a graph searching problem. The dynamic programming algorithm then finds the optimal path between a set of start nodes and a set of end nodes of this graph. Typical applications of the use of dynamic programming in boundary tracking problems are tracing borders of elongated objects like roads and rivers in aerial photographs and the segmentation of handwritten characters. Medical applications include the segmentation

of spine boundaries and tracing vessel borders.

In this study we apply dynamic programming to find the boundary of a mass. Most mass lesions are approximately circular in shape. We implement this circularity constraint by carrying out the calculations in polar space. We first determine the centre of the mass lesion (μ_x, μ_y). Then we define a circular region of interest (ROI) with centre (μ_x, μ_y) and radius R. The radius should be large enough to allow application of the algorithm to masses of different sizes. We choose a radius of 2.4 cm. Next we transform the circular ROI to a polar ROI where the x-axis represents the angle from $-\pi$ to π and the y-axis the radius r from 0 to R. Figure 3.3(a) and 3.3(b) show the coordinate transform. Finally the dynamic programming algorithm finds the *optimal path* from one of the pixels in the first column to one of the pixels in the last column of the polar ROI. We consider a path as optimal when the cumulative costs—that is the sum of the local costs of all pixels along the path—are minimal. The next section describes the local cost.

Local Cost

For each pixel in the polar image we calculate three cost measures, which represent characteristics of a good boundary. These three cost measures together form the local cost for each pixel $c(i, j)$:

$$c(i, j) = w_e e(i, j) + w_s s(i, j) + w_g g(i, j), \tag{3.1}$$

where e represents the edge strength, s the deviation from the expected size, and g the deviation from the expected grey level. The weights for the components are given by w_e, w_s and w_g. The cost measures are chosen such that pixels that possess many characteristics of the searched boundary are assigned low cost and vice versa. Below we describe each cost measure.

- Edge Strength $e(i, j)$ We assign pixels with strong edge features low cost as this may indicate the presence of the contour. To determine the edge strength we first determine for each pixel the gradient magnitude y' in the direction normal to the contour. In the polar image this corresponds with the gradient magnitude in vertical direction. Then we select the 99th percentile of all gradient magnitudes in the ROI. We call this value $\max(y')$. We obtain the relative edge strength by normalising the gradient values in the ROI with $\max(y')$. This normalisation ensures that subtle contours with low global but high local edge strength can be found as well. By taking the 99th percentile it is prevented that one outlier, for instance a very bright micro-calcification, decreases the relative edge strength of all other pixels in the ROI. We invert the normalised gradient value such that high gradients produce low costs and vice versa. The final gradient cost measure is:

$$e(i, j) = \frac{\max(y') - y'(i, j)}{\max(y')}.$$

- Deviation from expected size $s(i,j)$ We assign contours with a size that is common for masses a low cost value. On the other hand, masses that are very small or very large are assigned higher cost. Most masses have a radius between 5 mm and 15 mm, with a mean radius r of about 9 mm (Timp *et al.* 2002a). We use the following size measure in the cost function:

$$s(i,j) = \begin{cases} (j-r)^2 & : \quad j < m \\ (m-r)^2 & : \quad j \geq m \end{cases}$$

  where r is the mean radius of masses, that is 9 mm. In the polar image j represents the distance from pixel (i,j) to the centre of the mass (μ_x, μ_y). We set the maximal distance to m to prevent that the size component of the cost function completely determines the value of the cost function for large masses. We use $m = 15$ mm. Alternatively we could determine this cost measure by estimating the distribution function of the size of masses. In that case we could base the cost value for each pixel (i,j) on the relative frequency of masses with size j. To estimate this size distribution however we need a large representative database with benign and malignant masses of known size. Currently we use the first method as we do not have an independent database that we can use for this purpose.

- Deviation from expected grey level $g(i,j)$ Another predictable characteristic of the mass boundary is its grey level. We first estimate the grey level of the border and then calculate for each pixel the deviation from this expected grey level. A common assumption is that the border is located at the zero crossing of the second derivative of the edge profile. Claridge & Richter (1994) however found that in projective images the real edge is located more towards the darker side (background). Consequently, the grey value of the border will have a value closer to the background grey level than to the grey level of the mass region. Therefore we determine the preferred grey level of the border y_p as follows:

$$y_p = \alpha\, \overline{y}(\mathrm{M}) + (1-\alpha)\, \overline{y}(\mathrm{BG}),$$

  where $\overline{y}(\mathrm{M})$ and $\overline{y}(\mathrm{BG})$ are estimates of the mean grey level of the mass region (M) and the background tissue (BG). The value of α should be smaller than $1/2$ to ensure that the edge is located more towards the background level. We use two methods to estimate the grey level of the mass and the background tissue. In the first method we use a circle with centre (μ_x, μ_y) and radius 0.6 cm. We calculate the mean grey level inside and outside this circle and use these values as estimates for $\overline{y}(\mathrm{BG})$ and $\overline{y}(\mathrm{M})$. In the second method we use histogram analysis to estimate the grey level distributions of the mass and the background tissue. The histogram of the ROI contains pixels from both the inside and the outside region.

Therefore we can model this histogram reasonably well by a mixture of two Gaussian distributions, one narrow Gaussian in the low intensity range representing the fatty tissue, and a broader one in the middle/high intensity range representing the mass lesion. We estimate the parameters for the Gaussian distributions with the Levenberg-Marquardt method. We use the first peak in the histogram to estimate $\overline{y}(\mathrm{BG})$, and the second peak to estimate $\overline{y}(\mathrm{M})$. When the histogram can not be modelled by a mixture of two Gaussians we use the first method to estimate the preferred grey level. The grey level cost measure for each pixel is:

$$g(i,j) = \sqrt{|y(i,j) - y_p|},$$

where $y(i,j)$ is the grey value of the pixel (i,j).

Dynamic Programming Path Finding Algorithm

We first apply the cost function Eq. 3.1 to all pixels in the polar ROI. We then obtain the so-called *cost image*, as shown in Figure 3.3(c). The dynamic programming algorithm finds the optimal path in this image which corresponds with the best contour for this cost function. Pixels in the first column of the cost image ($\phi = -\pi$) represent the start nodes for the algorithm, whereas the end nodes are represented by the pixels in the last column of the image. The cumulative cost matrix C stores the cumulative cost of each path. Figure 3.3(d) shows the cumulative cost matrix. We construct this matrix in two steps. First we set the cumulative cost of pixels in the first column:

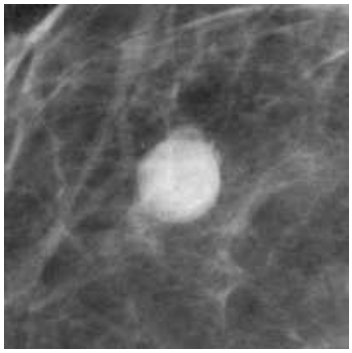
$$C(i,-\pi) = c(i,-\pi),$$

where $C(i,j)$ is the cumulative cost and $c(i,j)$ the local cost for pixel (i,j) in the polar image. For the other pixels we calculate the cumulative cost by a recursive step:

$$C(i,j+1) = \min_{-2 \le l \le 2} C(i+l,j) + c(i,j+1) + h(l), \tag{3.2}$$

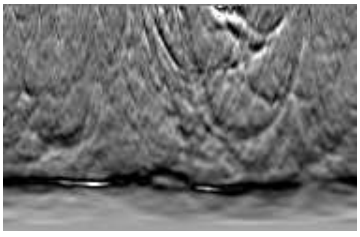
where l is the direction of the path. In this application the value of l falls inside the interval $[-2,\ldots,2]$. The function $h(l)$ is an increasing function as value of $|l|$ and controls the smoothness of the path. We use the path with the lowest cumulative cost as our final contour. The end point $C(i,\pi)$ of this contour is the pixel with the lowest cumulative cost value of all pixels in the last column. We find the optimal path by back tracing the path from the end pixel to one of the pixels in the first column. Figure 3.3(e) shows the final path in the cumulative cost matrix. Figure 3.3(f) shows the resulting segmentation in the original image.

Final Contour The dynamic programming algorithm does not guarantee that the final contour is closed. In our application we consider a contour as closed when the distance



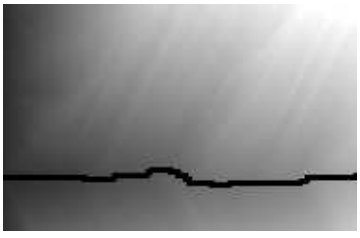
(a) ROI with a benign mass

(b) Polar ROI

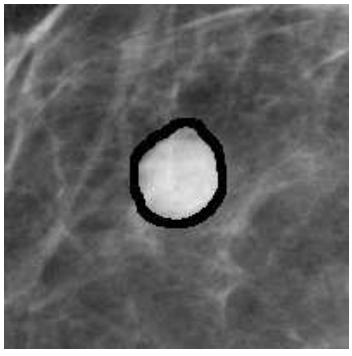


(c) Cost matrix

(d) Cumulative cost matrix



(e) Path found by the dynamic programming segmentation algorithm



(f) Final contour

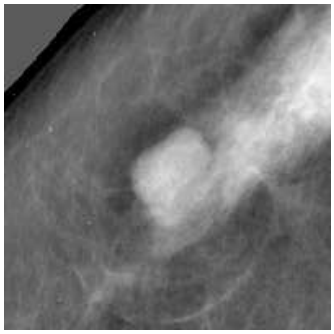
Figure 3.3: *Implementation of the proposed dynamic programming algorithm.*

between the start and the end point is less than 3 pixels. This is conform the interval of the direction parameter l. In most cases, especially when the mass is clearly visible, the algorithm will find a closed contour. When the mass is ill-defined or when other structures obscure the mass boundary, the segmentation programme can fail to find a closed contour. Figure 3.4 illustrates this problem. The small mass in the middle of the image is surrounded by some dense tissue. Figure 3.4(b) shows the polar image, with the resulting contour plotted on top of it. In the beginning the path is attracted to an image structure and deviates from the true mass boundary. As a consequence the contour is not closed and contains some extra tissue. Figure 3.4(d) shows the final contour on the original image.

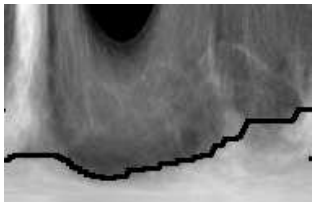
There are some methods to guarantee the contour to be closed. One of the methods is to calculate the optimal path for each radius $r = [0, \ldots, R]$ under the constraint that the start and the end point are $(r, -\pi)$ and (r, π) and thus have the same r-coordinate. The method works as follows. For a chosen value of r extra cost is added to all points in the first column of the cost matrix except to the point $(r, -\pi)$. Then the cumulative cost matrix is constructed. The optimal path is found by back tracing the path from the end point (r, π). The extra cost ensures that the path is back traced till the start point $(r, -\pi)$. This results in a path with start point $(r, -\pi)$ and end point (r, π). We call the cumulative cost associated with this path C_r. After having determined the cumulative cost for each value of r we select the path with the lowest cumulative cost C_r. This path represents the final contour. We call this the constraint algorithm. A disadvantage of this method is that the algorithm has to be applied once for each value of r which makes it computationally expensive.

We designed a more efficient method to ensure that the resulting contour is closed. Our solution uses an extended cost matrix where the cost matrix runs from $-\beta\pi$ to $\beta\pi$. The extension factor β determines the size of the extended cost matrix relative to the original cost matrix. We use the dynamic programming algorithm to find the optimal path in this extended cost matrix and extract the path from $-\pi$ to π as our final contour. In the original cost matrix the final contour depends strongly on the initial angle of the polar coordinate transform—in our case this is $-\pi$—and the resulting segmentation might be different for different initial angles. A disadvantage of this dependency is that image features near the boundaries of the interval $[-\pi, \ldots, \pi]$ can have undesirable effects on the resulting contour. In the new method we minimise the dependence on the initial angle. Consequently image features near the boundaries of the interval have less effect on the final contour. Another advantage of this method is that discontinuities at $-\pi$ and π are avoided which in turn may lead to more closed contours.

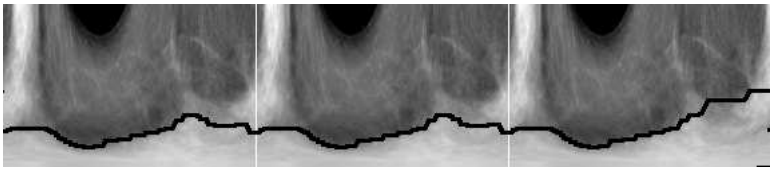
To determine the efficiency of this method we set up the following experiment. For each extension factor β we apply the dynamic programming algorithm to find the optimal contour. Afterward, we calculate for each extension factor the percentage of closed contours.



(a) Original image



(b) Polar image from $-\pi$ to π



(c) Polar image from -3π to 3π

(d) Contour as extracted from the polar image from $-\pi$ to π

(e) Contour as extracted from the extended polar image

Figure 3.4: *The original dynamic programming segmentation algorithm does not guarantee a closed contour. In this example the optimal path is attracted towards some dense tissue and the resulting contour is not closed. The new algorithm extracts the contour from an extended cost matrix resulting in a closed contour.*

3.2.2 Region Growing

Region growing is one of the most popular segmentation methods. Region growing takes an image and a seed point (s_x, s_y) as input. The seed point (s_x, s_y) is defined to be within the suspect region $\mathcal{R}$. Region growing then grows the seed regions in an iterative fashion. At each iteration the pixels that border the growing regions are examined. Conventional region growing defines several region partitions $\mathcal{R}_i$ based solely on grey level information in the image:

$$\mathcal{R}_i = \{y(i,j) > t_i\},$$

where $y(i,j)$ is the pixel grey level and t_i is the grey level threshold for partition $\mathcal{R}_i$. For each partition features are calculated such as circularity and size. Based on these features the partition that best characterises a mammographic lesion is selected as the final segmentation.

We implemented an extended version of the algorithm developed by Kupinski & Giger (1998). In this method the partitions are created using grey level information as well as prior knowledge about the shape of typical mass lesions. To include information about the shape of mammographic lesions, the region is pre-processed by multiplication with a Gaussian centred at (s_x, s_y). The partitions returned by thresholding are now more compact than before because distant pixels are suppressed. To determine which partition best delineates a mammographic lesion a likelihood measure is used. This measure estimates the grey level distribution for grey levels inside and outside the region for each partition. The partition that maximises this likelihood is selected as the final segmentation.

3.2.3 Discrete Contour Model

The active contour model (or snake) formulates the boundary detection issue as an energy function minimisation problem (McInerney & Terzopoulos 1996). We implemented a discrete version developed by Lobregt & Viergever (1995), the discrete contour model. Starting from an initial shape the discrete contour model actively modifies its shape approximating some desired contour. Internal and external forces together determine the final shape deformation.

The basic structure of the model is a set of vertices v_i which are connected by edges d_i, see Figure 3.5. The unit vector $\hat{d}_i$ describes the direction of d_i. For each vertex i with connecting edges d_i and d_{i-1} a local coordinate system is constructed represented by a tangential unit vector $\hat{t}_i$

$$\hat{t}_i = \frac{\hat{d}_i + \hat{d}_{i-1}}{\|\hat{d}_i + \hat{d}_{i-1}\|},$$

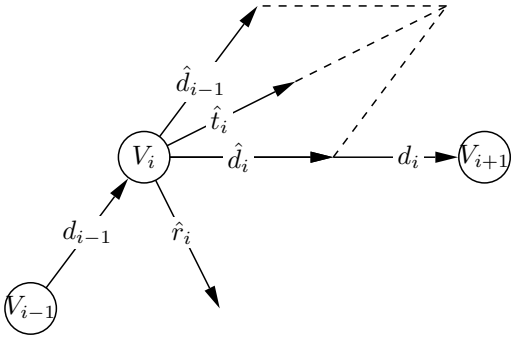


Figure 3.5: *Part of the discrete dynamic contour model. The discrete dynamic contour model consists of a set of vertices v_i that are connected by edges d_i.*

and a radial unit vector $\hat{r}_i$

$$\hat{r}_i = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \hat{t}_i.$$

The internal force is based on the local shape of the contour, and aims at minimising local curvature. The local curvature l_i for vertex v_i is the difference between the directions of the two edge segments that join at that location:

$$l_i = \hat{d}_i - \hat{d}_{i-1}.$$

Local curvature therefore, has direction equal or opposite to the radial vector $\hat{r}_i$. To prevent the contour from imploding, vertices in parts of the contour with constant curvature should have an internal force of zero. To achieve this, the internal force of vertex v_i is computed by combining its local curvature with the local curvature of the two neighbour vertices in a local coordinate system:

$$f_{in,i} = \{-\frac{1}{2}(l_{i-1}\hat{r}_{i-1}) + l_i\hat{r}_i - \frac{1}{2}(l_{i+1}\hat{r}_{i+1})\}\hat{r}_i.$$

The external force $f_{ext,i}$ is based on the image gradient magnitude. This force moves the vertices to locations in the image with strong gradients: the edges of the mass. Computation of the external force is done in the radial direction, as this prevents vertices from moving along the contour.

The total force f_i acting on vertex v_i is a weighted combination of external and internal forces. As a result of this force the vertex v_i will start to move and change its position. The deformation process stops when the system reaches a stable state.

3.3 Experiments to Evaluate Segmentation Methods

We performed four experiments to evaluate the performance of the dynamic programming method. In the first experiment we estimate the value of the extension factor that is needed to guarantee a closed contour. In the second experiment we quantitatively analyse the segmentation performance of the new method compared with the other two methods—region growing and the discrete contour model—using an overlap criterion. We did two additional experiments to evaluate the effect of the segmentation on the ability of the CAD system to detect and characterise mass lesions. The next subsection first describes the dataset used for the experiments. The other subsections explain each experiment in more detail.

3.3.1 Database

The mammograms used in this study all came from the Dutch Breast Cancer Screening Programme. All women aged 50-70 are invited bi-annually to participate in this programme. Two mammographic projections—medio lateral oblique (MLO) and cranio caudal (CC)—are obtained at the initial screening in this programme. At subsequent screenings only medio lateral views are obtained, unless there is an indication that additional cranio caudal views would be beneficial. The mammograms were digitised with a Canon laser scanner at a pixel resolution of 50 μm, and averaged down to a resolution of 200 μm maintaining the original grey value resolution of 12 bits.

The total dataset consisted of 1427 two view and four view mammograms, resulting in a total of 4295 images. We excluded images with only micro-calcifications. The remaining set consisted of 1152 images each containing at least one biopsy proven mass, called the *mass dataset*, and 2822 normal images without pathology, called the *normal dataset*. The *mass dataset* contained a total of 1210 masses, 551 malignant and 659 benign, including spiculated, circumscribed, and ill-defined masses, ranging from obvious to very subtle. An expert radiologist manually segmented all 1210 masses on a dedicated mammographic review station. We used these annotations as the ground truth for our experiments. The centre of each annotation (μ_x, μ_y) was used as seed point for the dynamic programming and region growing segmentation algorithm. The discrete contour model was initialised with a circular region with centre (μ_x, μ_y) and radius 0.6 cm.

3.3.2 Extension Factor for Closed Contours

In the first experiment we estimate the minimum value of the extension factor β needed to guarantee that almost all contours are closed. For this purpose we vary β and apply the dynamic programming algorithm to the extended cost matrix from $-\beta\pi$ to $\beta\pi$. We then extract the path from $-\pi$ to π as final contour. For each extension factor β we

calculate the percentage of closed contours. We consider a contour as closed when the distance $|r_1 - r_2|$ between the start point $(-\pi, r_1)$ and the end point (π, r_2) is less than three pixels. To estimate an appropriate value for β we calculate the percentage of closed contours for each value of β for all images that contain a mass lesion. We then select the minimum value of β for which almost all contours are closed.

3.3.3 Segmentation

In the second experiment we evaluate the segmentation performance of all three segmentation methods using an area overlap criterion. The used dataset for this experiment is the *mass dataset*, that is the set of images that contain at least one mass lesion. The segmentation performance of each method is evaluated with the following overlap criterion:

$$O = (S \cap T)/(S \cup T),$$

where O is the overlap fraction, S the region obtained by one of the segmentation algorithms and T the manually segmented region. An overlap fraction close to one means a good match between the two regions. We use the two-sided Wilcoxon test with confidence level 0.95 to asses the difference in overlap fraction between two segmentation methods.

3.3.4 Mass Detection

The third experiment was done to study the influence of the segmentation method on the mass detection performance. The dataset for this experiment consists of normal images and images with at least one malignant lesion, that is the *normal dataset* and the *malignant mass dataset*. In this experiment we first apply the single view CAD algorithm to each image to find the most suspicious locations inside the breast area, see Section 2.2. The coordinates of the selected locations are used as seed points for the segmentation algorithms. After segmentation several features are determined to classify each segmented region as normal or malignant.

We use cross-validation to randomly partition the dataset into a training set and a test set on a 10:1 ratio under the constraint that the images from the same patient are grouped into the same subset. The training set is used for feature selection and classifier training, the test set for classifier validation.

For feature selection we use a k-nearest neighbour (KNN) algorithm in a leave-one-out basis to select the most useful features from the entire feature space. Selection was done with a sequential forward procedure, which means that new features are added when they increase the performance of the classifier. A KNN classifier then combines the selected features into a malignancy score, representing the likelihood that a region is

malignant. FROC analysis was done to determine the performance of the CAD system for the different segmentation methods.

3.3.5 Benign/Malignant Classification

In the last experiment we investigate whether the segmentation methods influence the performance of the CAD system in classifying lesions as benign or malignant. For this experiment we use the *mass dataset*, which consists of benign and malignant lesions with known ground truth. For each segmented lesion several features are calculated. For classification we use the same procedure as described above for the detection experiment: KNN based feature selection and classification. ROC analysis is done to evaluate the classification performance for each segmentation method. We used the LABROC programme to determine ROC curves, and the CLABROC programme to evaluate the statistical significance between the different methods (Metz *et al.* 1998a; Metz *et al.* 1998b).

3.4 Results

3.4.1 Percentage of Closed Contours

In the first experiment we determined an optimal value for the extension factor β. Figure 3.6 shows the percentage of closed contours for several values of β, ranging from one to three. This figure shows that about 40% of the contours is immediately closed. This percentage increases and reaches 98% for $\beta = 2$. The maximum number of closed contours is reached for $\beta = 3$. At that time there was only one contour not closed and another contour was found closed but instable. Figure 3.7 shows both cases. The top row shows the case where the dynamic programming algorithm was unable to reach closure. Figure 3.7(a) shows the manual segmentation of this mass, Figure 3.7(b) the contour of the proposed algorithm. For this contour we applied the constraint dynamic programming algorithm to force a closed contour, see Section 3.2.1. Figure 3.7(c) shows the final contour obtained with the constraint algorithm. The bottom row of Figure 3.7 shows the case where the final contour was closed but not stable. The contour alternated between two states for different values of β and thus depended on the initial angle of the coordinate transform. Figure 3.7(d) shows the manual segmentation of the mass. It is a benign mass embedded in dense tissue. Figure 3.7(e) and 3.7(f) show the different states of the contour. The contour in Figure 3.7(e) is too large and contains some dense tissue around the mass. The other state, shown in Figure 3.7(f), gives a correct segmentation.

For the experiments described below we applied the algorithm with $\beta = 3$ and thus used an extended cost matrix from -3π to 3π.

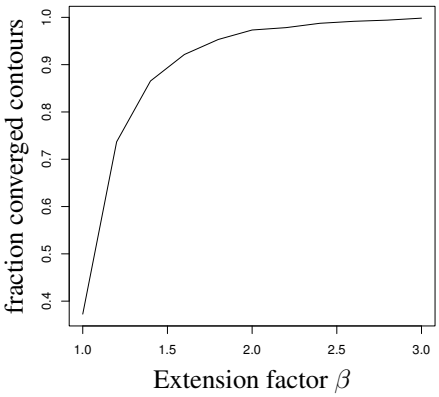


Figure 3.6: *The percentage of closed contours is plotted against the extension factor β. In the original dynamic programming algorithm (with $\beta = 1.0$) 40% of the contours is closed. In the improved algorithm where the path is calculated over a larger area ($\beta = 3.0$), more then 99% of the contours is closed.*

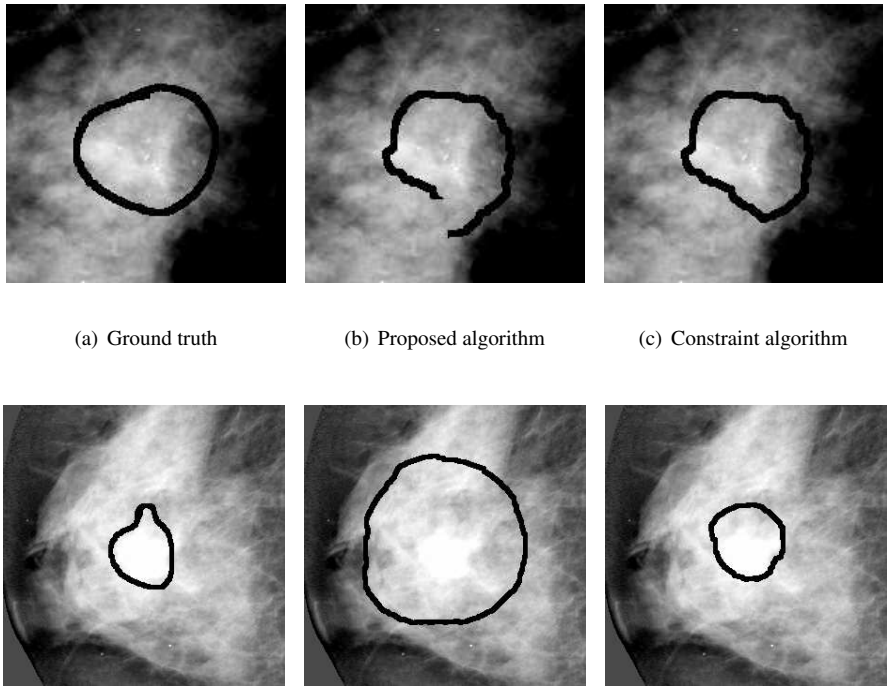
3.4.2 Segmentation Performance

Table 3.1 gives the segmentation performance for each method measured as the overlap fraction with the ground truth. The average overlap fraction for dynamic programming

Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
dynamic programming	0.005	0.608	0.747	0.687	0.825	0.940
discrete contour model	0.006	0.494	0.633	0.599	0.743	0.913
region growing	0.034	0.460	0.641	0.586	0.748	0.914

Table 3.1: *Summary statistics for the performance of the three segmentation methods based on an area overlap criterion measuring the overlap between the automated segmentation and the manual segmentation.*

was 0.69, for the discrete contour model 0.60 and for region growing 0.59. These results indicate that the dynamic programming method is more suited to segment mammographic masses than the other two methods. Figure 3.8 displays the overlap fractions for the different methods. The figure shows that not only the mean overlap fraction is higher, but also that the percentage of masses with poor overlap is smaller for dynamic programming than for the other two segmentation methods.



(a) Ground truth (b) Proposed algorithm (c) Constraint algorithm

(d) Ground truth (e) State 1 of the contour (f) State 2 of the contour

Figure 3.7: *Cases where the dynamic programming algorithm did not find a closed or stable contour. The upper row shows the case where the dynamic programming algorithm was unable to find a closed contour. Figure 3.7(a) shows the manual segmentation, 3.7(b) the proposed algorithm and 3.7(c) the constraint algorithm. The bottom row shows the case where the final contour alternated between two states 3.7(e) and 3.7(f).*

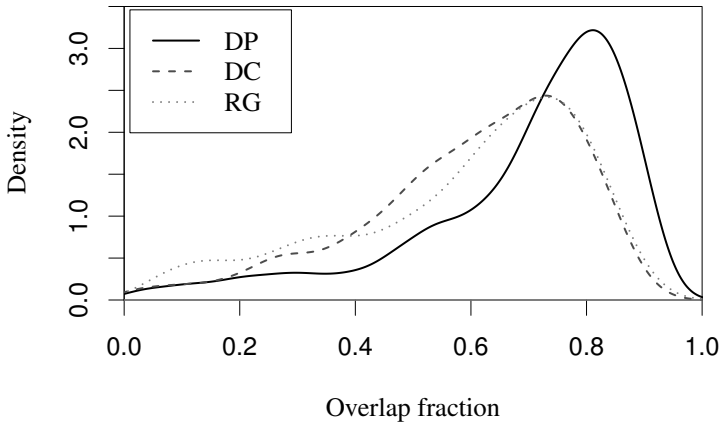


Figure 3.8: *Distribution of the overlap fractions for the different segmentation methods. The average overlap fraction for dynamic programming (DP) is higher then for region growing (RG) and the discrete contour model (DC)*

We used the two-sided Wilcoxon test to determine whether the difference in overlap fraction was statistically significant. Table 3.2 shows the results. The difference in overlap fraction between the proposed method and the other two methods was statistically significant ($P \ll 0.05$). Although the discrete contour model had better results than the region growing method, these results were not statistically significant.

Methods	**P-value**	**Conf. Interval**
dynamic programming - discrete contour model	$\ll 0.05$	$[0.081, 0.11]$
discrete contour model - region growing	0.6507	$[-0.01, 0.01]$
dynamic programming - region growing	$\ll 0.05$	$[0.08, 0.11]$

Table 3.2: *Results of the Wilcoxon's test for the statistical difference in overlap fraction between the existing methods and the proposed method. The second column gives the P-value, and the last column the 95% confidence interval for the difference in the means.*

3.4.3 Mass Detection Performance

The FROC curve in Figure 3.9 shows the case based detection performance for the different segmentation methods. The horizontal axis gives the number of false positive detections per image, the vertical axis the sensitivity. In case based evaluation a lesion is considered detected if it is detected on either view. This figure shows that the detection performance is nearly identical for all three segmentation methods.

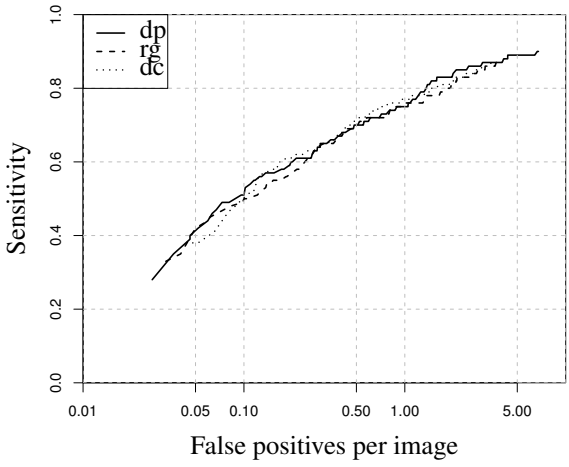


Figure 3.9: *FROC curves for the different segmentation methods: dynamic programming (dp), region growing (rg) and discrete contour model (dc)*

3.4.4 Benign/Malignant Classification Accuracy

In the last experiment we studied whether the used segmentation method influenced the ability of the CAD system to discriminate between benign and malignant lesions. For each segmentation method we constructed an ROC curve with the freely available LABROC programme (Metz *et al.* 1998b). As performance measure we used the area under the ROC curve (A_z value). Figure 3.10 shows case based ROC curves for the different segmentation methods. Table 3.3 summarises the corresponding A_z values.

For image based evaluation the A_z value was 0.74 for dynamic programming, 0.67 for region growing and 0.71 for the discrete contour model. The average A_z value for case based evaluation, where different views of the same lesion are combined by the classifier, was 0.74 for dynamic programming, 0.67 for region growing and 0.72 for

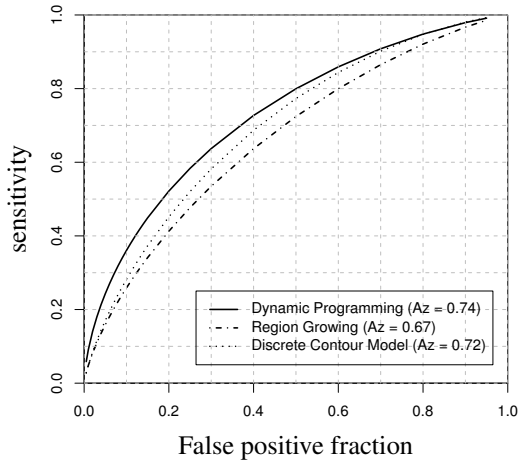


Figure 3.10: *Case based ROC curves for the three different segmentation methods: dynamic programming, region growing and the discrete contour model. The horizontal axis gives the false positive fraction, the vertical axis the true positive fraction (sensitivity).*

Method	**Image Based** A_z	**Case Based** A_z	**P-value**
dynamic programming	0.72	0.74	
discrete contour model	0.71	0.72	0.20
region growing	0.67	0.67	0.0044

Table 3.3: *A_z value that indicates the area under the ROC curve for the different segmentation methods. The last column gives the results of the CLABROC programme that measures the difference in A_z value between the proposed method and existing methods.*

the discrete contour model. We used the CLABROC programme ((Metz *et al.* 1998a) to evaluate the statistical significance of differences in classification performance. We found that the difference in A_z values between the region growing and the dynamic programming method was statistically significant ($P = 0.0044$, two-tailed). The difference between the dynamic programming method and the discrete contour model was not significant ($P = 0.20$), neither was the difference between the discrete contour model and region growing ($P = 0.08$).

3.5 Discussion

In this work we developed a segmentation algorithm based on dynamic programming to accurately extract mammographic mass contours. In addition we developed a method to obtain closed contours. We found that with our proposed method 99.9% of the contours was closed when we used an extended cost matrix with $\beta = 3.0$. As this percentage did not change for larger values of β, we used $\beta = 3.0$ in the other experiments. Another option would be to set $\beta = 2.0$ and increase β for contours that are not closed. A disadvantage of this is that some contours might be closed but instable for $\beta = 2.0$, that is the contour might change for other values of β. We consider these contours as suboptimal. It depends on the application to make a balance between optimality and speed. In our application, as the computational burden of applying the algorithm with $\beta = 3.0$ is minimal, we choose for optimality.

We compared the proposed method with two other methods: region growing and the discrete contour model. We determined the accuracy of each method by comparing the automated segmentation with a manual segmentation using an area overlap criterion. The mean overlap fraction for dynamic programming was 0.69, for region growing 0.59 and for the discrete contour model 0.60. The difference in overlap fraction between the dynamic programming method and the other methods was statistically significant ($P \ll 0.05$).

Besides analysing all segmentations quantitatively we also performed a case review in which we judged the automated segmentations and decided whether the automated segmentation was visually in agreement with the manual segmentation. In this case review we found that in general segmentations with an overlap fraction of at least 0.70 are appropriate. For dynamic programming 52% had an overlap fraction larger than 0.70, for region growing 39% and for the discrete contour model 38%. These results demonstrate that the segmentations obtained with dynamic programming more closely match visually acceptable segmentations than the other two automated segmentation methods. All three methods rarely achieved more than 90% overlap. One reason for this is that the accuracy of the manual segmentation is limited. Often the manual segmentations are somewhat large to make sure the whole tumour is inside the annotation. Another reason is that

when the radiologists' annotation and the automatically segmented area are not identical the chosen overlap criterion quickly decreases. In this case review we also found that in cases where the overlap fraction is less than 0.50, the segmentation method often failed to find the right contour. Figures 3.1 and 3.2 show examples where the two existing methods fail to find the right contour and where the resulting overlap fraction is low. In Figure 3.1 the contrast of the mass is low and the resulting segmentation is too large. In Figure 3.2 the contour is attracted towards the strong edge of the pectoral muscle. In both cases the dynamic programming method correctly segments the mass. The advantage of the dynamic programming method is that both global and local cost are combined with different weights to determine the optimal contour.

In this study we only had one radiologist to do the manual segmentations. When another radiologist would have done the segmentations, both the ground truth and the centre of mass would have changed. We however believe that this would not have influenced the results for the following reasons. First, from the literature we know that region growing and the discrete dynamic contour model are not very sensitive to small changes in the seed point (Kupinski & Giger 1998; Lobregt & Viergever 1995). To determine whether the dynamic programming method depends on the initial seed point, we applied the method with two different types of seed points: the centre of the manually segmented mass (μ_x, μ_y) and the most suspicious site in a neighbourhood of (μ_x, μ_y). The results for both seed points were comparable. This indicates that small changes in the seed point do not influence the segmentation performance. Second, we took several measures to minimise uncertainty of the ground truth due to intra-observer variation. The radiologist used a dedicated mammographic review station to outline the contour of each lesion. We gave the radiologists clear instructions how to outline certain mass types. For example, for architectural distortions and spiculated masses, only the central tumour had to be annotated, and not the individual spicules. The main reason for this is that the outlining of spicules is very subjective. By taking these measures we expect that both intra- and inter-observer variability will remain small. Finally we believe that segmentation differences caused by inter-observer variability will be rather small compared to differences between the segmentation methods. Differences between radiologists are often of a subtle nature, and mainly concern slight variations in outlining tumours with vague boundaries and architectural distortions. Differences between the segmentation methods may be quite large, for instance when one of the methods fails to find the right contour. From Figure 3.8 we see that there is indeed a considerable number of cases where the overlap fraction is less then 0.50, which corresponds with an incorrect segmentation. Most of these cases would also have had a low overlap fraction when another radiologist had done the segmentations.

The detection experiment showed that the improvement in segmentation performance did not result in a better detection of malignant masses. This may be understandable as we did not yet focus on the design of special contour features for mass detection. Instead

we used contour features that we initially developed for the characterisation of mass lesions such as the sharpness of the boundary. For mass detection the CAD programme should discriminate between malignant masses and patches of normal tissue. These often have similar border characteristics. Future research may aim at developing contour features that capture relevant aspects of the mass boundaries to discriminate between malignant masses and false positive detections. With better contour features we expect that the mass detection performance might benefit from a more accurate segmentation.

In the classification experiment we found that the proposed dynamic programming method improved the classification accuracy to discriminate between benign and malignant masses. The A_z value for case based performance was 0.74 for the proposed segmentation method, 0.67 for the region growing method and 0.72 for the discrete contour model. The difference in A_z value between the proposed method and the region growing method was statistically significant. Differences between other methods were not statistically significant. The classification results were in agreement with the ranking of the segmentation results. The best segmentation method—the dynamic programming algorithm—also performed best in classification. Region growing—the method with the lowest segmentation performance—also showed the lowest classification performance. These results contradict the study from Sahiner *et al.* (2001). They compared a mass segmentation method based on an active contour model with the manual segmentations from two expert radiologists. Even when the radiologist and the computer had high disagreement they observed no difference in classification accuracy. Our experiments however indicate that a more accurate segmentation may result in an improved characterisation of mass lesion.

Chapter 4

Temporal Changes in Masses [1]

In this chapter we study temporal changes in mammographic masses. For this purpose we use a set of malignant cases from the Dutch Breast Cancer Screening Programme. Each case consists of the current mammogram and the mammograms from the previous two screening rounds. We first calculate several features for each mass lesion on the current view. When the mass is visible on the prior view, we also determine features for this mass lesion. When the mass is not visible in retrospect, we determine the location on the prior view where the mass most likely developed and calculate features at this location. We then determine the change in feature values extracted from the prior and the current region. The goal of this study is twofold. First to get insight into the temporal behaviour of masses and second to study variations in this temporal behaviour. We can use this information to estimate the benefit of using previous mammograms for radiologists and computer aided diagnosis and detection (CAD) systems. When the analysis of temporal changes provides useful information we can incorporate this into a CAD programme to improve the detection or characterisation of masses. Chapter 6 and 7 concern the development of such a CAD system.

4.1 Dataset

For the experiment we use a total of 250 biopsy proven breast cancer cases from the Dutch Breast Cancer Screening Programme. These cases are either screen detected or interval carcinoma. Table 4.1 summarises the histological class of all cases. Most cancers are invasive carcinoma (90%). For our study we exclude cases with in situ cancers, cases that consist solely of micro-calcifications (28 cases), and seven additional cases for different reasons, for instance an incomplete mammogram. We use the remaining 215

[1] The content of this chapter has been published previously in Timp *et al.* (2002a)

cases to study temporal changes.

Histology	**No. (Percentage)**
ductal carcinoma in situ	17 (7%))
invasive ductal carcinoma	179 (72%)
lobular carcinoma in situ	0
invasive lobular carcinoma	27 (11%)
papillary carcinoma in situ	3 (1%)
papillary carcinoma infiltrative	2 (1%)
medullary carcinoma	2 (1%)
mucinous/colloid carcinoma	4 (2%)
tubular carcinoma	6 (2%)
invasive carcinoma NOS	8 (3%)
unknown	2 (1%)

Table 4.1: *Histology for the 215 breast cancer cases that we used in our experiment.*

For each case we collected the mammograms at three different points in time: the diagnostic mammogram and the mammograms from the previous two screening rounds. Figure 1.3 shows an example of a case consisting of three consecutive mammographic exams. The diagnostic exam is either a clinical mammogram for interval cancers or a screening mammogram for screen detected cancers. In every screening round medio lateral oblique (MLO) views are made. The following guidelines exist when to make additional cranio caudal (CC) views. When attending the screening programme for the first time CC views are always taken. In subsequent screening rounds the radiographer decides whether additional views are useful. The images from two consecutive screening rounds form a temporal image pair. All images were digitised at a pixel resolution of 50 μm.

An expert radiologist identified the mass lesion on each view. The radiologist also determined whether the tumour was visible on previous mammograms. When the mass was visible on prior views, the radiologist rated its visibility as *clearly visible* or *minimal sign*. The indication *minimal sign* means that only very subtle tumour characteristics were present that probably would have been overlooked when the diagnostic mammogram was not available. When the tumour was not visible we estimated the location on the prior view where the mass most likely developed. This makes it possible to determine temporal changes for lesions that are visible and for lesions that are not visible on the prior view.

Table 4.2 summarises the percentage of mass lesions that was visible on the prior

view. From this table we see that in 54% of the cases the mass was visible in retrospect on the prior I mammogram, that is the most recent prior mammogram (see Figure 1.3), 31% was rated as clearly visible and 24% as minimal sign. In 25% of the cases the tumour was visible on both prior I and prior II mammograms. All mass lesions that were visible on the prior II mammograms were rated as minimal sign.

	Total	**Clearly Visible**	**Minimal Sign**
visible on prior I	117 (54%)	66 (31%)	51 (24%)
visible on prior II	53 (25%)	-	53 (25%)

Table 4.2: *Percentage of cases that were visible on previous mammograms. In 54% of the cases the lesion was already visible on the prior I screening mammogram, in 25% of the cases the lesion was visible on both prior I and prior II screening mammograms.*

For each image we determined the corresponding mammographic exam (diagnostic, prior I or prior II, see Figure 1.3) and whether the image contained a visible mass lesion. Based on these characteristics we divided each image into one of the subsets of Table 4.3. The diagnostic set contains 360 images. The priors of these images are classified as *prior I masses* (138 images) or *prior I normals* (94 images), depending on whether the mass lesion was visible or not. The remainder of the diagnostic images (128 images) did not have prior views. From this table we can deduce that about $59\%(\frac{138}{94+138})$ of the diagnostic masses was already visible on the previous screening mammogram. This percentage slightly differs from Table 4.2 because Table 4.2 calculates the visibility for each case and Table 4.3 for each single image.

4.2 Temporal Change Analysis.

We study temporal changes for the whole dataset and for the above mentioned subsets of the whole dataset. For this purpose we first segment corresponding regions on prior and current views. Then we determine features for each segmented region. The difference in feature values extracted from the prior and current region gives us information about temporal changes. Below we describe both steps.

4.2.1 Segmentation

For all images with a visible mass lesion we calculate the centre of the annotated lesion (μ_x, μ_y). Our dynamic programming based segmentation algorithm uses these coordinates as starting point to determine a contour for the *current region*. Chapter 3 describes

Image Subset	Description Image Subset	No.
experiment set	all images used for the experiment	784
mass dataset	images with mass lesion	576
-diagnostic masses	diagnostic images with a visible mass lesion	360
-prior I visible masses	prior I images with a visible mass lesion	138
-prior II visible masses	prior II images with a visible mass lesion	78
normal dataset	images without a visible mass lesion	208
-prior I normals	prior I images without a mass lesion	94
-prior II normals	prior II images without a mass lesion	114

Table 4.3: *Description of different image subsets. The first column gives the image name, the second column the description of the subset. The last column gives the number of images in each subset.*

this segmentation algorithm in detail. Then we determine a contour for the corresponding *prior region*. When the lesion is visible on the prior view we use the centre of mass of the radiologists annotation of the prior lesion as starting point for the segmentation algorithm. Otherwise, when the lesion is not visible in retrospect, we estimate the location on the prior view where the mass most likely developed. The dynamic programming algorithm uses this location as starting point to determine a contour for the *prior region*. The segmented region on the prior view and the segmented region on the current view form a temporal region pair.

4.2.2 Feature Calculation

After segmentation we apply the single view CAD programme (see Chapter 2) to calculate the following six features for all prior and current regions:

- $\overline{f1}$ & $\overline{f2}$: indicate the presence of a stellate pattern.
- $\overline{g1}$ & $\overline{g2}$: indicate the presence of a focal mass.
- $C1$: contrast difference between the segmented region and its surround.
- size: area of the segmented region (in cm^2). We calculate this feature only for visible masses.

For each temporal region pair we then determine the relative change in feature values between two consecutive mammographic exams:

$$f' = (f_c - f_p)/f_p,$$

where f' is the difference feature, f_c the feature value of the current region and f_p the feature value of the prior region. Difference features might be useful to measure temporal changes between two consecutive mammographic exams. We investigate whether these temporal changes depend on 1) the exam of the current mammogram—that is diagnostic or prior I—and 2) the visibility of the lesion on the prior view. For this purpose we construct different sets of temporal region pairs. The first set—temporal set I—consists of temporal region pairs with a visible lesion on both prior and current views. This set is divided into temporal set Ia and set Ib. Set Ia contains temporal region pairs in which the current mammogram is a diagnostic mammogram, in set Ib the current mammogram is a prior I mammogram. Temporal set II contains temporal region pairs with a visible mass on the current view that is not visible on the prior view. In set IIa the current mammogram is a diagnostic mammogram, in set IIb a prior I mammogram.

4.3 Results

4.3.1 Average Feature Values

Table 4.4 summarises the average feature values for the different image subsets. From

Subset	**No.**	$\overline{f1}$	$\overline{f2}$	$\overline{g1}$	$\overline{g2}$	$C1$	**size**
total	784	1.161	1.046	1.282	1.241	0.950	0.653
mass dataset	576	1.182	1.058	1.331	1.284	1.107	0.574
diagnostic masses	360	1.189	1.064	1.348	1.304	1.316	0.788
prior I visible masses	138	1.182	1.051	1.352	1.290	0.996	0.367
prior II visible masses	78	1.169	1.051	1.274	1.238	0.801	0.352
normal dataset	208	1.091	1.010	1.127	1.104	0.329	
prior I normals	94	1.093	1.008	1.146	1.133	0.419	
prior II normals	114	1.089	1.011	1.107	1.075	0.238	

Table 4.4: *Mean feature values for the different subsets.*

this table we see that features extracted from different mammographic exams often have similar values. Features extracted from normal regions differ considerable from features values extracted from mass regions.

Set	Description	$\Delta\overline{f1}$	$\Delta\overline{f2}$	$\Delta\overline{g1}$	$\Delta\overline{g2}$	$\Delta C1$	**Δsize**
I	mass that is visible on prior	2.8	1.4	2.6	2.9	19.3	140.7
Ia	diag mass & prior I visible	2.8	1.3	1.7	2.6	13.4	151.3
Ib	prior I mass & prior II visible	2.9	1.8	4.6	3.9	32.8	116.3
II	mass with normal prior	5.6	3.3	12.5	12.1	80.4	
IIa	diag mass & prior I normal	7.4	4.1	17.7	15.2	108.2	
IIb	prior I mass & prior II normal	8.7	6.1	17.5	17.2	94.9	

Table 4.5: *We construct different temporal image sets based on whether the mass lesion is visible in retrospect and on the screening round from which the current mammogram is taken. For each set we calculate the difference in feature values between segmented regions on prior and current views.*

4.3.2 Difference Features

Table 4.5 summarises the average value of difference features for each set of temporal region pairs. This table is useful to study temporal changes between consecutive mammographic exams and to evaluate whether these changes depend on the kind of exam of the current mammogram (diagnostic or prior I) or on the visibility of the lesion on the prior view. We see that temporal changes are more prominent when the mass lesion is not visible on the prior view. Otherwise, when the lesion is already visible on the prior view, changes are rather small. Furthermore, concerning the kind of exam of the current mammogram, we conclude that almost all features increase more when the current mammogram is a prior I mammogram than when the current mammogram is a diagnostic mammogram. The size of a lesion however increased most between the prior I and the diagnostic exam.

4.4 Discussion

In this chapter we studied the behaviour of masses during time. We found that on average features change between two consecutive mammographic exams. Changes were largest when the lesion was not yet visible on the prior view. For these lesions we obtained an artificial region on the prior view at the location where the mass most likely developed. We then calculated several single view features for this artificial region. This region will not display tumour characteristics because the tumour is not yet visible resulting in low values of the respective features $\overline{f1}$, $\overline{f2}$, $\overline{g1}$, $\overline{g2}$ and contrast. Consequently the difference between the feature value of a lesion on the current view and an artificial region on the prior view will be large.

For lesions that were visible on all three consecutive screening mammograms, we found the largest change in feature values from the prior II screening round to the prior I screening round. The change from the prior I screening round to the diagnostic screening round was rather small.

We studied changes in the size of a lesion in more detail. Figure 4.1 shows the distribution of the feature *difference in size*. From this figure we see that most masses increase in size during time. A considerable part (about 25%) of all masses however decreased in size or remained unchanged. Further inspection of these masses showed that the largest part (27%) concerned architectural distortions that change into more focal mass lesions with smaller size and more contrast. Other reasons for a decrease in size were inaccurate segmentations (25%), masses with a similar appearance on two consecutive mammographic exams (20%), masses that really decreased in (projected) size (15%), masses located on the border of the mammogram (8%), and other (8%). Histograms of other difference features also show large variations.

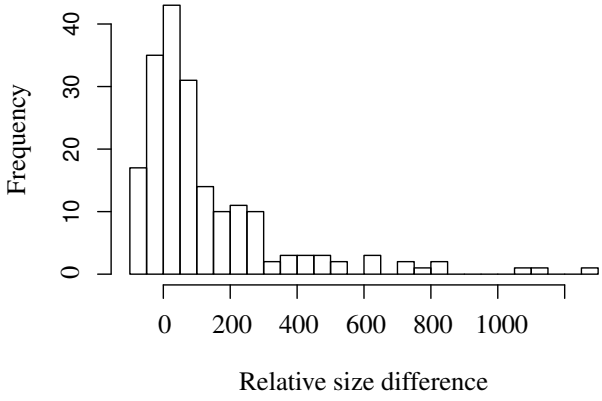


Figure 4.1: *Histogram of the feature difference in size. The x-axis indicates the relative difference in size between the current and the prior screening round. The y-axis represents the number of masses.*

From this study we conclude that temporal features might improve the performance of a CAD system. From Figure 4.1 we learn that difference features may show large variations, making it difficult to draw strong conclusions from temporal change information. Radiologists and CAD systems can use this knowledge when discriminating between normal tissue, malignant lesions, and benign lesions.

Chapter 5

Regional Registration to find Corresponding Masses in Temporal Images [1]

In the previous chapter we found that malignant masses on average change between two consecutive mammographic exams. Using information about temporal changes may therefore be useful to improve the detection and characterisation of mass lesions. Before we can compare prior and current regions we should link each current region to a corresponding region on the prior view. In this chapter we develop a regional registration technique to accomplish this. Starting from a current image containing a mass lesion, this registration technique aims at locating the same mass lesion on the prior image.

5.1 Introduction

Studies report a positive effect on either recall rate or an improvement in mass detection performance when using multiple views in mammography screening compared to single-view mammography, cf. (Wald *et al.* 1995; Sickles *et al.* 1986; Thurfjell *et al.* 2000; Callaway *et al.* 1997). Given the positive effect of multi view systems on radiologists' performance we expect that fusion of information from different views might improve CAD systems as well. A first step towards a multi view approach is the development of programmes to link corresponding lesions.

Few studies have been done to find corresponding regions in different mammographic views. These studies aim at finding similar structures in either different projections of the

[1] The content of this chapter has been published previously in Timp *et al.* (2005).

same breast (Good *et al.* 1999; Paquerault *et al.* 2002) or mammograms obtained at different points in time (Sanjay-Gopal *et al.* 1999; Hadjiiski *et al.* 2001a; Filev *et al.* 2005; Timp & Karssemeijer 2006). The first three studies on temporal registration (Sanjay-Gopal *et al.* 1999; Hadjiiski *et al.* 2001a; Filev *et al.* 2005) first localise the mass on the current mammogram in a polar coordinate system with the nipple as the origin. Based on these coordinates they estimate the location of the mass on the prior mammogram. This predicted location of the mass centroid on the prior mammogram determines a fan-shaped search region. A similarity measure then determines the best matching location inside this fan-shaped search region. Hadjiiski *et al.* (2001a) investigated the usefulness of correlation and mutual information as registration measures. In a recent study Filev *et al.* (2005) compared twelve different similarity measures for the task of template matching. That study shows that the best performing similarity measures for matching corresponding regions in temporal mammogram pairs are Pearson's correlation, the cosine coefficient, and Goodman and Kruskal's Gamma coefficient. In a previous study we developed a regional registration method in which the search for correspondence is done in a feature space (Timp & Karssemeijer 2006). We constructed this feature space by estimating at each location inside a circular search area the likelihood that a mass is present, called the *mass likelihood*. Then we selected the location with the highest *mass likelihood* inside this search area as match for the lesion on the current view.

Both registration methods have some disadvantages. A problem with methods based on template matching is that these only work when both regions are more or less similar in appearance. This might be true for some—especially benign—lesions that stay more or less constant in time, but is obviously not true for malignant lesions that change considerably in time. Registration methods that work in a feature space and use the *mass likelihood* or a comparable registration measure will work well in cases with relatively few potential lesion candidates. These methods may fail when the prior region does not display enough mass characteristics—resulting in a low *mass likelihood*—or when the search area contains more than one region with a high *mass likelihood*. In this chapter we therefore develop a new registration technique that combines the above mentioned methods.

Method Our combined regional registration method comprises three steps. In the first we align both images. Then, in the second step, we define for each mass lesion on the current view a search area on the prior view in which the mass lesion is most likely located. In the third step we combine three registration measures to determine the best location inside the search area. Finally we select this location as estimate for the centre of the prior mass lesion.

More specifically, in the third step we apply the following three registration measures. The first measure represents the likelihood that a mass is present, i.e. the *mass likelihood*. As second measure we use Pearson's correlation coefficient to measure the similarity be-

tween the mass on the current view and a candidate region for the corresponding mass on the prior view. We evaluate the effect of different template shapes on the performance of this correlation measure. The last measure is a distance criterion that gives preference to locations near an initial estimate. Pertaining to the running time we provide a fast variant of the combined registration method in which the measures are applied sequentially.

We compare the performance of the combined method with techniques that use only one registration measure. For this purpose we use a dataset consisting of 389 temporal mammogram pairs that contain a mass lesion that is visible on the prior and the current view. Finally we investigate possible shortcomings of each method by comparing the registration performance for different sets of lesions including benign and malignant masses, and masses that are subtle or obvious on the prior view.

Structure The chapter has the following structure. In Section 5.2 we explain the registration methods in more detail. Section 5.3 describes the experiments to evaluate the different registration methods. In Section 5.4 we present the results with a discussion in the last section.

5.2 Registration Procedure

In this section we present the general procedure we follow to register temporal mass pairs. First, in Section 5.2.1, we describe pre-processing and global registration. Section 5.2.2 describes the definition of a search area. Then, in Section 5.2.3, we explain each of the applied registration measures.

5.2.1 Pre-processing and global registration

Before we can globally register prior and current views we have to pre-process both images. To this end, we first segment each image into breast region, background tissue and pectoral muscle, using a breast boundary and pectoral muscle segmentation algorithm developed previously in our group. We subsequently apply an algorithm that removes additional attenuation from the pectoral muscle. This pectoral equalisation method makes the border region more homogeneous, which is advantageous when dealing with masses that develop on the pectoral boundary. Finally we apply a peripheral enhancement algorithm to the breast area to correct for differences in tissue thickness. Section 2.1 describes these algorithms.

Following pre-processing we use a simple procedure based on a centre of mass alignment to globally register both images. For this alignment we first determine the mathematical centre of mass of the prior and the current image. We can determine the centre of mass using the whole breast area including the pectoral muscle or using the breast area

with the pectoral muscle excluded. In our experiments we exclude the pectoral muscle as this improves the registration accuracy (see Table 5.2 and (Van Engeland *et al.* 2003)). Then we horizontally and vertically shift the prior image such that its centre of mass coincides with the centre of mass of the current image. Figure 5.1 illustrates this alignment after a vertical shift of ty and a horizontal shift of tx. After alignment of both images we use the centre coordinates of the lesion on the current image (μ_x, μ_y) as initial estimate of the location of the lesion on the prior image.

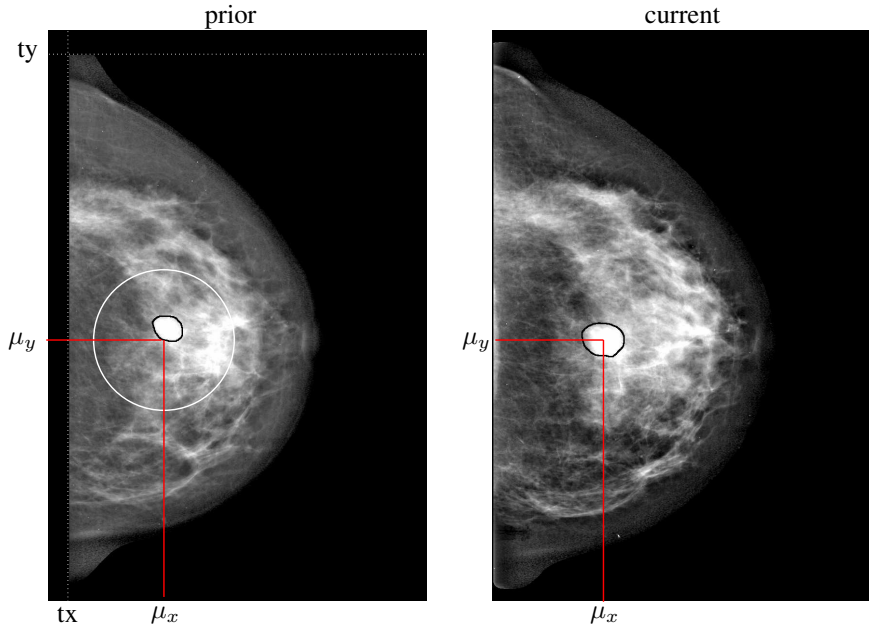


Figure 5.1: *Global alignment and definition of the search area. First both images are aligned by shifting the prior image with tx and ty. The centre coordinates of the lesion on the current image (μ_x, μ_y) then form the initial estimate for the lesion on the prior image. This initial estimate is the centre of a circular search area (white circle) with radius r. We calculate the registration measures at each location inside this search area.*

5.2.2 Definition of the search area

After the images have been aligned we define for each mass lesion on the current image a search area on the prior image. In the literature two different shapes of a search area have been proposed. Timp & Karssemeijer (2006) used a circular search area and Sanjay-Gopal *et al.* (1999) a fan-shaped search area. As prerequisite we consider it important that

the definition of a search area can be done completely automatic. A fan-shaped search area, as proposed in (Sanjay-Gopal *et al.* 1999), requires the location of the nipple. This is a disadvantage as it is sometimes difficult to identify the nipple, in particular if modern high contrast film screen combinations are used. Furthermore, we assume that the shape of the search area will have little influence on the final registration performance when the search area is large enough. Therefore we decide to use a circular search. As centre for this search area we use the initial estimate of the location of the mass lesion on the prior view (μ_x, μ_y).

To ensure that the search area includes most masses we use a large radius of 30 mm. The size of the search area is based on a comparative study Van Engeland *et al.* (2003) performed. They compared several registration methods and found that the maximum error—i.e. the maximum distance between the estimated mass location and the real mass location—was 30 mm for a centre of mass alignment. Figure 5.1 shows the final search area on the prior view with centre (μ_x, μ_y) and a radius of 30 mm. After defining the search area we calculate at each location inside this area the regional registration measures.

5.2.3 Registration Methods

In this section we first describe the individual registration measures *mass likelihood* and *correlation*. Then we explain our proposed registration methods that combine different registration measures.

Registration based on Mass Likelihood

These methods determine at each location inside the circular search area on the prior view the likelihood that a mass is present, that is a *mass likelihood* measure. A high mass likelihood on the prior view may indicate the presence of a lesion. Registration methods based on mass likelihood assume that the location with the highest mass likelihood corresponds with the lesion on the current view. As mass likelihood measure we use the outcome of our pixel level mass detection algorithm. Section 2.2 describes this algorithm in detail. Shortly the algorithm works as follows. At each location inside the breast area two features for the detection of stellate lesions and two features for the detection of focal mass lesions are calculated. These features are used as input for a 3-layer feed-forward neural network trained on known abnormalities. Next we construct the likelihood image by assigning each pixel inside the breast area the corresponding classifier output. Then we slightly smooth this image. The middle row images in Figure 6.2 and 6.3 show examples of likelihood images. We define the *mass likelihood* as the smoothed classifier output at each location in the breast area. For each current mass lesion we select the location inside the search area with the highest *mass likelihood* as estimate for the location

of the mass lesion on the prior view.

Registration based on Grey Scale Correlation

Registration methods based on correlation select a region on the prior view that is similar to the lesion on the current view and assume that this region represents the same mass lesion. Below we explain this method and describe different templates that we tested for the correlation method.

Method Registration methods based on grey scale correlation calculate the pixel correlation between a template image of the current mass—the current mass template—and a candidate region on the prior image. We first select one of the templates described below and put this template over the current mass lesion to obtain the current mass template. Then we obtain candidate regions for the prior mass by putting the template at each location inside the search area on the prior image. Finally we calculate Pearson's correlation measure between the current mass template and the candidate region centred at (i, j) on the prior image:

$$C(i,j) = \frac{\sum_{(m,n)}(y_c(m,n) - \overline{y}_c)(y_p(m',n') - \overline{y}_p)}{\sqrt{\sum_{(m,n)}(y_c(m,n) - \overline{y}_c)^2 \sum_{(m,n)}(y_p(m',n') - \overline{y}_p)^2}} \tag{5.1}$$

The grey level at location (m, n) in the current mass template is given by $y_c(m, n)$ and the grey level of the candidate region with centre (i, j) at the same relative location by $y_p(m', n')$. The summation is performed over all locations (m, n) inside the current mass template. The average grey level in the mass template and the candidate region is given by $\overline{y}_c$ and $\overline{y}_p$ respectively. We select the location with the highest correlation as estimate for the location of the mass on the prior. The next paragraph describes different mass templates.

Mass Templates We designed different templates for the registration method based on correlation: an inner mass template, an outer mass template, and three extended templates. These templates cover different parts of the underlying mass lesion and its surrounding tissue. Figure 5.2 illustrates the templates for a benign mass. Before constructing a template we first use the dynamic programming based segmentation algorithm from Chapter 3 to determine the contour of the current mass lesion.

The inner mass template, as illustrated in Figure 5.2(b), consists of all pixels inside the contour and exactly represents the underlying mass lesion. A candidate region on the prior image highly correlates with this mass template when the mass lesion is similar in appearance on prior and current views. On the other hand, when the mass changes considerably between two consecutive mammographic exams, for example in size or contrast, the correlation between both regions will be low.

Figure 5.2(c) illustrates the outer mass template. This template consists of all pixels outside the mass lesion that have a distance of less than 6 mm from the border. Consequently the correlation only depends on the similarity between the outer border region of the current mass and a similar region on the prior view. This can be an advantage for masses that change significantly in appearance. For these masses the outer border region will stay more or less similar in appearance between both views. On the other hand, this template can cause problems when the grey level characteristics of the outer border region are not unique. An example is a tumour completely embedded in fatty tissue. The outer mass template will represent grey level characteristics of fatty tissue and thus show little variation. Consequently, the correlation between this template and a candidate region on the prior image will be high when the candidate region is homogeneous as well. In uniform breasts this may result in many candidate regions all correlating equally well with the current mass template.

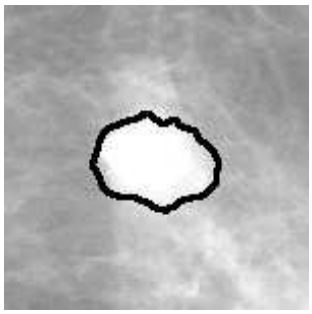
Figure 5.2(d), 5.2(e) and 5.2(f) show the extended templates. These templates consist of a part of the inside region—the inner part—and an outer border region. The first extended template, see Figure 5.2(d), is a simple extended template that consist of the whole inside region and an outer border region. The second one—the growing mass template—only contains the central part of the inside region and an outer border region. We designed this template for masses that grow between two mammographic exams. We assume that for these masses the most inner part and the outside border region are more or less similar on the prior and the current view. As inner part we use the most central region with a size of $\frac{1}{2}\,A$ where A is the area of the whole inside region. We base the size of the inner part on the observation that most masses in our database at most double in (projected) size between two consecutive screening rounds. The last extended template is the circular template. For this template we first determine the effective radius $R = \sqrt{(}A/\pi)$ of the inside region. The circular template then simply is a circular region with radius $R + b$ where b is the size of the outer border region. For all three extended templates the size of the outer border b region is 3 mm.

Combined Registration Methods The last registration method combines the mass likelihood with a correlation measure and a distance criterion. We develop two variants of this combined method. In both methods we first determine the individual registration measures at selected locations inside the search area. After calculating the individual measures at each location we normalise each measure v using the minimum and maximum values found in the dataset:

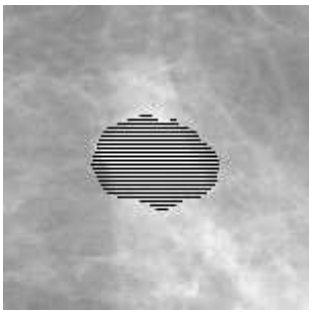
$$\tilde{v} = \frac{v - \min(v)}{\max(v) - \min(v))} \tag{5.2}$$

and then linearly combine them:

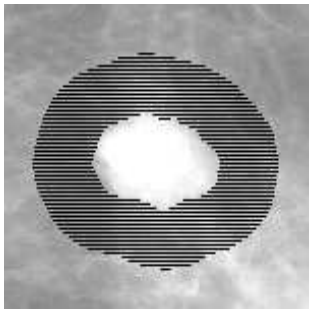
$$R(i,j) = w_c C(i,j) + w_l l(i,j) - w_d d(i,j), \tag{5.3}$$



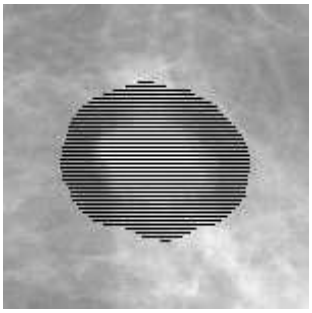
(a) Benign mass lesion



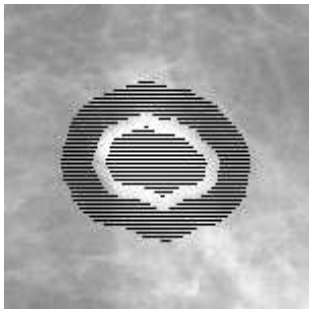
(b) Inner mass template



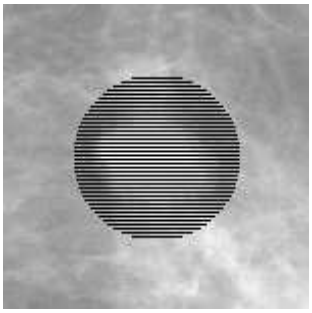
(c) Outer mass template



(d) Simple extended template



(e) Growing mass template



(f) Circular mass template

Figure 5.2: *Figure 5.2(a) shows the segmentation of a benign mass. The other figures show the different mass templates.*

where $R(i,j)$ is the combined registration measure, $C(i,j)$ the grey scale correlation measure for the best performing template shape, $l(i,j)$ the mass likelihood and $d(i,j)$ the distance to the initial estimate. We use the whole dataset to determine the weights w_c, w_l and w_d to achieve maximum registration performance. To find the optimal weights we vary the coefficients w_c and w_l between zero and 100 and keep w_d fixed at 51.

The difference between both combination methods concerns the selection of the locations where the measures are calculated. The first variant simply calculates the three registration measures at each location inside the search area. As we calculate all measures simultaneously we call this method the simultaneous combination method. To reduce the computational effort we developed a second variant in which we calculate the registration measures sequentially. This method first selects all locations inside the search area with a mass likelihood above a certain threshold. If two selected locations are less then one millimetre apart we remove the one with the lowest mass likelihood. This procedure results in an average of 100 selected locations for each search area. We then determine the other two registration measures—correlation and distance to initial estimate—for the selected locations. We call this method the sequential combination method as we calculate the registration measures sequentially.

We compare both variants with respect to registration performance and computational efficiency. An important difference between both methods is that the sequential method only processes locations that show mass characteristics. This can have negative and positive consequences for the registration performance. A negative consequence is that a correct location will be missed when its mass likelihood is below the threshold, independent of the value of the correlation measure. This may result in a decrease of the registration performance for (benign) lesions with few mass characteristics. A positive consequence is that the sequential method will skip locations with accidentally high correlation when they display not enough mass characteristics. This may increase the probability that a correct match occurs. Considering the computational efficiency we notice that this mainly depends on the number of locations where the correlation measure is calculated. For the sequential method this corresponds with on average 100 locations for each search area. The simultaneous method calculates the correlation measure at each location inside the search area. This amounts with almost 71,000 locations for a search area with radius 30 mm and a pixel resolution of 200 μm. The computational effort is thus reduced about a thousandfold by using the sequential method compared to the simultaneous method. For the sequential registration method, the whole procedure, including the calculation of the mass likelihood, takes less than one minute per image. As we determined the mass likelihood already in our single view CAD programme—see Chapter 2—the extra time needed for the registration is based solely on the calculation of the correlation measure. This takes a few seconds per image in the sequential registration method. This means that the method can be implemented into a CAD system without much additional time costs.

5.3 Experiments to Evaluate Regional Registration

In this section we first describe the dataset and the subsets that we used for the experiments. Then we describe the evaluation measure we used to quantify the performance of each registration method.

5.3.1 Dataset

The mammograms used in this study all came from the Dutch Breast Cancer Screening Programme. We constructed the dataset for the experiments by collecting all temporal image pairs with a visible mass lesion on the prior and the current view. Each temporal image pair consisted of the mammograms from two consecutive mammographic exams. We call the most recent image in a temporal pair the current mammogram and the image obtained in the previous screening round the prior or previous mammogram. The images came from two different sets. Table 5.1 summarises information about each set. The first dataset consisted of 155 image pairs with a malignant mass on prior and current views. This dataset contained 281 images from 87 patients. The images were digitised with a Lumisys 85 digitiser at a pixel resolution of 50 μm.

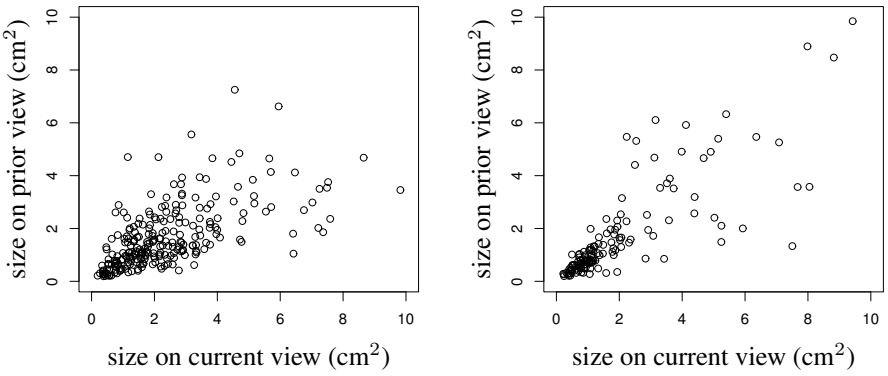
The second dataset consisted of 234 image pairs, 94 with a malignant mass and 140 with a benign mass. This dataset contained 434 images from 155 patients. The images were digitised with a Canon CFS300 laser scanner at a pixel resolution of 50 μm. A radiologist rated all masses in this dataset for their visibility on a scale from 1 to 5. A rating of 1 corresponds to masses that are clearly visible. A rating of 5 corresponds with subtle masses that are difficult to see. Most of these can only be detected in retrospect.

Combination of the two sets resulted in 389 temporal image pairs, 140 benign and 249 malignant. The number of temporal pairs is larger than half of the number of the images since for some women the mammograms of three consecutive mammographic exams were available. For the experiments we used a spatial resolution of 200 μm maintaining the original grey value resolution of 12 bits.

We annotated all mass lesions on prior and current views under supervision of an expert radiologist. For this purpose we used specially designed software on a dedicated mammographic review station. We determined the size of each annotated mass lesion on both the prior and the current view. Figure 5.3 shows the distribution of the mass size for benign and malignant masses. The mean size of benign masses was 2.2 cm^2 on the current mammogram versus 1.8 cm^2 on the prior mammogram. The average growth of the benign masses, defined as the ratio between the current and the prior mass size, was 1.4. The mean size of malignant masses was 2.4 cm^2 on the current mammogram versus 1.7 cm^2 on the prior mammogram. The average growth of malignant masses was 1.66.

	Dataset I	**Dataset II**
no. of image pairs	155	234
no. of images	281	434
no. of patients	87	155
malignant image pairs	155	94
benign image pairs	0	140
MLO views	124	194
CC views	31	40

Table 5.1: *Composition of the datasets used for the experiments.*



(a) Malignant lesions

(b) Benign lesions

Figure 5.3: *Size of masses on prior and current views.*

5.3.2 Subsets

We tested the performance of each registration method on the whole dataset and on several subdivisions of the original dataset. These subdivisions contain different mass types and the performance on these subsets informs us about specific shortcomings of each method. The first subdivision is between benign and malignant masses. We use this subdivision to test our assumption that correlation measures are more suited for benign masses and measures based on mass likelihood for malignant masses. We base this assumption on the fact that correlation measures work best for lesions that stay more or less constant in time, which is often the case for benign masses. Malignant masses on the other hand can change considerably in time, not only in size, but also in contrast and overall appearance. We use the set of malignant masses that have been rated for their visibility to make a subdivision between masses that are clearly visible on the prior view and masses that are very subtle on the prior view. To this end we put all masses with a visibility rating of 5 in the group of *subtle priors* and all other masses in the group of *obvious priors*. We expect that most masses in the group of subtle priors will have a different appearance on the prior and the current view. These masses may thus be less suited for methods based on correlation.

5.3.3 Validation

As evaluation measure for the registration methods we use the fraction of correctly matched lesions. We count a match as correct when the selected location is inside the annotation of the radiologist. Besides evaluating the performance of each registration method we also determine the optimal search radius by varying the radius of the search area between 0 and 30 mm.

5.4 Results

In the first two paragraphs we present the results for the global and regional registration methods. For this purpose we used the complete dataset of 389 temporal image pairs. In the third paragraph we give the performance each registration method for different subsets of the original dataset. In the last paragraph we describe cases where the proposed registration method failed to establish a correct link.

5.4.1 Global Registration Performance

We tested the accuracy of global alignment for two implementations. In the first we used the whole breast—including the pectoral muscle—to determine the centre of mass of the breast area. In the second we excluded the pectoral muscle when determining

the centre of mass of the breast. After alignment we used the centre coordinates of the current mass lesion as estimate for the location of the mass lesion on the prior view. We then determined the fraction of correctly linked masses. We counted a link as correct when the initial estimate fell inside the manual segmentation of the lesion on the prior view. Furthermore we determined for each lesion the distance between the centre of the current mass lesion—i.e. the initial estimate—and the centre of the prior mass lesion. Table 5.2 presents the results. From Table 5.2 we see that the global registration method improves when we exclude the pectoral muscle for determining the centre of mass. In the other experiments we therefore use the implementation in which the pectoral muscle is excluded.

	Fraction Correct	**Mean Distance to Ground Truth (mm)**
with pectoral muscle	0.30	11.9
without pectoral muscle	0.37	9.9

Table 5.2: *Results for the global registration procedure where we determined the centre of mass of the breast area with and without the pectoral muscle. The first column gives the fraction of correctly linked masses. The second column gives the mean distance from the centre of the current mass lesion to the centre of the prior mass lesion.*

5.4.2 Performance Registration Measures

Table 5.3 and Figure 5.4 show the results for the different registration measures. The best performance for the measure based on mass likelihood is 0.71 for a search radius of 12 mm. Considering the correlation measure we find that the inner and outer mass templates have a significantly lower performance than the extended mass templates. The best performing extended mass template is the growing mass template, although the difference with the other extended templates is not statistically significant. We furthermore studied the influence of the outer border region by varying the size of this region in the simple extended template between 0 and 6 mm. From Table 5.4 we see that the fraction of correctly linked masses is 0.60 for the simple extended template without an outer border region, that this fraction increases up until 0.68 for an outer border region of 1.4 mm and then stays more or less constant.

We select the growing mass template for the grey scale correlation measure in the combined registration methods. The difference between the performance of the combined registration methods and the individual registration measures is statistically significant. Figure 5.4 shows that the performance of both combined methods increases up

Registration Measure	Fraction Correct	Radius	Distance
mass likelihood	0.71 ± 0.02	12	3.6
inner mass template	0.60 ± 0.02	16	4.2
outer mass template	0.48 ± 0.03	8	4.6
simple extended mass template	0.69 ± 0.02	20	3.6
growing mass template	0.71 ± 0.02	20	3.5
circular mass template	0.69 ± 0.02	16	3.7
simultaneous combination	0.82 ± 0.02	20	2.6
sequential combination	0.82 ± 0.02	20	2.8

Table 5.3: *Registration results for the different methods. The first column shows the registration measure. The second column gives the fraction correctly linked masses and the standard deviation. The third column shows the radius of the search area where the maximum performance has been obtained and the last column the mean distance to the ground truth.*

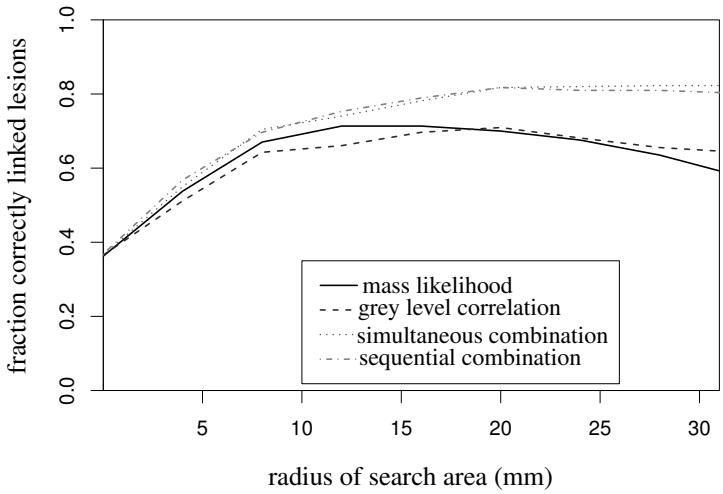


Figure 5.4: *Overview of regional registration measures. For each method the fraction of correctly linked lesions is plotted against the radius of the search area.*

Outer Border Size (mm)	0	0.6	1.0	1.4	2.0	3.0	4.0	6.0
Fraction Correct	0.60	0.63	0.66	0.68	0.68	0.69	0.69	0.69

Table 5.4: *Fraction correctly linked masses where we varied the outer border size of the simple extended template.*

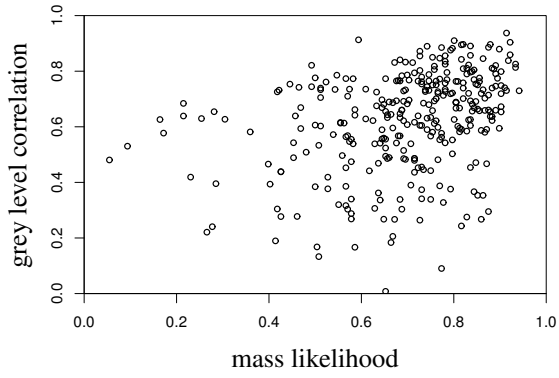
until 0.82 for a search radius of 20 mm and then stays more or less constant. The weights for w_c, w_l, and w_d were 33, 29 and 51 for the simultaneous combination method and 51, 75, 51 for the sequential combination method. From these coefficients we see that the distance measure is more important for the simultaneous method than for the sequential method. For the sequential method the mass likelihood measure has a lower weight than the correlation measure. This was expected as all processed locations in the sequential method already have a relatively high mass likelihood. The choice for a location then mainly depends on the grey scale correlation measure. For both methods we find that small variations in the coefficients have little influence on the results. For example, when the coefficients w_c, w_l and w_d are equal, the performance of both methods is 0.80.

Figure 5.5 shows a scatter plot for the measure based on mass likelihood versus grey scale correlation for correctly linked masses and masses that were linked incorrect. The correlation between both measures is 0.34 for correctly linked masses and 0.22 for incorrect matches. From the figure we see that most correctly linked masses have a high correlation measure and a high mass likelihood. However, there is also a large number of masses with either a low correlation or a low mass likelihood. This explains the increased performance of the combination methods compared to the performance of the individual measures.

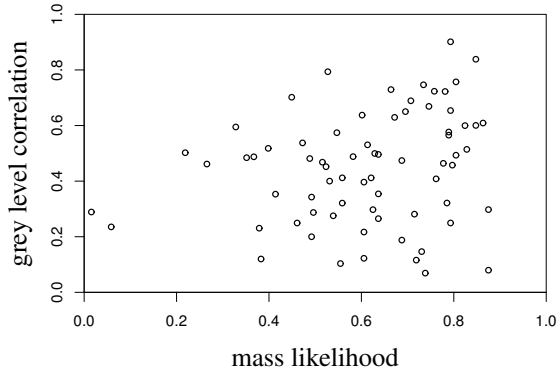
Figure 5.6 shows the histogram of the distance between the selected location and the centre of the ground truth. The mean distance for correctly linked masses is 1.2 mm. For incorrect matches the mean distance is 10.0 mm. This is more or less equal to the mean distance measured after the global registration step. There are a few outliers among the incorrect links. In these cases the global registration failed and the true mass lesion was located outside the search area.

5.4.3 Registration Performance for Subsets

Table 5.5 gives the fraction of correctly linked masses for different subsets. This table shows that the mass likelihood performs best on malignant masses and the grey scale correlation measure on benign masses. The combination methods perform satisfactory on both subsets. Table 5.5 also shows that the individual registration measures perform similarly for masses that are subtle and masses that are obvious on the prior view. Furthermore we find that the sequential combination method performs better than the simul-

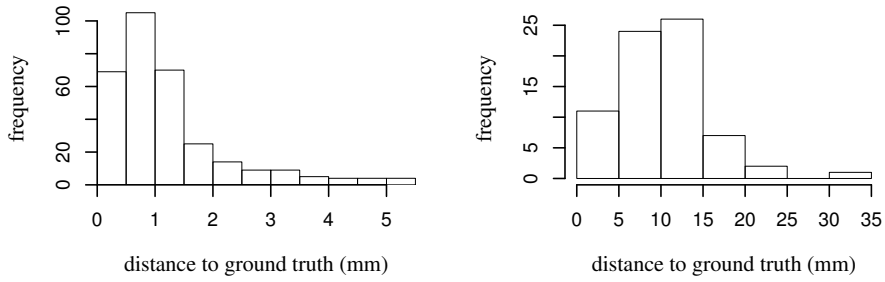


(a) correct link



(b) incorrect link

Figure 5.5: *Scatter plot for correct and incorrect links.*



(a) Correct links (b) Incorrect links

Figure 5.6: *Histogram for the distance from the selected location to the ground truth for correct and incorrect links.*

taneous combination method on the group of masses that are subtle on the prior view.

5.4.4 Link Errors

One of the best performing methods—the simultaneous combination method—links 18% of all lesions incorrect. To obtain insight into the possible causes of these link errors we compare each of the three registration measures—correlation, mass likelihood, and distance—at the selected location with the same measures at the correct location. When one measure performs substantially better at the selected location, we choose failure of this measure as the most important cause of the incorrect match. Table 5.6 shows that a combination of a low correlation and a low mass likelihood is the most common cause of an incorrect match. In most of these cases the mass on the prior mammogram is very subtle, which has consequences for both the correlation measure and the mass likelihood. The second most important cause of link errors is a large distance to the initial estimate. In these cases the global registration method did not work very well. We find a low mass likelihood as cause for the link errors for benign masses that are subtle on the prior view. Finally, we see that masses with a low correlation often change considerably between two consecutive mammographic exams. Figure 5.7 gives some examples where the combined registration method failed. The letter C indicates the correct location, S the selected location. Figure 5.7(a) shows a very subtle mass on the prior view, in Figure 5.7(b) the selected location shows a spiculation pattern resulting in a higher mass likelihood than

Subset	No. of Lesions	Mass Likelihood	Gray scale Correlation	Sim Comb	Seq Comb
original dataset	389	0.71	0.71	0.82	0.82
benign masses	140	0.66	0.74	0.82	0.79
malignant masses	249	0.75	0.69	0.82	0.84
subtle on prior view	37	0.76	0.68	0.78	0.84
obvious on prior view	57	0.74	0.67	0.84	0.84

Note.—Sim = simultaneous, Seq = sequential, Comb = combination.

Table 5.5: *Registration performance for different subsets. The second column gives the number of temporal image pairs in each subset. The other columns give the registration performance for each measure.*

Reason of Incorrect Link	Percentage
combination of low correlation and low mass likelihood	38%
far from initial estimate	25%
low mass likelihood	21%
low correlation	17%

Table 5.6: *Summary of the most important causes of incorrect links.*

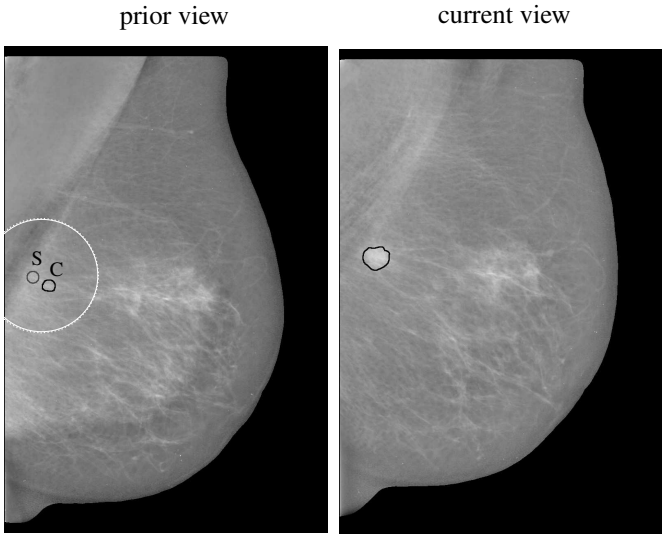
the correct location.

5.5 Discussion

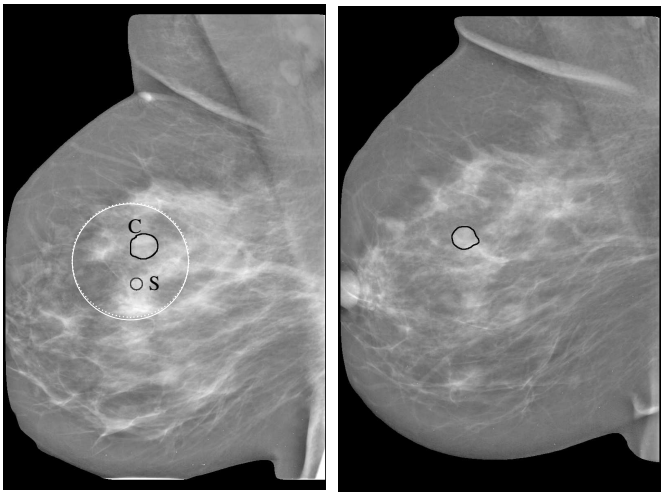
In this chapter we presented an automatic regional registration method that finds corresponding mass lesions in temporal mammogram pairs. This method combines three registration measures: a measure based on correlation, a mass likelihood measure, and a distance criterion. On the complete set of masses this combined method linked 82% of the masses correctly, compared with 71% for both individual measures.

For the measure based on correlation we designed different template shapes and investigated the influence of the template shape on the registration performance. For all shapes we used Pearson's correlation coefficient as similarity measure. In a recent study Filev *et al.* (2005) found that this measure works best among a selection of twelve different similarity measures. Results for the different template shapes show that the best performing template is the growing mass template. We designed this template for masses that either grow or stay constant in time. The registration performance of the other two extended mass templates, the simple extended template and the circular template, was only slightly lower. This shows that the correlation measure is not very sensitive for small changes in template shape. The low performance obtained with the inner and outer mass templates shows that both regions are necessary to obtain good registration results. We furthermore found that the minimal size of the outer border region is 1.4 mm and that the registration performance is similar for larger outer border regions. This observation differs from the study from Sanjay-Gopal *et al.* (1999). They found a decrease in registration performance when the size of the outer border region increased. The main difference between both studies is that Sanjay-Gopal *et al.* (1999) used a bounding box as template whereas our template depends on the contour of the mass. As most masses are more or less circularly shaped a bounding box will always contain surrounding tissue along some—but not all—parts of the contour. This might influence the registration results.

To obtain information about the performance of each method on specific mass types we divided the original dataset into several subsets. The first subdivision of the original dataset was between benign and malignant masses. This subdivision shows that a correlation measure is more suited for benign masses. This may be explained by the fact that benign masses stay more or less constant over time resulting in a good correlation between both views. From the results we conclude that a measure based on mass likelihood is more suited for malignant masses. This measure will select a correct location on the prior view even when the mass has changed considerably, provided that the mass lesion on the prior is at the location with the highest mass likelihood. Furthermore, this measure also takes spiculation into account, which is a frequent sign of malignant masses.



(a) Subtle mass on prior view



(b) Selected location has high mass likelihood

Figure 5.7: *Examples of link errors. The white circle indicates the search area. C is the correct location, S the selected location.*

The second subdivision was between masses that are very subtle and masses that are obvious on the prior view. Concerning the combined registration methods we found that the sequential method was more suited to find subtle priors than the simultaneous method. This is in agreement with the observation that the correlation between a subtle masses on the prior view and its corresponding mass on the current view is often quite low. Consequently, making a pre-selection of locations with a high mass likelihood increases the probability that the correct mass location is selected, on condition that this location has enough mass characteristics to be selected. When all locations are processed—like in the simultaneous combination method—some incorrect locations accidentally may have a high correlation, increasing the probability that an incorrect match occurs.

In summary we found that methods that combine several registration measures perform better than methods that use only one registration measure. The choice between both combination methods depends 1) on the number of regions initially detected by a CAD programme and 2) on whether the CAD programme aims at detecting all kinds of masses or only malignant ones. When the number of initial regions is quite large, what is common for CAD programmes, the sequential combination method is preferred because it is very fast compared to the simultaneous method. The sequential method also performs better on the subset of malignant masses. On the other hand, we might choose the simultaneous method when the CAD programme mainly aims at detecting benign lesions.

A computer aided detection (CAD) system that includes temporal information can use this regional registration method to link selected regions on the current mammogram to corresponding locations on the prior mammogram. Combination of features from linked regions gives information about temporal changes. In the next chapters we will build such a CAD system and evaluate the effect of temporal features on the detection and characterisation performance.

Chapter 6

Interval Change Analysis for the Detection of Masses [1]

In this chapter we include temporal information in our CAD programme to improve the detection of malignant masses. For this purpose we first use a simplified version of the registration method described in Chapter 5. Following the linking process we calculate several features for the current and prior region. We then obtain temporal features by combining the feature values from both regions. Finally we evaluate the effect of temporal features on the detection performance of our CAD programme.

6.1 Introduction

At the moment most CAD (computer aided diagnosis) programmes in mammography use a single view to detect abnormalities. However, when mammograms from multiple examinations are available, and CAD makes use of correlations between exams, a higher accuracy may be achieved in detecting malignancies. In this study we concentrate on using information from previous and current views. In Chapter 1 we already mentioned some advantages of using previous mammograms. Despite these advantages the development of CAD systems that include temporal information has not yet received much attention.

We can divide previous work into two main categories: (1) methods that compare current images with priors to detect subtle changes in the breast and (2) methods that compare suspicious regions in current mammograms with corresponding regions in priors. Vujovic *et al.* (1995) used the first method to detect abnormalities. They first divided

[1] This chapter is based on Timp & Karssemeijer (2004b) and Timp & Karssemeijer (2006)

the current and prior mammogram into several regions using internal control points. Then they used these control points to define circular regions inside each mammogram. Next they compared corresponding regions from prior and current views using texture and contrast measures. They found that the intensity histogram carried useful information in separating normal from abnormal tissue. Kok-Wiles *et al.* (1998) and Timp & Karssemeijer (2004b) used the second method. Kok-Wiles *et al.* (1998) represented the breast as a nested structure of salient regions and used this representation to compare prior and current regions. In a previous study (Timp & Karssemeijer 2004b) we compared the contrast and size of corresponding regions on prior and current mammograms and found that the detection performance improved by adding this information to the CAD system

Both methods depend more or less on the accuracy of the temporal registration. Temporal registration includes global registration and regional registration. Global techniques register current and prior mammograms. In the literature some approaches have been described for global mammogram registration, cf. (Sallam & Bowyer 1994; Vujovic & Brzakovic 1997; Richard & Graffigne 2000). A comparative study for global registration methods in mammography has been done by Van Engeland *et al.* (2003). They compared four methods for mammogram registration: alignment based on nipple position, alignment based on the centre of mass of the breast tissue, warping, and registration based on mutual information. They measured the performance of all methods by comparing the distance between the centre of the manual segmentation of abnormalities on the previous and the current view before and after registration and found that the method based on mutual information worked best. The method based on centre of mass alignment worked reasonably well, in particular if the pectoral muscle was excluded for centre of mass calculation. The method based on nipple alignment only worked if the nipple was visible in profile. The method based on warping performed worst and could cause unrealistic deformations inside the breast area.

In this study we develop a temporal CAD method and investigate the effect of temporal features on the detection performance. Figure 6.1 summarises the different steps. The method starts with the mammograms of two consecutive mammographic exams: the prior and the current mammogram. On both mammograms the breast area and the pectoral muscle are segmented. A global registration method based on centre of mass alignment registers the current and the prior images. Next a pixel level mass detection algorithm assigns each pixel inside the breast area a measure of suspiciousness, the so-called mass likelihood. This measure represents the likelihood that a malignant mass is present at that location. We then select the most suspicious locations on the current image and link these to a corresponding location on the prior view. After linking we calculate features for prior and current regions. The combination of features from both regions results in the so-called temporal features. We use FROC analysis to evaluate the detection performance with and without the use of temporal features.

The remainder of this chapter is organised as follows. In Section 6.2 we first briefly

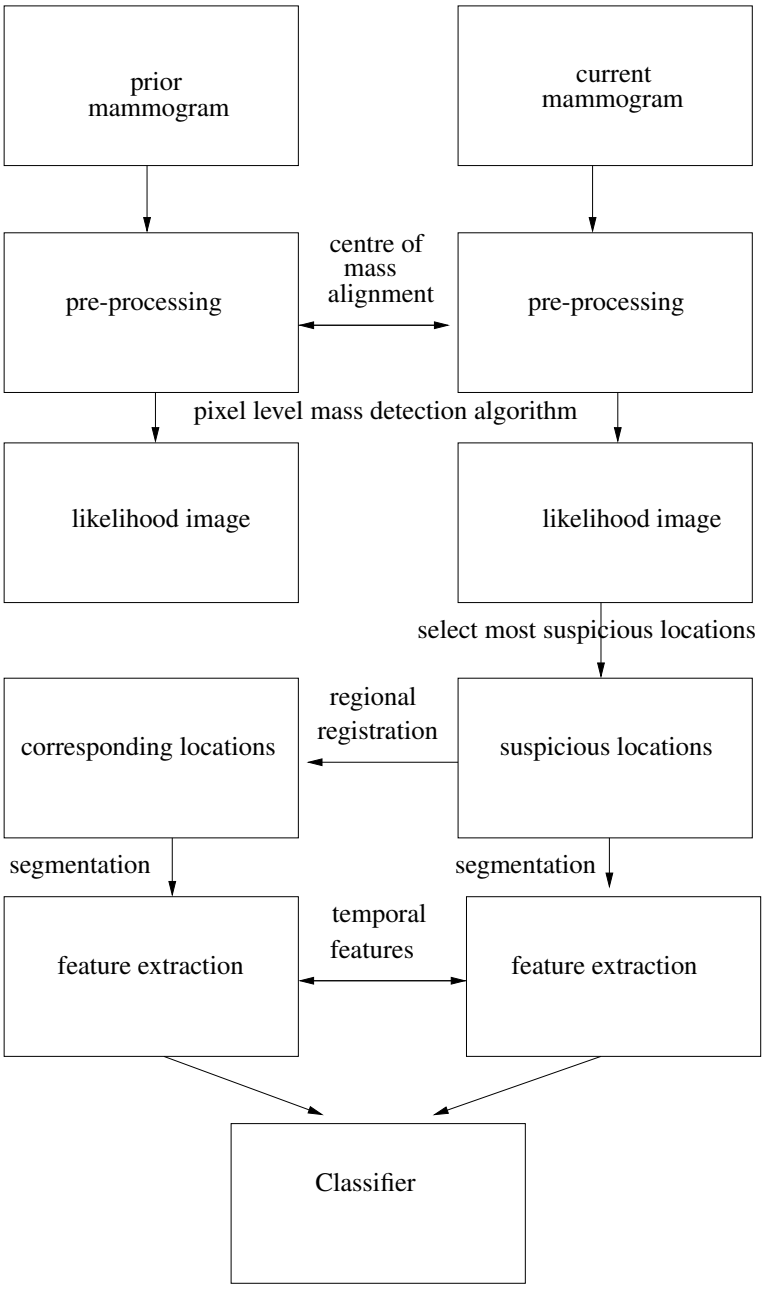


Figure 6.1: *Outline of the temporal CAD method.*

discuss the single view CAD programme and then explain the proposed temporal CAD programme in more detail. Section 6.3 describes the experiments to evaluate the regional registration technique and the temporal CAD programme. In Section 6.4 we present the results of our experiments. Section 6.5 includes some discussion and comparison with a conclusion in the last section.

6.2 Single View and Temporal CAD programme

The CAD programme consists of the following three components: initial CAD programme, single view part, and temporal part. Figure 6.1 gives the outline of the complete method. Subsection 6.2.1 describes the initial CAD programme that includes some pre-processing steps and a pixel level mass detection algorithm. This algorithm detects tumour characteristics and assigns each location in the breast area a score that indicates how likely it is that a lesion is present. Subsection 6.2.2 describes the single view part that includes segmentation of the current image at suspicious locations and feature extraction. Subsection 6.2.3 explains the three steps of the temporal CAD program: global registration, regional registration and feature combination. In the last subsection we describe feature selection and classification for both single view and temporal CAD methods.

6.2.1 Initial CAD programme

Below we shortly review the initial CAD programme, for details see Chapter 2. We apply the initial CAD programme to all prior and current images. We start with pre-processing all images: segmentation of the image into breast tissue, background tissue and pectoral muscle (Karssemeijer 1998), peripheral enhancement to correct for differences in tissue thickness, and removal of the sharp transition in grey level from the breast area to the pectoral region (Timp & Karssemeijer 2006). We then apply a pixel level mass detection algorithm that estimates the potential presence of a tumour at each location inside the breast area. For this purpose we calculate at each location two features for the detection of a spiculation pattern or architectural distortion and two features for the detection of a focal mass. A neural network classifier combines these features into a single score that represent the likelihood that a mass is present at that location. Therefore we call this classifier output score the *mass likelihood*.

6.2.2 Single view CAD method

The single view CAD programme selects locations with a high *mass likelihood* for further processing. First a segmentation algorithm segments the current image at the selected locations. For segmentation we use an algorithm based on dynamic programming, see Chapter 3. We then calculate features for each segmented region resulting in a total of 39

Group Name	No.	Temp	Description
Dense Tissue	5		Features that determine the location of a region with respect to the dense tissue
Spiculation	4	*	Features that detect spiculated lesions
Focal Mass	4	*	Features that detect a focal mass
Mass Likelihood	3	*	Mass likelihood measures
Intensity	1	*	Mean grey value inside the contour
Contrast	5	*	Difference between the grey level histograms of a region and its surround
Variance	4	*	Variance in grey level histogram of a region and its surround
Linear Texture	6	*	Presence of linear texture
Iso-denseness	1	*	Iso-denseness of the segmented region
Location	3		Features that determine the location of a region relative to the pectoral muscle and the skin
Size	1	*	Size of the segmented region
Circularity	1	*	Circularity of the segmented region
Wolfe	1		Estimated Wolfe class

Table 6.1: *Feature description. We divide the basic features into twelve different groups. The first column gives the group name, the second column the number of features in each group, and the last column a description of the group. A star indicates that we calculate these features both for prior and for current regions to obtain temporal information.*

features. Chapter 2 describes these features in detail. We call these single view features the *basic feature set* and divide each feature into one of 12 different categories according to the type of characteristic it represents. Table 6.1 lists the different feature groups.

6.2.3 Temporal CAD method

In the temporal CAD part we first globally register previous and current views. Then we apply a regional registration technique to link each suspicious site on the current view to a corresponding site on the prior view. After completion of the linking procedure the prior image is segmented at the selected location and features are calculated for the segmented region on the prior view. Combining features from both views provides temporal information. Figure 6.2 and 6.3 show temporal image pairs and the corresponding likeli-

hood images. Figure 6.2 shows a newly developed mass, whereas the mass on Figure 6.3 was already visible on the previous mammogram. The three most suspicious sites are indicated with their corresponding numbers.

Global Registration

To register current and the prior views we use a simple procedure based on centre of mass alignment. A problem with centre of mass alignment is the varying proportion of the pectoral muscle that is visible. Van Engeland *et al.* (2003) found that the registration improved considerably by excluding the pectoral muscle in the centre of mass calculation. Therefore we first segment the pectoral muscle using the Hough transform, see (Karssemeijer 1998). Then we calculate the centre of the breast area for both prior and current views with the pectoral muscle excluded. Next we register both views using vertical and horizontal translations.

Regional Registration

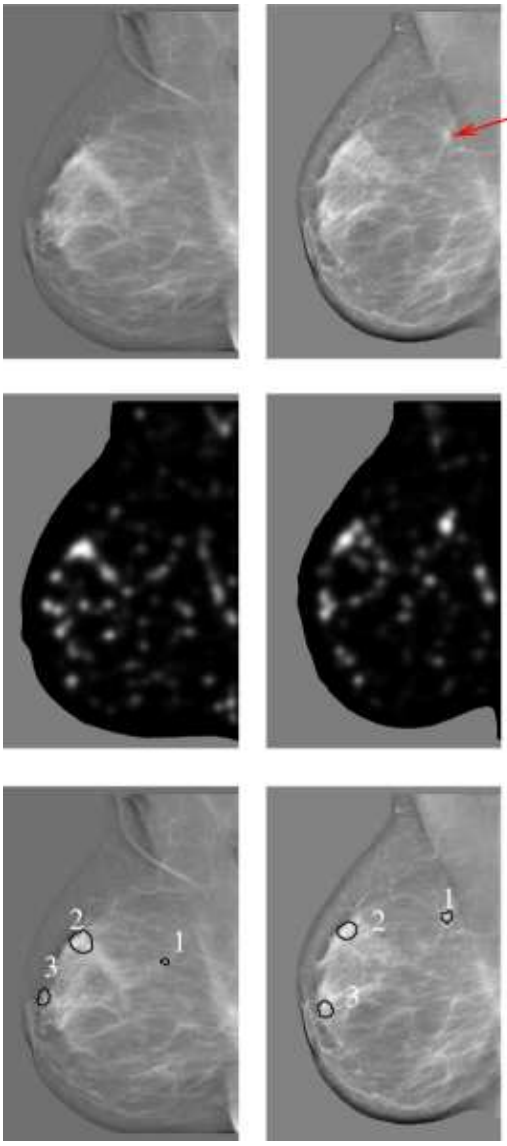
The next step in the temporal CAD programme is regional registration to find for each suspicious site on the current view a corresponding site on the prior view. In the previous chapter we described a registration method that combines three different measures. In this chapter we use a simplified version of this method that only uses two different registration measures: the mass likelihood and a distance criterion. To this end we first define a search area on the prior view in which the mass is likely to be located. As both mammograms are globally aligned we can use the coordinates of the centre of the current lesion (μ_x, μ_y) as initial estimate for the location of the lesion on the prior mammogram. This initial estimate defines the centre of a circular search area with radius r as illustrated in Figure 6.4. We calculate both registration measures at each location inside this search area. The combined registration measure is proportional to the mass likelihood and inversely proportional to the distance from the pixel to the initial estimate:

$$R(i,j) = l(i,j) - w_d d(i,j), \tag{6.1}$$

where $l(i,j)$ is the value of the mass likelihood at location (i,j) and $d(i,j)$ the distance to the initial estimate (μ_x, μ_y). The factor w_d is a weight factor that determines the relative importance of the distance criterion. We select the location with the highest registration measure as match for the location on the current view. In Subsection 6.4.1 we evaluate the performance of this regional registration technique for different values of r and w_d.

Segmentation and Feature Extraction

The last step in the temporal CAD programme concerns segmentation of the prior image at the selected locations and extraction of features from these segmented regions. For

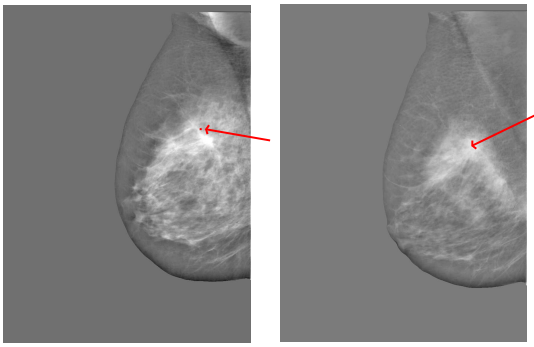


Prior (left) and current image of a newly developing mass.

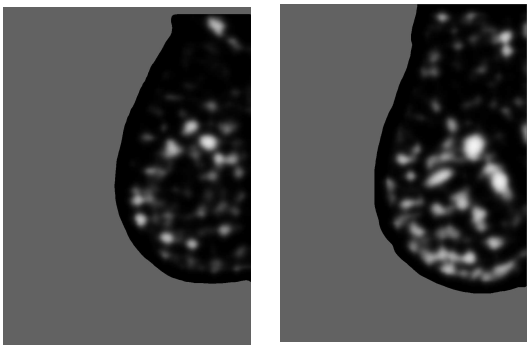
Likelihood image: the most suspicious locations are the white spots.

A regional registration technique links each selected site on the current image to a corresponding location on the prior image.

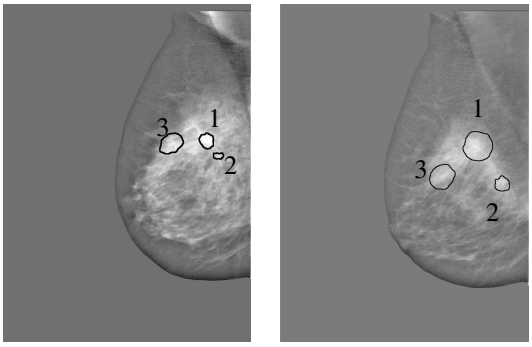
Figure 6.2: *The upper row shows the prior (left) and current image of a newly developing mass. The middle row shows the likelihood images for both prior and current mammograms. The most suspicious locations on the current likelihood image are selected. A regional registration technique then links each selected site to a corresponding location on the prior image.*



Prior (left) and current image of a growing mass.



Likelihood image: the most suspicious locations are the white spots.



A regional registration technique links each selected site on the current image to a corresponding location on the prior image.

Figure 6.3: *The upper row shows the prior and the current image of a growing mass. The middle row shows the likelihood images for both prior and current mammograms. The most suspicious locations on the current likelihood image are selected. A regional registration techniques then links each selected site to a corresponding location on the prior image.*

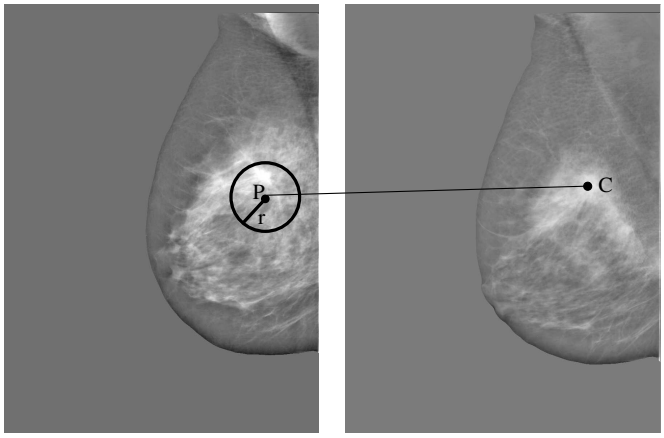


Figure 6.4: *After global registration each suspicious site $C = (\mu_x, \mu_y)$ on the current view defines a circular search area on the prior view with centre $P = (\mu_x, \mu_y)$ and radius r. The best matching site inside this area is linked to location C.*

each region pair we obtain temporal features by subtracting the prior feature value from the current feature value. We determine temporal features for all single view features except for the location features, the dense tissue features, and the estimated Wolfe class. This results in a total of 30 temporal features. We call the set containing both single view and temporal features the *temporal feature set*. Table 6.1 summarises these features.

6.2.4 Classification

Before classification we normalise each feature to zero mean and unit variance:

$$f' = \frac{f - \overline{f}}{\sigma(f)},$$

where we used the whole dataset to determine the mean $\overline{f}$ and standard deviation $\sigma(f)$ of each feature f. The classifier design consists of the following two stages: feature selection and classifier training. Both parts are done completely independent from the evaluation of the classifier. We use a cross-validation scheme to randomly partition the dataset into a training set and a test set on a 10:1 ratio under the constraint that images from the same patient are grouped into the same subset.

In the first stage we use the training set to select the best subset of features. As feature selector we use sequential forward floating selection (SFFS) (Pudil *et al.* 1994). In the second stage we use the training set to train a simple 3-layer feed-forward neural network classifier. After training the neural network assigns all regions in the test set a

score that indicates whether the region is malignant or not, called the *malignancy score*. By making both feature selection and classifier design independent of the test set, we aim at improving the generalisability of our classification results to unknown cases in the patient population.

6.3 Mass Detection Experiments

6.3.1 Dataset

The dataset for this study consisted of 4871 single view images obtained from 938 women. All images used in this study came from the Dutch Breast Cancer Screening Programme. From these 4871 single view images we constructed 2873 temporal image pairs. The number of temporal pairs was larger than half of the number of the images since for some women the mammograms from three consecutive exams were available: the diagnostic mammogram, the most recent prior mammogram (prior I) and the second most recent prior mammogram (prior II), see Figure 1.3. The images were digitised with either a Canon CFS300 or a Lumisys 85 scanner at a pixel resolution of 50 μm, and were averaged to a resolution of 200 μm maintaining the original grey value resolution of 12 bits.

In 589 image pairs the current view contained exactly one malignant mass. We call this the malignant dataset. In 44% the mass was also visible on the prior view. This resulted in 262 temporal images pairs with a visible mass on both current and prior views, and 327 image pairs with a newly developing abnormality on the current view. No pathology was present in 2284 images. We call these images normal. We manually outlined all malignant masses under supervision of an expert radiologist on a dedicated mammographic review station.

For the experiments we made two subdivisions of the malignant dataset. The first subdivision was between masses that were visible on the prior view and masses that were not visible on the prior view, that is between visible priors and normal priors. We made this subdivision to study whether temporal features are as useful for new lesions as for existing lesions. The second subdivision was between image pairs in which the current mammogram was a diagnostic mammogram and pairs in which the current mammogram was a prior I screening round mammogram. Table 6.2 summarises the different sets.

6.3.2 FROC Analysis

We use Free Response Operating Characteristic (FROC) methodology to evaluate the detection accuracy of the total dataset and the different subsets for both the basic feature set and the temporal feature set. We consider a tumour as detected when the initial detection location is inside the ground truth. If multiple detections are found inside the

Subset	No. of Image Pairs	Diag	Prior I
malignant dataset	589	407	182
subset with visible priors	262	195	67
subset with normal priors	327	212	115

Table 6.2: *Description of different subsets: malignant dataset, subset with visible priors and subset with normal priors.*

same ground truth region they are considered as a single hit. We count detections outside the ground truth areas as false positive signals. We only perform image based analysis as the number of temporal cranio caudal image pairs is too low to perform a case based analysis.

Furthermore we calculate for each partition, obtained by ten fold cross-validation of the original dataset, the area under the FROC curve. We are mainly interested in the detection performance obtained for a low number of false positives per image as this corresponds with normal screening situations. Therefore we use a logarithmic scale for the number of false positives per image and calculate the area under the FROC curve from 0.05 FP/image to 1.0 FP/image. We use the two-sided paired Wilcoxon test with 0.95 confidence level to asses the difference in performance between the basic feature set and the temporal feature set.

6.4 Results

In this section we describe the results of the experiments. First we give the performance of the regional registration method. Then we describe the features chosen by the feature selector. Finally we give the detection performance for the total dataset and the different subsets.

6.4.1 Regional Registration

We evaluated the regional registration performance on a set of malignant lesions with known ground truth. In this set all lesions were visible on current and prior mammograms. As evaluation measure we used the percentage of correctly linked locations. We considered a match as correct when the location selected by the regional registration method was inside the ground truth area.

Figure 6.5 shows the performance of the regional registration performance. The y-axis plots the fraction of correctly matched regions. The x-axis indicates the radius r of

the circular search area. The size of this radius depends on the accuracy of the global registration method. A small search area would suffice for an almost perfect global registration. However, as registration is a difficult task in mammography, a large search area in combination with a proper regional registration technique might be preferred. We compared our combined registration method with the correlation measure from Sanjay-Gopal *et al.* (1999). This correlation measure indicates the similarity between regions on prior and current views. Our proposed combined registration uses both the mass likelihood and a distance criterion to select the best match. The highest number of correctly matched regions for the proposed method was 72% for $w_d = 2.0$ and $r = 20$ mm. The method based on correlation linked 69% correct for $r = 16$ mm.

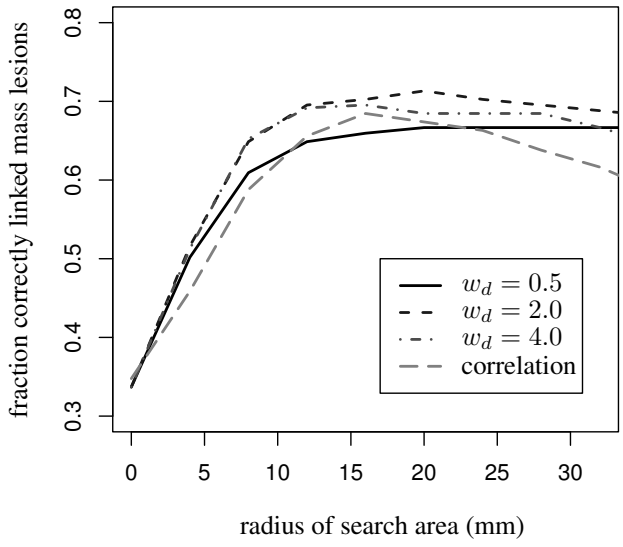


Figure 6.5: *Regional registration results for different values of w_d compared to the registration method based on template matching. On the horizontal axis the radius of the search area is plotted. The vertical axis shows the fraction of tumours that were correctly linked by each of the regional registration methods.*

6.4.2 Feature Selection

During the first stage of the classification procedure the best features were selected based on the training set. The feature set was either the *basic feature set* or the *temporal feature*

Group Name	No. from basic set	No. from temporal set
Mass Likelihood	20	20
Location	18	20
Dense Tissue	17	12
Contrast	16	7
Spiculation	12	10
Focal Mass	10	6
Circularity	8	7
Size	9	1
Iso-denseness	7	2
Linear Texture	3	9
Variance	0	2
Contrast Difference	-	10
Size Difference	-	9
Mass Likelihood Difference	-	3
Spiculation Difference	-	2

Table 6.3: *Results of the feature selection process. The first and the second column list the number of features that have been selected from the basic feature and the temporal feature set respectively.*

set. For both sets the feature selection procedure resulted in a subset of features from the total feature set. By ten fold cross-validation we obtained ten different subsets of features. Table 6.3 lists the number of selected features for the *basic feature set* and the *temporal feature set*.

From Table 6.3 we see that the most frequently selected temporal features are difference in contrast, difference in size, and difference in mass likelihood. At the same time corresponding features from the basic feature set were selected less frequently. So the temporal features were selected instead of their corresponding basic features. For example the basic feature selector almost always selects the feature size, while the temporal feature selector instead chose the feature *difference in size*. An explanation might be that difference features contain both information about the current region as well as temporal information. Table 6.4 lists some information about the selected temporal features. We calculated the mean and the standard deviation for both the current feature and for the corresponding difference feature. We furthermore studied the difference between false positive regions and true mass lesions. The table shows that difference features for the

Feature	False Positives				True Positives				A_z
	Basic		Temporal		Basic		Temporal		
	mean	sd	mean	sd	mean	sd	mean	sd	
Size	0.26	0.10	-0.02	0.16	0.54	0.28	0.21	0.34	0.67
Likelihood	1.58	0.11	0.04	0.27	2.04	0.11	0.33	0.25	0.66
Contrast	0.35	0.03	0.01	0.04	0.60	0.09	0.21	0.10	0.75

Table 6.4: *Mean and standard deviation (sd) of selected temporal features for false positives and true positives. The most frequently selected temporal features were difference in size, difference in contrast and difference in mass likelihood. The last column shows the A_z value for the selected temporal features.*

false positive regions have small values indicating that on average the features stay more or less constant during time. For the true positives we find that most feature values change during time. On average true positives are larger, have a higher mass likelihood, and have more contrast compared to lesions one screening round earlier. We evaluated the individual performance of each selected temporal feature by calculating the area under the individual ROC curve. For this purpose we first applied our initial detection algorithm to select the most suspicious region in each image. This resulted in 200 false positive regions and 389 true positive regions. We used these regions to construct an ROC curve. The last column in Table 6.4 gives the area under the ROC curve (A_z value) for each selected temporal feature.

6.4.3 FROC analysis

Figure 6.6 shows the mass detection performance for the basic feature set and the temporal feature set. The figure shows that temporal features improve the detection performance, especially at a low number of FP detections per image. Table 6.5 gives the results of the Wilcoxon statistic for the total dataset and the two subsets. The difference in performance between both feature sets is statistically significant.

We furthermore calculated FROC curves for the different subsets. Figure 6.7 shows the results for masses with visible and normal priors. We see that masses that are visible on the prior profit more from temporal features than masses with normal priors.

Figure 6.8 shows the results for mammogram pairs in which the current mammogram is the diagnostic mammogram and pairs in which the current mammogram is a prior I mammogram. The detection performance for diagnostic mammograms is better than for prior I mammograms. Both subsets show an improvement when temporal features are used. These improvements however are not statistically significant.

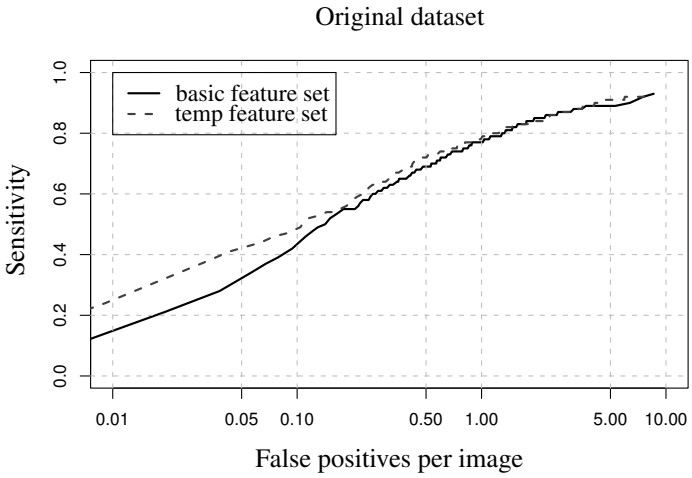
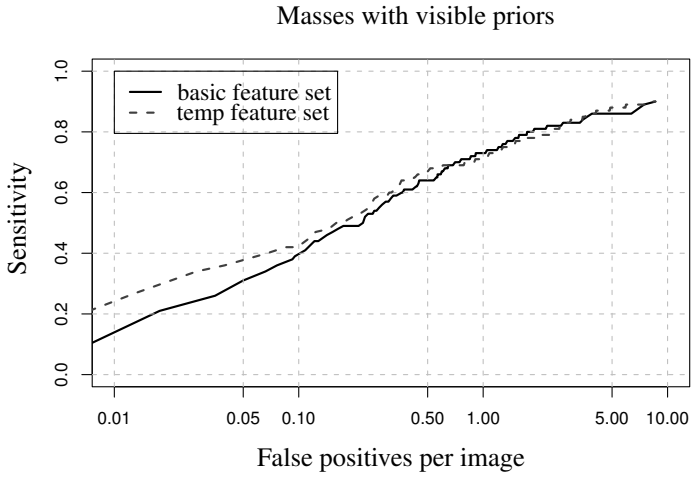


Figure 6.6: *Image based FROC detection results for the basic feature set and the temporal feature set.*

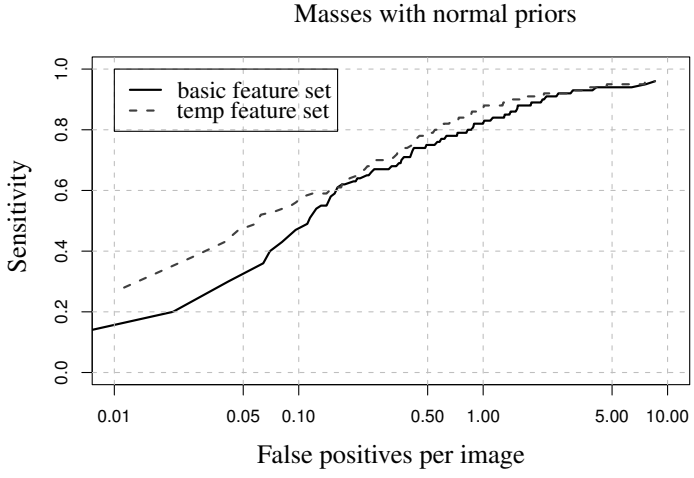
	Total Dataset	**Visible Priors**	**Normal Priors**	**Diag**	**Prior I**
Basic	0.706	0.762	0.656	0.785	0.45
Temp	0.721	0.785	0.669	0.797	0.46
P-value	0.05	0.05	0.13	0.19	0.38
CI	(0.00,0.03)†	(0.02,0.08)†	(-0.01,0.06)	(-0.01,0.03)	(-0.01,0.04)

Note:—† Statistically significant. CI = Confidence Interval

Table 6.5: *Results of the Wilcoxon's test for the statistical difference in area under the FROC curve for different datasets. The first two rows gives the mean area under the FROC curve for the basic and the temporal feature set. The third row gives the p-value for Wilcoxon's statistic. The last row gives the 95% confidence interval.*

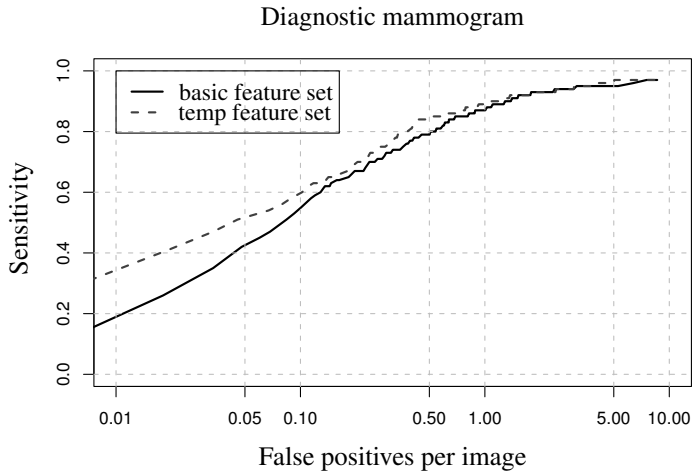


(a)

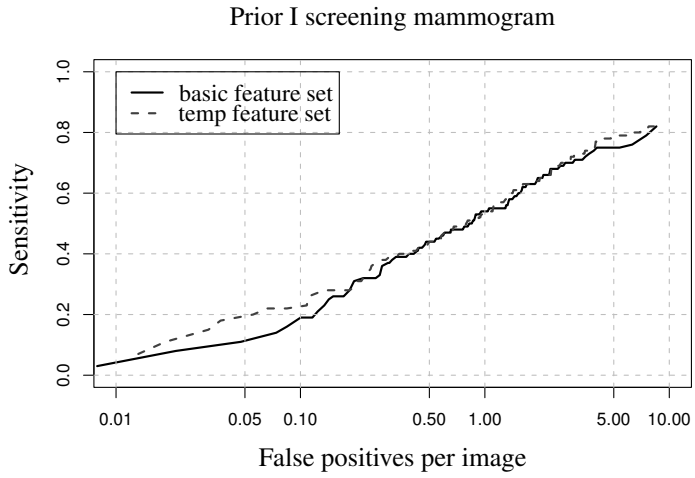


(b)

Figure 6.7: *Image based FROC detection results for the subsets in which the mass was visible cq. not visible on the prior mammogram.*



(a)



(b)

Figure 6.8: *FROC detection results for the subsets in which the current mammogram is the diagnostic mammogram and the prior I mammogram respectively.*

6.5 Discussion

In this study we investigated the additional value of temporal features for our detection scheme. For this purpose we used a simplified version of the registration method described in Chapter 5. This registration method combines the mass likelihood and a distance criterion to determine for each region on the current view the best matching location on the prior view. This technique correctly linked 72% of all mass lesion. This method is fast as we already calculated the mass likelihood in the single view CAD programme. Furthermore the method is completely automatic.

We used feature selection to find the best features in the basic and the temporal feature set. The feature selection method most frequently selected the following three temporal features: contrast difference, size difference, and difference in mass likelihood. The mass likelihood is a feature from the first detection step and indicates the likelihood that a focal mass lesion or a spiculation pattern is present. In Chapter 4 we examined mammographic changes in masses regarding to size and contrast. In that study we also found that the features contrast and size both increased in time. These features can thus be used as tumour markers.

Figure 6.6 gives the detection performance of the total dataset with and without the use of temporal features. The detection performance significantly improved when using temporal features. In the current study we calculated temporal features for all regions, regardless of whether they were visible on the prior or not. Figure 6.7 shows the results for existing and new lesions. We see that masses that are visible on the prior profit more from the use of temporal features than masses than new masses, although these also show a small—not significant—improvement when temporal features are used. This indicates that temporal features have a different effect on both groups. Therefore it might be better first to classify all regions on the current as new or existing and then decide which features to calculate for each group. Some features can be useful for both new and existing lesions. An example is contrast. If a tumour is not visible on the prior we can define an artificial region at the location selected by the regional registration programme and calculate contrast measures inside this region. Then we can compare the contrast of this region with the contrast of the region on the current. For the feature size we can not use the size of the artificial region, as nothing is visible. Instead we can for instance set the size to zero when nothing is visible on the prior view. From the above mentioned examples we conclude that it might be useful to take into account whether a region is new or already existed. Calculating different features for both groups may lead to a better detection performance.

In summary, we performed a study in which we obtained temporal difference features by subtracting the prior feature value from the current feature value for corresponding regions on both views. We observed an improvement in detection performance when using these temporal difference features. In the next chapter we focus on developing

specific temporal measures that determine whether prior and current regions are similar in appearance. When both regions are similar, it is likely that the region represents a false positive detection or a slowly growing benign mass. On the other hand, if a region has changed considerably, this is more suspect for a malignant lesion. These features therefore might be useful to discriminate between benign and malignant lesions.

Chapter 7

Interval Change Analysis for the Characterisation of Masses [1]

In this chapter we investigate the use of temporal features to improve the characterisation of masses. For this purpose we first apply the regional registration technique described in Chapter 5 that finds for each mass lesion on the current view a location on the prior view where the mass most likely developed. For the task of interval change analysis we use two kinds of temporal features: difference features and similarity features. Difference features indicate the (relative) change in feature values determined on prior and current views. These features may be especially useful for lesions that are visible on both views. Similarity features measure whether two regions are comparable in appearance and may be useful for lesions that are visible on the prior view as well as for newly developing lesions. We evaluate the effect of these features on the performance of a CAD system that discriminates between benign and malignant lesions.

7.1 Introduction

An important task of radiologists in mammography is to discriminate between benign and malignant lesions. In clinical practice a radiologist carefully analyses all detected lesions and classifies each lesion as benign, probably benign, suspicious, or highly suggestive of malignancy. The BI-RADS reporting system provides criteria on which radiologists should make this classification (D'Orsi & Kopans 1997; Orel *et al.* 1999). The subsequent management of lesions mainly depends on this classification. For probably benign findings short-interval follow-up is suggested. For suspicious abnormalities

[1] This chapter is based on Timp *et al.* (2006b)

biopsy should be considered. A good decision increases the number of correctly detected malignancies and reduces unnecessary additional examinations.

Mass lesions have some characteristics that can be used to discriminate between benign and malignant lesions (Friedrich & Sickles 2000; Homer 1997). An important characteristic is the margin type of a lesion. Most benign masses possess well-defined sharp borders, while malignant tumours often have ill-defined, micro-lobulated, or spiculated borders. Especially a spiculation pattern is strongly associated with the presence of a malignant lesion. The differential diagnosis of a spiculated lesion is short and includes a postoperative scar, a radial scar, fat necrosis, or any process resulting in marked fibrosis. Another characteristic that may be helpful in discriminating between benign and malignant lesions is the shape of a lesion. The shape of benign lesions is often round and oval, compared to a more irregular shape of most malignant lesions. The last difference between benign and malignant lesions is the *tumour behaviour* over time. Benign masses tend to change slowly and have a more or less similar appearance on two consecutive screening mammograms. Malignant masses on the other hand may change considerably and become more suspicious during time. This chapter focuses on the design of features that capture temporal changes to improve the characterisation of mass lesions.

Some studies have been done to evaluate the effect of using temporal information on either the detection (Bassett *et al.* 1994; Thurfjell *et al.* 2000; Callaway *et al.* 1997) or characterisation (Varela *et al.* 2005; Hadjiiski *et al.* 2004) of mass lesions. The last two studies are observer studies that evaluate the effect of prior views on the ability of radiologists to discriminate between malignant and benign lesions. Varela *et al.* (2005) did a study with six radiologists and found that the performance of each radiologist improved when using prior mammograms. Hadjiiski et al. (Hadjiiski *et al.* 2004) did a study with eight radiologist and two breast imaging fellows and also found a significant improvement when the radiologists used prior views.

To our knowledge only one study compared the performance of a CAD system with and without using prior views (Hadjiiski *et al.* 2001b). The dataset for that study consisted of mammograms from two consecutive screening rounds with a visible mass lesion on the current and prior view. A radiologist first identified the mass lesion on current and prior mammograms after which a CAD programme calculated single view and temporal features. On a dataset consisting of 140 temporal image pairs the A_z value significantly increased from 0.82 to 0.88 when temporal features were added to the CAD system.

In this chapter we develop a CAD programme for temporal change analysis to improve the characterisation of breast masses. This programme combines single view and temporal features to determine a likelihood of malignancy for each mass lesion. Our proposed method has some advantages. First, our method is almost completely automatic. It only requires manual identification of the mass on the *current* view, after which a regional registration programme is applied to identify a location on the prior view that best corresponds with the current mass lesion. Existing methods require manual identification

of the mass on both *prior* and *current* views. Second, our method is not only suited for masses that are visible on the prior view but also for masses that are new. This corresponds with normal screening situations where only some lesions are visible on the prior view. Third, besides using difference features we include temporal features that measure changes in appearance between a mass region on the current view and a similar region on the prior view. These features discriminate between benign lesions that stay more or less constant and malignant lesions that change between two consecutive screening rounds.

Radiologists can use this programme as an aid to characterise mass lesions. When a radiologist uses this method he should provide the coordinates of the lesion on the current view. The programme then automatically finds a corresponding location on the prior view and determines single view and temporal features to estimate the likelihood that the lesion is malignant. Studies in the literature suggest that a radiologist can use this *likelihood of malignancy* to improve interpretation of lesions (Hadjiiski *et al.* 2004; Huo *et al.* 2002; Chan *et al.* 1999).

We evaluate the performance of our method on a dataset consisting of 238 benign and 227 malignant temporal mammogram pairs. Furthermore we split the dataset into two subsets. The first subset consists of masses that are visible on the prior view and the second subset of masses that are not visible on the prior view. We study which features are useful for each subset and determine the classification performance for each subset.

The remainder of this chapter is organised as follows. Section 7.2 explains the proposed CAD method for characterisation of mass lesions. Section 7.3 describes the experiments, including the dataset in Section 7.3.1, and the classification results in Section 7.3.2 and 7.3.3. The last section contains a discussion and conclusion.

7.2 Single View and Temporal CAD Programme

This section describes our CAD programme that processes mammograms from consecutive mammographic exams in which the most recent mammogram contains a visible lesion. This lesion has been annotated by or under supervision of an expert radiologist. Figure 7.1 shows an example of a case that consists of three consecutive mammograms. In this example we see that priors are not always available for CC views. The CAD programme consists of the following three components: initial CAD programme, single view part and temporal part. Subsection 7.2.1 describes the initial CAD programme that includes some pre-processing steps and a pixel level mass detection algorithm. This algorithm detects tumour characteristics and assigns each location in the breast area a score that indicates how likely it is that a lesion is present. Subsection 7.2.2 describes the single view part that is applied to all current images. In brief a segmentation programme determines a contour for each current lesion after which several features are calculated to discriminate between benign and malignant lesions. A Support Vector Machine clas-

sifier combines these features into a single malignancy score that indicates whether the lesion is malignant or not. Subsection 7.2.3 describes the temporal CAD programme that is applied to images for which prior views are available. This programme works as follows. A regional registration method finds for each segmented lesion on the current view a corresponding location on the prior view where the mass most likely developed. We calculate two kinds of temporal features at this location to measure interval changes between the current lesion and a corresponding region on the prior view. We add these temporal features to the single view features to improve the characterisation performance.

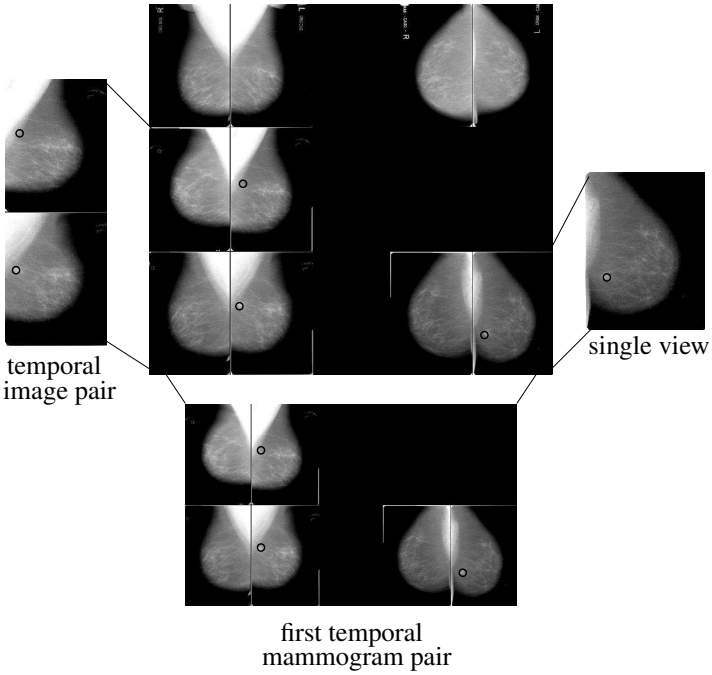


Figure 7.1: *Example of three consecutive mammograms of the same woman. Mammograms are displayed in chronological order. The bottom row represents a referral mammogram. A malignant lesion is present in the left MLO image of the referral and its corresponding prior mammogram. The mammograms from two consecutive screening rounds form a temporal mammogram pair. This case provides two temporal mammogram pairs. The bottom and middle rows show the first mammogram pair, in which the referral mammogram represents the current mammogram. This mammogram pair consists of two temporal image pairs (left and right MLO current-prior) and two single views (left and right CC). The top and middle rows form the second mammogram pair, in which the mammogram prior to referral represents the current mammogram. This mammogram pair contains two temporal image pairs (left and right MLO current-prior).*

7.2.1 Initial CAD programme

Below we shortly describe the initial CAD programme, for details see Chapter 2. The initial CAD programme is applied to all prior and current images. We start with pre-processing all images: segmentation of the image into breast tissue, background tissue, and pectoral muscle (Karssemeijer 1998); peripheral enhancement to correct for differences in tissue thickness; and removal of the sharp transition in grey level from the breast area to the pectoral region (Timp & Karssemeijer 2006). We then apply a pixel level mass detection algorithm that estimates the potential presence of a tumour at each location inside the breast area. For this purpose we calculate at each location two features for the detection of a spiculation pattern or architectural distortion and two features for the detection of a focal mass. A neural network classifier combines these features into a single score that represent the likelihood that a mass is present at that location. Therefore we call this classifier output score the *mass likelihood*.

7.2.2 Single View CAD

After pre-processing the single view CAD programme processes all current images. First the mathematical centre of mass of the radiologists' annotation is determined for each mass lesion. A segmentation algorithm based on dynamic programming—for details see Chapter 3—uses this location as starting point to determine a contour for each lesion. For each segmented lesion several single view features are determined that are useful for characterisation of mass lesions. A Support Vector Machine classifier combines these features into a single score that represents the probability that the lesion is malignant. Table 7.2 summarises the single view features that include spiculation measures, border features, location features, morphological features, and a feature that indicates the presence of micro-calcifications. For a description of these features see Chapter 2.

7.2.3 Temporal CAD

The temporal CAD part consists of three steps. In the first step prior and current images are globally registered using a centre of mass alignment (Van Engeland *et al.* 2003). After alignment we use the centre coordinates of the current lesion (μ_x, μ_y) as midpoint of a circular search area on the prior view with radius 2 cm. Inside this search area we use a regional registration programme to select the location on the prior view where the mass most likely developed. This registration method has been described in detail in Chapter 5. Shortly the method works as follows. At each location (i, j) inside the search area we calculate three registration measures: mass likelihood, distance and grey level correlation. The mass likelihood indicates the potential presence of a mass at each location. The distance measure indicates the distance from (i, j) to the centre of the search area (μ_x, μ_y). The last measure is Pearson's correlation between the current region and

a similar region on the prior view centred at (i, j). A linear discriminant analysis (LDA) classifier combines these measures—mass likelihood, distance and correlation—into a single score: the *registration score*. We select the location with the highest registration score (i_s, j_s) as match for the current mass lesion. Figure 7.2 shows some examples of temporal image pairs and the location selected by the regional registration programme.

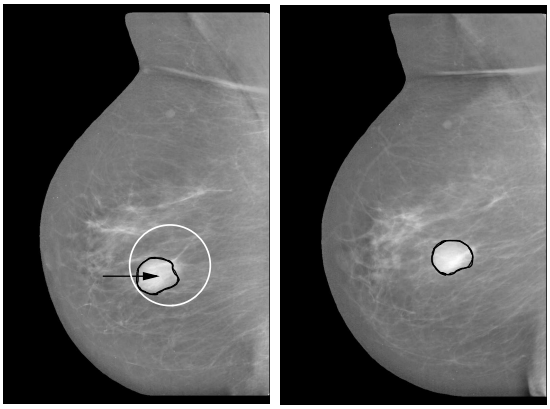
We then use the location (i_s, j_s) as starting point for our segmentation algorithm. This algorithm determines a contour of the region on the prior view, independent of whether the lesion is visible or not. We extracted single view features from the segmented region on the prior view and calculate two kinds of features that measure temporal changes: difference features and similarity features.

Difference Features Difference features measure changes in feature values between the prior and the current region. In our experiments we determine difference features for all single view features except for the location features. For the feature "size" we use the relative change between the feature value of the current region and the feature value of the prior region. For the other features we use the absolute change as these features are already normalised measures. Difference features may be especially helpful when the tumour is already visible on the prior view. When the lesion is not yet visible on the prior view the contour defined by our segmentation programme is not meaningful. Features that depend on the contour such as the size of a region are not useful in that case.

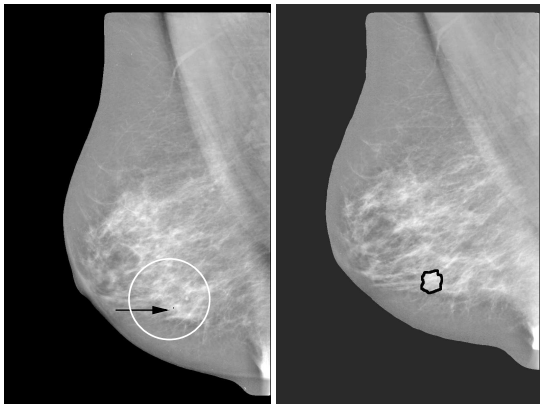
Similarity Features The second group of temporal features measure the similarity between the current region and the selected region on the prior view.

- Regional Registration Score. The first similarity feature is the output from the regional registration programme. This feature corresponds with the likelihood that a correct link has been established. A low registration score therefore may indicate that the lesion is not visible on the prior view. The classifier might use this information to determine the relative usefulness of temporal difference features. The registration score on its own may also help to characterise mass lesions. A high registration score for example may indicate the presence of a benign mass when the mass is obvious on the prior view—resulting in a high mass likelihood—and similar on prior and current views—resulting in a high correlation measure. A low registration score on the other hand may suggest the presence of a malignant lesion as malignant lesions often change more between two consecutive screening rounds.

- Relative Grey Level Change. The second similarity feature calculates the relative difference in grey level between the current and prior region. For this purpose we transform the current image such that its grey level histogram matches that of the

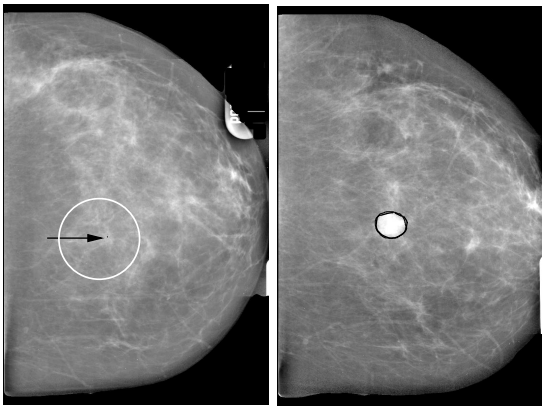


(a) On prior and current views a benign mass is present that is similar in appearance on both views.



(b) On the current view a benign mass is present. On the prior view a similar region is selected.

Figure 7.2: *Pairs of temporal images. Left and right images correspond to prior and current views. In each prior view the arrow indicates the location selected by the regional registration programme. Fig. 7.2(a) shows a benign mass that is similar on the prior and the current view. The benign mass in Fig. 7.2(b) is not yet visible on the prior view. The registration programme selects the most probable location on the prior view.*



(c) On the current view a malignant mass is present. The prior view shows no abnormality.

Figure 7.2: *(cont.) Pairs of temporal images. Fig. 7.2(c) shows a malignant mass on the current view. On the prior view no abnormality is discernible. The registration programme selects the most probable location on the prior view.*

prior image. We first calculate for the prior and the current image the cumulative histograms of the grey values inside the breast area. For each grey level y the cumulative histograms are

$$f_C(y) = \sum_{i=0}^{y} H_C(i) \qquad\qquad f_P(y) = \sum_{i=0}^{y} H_P(i),$$

where H_C (H_P) is the histogram of the grey value inside the current (prior) breast area. We then transform each grey level y of the current image

$$\tilde{y} = f_P^{-1}(f_C(y)).$$

After histogram matching we determine the relative grey level change between a similar region on the prior and the current view. We use the segmented region on the current view as a template and put this template over the selected location (i_s, j_s) on the prior image. The relative grey level change between both regions is

$$RGLC = \frac{1}{N} \sum_{(m,n)\in C} (\tilde{y}_c(m,n) - y_p(m',n')),$$

where the summation is performed over all locations (m, n) inside the current region C. N denotes the number of pixels inside C, $\tilde{y}_c(m, n)$ the transformed

grey level at location (m, n) in C and $y_p(m', n')$ the grey level at the same relative location in the prior region with centre (i_s, j_s).

7.2.4 Case Based Classification

As classifier we use a Support Vector Machine (Cristianini & Shawe-Taylor 2000) where we use the implementation provided in the freely available package from CRAN (Hornik 2005). We use the radial basis kernel for training and testing. For testing we use the probability model for classification assuming equal priors. The probability model for classification fits a logistic distribution using maximum likelihood to the classifier outputs. The probabilistic regression model assumes (zero-mean) Laplace-distributed errors for the predictions, and estimates the scale parameter using maximum likelihood (Hornik 2005).

As not all images in our dataset have prior views (see Figure 7.1) we train two different classifiers: a single view classifier and a temporal classifier. For both classifiers we apply a 20-fold cross-validation scheme to partition the dataset into a training set and a test set. The *single view classifier* estimates for each image the posterior probability $p(m|\mathbf{x_{sv}})$ that a lesion is malignant given the feature vector $\mathbf{x_{sv}}$ with single view features extracted from the current region . The case based malignancy score $\zeta(l)$ combines the posterior probabilities from available MLO and CC views to estimate the likelihood that a lesion is malignant. When both MLO and CC views are present we use the sum rule to determine this case based malignancy score (Kittler *et al.* 1998):

$$\zeta(l) = \frac{1}{2}(p_{mlo}(m|\mathbf{x_{sv}}) + p_{cc}(m|\mathbf{x_{sv}})).$$

When only the MLO image is available the case based malignancy score is equal to the posterior probability from the MLO view:

$$\zeta(l) = p_{mlo}(m|\mathbf{x_{sv}}).$$

To include temporal information we train a second (temporal) classifier that determines the posterior probability $p(m|\mathbf{x_t})$ that a lesion is malignant given a temporal feature vector $\mathbf{x_t}$ containing single view, difference and/or similarity features. The case based malignancy score $\zeta(l)$ indicates the likelihood that a lesion is malignant and depends on the available views of the current and the prior mammogram. We distinguish the following situations.

- The temporal mammogram pair only contains a temporal MLO image pair. For the current mammogram no CC views are available. This situation corresponds with the second mammogram pair in Figure 7.1. We use the posterior probability from the MLO view as the case based malignancy score:

$$\zeta(l) = p_{mlo}(m|\mathbf{x_t}).$$

- The temporal mammogram pair consists of a temporal MLO image pair and single view CC images. Prior CC views are not available. For example see the first mammogram pair in Figure 7.1. To determine the case based malignancy score we use the sum rule to combine the posterior probabilities from the temporal MLO classifier and the single view CC view classifier:

$$\zeta(l) = \frac{1}{2}(p_{mlo}(m|\mathbf{x_t}) + p_{cc}(m|\mathbf{x_{sv}})).$$

- The temporal mammogram pair consists of a temporal CC image pair and a temporal MLO image pair. We use the sum rule to combine the posterior probabilities for the MLO and CC view, both obtained with the temporal classifier:

$$\zeta(l) = \frac{1}{2}(p_{mlo}(m|\mathbf{x_t}) + p_{cc}(m|\mathbf{x_t})).$$

For the evaluation of the single view and the temporal CAD system we use Receiver Operating Characteristic (ROC) methodology (Metz 1986; Metz *et al.* 1998b). We quantify the classification accuracy as the area under the case based ROC curve (A_z value). To test whether temporal features improve the performance we perform a paired comparison of both conditions—CAD with and without the use of temporal features—with regard to differences in the area under the two estimated ROC curves. For this purpose we use the freely available CLABROC software (Metz *et al.* 1998a).

7.3 Mass Characterisation Experiments & Results

7.3.1 Dataset

The mammograms used in this study all came from the Dutch Breast Cancer Screening Programme. All women aged 50-75 are invited bi-annually to participate in this programme. Two mammographic views—medio lateral oblique (MLO) and cranio caudal (CC)—are obtained at the initial screening. At subsequent screenings only medio lateral views are obtained, unless there is an indication that additional cranio caudal views would be beneficial. All images were digitised with a Canon CFS300 laser scanner at a pixel resolution of 50 μm and averaged to a resolution of 200 μm maintaining the original grey value resolution of 12 bits. All visible masses were annotated by or under supervision of an expert radiologist.

For the experiments we used consecutive mammograms from a collection of cases that were referred between 1996 and 2000. These cases consist of mammograms at referral and mammograms from up to two previous screening rounds. All images from two consecutive screening rounds form a temporal mammogram pair. In a temporal mammogram pair we call the most recent mammogram the current mammogram and

Name Set	No of Cases (Images)	Benign Cases (Images)	Malignant Cases (Images)
total dataset	465 (720)	238 (356)	227 (364)
temporal dataset	465 (542)	238 (279)	227 (263)
single view dataset	178 (178)	77 (77)	101 (101)
subset with visible priors	202 (246)	108 (133)	94 (113)
subset with normal priors	263 (296)	130 (146)	133 (150)

Table 7.1: *Information about the subsets.*

the mammogram from one screening round earlier the previous or prior mammogram. Figure 7.1 shows an example of a case that contains two temporal mammogram pairs. The first mammogram pair consists of the referral mammogram and the mammogram from the screening round prior to referral. In this temporal pair we call the referral mammogram the current mammogram. This case contains a second mammogram pair because the mass lesion is also visible on the mammogram prior to referral. In this second mammogram pair the mammogram prior to referral represents the current mammogram and the mammogram obtained one screening round earlier the prior mammogram. This means that at the time the current mammogram was taken the woman was not referred for further examination. These mammogram pairs make up 35% of the total dataset and often contain subtle lesions that are difficult to characterise.

We constructed the dataset for the experiments by collecting all temporal mammogram pairs in which the current MLO view contained exactly one visible mass lesion. This resulted in 465 temporal mammogram pairs, 238 benign and 227 malignant. The temporal mammogram pairs consists of 542 temporal image pairs—465 MLO and 77 CC—and 178 single view CC images with a visible mass lesion. The single view images form the *single view dataset*, the temporal image pairs the *temporal dataset*. We constructed two different subsets of the temporal dataset. The first subset consists of masses that are also visible on the prior view and is called the *subset with visible priors*. This set contains 202 cases. The second subset consists of masses that were not visible on the prior view and is therefore called the *subset with normal priors*. This set contains 263 cases. Table 7.1 summarises information about the subsets. We evaluated the benefit of using temporal features on the total dataset as well as on different subsets.

7.3.2 Single View Classification

This section presents the results of our single view CAD system. Table 7.2 gives the performance of the individual features measured as the area under the ROC curve (A_z

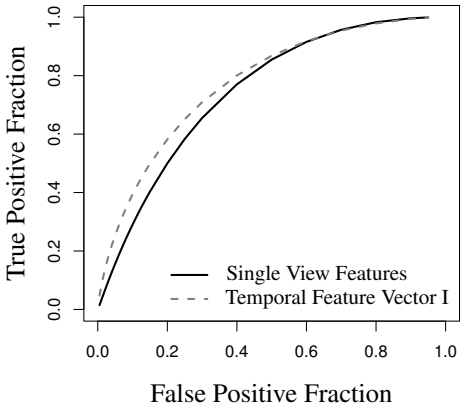
Feature Name	Description	A_z
$f1$	spiculation	0.68
$f2$	spiculation	0.68
$\overline{f1}$	mean value of f1 inside the segmented region	0.68
$\overline{f2}$	mean value of f2 inside the segmented region	0.66
size	size of the segmented region	0.58
circularity	ratio between perimeter and size	0.52
calcification	number of calcifications	0.54
locx	relative x-location	0.56
locy	relative y-location	0.50
BC	continuity of the contour	0.62

Table 7.2: *Summary of the single view features. For each feature we calculated the individual A_z value for the total dataset consisting of 238 benign and 227 malignant mass lesions.*

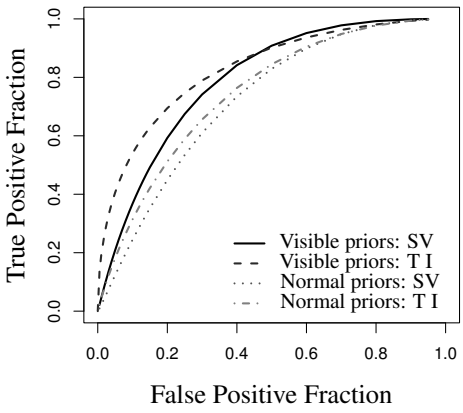
value) for the total dataset consisting of 238 benign and 227 malignant mass lesions. For each region we constructed a single view feature vector that contained all single view features as described in Table 7.2. Table 7.5 gives the performance obtained with this feature vector for the total dataset, the subset with visible priors and the subset with normal priors. This table shows that there is a large difference in classification performance between the subset with visible priors and the subset with normal priors. For the former the average A_z value is 0.79, for the latter 0.72. This may be explained by the observation that masses in the set with visible priors are often quite obvious on the current view. This may result in more distinct tumour characteristics making it easier to characterise these lesions. The set with normal priors on the other hand consists of masses that are only visible on the current view. This set therefore also contains subtle masses which are harder to classify.

7.3.3 Temporal Classification

This section describes the results of our temporal CAD programme. This programme uses single view features extracted from the current region, similarity features, and the four best performing difference features. The best performing difference features were relative difference in size, difference in border continuity, and two features that represent the difference in spiculation. Table 7.3 summarises the individual performance of the



(a) Total dataset



(b) Subsets visible and normal priors

Figure 7.3: *ROC curve for the single view (SV) feature vector and the temporal feature vector I (T I). Fig. 7.3(a) gives the ROC curve for the total dataset, Fig. 7.3(b) gives the ROC curve for the subsets with normal and visible priors. For each set the performance improves when temporal features are used.*

Feature Name	**Description**	A_z
Similarity Features		
registration prob	probability that match is correct	0.60
RGLC	relative grey level change	0.63
Difference Features		
size_diff	relative difference in size	0.61
$f1$_diff	absolute difference in $f1$	0.62
$f2$_diff	absolute difference in $f2$	0.62
BC_diff	difference in continuity of the border	0.61

Table 7.3: *Summary of the temporal features. For each feature we calculated the individual A_z value.*

selected temporal features on the temporal dataset.

For each region we constructed three different temporal feature vectors, see Table 7.4. The first temporal feature vector contains single view and similarity features. The second temporal feature vector contains single view and difference features. The last temporal feature vector contains single view, similarity, and difference features. Table 7.5 gives the classifier performance obtained with the different feature vectors for each dataset. We used the CLABROC programme to compare the performance obtained with the *single view feature vector* with the performance obtained with different temporal feature vectors. We found that the use of *temporal feature vector I* significantly improved the classification performance for the total dataset ($P = 0.005$, two-tailed) and for the subset with visible priors ($P = 0.02$, two-tailed). The improvement for the subset with normal priors however was not significant ($P = 0.11$, two-tailed). Figure 7.3 shows ROC curves obtained with the *single view feature vector* and *temporal feature vector I*. For *temporal feature vector II*—containing single view and difference features—the classification performance only improved for the set with visible priors. This improvement was not significant ($P = 0.22$, two-tailed). For the set with normal priors the performance even decreased, indicating that difference features may not be useful for lesions that are not visible on the prior view. The last temporal feature vector—*temporal feature vector III*—contained single view, difference, and similarity features. The use of this feature vector improved the classifier performance for the total dataset ($P = 0.05$, two-tailed) and for the subset with visible priors ($P = 0.06$, two-tailed).

To estimate the usefulness of difference features we investigated the difference between the classification performance obtained with *temporal feature vector I* and the performance obtained with *temporal feature vector III*. For the subset with visible pri-

Name	Description	No.
single view feature vector	single view features	10
temporal feature vector I	single view and similarity features	12
temporal feature vector II	single view and difference features	14
temporal feature vector III	single view, difference and similarity features	16

Table 7.4: *Summary of the different feature vectors. The first column gives the name of the feature vector, the second column the features that each vector contains, and the last column the number of features in each vector. The first set only contains single view features extracted from current lesions. The temporal feature vectors contain single view and temporal features.*

Dataset	Single View FV	Temporal FV I	Temporal FV II	Temporal FV III
total dataset	0.74±0.02	0.78±0.02 †	0.74±0.02	0.77±0.02 †
visible priors	0.79±0.03	0.83±0.03 †	0.81±0.03	0.83±0.03
normal priors	0.72±0.03	0.75±0.03	0.70±0.03	0.73±0.03

† Statistically significant.

Table 7.5: *A_z value and standard deviation for different feature vectors and different subsets. The single view feature vector consists of features extracted from the current mass lesion. Temporal features contain information of both prior and current regions.*

ors both feature vectors had an equal performance, indicating that difference features do not have an additional effect when similarity features are already used. For the subset with normal priors the addition of difference features even lead to a decrease in performance. These results suggest that similarity features are preferred over a combination of similarity and difference features.

7.4 Discussion

In this chapter we present a completely automated temporal CAD programme for the characterisation of mass lesions. This CAD programme uses two two kinds of temporal features: difference and similarity features. We first discuss the use of difference features. These features only improved the performance when the mass lesion was visible on the prior view. In a previous study Hadjiiski et al. (Hadjiiski *et al.* 2001b) evaluated

the effect of difference features. This work was restricted to cases which had visible masses on the prior views. They found an improvement in A_z value from 0.82 to 0.88 when adding difference features. In our study the A_z value improved from 0.79 to 0.81 when using difference features on the set with visible priors. There are some differences between both studies. In our experiments we used an automated registration programme to determine the location of the mass lesion on the prior view while in their study a radiologist indicated this location. To investigate whether this influenced the results we did an experiment in which we used the centre of the manual segmentation on the prior view instead of the location selected by the registration programme. In this experiment the classification performance increased to 0.82 for the set of visible masses. This result differs not much from the proposed CAD method indicating that we can use our automated registration programme instead of manual annotations. Another difference between both experiments concerns the used difference features. Hadjiiski et al. used texture difference features, while this study included only spiculation and morphological difference features. Texture features may be useful when the mass lesion is subtle on the prior view. The last difference between both experiments concerns the dataset. In the Netherlands the referral percentage is very low, about 1.0%. We believe that because of this low referral rate benign cases in our dataset are biased towards cases that show temporal changes, as these cases look more suspicious and are thus earlier referred. Benign cases that remain constant between two screening rounds are often not referred in the Netherlands. Therefore the dataset we use may have been more difficult to improve by adding features that capture temporal changes.

The classification performance for masses with normal priors decreased when using difference features. We can explain this because the CAD programme always follows the same procedure and does not distinguish between lesions that are visible and lesions that are not visible on the prior view. When the lesion is not visible it is not possible to determine an appropriate contour of the prior region and the segmentation programme will use accidental mammographic structures to determine a contour of the prior region. Consequently, features that depend on this contour will not be meaningful. The addition of these valueless features may result in a lower classification performance.

The second group of temporal features–the so-called similarity measures—determine whether the region on the current view and a corresponding region on the prior view are similar in appearance. These features improved the classification performance for masses with visible priors and for masses with normal priors. These features therefore seem more promising than difference features.

The first similarity feature is the registration score. This feature combines three registration measures including the correlation between the current region and a similar region on the prior. Results show that a very low registration score is more often seen for malignant masses than for benign masses and vice versa. The second similarity feature is the relative grey level change. Causes for an increase in relative grey level are twofold.

First, a mass that becomes more dense between two screening rounds will have a higher relative grey level on the current view than on the prior view. Second, an increase in size on average also results in a corresponding change in relative grey level as mass tissue on average is more dense than normal breast tissue. This feature is thus a measure of the change in contrast as well as the change in size. An advantage of this feature is that it is useful for masses with visible and for masses with normal priors because it does not depend on the contour of the prior region.

Difference features on the other hand are only useful for masses that are visible on both the prior and the current view. Therefore the relative grey level change is preferred above the temporal features that measure the difference in size or contrast.

Figures 7.2, 7.4, and 7.5 show some examples to illustrate potential benefits and drawbacks of including temporal change information into a CAD system. For these examples we compared the malignancy score from the single view classifier with the score from the temporal classifier. For the temporal classifier we used *temporal feature vector I* containing single view and similarity features. Figures 7.2(a) and 7.2(c) show examples where the temporal classifier performed better than the single view classifier. Figure 7.2(a) concerns a benign mass that is almost identical in appearance on the prior and the current view. Therefore the use of temporal features resulted in a better characterisation of the lesion. Figure 7.2(c) shows a malignant mass that is not visible on the prior view. The whole area on the prior view is "empty" resulting in a low registration score and a high grey level change. Consequently the temporal classifier assigned this lesion a higher malignancy score than the single view classifier.

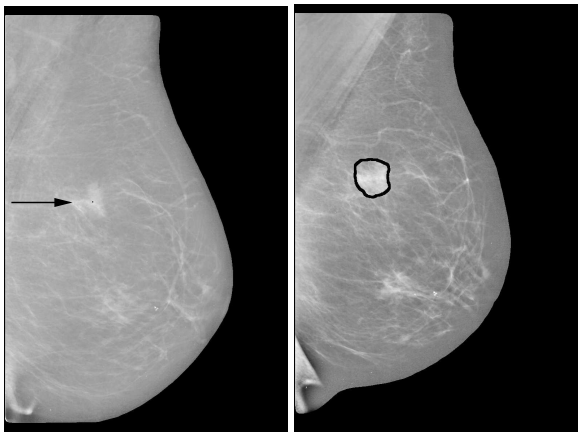


Figure 7.4: *Example where the temporal classifier performed worse than the single view classifier. The image shows a new benign mass that was not yet present on the previous screening mammogram.*

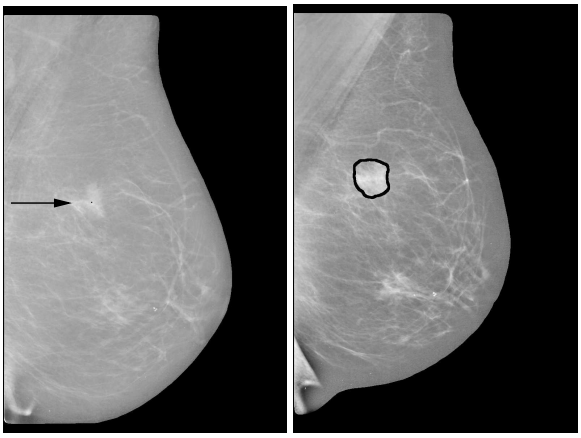


Figure 7.5: *Example where the temporal classifier performed worse than the single view classifier. The image shows a malignant mass that is similar in appearance on prior and current views.*

Figure 7.2(b) shows an example of a benign mass where both classifiers had equal performance. As the mass has a suspicious appearance both single view and temporal classifier assigned a high malignancy score to this lesion. The malignancy score did not change by adding temporal features. Figure 7.4 and 7.5 show two examples where the single view classifier performed better than the temporal classifier. Figure 7.4 shows a malignant mass that is also visible on the prior view and similar in appearance on both views. Therefore both temporal features were suggestive for a benign lesion resulting in a lower malignancy score. Figure 7.5 concerns a benign mass that is subtle on the prior and obvious on the current view. Therefore both temporal features suggested the presence of a malignant lesion resulting in a high malignancy score. These examples illustrate that benign and malignant masses sometimes show similar temporal behaviour. Temporal features should therefore always be used in conjunction with single view features.

In summary we designed a CAD system that includes temporal information for the characterisation of mass lesions. The classification performance significantly improved when adding temporal features compared to a single view CAD system. Results obtained in this study suggest that similarity features are more useful than difference features, both for masses that are visible on the prior view and for masses that are new.

Chapter 8

Effect of Temporal CAD on Radiologists' Performance [1]

In this chapter we investigate the use of a temporal CAD programme to help radiologist with the characterisation of mass lesions. For this purpose we set up an observer study with six radiologists. Each radiologist rated 198 cases, 99 containing a benign mass and 99 containing a malignant mass. For each case the mammograms from two consecutive screening rounds were available. The mass was visible on the prior view in 40% of the cases. Independently a CAD programme also rated each mass lesion making use of information from prior and current views. The following reading situations were compared: single reading, independent reading with CAD, and independent double reading.

8.1 Introduction

At this moment mammography is the method of choice for breast cancer screening. An important task of radiologists is to classify mammographic abnormalities as benign or malignant. The BI-RADS system provides criteria to classify each abnormality as benign, probably benign, suspicious, or highly suggestive of malignancy (D'Orsi & Kopans 1997). Despite these criteria interpretation and subsequent classification of abnormalities remains a difficult task. Studies show that misinterpretation is an important cause of missing breast cancer (Bird *et al.* 1992; Harvey *et al.* 1993). Furthermore, interpretation errors lead to unnecessary additional examinations. Only about 20%-50% of patients referred for biopsy are found to have a malignancy. The effect of an improvement in classification accuracy will thus be twofold. First, the cancer detection rate will increase.

[1] The content of this chapter has been published previously in Timp *et al.* (2006a).

Second, the positive predictive value of mammography will increase. Therefore we consider it important to investigate whether an automated CAD system can improve the characterisation accuracy of radiologists.

There have been some studies that evaluate the additional effect of CAD systems on radiologists' assessment of mass lesions. Most studies compare the performance of radiologists with and without using CAD. Chan *et al.* (1999) studied the effect of using a CAD system on radiologists' ability to characterise mass lesions and found that the performance significantly ($P < 0.001$) increased when using CAD. Huo *et al.* (2002) performed an observer experiment to evaluate the effect of a CAD system on the characterisation of benign and malignant masses using multiple views from the same examination. In their study the performance of the radiologists significantly ($P < 0.001$) improved when using computer aid. These CAD systems however only used the current view to characterise mass lesions.

Radiologists on the other hand routinely compare the current view with previous screening examinations when assessing mass lesions. Studies show that the performance of radiologists improves when using prior views (Varela *et al.* 2005; Hadjiiski *et al.* 2001b). In a recent study Hadjiiski *et al.* (2004) evaluated the effect of a CAD system that includes information from prior views. This study was restricted to mass lesions that were visible in retrospect. They found that the performance of radiologists significantly ($P = 0.005$) increased when using CAD. It should be remarked that in this study radiologists read digitised regions of interest extracted from temporal image pairs. This differs from the usual clinical setting where radiologists read a complete mammogram to estimate the malignancy of a lesion.

The purpose of this study is to evaluate the potential benefit of a temporal CAD system on radiologists' interpretation of mass lesions. In the previous chapter we evaluated the performance of this temporal CAD system and found that the characterisation accuracy significantly improved when using this temporal CAD system compared to a single view CAD system. Unlike other observer experiments with CAD this experiment more closely resembles clinical practice as 1) both radiologist and the CAD programme use prior views and 2) the dataset consists of masses that are visible in retrospect as well as masses that are new, and 3) radiologists read the mammograms as in the usual clinical setting on a dedicated mammography workstation. Furthermore we simulate double reading and compare this with independent reading with CAD.

The chapter is organised as follows. First we describe the dataset used for the experiments. In Subsection 8.2.2 we shortly summarise the single view and temporal CAD programme. Subsection 8.2.3 describes the observer study. Section 8.3 presents the results with a discussion in the last section.

8.2 Description CAD Programme and Observer Experiment

8.2.1 Dataset

The mammograms used in this study all came from the Dutch Breast Cancer Screening Programme. All women aged 50-75 are invited bi-annually to participate in this programme. Two mammographic projections—medio lateral oblique (MLO) and cranio caudal (CC)—are obtained at the initial screening in this programme. At subsequent screenings only medio lateral views are obtained, unless there is an indication that additional cranio caudal views would be beneficial.

At our institution we have a collection of consecutive cases with suspect abnormalities that were referred in the screening programme between 1996 and 2000. These cases contain mammograms at referral and mammograms of all previous screening rounds. Figure 8.1 shows an example of three consecutive mammograms. From the collection

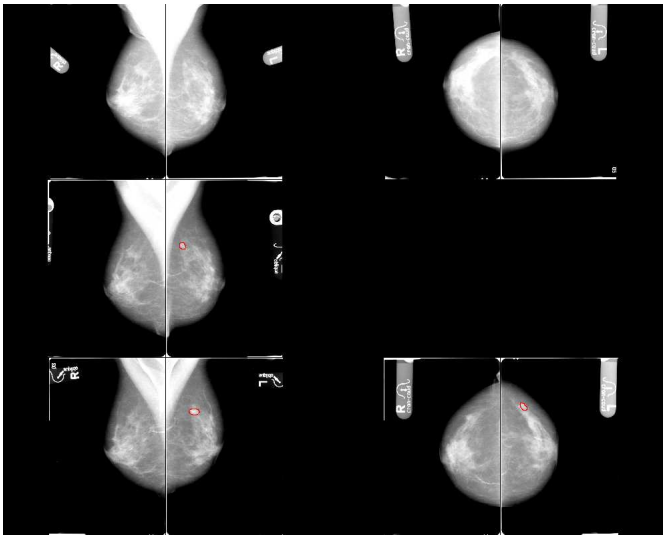


Figure 8.1: *Example of three consecutive mammograms for the same women. The last mammogram is the mammogram at time of referral. The other mammograms are obtained at previous screening rounds.*

of cases we composed temporal mammogram pairs that consist of all images from two consecutive screening rounds. We call the most recent mammogram the current mammogram and mammograms from earlier screening rounds previous or prior mammograms.

Characteristic	No.
Biopsied	99
Invasive ductal carcinoma	71
Invasive lobular carcinoma	18
Tubular carcinoma	3
Mucineus/colloid carcinoma	2
Intracystic carcinoma with invasion	1
Intracystic carcinoma without invasion	2
Ductal carcinoma in situ	2
Mammographic lesion size	99
<11 mm	14
11–20 mm	59
>20 mm	26
Lesion type	99
Mass	90
Architectural distortion	7
Asymmetry	2

Table 8.1: *Histopathologic and mammographic characteristics of malignant cases. Lesion size corresponds to the mammographic annotation made by the study radiologist.*

We then selected mammograms that fulfilled the following requirements: 1) the current mammogram contained both MLO and CC views, and 2) at least one view of the current mammogram contained a mass, asymmetry, or architectural distortion, all referred to as mass (lesion) in the sequel. Of all referral mammograms that met these criteria we randomly selected 171 cases, 87 malignant and 84 benign. In addition, we selected all cases in which the last mammogram before referral met these criteria. These were 27 cases, 12 malignant and 15 benign. Combining the two selections we obtained 198 cases, 99 malignant and 99 benign.

Table 8.1 and 8.2 show histopathologic and mammographic characteristics for benign and malignant cases. Fourteen cases contained micro-calcifications in addition to the mass, of which ten were malignant and four benign.

Because of the selection criteria, current mammograms always had MLO and CC views. Prior mammograms always had MLO views, while CC views were only available for 21.7% of the cases (43/198, 22 malignant and 21 benign). All 99 malignant cases

Characteristic	No.
Biopsied	39
Solitary cyst	17
Fibroadenoma	4
Fibrocystic change	2
Atypical ductal hyperplasia	5
Other benign lesion	8
Normal tissue	3
Mammographic lesion size	99
$<$11 mm	20
11–20 mm	54
$>$20 mm	25
Lesion type	99
Mass	90
Architectural distortion	7
Asymmetry	2

Table 8.2: *Histopathologic and mammographic characteristics of benign cases. Lesion size corresponds to the mammographic annotation made by the study radiologist.*

were biopsy proven. Of the benign cases 39 were histologically confirmed, while the remaining 60 cases had at least 6-month follow-up.

All visible abnormalities were annotated by or under supervision of an experienced radiologist, called the study radiologist, using all available information such as pathology results when a biopsy had been performed. In about 40% of the cases the lesion was also visible on previous screening mammograms. The images were digitised with a laser scanner (CFS300, Canon) at a pixel size of 50 μm and subsequently down sampled to a final resolution of 100 μm maintaining the original grey scale resolution of 12 bits.

8.2.2 Computer Aided Diagnosis System

Our CAD programme consists of a single view part and a temporal part. Both parts are described in detail in Chapter 7. Below we shortly summarise both parts.

Single View CAD Part The single view CAD programme starts with the segmentation of all current lesions that have been manually outlined by the study radiologist. The segmentation algorithm we use for this purpose requires a seed point in the centre of the mass. For details about the segmentation algorithm see Chapter 3. As seed point we use the mathematical centre of the contour delineated by the study radiologist. For each segmented region different single view features are calculated: spiculation measures, border features and morphological features. The total number of calculated features for each lesion is ten. As classifier we used a Support Vector Machine. Training and testing of the classifier were done completely independent using a 20 fold cross-validation scheme. For each view the classifier output represents the image based malignancy score, where small values correspond with benign ratings and large values with malignant ratings. The malignancy scores from both MLO and CC views are averaged to obtain a case based malignancy score that indicates the probability that the lesion is malignant. This score is called the *current view malignancy score* as it is based on features extracted from the current lesion.

Temporal CAD Part The temporal CAD part consists of two steps. In the first step the regional registration technique described in Chapter 5 finds for each lesion on the current view a corresponding location on the prior view. This location on the prior view is used as a seed point for our segmentation algorithm (see Chapter 3). This segmentation algorithm determines a contour for the prior region. Then several features are determined for the segmented region on the prior view. For each current region two temporal features are determined that indicate whether the current region and the corresponding region on the prior view are similar in appearance. These features are designed in a way such that they can be used for masses that are visible on the prior view and for masses that are new. In total we calculate twelve features for each region. As classifier we use a Support Vector Machine. Training and testing of the classifier are implemented as described above for the single view classifier. The temporal classifier uses single view and temporal features to determine for each lesion an image based malignancy score. The malignancy scores from available CC and MLO projections are averaged to obtain for each lesion a case based malignancy score that indicates the probability that a lesion is malignant. This score is called the *temporal malignancy score*.

8.2.3 Observer Study

A panel of six radiologists, not including the study radiologist, participated in the observer study. Each of the six radiologists rated 99 benign and 99 malignant cases with and without the use of prior views. The reading sessions were structured as follows. First only the current mammogram was shown. By pressing a key the radiologist could see the contour drawn by the study radiologist. The radiologist then rated each lesion on a scale

between 0 and 100, where a value near 100 indicates a high likelihood of malignancy. Because of our study design all current mammograms had MLO and CC projections. Each radiologist therefore had access to information from both CC and MLO views. After the radiologist had recorded his score the prior mammogram was displayed, and the radiologist could change his rating accordingly. In this study we only used the second rating that the radiologists gave when both prior and current views were available. All cases were presented in a randomised order.

The mammograms were displayed on this dedicated mammography workstation (Mevis BreastCare, MBC-SCR1, Bremen, Germany). In a recent study Roelofs *et al.* (2005) found that radiologists perform equally well reading digitised mammograms on a dedicated workstation as reading the original films. The workstation was equipped with two high-resolution CRT monitors (BARCO, MGD 521, 300 Cd/m^2, using BarcoMed 5MP1H 12 bit graphics boards), and a dedicated key-pad to access the main functions with a single keystroke. The CRT displays had a spatial resolution of 2,048 x 2,560 pixels each, which is sufficient to display one image at 100 μm. Initially, images were displayed at low spatial resolution (200 μm), in such a way that all images included in a case could be displayed simultaneously. Current MLO and CC views appeared in the lower half of the left and right monitor. Prior views were displayed in the same way in the upper half of both monitors. Full spatial resolution (100 μm) images could subsequently be displayed by pressing a key of the dedicated key-pad. Images from the same breast but different views could also be displayed on both monitors at the same time. The same was possible for images of the same breast and same view but different screening rounds. This made it possible to analyse temporal changes in a simple and user-friendly way. Images were preprocessed using an unsharp-masking technique (Roelofs *et al.* 2003) to compensate for the decrease of sharpness with respect to the original films due to digitisation and electronic display.

Five of the six readers in this study were attending radiologists with breast cancer screening experience. The other radiologist was a radiology resident in her last year who had specialised in mammography. All participants received a training session before the observer study started to become familiar with the soft-copy reading system and the design of the experiment. The training set consisted of 25 cases. The true diagnosis was given immediately after each training case. During training, radiologists were encouraged to make use of the different tools provided by the soft-copy reading system. General information about the data set was provided to the radiologists. They were informed that the number of benign and malignant cases was about equal, and that all cases were referrals from a screening programme.

8.2.4 Data Analysis

The performance of each radiologist was evaluated for the three different reading modes using ROC methodology. The classification accuracy was quantified with the area under the ROC curve (A_z value).

The A_z values for single reading and independent reading with CAD were estimated by using the Dorfman-Berbaum-Metz (DBM) method for analysis of multi-reader multi-case data (Dorfman *et al.* 1992). In this method the maximum likelihood estimation of the binormal distributions is fitted to the observer ratings to obtain an ROC curve. This method has been widely adopted in recent years for analysing experimental data obtained in a multi-reader multi-case (MRMC) study design (Hadjiiski *et al.* 2004; Beiden *et al.* 2002). It has the great advantage that both reader and case variability are taken into account in a proper way, such that generalisation to both the population of readers and cases is permitted. We used the publicly available LABMRMC software (Metz *et al.* 1998a) for MRMC computations. To estimate the potential benefit of using the CAD system we independently combined the malignancy ratings of each radiologist and the CAD system. We first linearly scaled the ratings of each radiologist and the CAD system between zero and one hundred. We then assigned each lesion a combined rating computed as the arithmetic mean of the CAD malignancy rating and a radiologists malignancy rating. The multi-reader multi-case DBM method analysed the average scores to estimate A_z values for independent reading with CAD. The statistical significance of the difference in A_z between reading without CAD and independent reading with CAD was also estimated by using the DBM method (Metz *et al.* 1998a).

To simulate double reading we combined for each lesion the malignancy ratings of two radiologists. For each radiologist this resulted in five different double reading results. For each double reading result we calculated an ROC curve. This resulted in five different ROC curves for each radiologist. Each ROC curve is completely described by two parameters that characterise the underlying normal distributions. For each radiologist we used these parameters to determine an average ROC curve and the corresponding A_z value (Obuchowski 2005). The Student's t-test for paired data was used to assess the significance of differences between A_z values of the double reading mode on one side and the single reading mode and reading with CAD on the other side.

8.3 Results Reading with CAD and Double Reading

The performance, measured as area under the ROC curve, was calculated for each radiologist for the different reading modes: single reading, independent reading with CAD, and independent double reading. Table 8.3 lists the individual and the mean performances of the radiologists for the three reading modes. The A_z value of the stand alone CAD programme was 0.81. The average A_z value for the radiologists was 0.80 for single reading.

For all radiologists the performance improved for independent reading with CAD. The average A_z value significantly increased to 0.83 for reading with CAD ($P < 0.05$, DBM method; $P = 0.02$, Student's paired t-test). Table 8.3 shows that the performance of radiologists with a better performance than the CAD system (number two and number five) improved as much as the performance of radiologists with a lower performance than the CAD system.

The average performance for independent double reading was 0.81. The difference between single reading and independent double reading was not significant ($P = 0.12$, Student's paired t-test). For independent double reading the performance increased for all radiologists except for the best performing radiologist (number two). The least performing radiologist, number three, benefited most from independent double reading. These results suggest that the benefit that can be obtained with double reading depends on the performance of the individual radiologist. The difference between reading with CAD and independent double reading was not significant ($P = 0.08$, Student's paired t-test).

Figure 8.2 shows average ROC curves for the different reading modes. For each reading mode the average ROC curve is obtained by averaging the fitted parameters of the individual ROC curves of each radiologist (Obuchowski 2005).

Radiologist	Single Reading	Ind. Reading with CAD	Ind. Double Reading
1	0.792 $\pm$0.034	0.805$\pm$0.031	0.800 $\pm$0.023
2	0.825 $\pm$0.029	0.835$\pm$0.028	0.820 $\pm$0.016
3	0.749 $\pm$0.034	0.809$\pm$0.030	0.798 $\pm$0.025
4	0.801 $\pm$0.031	0.839$\pm$0.028	0.809 $\pm$0.024
5	0.813 $\pm$0.031	0.855$\pm$0.026	0.825 $\pm$0.021
6	0.790 $\pm$0.032	0.829$\pm$0.029	0.814 $\pm$0.019
average A_z	0.796	0.829	0.811

Table 8.3: *$A_z \pm$standard deviation for each radiologist for the three different reading conditions: single reading, independent reading with CAD and independent double reading.*

8.4 Discussion

In this study we investigated the effect of three different reading modes on the characterisation of mass lesions on serial mammograms: single reading, independent reading with CAD, and independent double reading. We implemented independent reading with

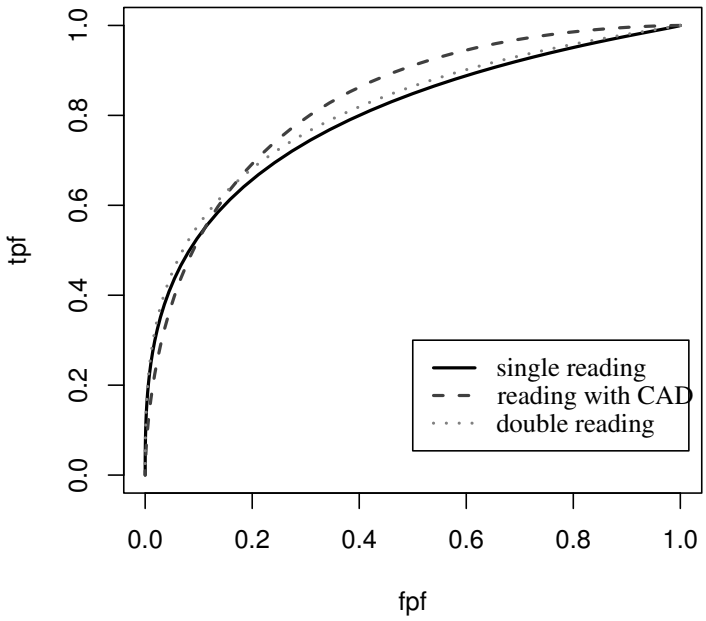


Figure 8.2: *ROC curves for the three reading modes: single reading, independent double reading and independent reading with CAD.*

CAD by averaging the score of each radiologist with the score from the CAD system for all mass lesions. Using the CAD system in this way improved the classification performance for each radiologist, also for the best performing one. The average A_z value significantly increased from 0.80 to 0.83 when CAD was used. From Figure 8.2 we see that this improvement mainly concerns a better characterisation of benign lesions. Using CAD in this way may thus lead to a decrease in the number of false positives at the same sensitivity level. Furthermore, we found that for each radiologist the performance of independent reading with CAD was equal or higher than the performance of the standalone CAD system, which was 0.81.

Some studies have been done to evaluate the effect of using a CAD system on the ability of radiologists to discriminate between benign and malignant lesions (Chan *et al.* 1999; Huo *et al.* 2002; Hadjiiski *et al.* 2004). These studies all show an improvement in characterisation accuracy when a CAD system is used. An important difference between

this study and existing studies is that the present CAD system also uses information extracted from prior views. To our knowledge there is only one other study where the CAD system used temporal information as well (Hadjiiski *et al.* 2004). In that study they found that using a CAD system with temporal analysis improved the classification performance of radiologists. In the current study we tried to resolve some of the limitations of (Hadjiiski *et al.* 2004). First, in (Hadjiiski *et al.* 2004) the observers read ROI's extracted from temporal image pairs. In the current study observers read whole mammograms, including views from the left and right breast and from different projections. Second, in (Hadjiiski *et al.* 2004) the performance of the standalone CAD system was better than the performance of radiologists using CAD. An explanation may be that the reading conditions for the radiologists were not optimal, resulting in a significantly lower performance for each individual radiologist than for the CAD system. In the current study we found that the performance of each individual radiologist using CAD was higher than the performance of the standalone CAD system. This might be caused by the fact that the performance of radiologists and the CAD system were comparable. It should also be noted that in our study the ratings from each radiologist and the CAD system were independently combined while in (Hadjiiski *et al.* 2004) radiologists used CAD to adjust their own assessment. Last, the CAD system used in (Hadjiiski *et al.* 2004) was restricted to mass lesions that were visible in retrospect. The CAD system used in our study is suited for masses that are visible on retrospect as well as for masses that are new.

A CAD system for mass characterisation may also be helpful for screening purposes. During screening two types of perception errors are made: search errors and interpretation errors. Search errors are defined as lesions that are overlooked or only briefly fixated. Interpretation errors conern lesions that are missed due to wrong decisions. Some studies suggest that the majority of errors in radiological detection tasks may be due to incorrect interpretation of lesions (Karssemeijer *et al.* 2003; Manning *et al.* 2004). The use of CAD systems to characterise mass lesions may result in less interpretation errors. Ultrasound also has an important role in characterising lesions as benign or malignant. In many screening programmes however mammograms are read in afterwards in batches such that ultrasound can not be used. Radiologists then base their decision to refer a woman on mammography results only. Using CAD in these situations may lead to a decrease in the number of false positive referrals while maintaining the same cancer detection rate.

Additionally, in this study we compared the performance of independent reading with CAD with the performance of independent double reading. To implement independent double reading we combined the malignancy ratings from each pair of radiologists. We found that double reading improves the classification performance, although this difference was not significant. The statistical difference in performance between independent reading with CAD and independent double reading was also not significant. This suggests that the CAD system may be used as an independent additional reader. For instance, a single reader with CAD might be used to select cases with suspicious abnormalities for

further inspection by a second reader, who then makes the final decision whether referral is necesary. This approach resembles double reading with arbitration, which is common in screening in the United Kingdom, where a third reader assesses those mammograms for which two screening radiologists do not reach consensus (Smith-Bindman *et al.* 2003).

In summary we find that using a CAD system with temporal analysis can help radiologists to interpret mass lesions. Further studies are needed to investigate the performance of CAD systems in clinical settings.

Bibliography

ALEXANDER, F.E., T.J. ANDERSON, H.K. BROWN, A.P. FORREST, W. HEPBURN, A.E. KIRKPATRICK, B.B. MUIR, R.J. PRESCOTT, & A. SMITH. 1999. 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. *Lancet* 353(9168):1903–1908.

ANDERSSON, I., K. ASPEGREN, L. JANZON, T. LANDBERG, K. LINDHOLM, F. LINELL, O. LJUNGBERG, J. RANSTAM, & B. SIGFÚSSON. 1988. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. *BMJ* 297(6654):943–948.

——, & L. JANZON. 1997. Reduced breast cancer mortality in women under age 50: updated results from the Malmö Mammographic Screening Program. *J Natl Cancer Inst Monogr* (22):63–67.

BALLARD, D.H., & C.M. BROWN. 1982. *Computer Vision*, chapter 4. Prentice-Hall.

BASSETT, L.W., B. SHAYESTEHFAR, & I. HIRBAWI. 1994. Obtaining previous mammograms for comparison: usefulness and costs. *AJR* 163(5):1083–1086.

BEIDEN, S.V., R.F. WAGNER, K. DOI, R.M. NISHIKAWA, M. FREEDMAN, S.C. LO, & X.W. XU. 2002. Independent versus sequential reading in ROC studies of computer-assist modalities: analysis of components of variance. *Acad Radiol* 9(9):1036–1043.

BIRD, R.E., T.W. WALLACE, & B.C. YANKASKAS. 1992. Analysis of cancers missed at screening mammography. *Radiology* 184(3):613–617.

BIRDWELL, R.L., D.M. IKEDA, K.F. O'SHAUGHNESSY, & E.A. SICKLES. 2001. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology* 219(1):192–202.

BISHOP, C.M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

BJURSTAM, N., L. BJÖRNELD, S.W. DUFFY, T.C. SMITH, E. CAHLIN, O. ERIKSON, H. LINGAAS, J. MATTSSON, S. PERSSON, C.M. RUDENSTAM, & J. SÄWE-SÖDERBERG. 1997. The Gothenburg Breast Cancer Screening Trial: preliminary results on breast cancer mortality for women aged 39–49. *J Natl Cancer Inst Monogr* (22):53–55.

BOVIS, K., S. SINGH, J. FIELDSEND, & C. PINDER. 2000. Identification of masses in digital mammograms with MLP and RBF nets. In *Proceedings of the IEEE-INNS-ENN*, volume 1, pages 342–347.

BREM, R.F., J. BAUM, & M. LECHNER ET AL. 2003. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. *AJR Am J Roentgenol* 181(3):687–693.

CALLAWAY, M.P., C.R.M. BOGGIS, S.A. ASTLEY, & I. HUTT. 1997. The influence of previous films on screening mammographic interpretation and detection of breast carcinoma. *Clin Radiol* 52(7):527–529.

CAULKIN, S., S. ASTLEY, J. ASQUITH, & C. BOGGIS. 1998. Sites of occurrence of malignancies in mammograms. In *4th International Workshop on Digital Mammography, Nijmegen, the Netherlands*, ed. by N. Karssemeijer, M.A.O. Thijssen, J.H.C.L. Hendriks, & L.J.T.O. van Erning, pages 279–282. Kluwer, Dordrecht.

CHAN, H-P., B. SAHINER, M.A. HELVIE, N. PETRICK, M.A. ROUBIDOUX, T.E. WILSON, D.D. ADLER, C. PARAMAGUL, J.S. NEWMAN, & S. SANJAY-GOPAL. 1999. Improvement of radiologists's characterization of mammographic masses by using computer-aided diagnosis: An ROC study. *Radiology* 212(3):817–827.

CHU, K.C., C.R. SMART, & R.E. TARONE. 1988. Analysis of breast cancer mortality and stage distribution by age for the Health Insurance Plan clinical trial. *J Natl Cancer Inst* 80(14):1125–1132.

CLARIDGE, E., & J.H. RICHTER. 1994. Characterisation of mammographic lesions. In *2th International Workshop on Digital Mammography, York, UK*, ed. by A.G. Gale, S.M. Astley, D.R. Dance, & A.Y. Cairns, pages 241–250. Elsevier, Amsterdam.

CRISTIANINI, N., & J. SHAWE-TAYLOR. 2000. *An introduction to Support Vector Machines and other kernel based learning methods*. Cambridge University Press.

CUPPLES, T.E., J.E. CUNNINGHAM, & J.C. REYNOLDS. 2005. Impact of computer-aided detection in a regional screening mammography program. *AJR Am J Roentgenol* 185(4):944–950.

DORFMAN, D.D., K.S. BERBAUM, & C.E. METZ. 1992. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Invest Radiol* 27(9):723–731.

D'ORSI, C.J., & D.B. KOPANS. 1997. Mammography interpretation: the BI-RADS method. *American Family Physician* 55(5):1548–1552.

DUDA, R.O., R.E. HART, & D.G. STORK. 2001. *Pattern Classification*. John Wiley & Sons.

Dutch Cancer Registry, 2003. Internet. *www.ikcnet.nl*.

FILEV, P., L. HADJIISKI, B. SAHINER, H-P. CHAN, & M.A. HELVIE. 2005. Comparison of similarity measures for the task of template matching of masses on serial mammograms. *Med Phys* 32(2):515–529.

FRACHEBOUD, J., H.J. DE KONING, P.M. BEEMSTERBOER, R. BOER, J.H. HENDRIKS, A.L. VERBEEK, B.M. VAN INEVELD, A.E. DE BRUYN, & P.J. VAN DER MAAS. 1998. Nation-wide breast cancer screening in The Netherlands: results of initial and subsequent screening 1990-1995. National Evaluation Team for Breast Cancer Screening. *Int J Cancer* 75(5):694–698.

FREER, T., & M. ULISSEY. 2001. Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. *Radiology* 220(3):781–786.

FRIEDRICH, M., & E.A. SICKLES (eds.) 2000. *Radiological diagnosis of breast diseases*. Springer.

FRISELL, J., E. LIDBRINK, L. HELLSTRÖM, & L.E. RUTQVIST. 1997. Followup after 11 years–update of mortality results in the Stockholm mammographic screeningtrial. *Breast Cancer Res Treat* 45(3):263–270.

FUKUNAGA, K. 1990. *Introduction to statistical pattern recognition*. Academic Press.

GOERGEN, S.K., J.E. EVANS, G.P.B. COHEN, & J.H. MACMILLAN. 1997. Characteristics of breast carcinomas missed by screening radiologists. *Radiology* 204:131–135.

GOOD, W.F., B. ZHENG, Y-H. CHANG, X.H. WANG, MAITZ G.S., & D. GUR. 1999. Multi-image CAD employing features derived from ipsilateral mammographic views. In *Proc SPIE Medical Imaging*, volume 3661, pages 474–485.

GOTZSCHE, P.C., & O. OLSEN. 2000. Is screening for breast cancer with mammography justifiable? *Lancet* 355(9198):129–134.

GULIATO, D., R.M. RANGAYYAN, W.A. CARNIELLI J.A. ZUFFO, & J.E.L. DESAUTELS. 1998. Segmentation of breast tumors in mammograms by fuzzy region growing. In *Proc. 20th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*, pages 1002–1004.

GUR, D., J.H. SUMKIN, H.E. ROCKETTE, M. GANOTT, C. HAKIM, L. HARDESTY, W.R. POLLER, R. SHAH, & L. WALLACE. 2004. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J Natl Cancer Inst* 96(3):185–190.

HADJIISKI, L., H-P. CHAN, & B. SAHINER ET AL. 2001a. Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis – local affine transformation for improved localization. *Med Phys* 28(6):1070–1079.

——, H-P. CHAN, B. SAHINER, M.A. HELVIE, M.A. ROUBIDOUX, C. BLANE, C. PARAMAGUL, N. PETRICK, J. BAILEY, K. KLEIN, M. FOSTER, S. PATTERSON, D. ADLER, A. NEES, & J. SHEN. 2004. Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: An ROC study. *Radiology* 233(1):255–265.

——, B. SAHINER, H-P. CHAN, N. PETRICK, M.A. HELVIE, & M. GURCAN. 2001b. Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign masses. *Med Phys* 28(11):2309–2317.

HARVEY, J.E., L.L. FAJARDO, & C.A. INIS. 1993. Previous mammograms in patients with impalpable breast carcinoma: retrospective vs blinded interpretation. *AJR* 161:1167–1172.

HEINE, J.J., & P. MALHOTRA. 2002. Mammographic tissue, breast cancer risk, serial image analysis, and digital mammography. Part 2. Serial breast tissue change and related temporal influences. *Acad Radiol* 9(3):317–335.

HELVIE, M.A., L. HADJIISKI, E. MAKARIOU, H.P. CHAN, N. PETRICK, B. SAHINER, S.C. LO, M. FREEDMAN, D. ADLER, J. BAILEY, C. BLANE, D. HOFF, K. HUNT, L. JOYNT, K. KLEIN, C. PARAMAGUL, S.K. PATTERSON, & M.A. ROUBIDOUX. 2004. Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: pilot clinical trial. *Radiology* 231(1):208–214.

HOMER, M.J. 1997. *Mammographic interpretation*. The McGraw-Hill Companies, Inc.

HORNIK. 2005. The R FAQ. ISBN: 3-900051-08-9. Available online: http://CRAN.R-project.org/doc/FAQ.

HUO, Z., M.L. GIGER, C.J. VYBORNY, & C.E. METZ. 2002. Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms. *Radiology* 224(2):560–568.

JACKSON, V.P. 2002. Screening mammography: controversies and headlines. *Radiology* 225(2):323–326.

KARSSEMEIJER, N. 1998. Automated classification of parenchymal patterns in mammograms. *Phys Med Biol* 43(2):365–378.

——, J.D. OTTEN, A.L. VERBEEK, J.H. GROENEWOUD, H.J. DE KONING, J.H. HENDRIKS, & R. HOLLAND. 2003. Computer-aided detection versus independent double reading of masses on mammograms. *Radiology* 227(1):192–200.

——, & G.M. TE BRAKE. 1996. Detection of stellate distortions in mammograms. *IEEE Trans Med Imaging* 15:611–619.

——, & ——. 1998. Combining single view features and asymmetry for detection of mass lesions. In *4th International Workshop on Digital Mammography, Nijmegen, the Netherlands*, ed. by N. Karssemeijer, M.A.O. Thijssen, J.H.C.L. Hendriks, & L.J.T.O. van Erning, pages 95–102. Kluwer, Dordrecht.

KITTLER, J., M. HATEF, R.P.W. DUIN, & J. MATAS. 1998. On combining classifiers. *Trans Pattern Anal Machine Intell* 20(3):226–239.

KOK-WILES, S., M. BRADY, & R. HIGHNAM. 1998. Comparing mammogram pairs for the detection of lesions. In *4th International Workshop on Digital Mammography, Nijmegen, the Netherlands*, ed. by N. Karssemeijer, M.A.O. Thijssen, J.H.C.L. Hendriks, & L.J.T.O. van Erning, pages 103–110. Kluwer, Dordrecht.

KUPINSKI, M.A., & M.L. GIGER. 1998. Automated seeded lesion segmentation on digital mammograms. *IEEE Trans Med Imaging* 17(4):510–517.

LAU, T.K., & W.F. BISCHOF. 1991. Automated detection of breast tumors using the asymmetry approach. *Comp and Biomed Research* 24:273–295.

LEVI, F., F. LUCCHINI, E. NEGRI, & C. LA VECCHIA. 2004. Trends in mortality from major cancers in the European Union, including acceding countries, in 2004. *Cancer* 101(12):2843–2850.

LEWIN, J.M., R.E. HENDRICK, & C.J. D'ORSI ET AL. 2001. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. *Radiology* 218(3):873–880.

LOBREGT, S., & M. VIERGEVER. 1995. A discrete dynamic contour model. *IEEE Trans Med Imaging* 14(1):12–24.

MANNING, D.J., S.C. ETHELL, & T. DONOVAN. 2004. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *Br J Radiol* 77(915):231–235.

MCINERNEY, T., & D. TERZOPOULOS. 1996. Deformable models in medical image analysis: a survey. *Med Image Anal* 1(2):91–108.

METZ, C.E. 1986. ROC methodology in radiographic imaging. *Invest Radiol* 21(9):720–733.

——, B.A. HERMAN, & C.A. ROE. 1998a. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Med Decis Making* 18(1):110–121. Available online: http://www-radiology.uchicago.edu/krl/rocstudy.htm.

——, B.A. HERMAN, & J.H. SHEN. 1998b. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat Med* 17(9):1033–1053. Available online: http://www-radiology.uchicago.edu/krl/rocstudy.htm.

MILLER, A.B., C.J. BAINES, T. TO, & C. WALL. 1992a. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *CMAJ* 147(10):1459–1476.

——, ——, ——, & ——. 1992b. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *CMAJ* 147(10):1477–1488.

MILLER, P., & S. ASTLEY. 1993. Detection of asymmetry using anatomical features. *SPIE Medical Imaging* 1905:433–442.

OBUCHOWSKI, N.A. 2005. ROC analysis. *AJR Am J Roentgenol* 184(2):364–372.

OLSEN, O., & P.C. GOTZSCHE. 2001. Cochrane review on screening for breast cancer with mammography. *Lancet* 358(9290):1340–1342.

OREL, S.G., N. KAY, C. REYNOLDS, & D.C. SULLIVAN. 1999. BI-RADS categorization as a predictor of malignancy. *Radiology* 211(3):845–850.

OTTEN, J.D., N. KARSSEMEIJER, J.H. HENDRIKS, J.H. GROENEWOUD, J. FRACHEBOUD, A.L. VERBEEK, H.J. DE KONING, & R. HOLLAND. 2005. Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. *J Natl Cancer Inst* 97(10):748–754.

OTTO, S.J., J. FRACHEBOUD, C.W. LOOMAN, M.J. BROEDERS, R. BOER, J.H. HENDRIKS, A.L. VERBEEK, & H.J. DE KONING. 2003. Initiation of population-based mammography screening in Dutch municipalities and effect on breast-cancer mortality: a systematic review. *Lancet* 361(9367):1411–1417.

PAQUERAULT, S., N. PETRICK, H-P. CHAN, B. SAHINER, & M.A. HELVIE. 2002. Improvement of computerized mass detection on mammograms: fusion of two-view information. *Med Phys* 29(2):238–247.

PETRICK, N., H-P. CHAN, B. SAHINER, & M. HELVIE. 1999. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. *Med Phys* 26(8):1642–1654.

PICCOLI, C., S. FEIG, & J. PALAZZO. 1999. Developing asymmetric breast tissue. *Radiology* 211(1):111–117.

PISANO, E.D., C. GATSONIS, E. HENDRICK, M. YAFFE, J.K. BAUM, S. ACHARYYA, E.F. CONANT, L.L. FAJARDO, L. BASSETT, C. D'ORSI, R. JONG, & M. REBNER. 2005. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 353(17):1773–1783.

PUDIL, P., J. NOVOVICOVA, & J. KITTLER. 1994. Floating search methods in feature selection. *Pattern Recognition Letters* 15(11):1119–1125.

RAHBAR, G., A.C. SIE, G.C. HANSEN, J.S. PRINCE, M.L. MELANY, H.E. REYNOLDS, V.P. JACKSON, J.W. SAYRE, & L.W. BASSETT. 1999. Benign versus malignant solid breast masses: US differentiation. *Radiology* 213(3):889–894.

RICHARD, F., & C. GRAFFIGNE. 2000. An image matching model for the registration of time sequence or bilateral mammogram pairs. In *5th International Workshop on Digital Mammography, Toronto, Canada*, ed. by M.J. Yaffe. Med Phys Publishing, Madison.

RIPLEY, B.D. 1996. *Pattern recognition and neural networks*. Cambridge University Press.

ROELOFS, A.A., S. VAN WOUDENBERG, J.D. OTTEN, J.H. HENDRIKS, A. BODICKER, C.J. EVERTSZ, & N. KARSSEMEIJER. 2005. Effect of soft-copy display supported by cad on mammography screening performance. *In press: Eur Radiol* .

ROELOFS, T., S. VAN WOUDENBERG, J. HENDRIKS, & N. KARSSEMEIJER. 2003. Optimized soft-copy display of digitized mammograms. In *Proc. SPIE, Image Perception and Performance*, volume 5043, pages 10–19.

——, S. VAN WOUDENBERG, & N. KARSSEMIJER. 2006. Importance of comparison of current and prior mammograms in breast cancer screening. *In press: Radiology* .

SAHINER, B., N. PETRICK, H.P. CHAN, L.M. HADJIISKI, C. PARAMAGUL, M.A. HELVIE, & M.N. GURCAN. 2001. Computer-aided characterization of mammographic

masses: accuracy of mass segmentation and its effects on characterization. *IEEE Trans Med Imaging* 20(12):1275–1284.

SALLAM, M., & K.W. BOWYER. 1994. Registering time-sequences of mammograms using a two-dimensional unwarping technique. In *2th International Workshop on Digital Mammography, York, USA*, ed. by A.G. Gale, S.M. Astley, D.R. Dance, & A.Y. Cairns, pages 121–131. Elsevier, Amsterdam.

SANJAY-GOPAL, S., H-P. CHAN, T. WILSON, M. HELVIE, & N. PETRICK. 1999. A regional registration technique for automated interval change analysis of breast lesions on mammograms. *Med Phys* 26(12):2669–2679.

SHAPIRO, S., W. VENET, P. STRAX, L. VENET, & R. ROESER. 1982. Ten-to fourteen-year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 69(2):349–355.

SICKLES, E.A., W.N. WEBER, H.B. GALVIN, S.H. OMINSKY, & R.A. SOLLITTO. 1986. Baseline screening mammography: one vs two views per breast. *AJR Am J Roentgenol.* 147(6):1149–1153.

SMITH-BINDMAN, R., P.W. CHU, D.L. MIGLIORETTI, E.A. SICKLES, R. BLANKS, R. BALLARD-BARBASH, J.K. BOBO, N.C. LEE, M.G. WALLIS, J. PATNICK, & K. KERLIKOWSKE. 2003. Comparison of screening mammography in the united states and the united kingdom. *JAMA* 290(16):2129–2137.

SOLTESZ, D. LEE. 2006. Encircled breasts. Available online: http://www.deborah.ws.

STOUTJESDIJK, M.J., C. BOETES, G.J. JAGER, L. BEEX, P. BULT, J.H. HENDRIKS, R.J. LAHEIJ, L. MASSUGER, L.E. VAN DIE, T. WOBBES, & J.O. BARENTSZ. 2001. Magnetic resonance imaging and mammography in women with a hereditary risk of breast cancer. *J Natl Cancer Inst* 93(14):1095–1102.

TABÁR, L., C.J. FAGERBERG, A. GAD, L. BALDETORP, L.H. HOLMBERG, O. GRÖNTOFT, U. LJUNGQUIST, B. LUNDSTRÖM, J.C. MÅNSON, & G. EKLUND. 1985. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet* 1(8433):829–832.

——, G. FAGERBERG, H.H. CHEN, S.W. DUFFY, C.R. SMART, A. GAD, & R.A. SMITH. 1995. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. *Cancer* 75(10):2507–2517.

TE BRAKE, G.M., & N. KARSSEMEIJER. 1999. Single and multiscale detection of masses in digital mammograms. *IEEE Trans Med Imaging* 18(7):628–639.

——, ——, & J.H. HENDRIKS. 2000. An automatic method to discriminate malignant masses from normal tissue in digital mammograms. *Phys Med Biol* 45(10):2843–2857.

——, M.J. STOUTJESDIJK, & N. KARSSEMEIJER. 1999. A discrete dynamic countour model for mass segmentation in digital mammograms. In *Proc SPIE Medical Imaging*, volume 3661, pages 911–919.

THURFJELL, M.G., B. VITAK, E. AZAVEDO, G. SVANE, & E. THURFJELL. 2000. Effect on sensitivity and specificity of mammography screening with or without comparison of old mammograms. *Acta Radiol* 41(1):52–56.

TIMP, S., & N. KARSSEMEIJER. 2004a. A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography. *Med Phys* 31(5):958–971.

——, & ——. 2004b. Use of temporal features to improve mass detection. In *7th International Workshop on Digital Mammography, Chapel Hill*.

——, & ——. 2006. Interval change analysis to improve computer aided detection in mammography. *Med Image Anal* 10(1):82–95.

——, ——, & J.H.C.L. HENDRIKS. 2002a. Analysis of changes in masses using contrast and size measures. In *6th International Workshop on Digital Mammography, Bremen, Germany*, ed. by H. Peitgen, pages 240–242. Springer-Verlag.

——, ——, & J.H.C.L. HENDRIKS. 2002b. Comparison of three different mass segmentation methods. In *6th International Workshop on Digital Mammography, Bremen, Germany*, ed. by H. Peitgen, pages 218–222. Springer-Verlag.

——, S. VAN ENGELAND, & N. KARSSEMEIJER. 2005. A regional registration method to find corresponding mass lesions in temporal mammogram pairs. *Med Phys* 32(8):2629–2638.

——, C. VARELA, & N. KARSSEMEIJER. 2006a. Computer-aided diagnosis with temporal analysis to improve radiologists' interpretation of mammographic mass lesions. *Submitted to Eur Rad* .

——, ——, & ——. 2006b. Temporal change analysis for characterisation of mass lesions in mammography. *Submitted to IEEE Trans Med Imaging* .

UNDERWOOD, J.C.E. (ed.) 1992. *General and systematic pathology*, chapter 16. Churchill Livingstone.

VAINIO, H., & F. BIANCHINI (eds.) 2002. *IACR handbooks of cancer prevention*. Lyon: IARCPress.

VAN ENGELAND, S., N. KARSSEMEIJER, & J. HENDRIKS. 2002. Using information from two mammographic views to improve computer-aided detection of mass lesions. In *6th International Workshop on Digital Mammography, Bremen, Germany*, ed. by H. Peitgen, pages 377–381. Springer-Verlag.

——, P. SNOEREN, N. KARSSEMEIJER, & J. HENDRIKS. 2003. A comparison of methods for mammogram registration. *IEEE Trans Med Imaging* 22(11):1436–1444.

——, S. TIMP, & N. KARSSEMEIJER. 2006. Finding corresponding regions of interest in mediolateral oblique and cranio caudal mammographic views. *Submitted to Med Phys* .

VARELA, C., N. KARSSEMEIJER, J.H. HENDRIKS, & R. HOLLAND. 2005. Use of prior mammograms in the classification of benign and malignant masses. *Eur J Radiol* 56(2):248–255.

——, S. TIMP, & N. KARSSEMEIJER. 2006. Use of border information in the classification of mammographic masses. *Phys Med Biol* 51(2):425–441.

VUJOVIC, N., & D. BRZAKOVIC. 1997. Establishing the correspondence between control points in pairs of mammographic images. *IEEE Trans Med Imaging* 6(10):1388–1399.

——, ——, & K. FOGARTY. 1995. Detection of cancerous changes in mammograms using intensity and texture measures. *Proc SPIE 2434* pages 37–47.

WALD, N.J., P. MURPHY, P. MAJOR, C. PARKES, J. TOWNSEND, & C. FROST. 1995. UKCCCR multicentre randomised controlled trial of one and two view mammography in breast cancer screening. *BMJ* 311(7014):1189–1193.

WARREN BURHENNE, L.J., S.A. WOOD, C.J. D'ORSI, S.A. FEIG, D.B. KOPANS, K.F. O'SHAUGHNESSY, E.A. SICKLES, L. TABÁR, C.J. VYBORNY, & R.A. CASTELLINO. 2000. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 215(2):554–562.

WHITE, K., K. BERBAUM, & W.L. SMITH. 1994. The role of previous radiographs and reports in the interpretation of current radiographs. *Invest Radiol* 29(3):263–265.

YIN, F.F., M.L. GIGER, K. DOI, C.E. METZ, C.J. VYBORNY, & R.A. SCHMIDT. 1991. Computerized detection of masses in digital mammograms : Analysis of bilateral substraction images. *Med Phys* 18(5):955–963.

ZONDERLAND, H.M., E.G. COERKAMP, J. HERMANS, M.J. VAN DE VIJVER, & A.E. VAN VOORTHUISEN. 1999. Diagnosis of breast cancer: contribution of US as an adjunct to mammography. *Radiology* 213(2):413–422.

Summary

Breast cancer is the most common type of cancer in women, with about one in ten women developing the disease in her lifetime. It is also the leading cause of cancer deaths for women aged between 35 and 55. The key to curing breast cancer is early detection and prompt treatment. A physical examination, mammography, and breast self-examination make up the conventional early detection approach. Recommendations for breast cancer screening vary from country to country according to the views of different organisations who recommend the screening. In the Netherlands the screening programme offers all women between 50 and 70 years a biennial mammography screening examination.

Although mammography is the most effective technology presently available for breast cancer screening, it still has some important limitations. First, during screening about 20% of all malignant breast tumours are 'missed'. The most important causes of these false negative screening exams are detection and interpretation errors. Second, the number of false positive detections is rather high. More then half of the women who are referred for further examination turn out not to have breast cancer. Third, accurate mammography interpretation depends heavily on the reader. To overcome some of these limitations computer aided diagnosis and detection (CAD) programmes are being developed. These programmes help radiologists with the detection and interpretation of mass lesions. Studies have shown that CAD systems may improve the diagnostic accuracy of mammography.

At the moment most CAD programmes only use information from a single view to detect and characterise mass lesions. Chapter 2 describes our single view CAD programme that separately processes each image. This programme consists of two parts. In the first part an algorithm calculates at each location inside the breast area several features that measure the presence of either a spiculated lesion or a focal mass lesion. A classifier combines these features into a score that represents the likelihood that a mass is present at that location, the so-called *mass likelihood*. In the second part locations with a high mass likelihood are selected for further processing: segmentation and feature extraction. For segmentation we developed a new method that is described in detail in Chapter 3. The method uses an optimisation technique—dynamic programming—to

find the best contour for each suspicious region. The method proved to be a robust technique to segment mass lesions from surrounding tissue. Compared to existing methods the new method performed significantly better. After segmentation several features are calculated for each region. A second classifier combines these features into a *malignancy score* representing the likelihood that the region is malignant.

Although the single view CAD programme performs quite well, the number of false positive detections is still rather large. An improvement might be obtained by using information from several views such as images from different projections of the same breast, images of the right and left breast or images obtained at different points in time. In this thesis we investigate whether using previous screening examinations is beneficial for a CAD system. We expect that temporal information may improve the detection and classification performance of a CAD system for the following reasons. First, comparing the current mammogram with mammograms from previous screening rounds may bring to attention subtle signs of malignancy that might have been overlooked otherwise. Second, suspicious regions on the current view can be evaluated more precisely when the region is compared with the same region on the previous view. In Chapter 4 we first study how malignant masses change in time. In that study we find that on average malignant masses increase in size and contrast between two consecutive screening rounds. About one quarter of the masses however stays more or less constant or decreases in size. Further inspection of these masses shows that these can be classified in the following categories: architectural distortions that become more compact, masses that are situated on the border of the mammogram, and masses that indeed decrease in size. This suggests that malignant masses differ in temporal behaviour.

In the remainder of the thesis we develop and evaluate a temporal CAD programme. The temporal CAD programme consists of three steps: global registration, regional registration, and extraction of temporal features. Chapter 5 presents a new automatic regional registration method to find corresponding masses on prior and current views. The method starts with a segmented region on the current view. Based on the global registration we make an initial estimate of the location on the prior view where the lesion most likely developed. We define a search area around this initial estimate and calculate three registration measures at each location inside the search area to quantify how well this location matches the region on the current view. As registration measures we use the grey scale correlation between the region on the current view and a candidate region on the prior view, the mass likelihood of the location on the prior view, and the distance from the location on the prior view to the initial estimate. Based on these measures we select the best location. Our segmentation algorithm then determines a contour for the selected region on the prior view. After each current region has been linked to a region on the prior view temporal features are calculated. We designed two kinds of temporal features: difference features and similarity features. Difference features represent changes between feature values extracted from the prior and the current region. Similarity features

measure whether both regions are comparable in appearance.

In Chapter 6 we apply the temporal CAD programme to improve the detection of masses. As regional registration method we use a simple variant of the method that has been described in Chapter 5. As temporal features we only use difference features. FROC (free response operating characteristic) analysis shows a small improvement in detection performance when temporal features are used in addition to the single view CAD system.

In Chapter 7 we evaluate the use of a temporal CAD programme to classify lesions as malignant or benign. For this purpose we use the complete regional registration programme as described in Chapter 5. As temporal features we use both difference and similarity features. We find that the classification performance measured as the area under the ROC curve significantly improves when temporal features are used.

Finally, in Chapter 8, we investigate the effect of a temporal CAD programme on the characterisation performance of radiologists and compare this with independent double reading, where the scores of two radiologists are combined. A total of six radiologists participated in the observer study. Each radiologist rated 198 cases, 99 containing a benign mass and 99 containing a malignant mass. Similarly our temporal CAD programme rated each lesion. We then compared the following reading modes: single reading, independent reading with CAD—that is independent combination of the CAD score and a radiologists score—and independent double reading. Results show that the performance of radiologists significantly improves for independent reading with CAD and for independent double reading. The improvement obtained by reading with CAD however was larger than the improvement obtained by independent double reading. From this study we conclude that a temporal CAD programme may be useful to help radiologists with the interpretation of mass lesions. Further studies are needed to investigate the best way a CAD system can be used in clinical settings.

Samenvatting

Borstkanker is de meest voorkomende soort kanker bij vrouwen. Ongeveer 1 op de 10 vrouwen zal ooit in haar leven borstkanker krijgen. Daarnaast is borstkanker in Nederland de meest voorkomende vorm van kanker waaraan vrouwen overlijden. Een vroege detectie van borstkanker is belangrijk omdat dit de kans op genezing aanzienlijk vergroot. De meest gebruikte technieken voor vroegtijdige detectie zijn zelfonderzoek, klinisch onderzoek en mammografie. Richtlijnen voor vroegtijdige detectie verschillen van land tot land. In Nederland worden alle vrouwen van 50 tot 70 jaar elke twee jaar persoonlijk uitgenodigd om een screeningsmammogram te laten maken.

Alhoewel mammografie op dit moment de meest effectieve methode is voor screening op borstkanker, zijn er ook enkele beperkingen. Ten eerste wordt nog steeds circa 20% van de tumoren 'gemist' tijdens de screening. De belangrijkste oorzaken hiervan zijn detectiefouten en interpretatiefouten. Ten tweede is het aantal fout positieve detecties te hoog. Meer dan de helft van de vrouwen die doorverwezen worden blijkt uiteindelijk geen borstkanker te hebben. Ten derde hangt een goede beoordeling van een mammogram erg af van de betreffende radioloog. Om een aantal van deze problemen te verminderen zijn er computer programma's ontwikkeld met als doel de detectie en interpretatie van tumoren te verbeteren, de zogenoemde *computer aided detection/diagnosis* (CAD) programma's. Studies tonen aan dat het gebruik van CAD programma's kan leiden tot een verbetering van de diagnostische accuraatheid van mammografie.

Op dit moment maken de meeste CAD programma's slechts gebruik van één enkele opname. Hoofdstuk 2 beschrijft ons CAD programma waarvoor de input bestaat uit een enkele afbeelding. Dit programma bestaat uit twee delen. Eerst wordt op iedere locatie in het borstgebied een aantal tumorkenmerken uitgerekend zoals de aanwezigheid van een verdacht lijnenpatroon (spiculation) en de aanwezigheid van een heldere densiteit. Een classifier combineert deze kenmerken in een score die aangeeft hoe waarschijnlijk het is dat er een tumor op de betreffende locatie aanwezig is, de zogeheten *mass likelihood*. In het tweede deel van het programma worden de meest verdachte locaties geselecteerd voor verdere bewerking: segmentatie en extractie van tumorkenmerken. Voor segmentatie hebben we een methode ontwikkeld die in staat is zeer nauwkeurig en snel

de contour van een verdachte regio te bepalen. Hoofdstuk 3 beschrijft deze methode in detail. De methode gebruikt een optimalisatietechniek—dynamic programming—om de beste contour voor iedere regio te vinden. In vergelijking met andere segmentatiemethoden presteert de nieuwe methode significant beter. Voor iedere gesegmenteerd regio worden vervolgens diverse kenmerken bepaald. Een tweede classifier combineert deze kenmerken in een *malignancy score*, die aangeeft hoe waarschijnlijk het is dat de regio een maligniteit bevat.

Een probleem met huidige CAD programma's is het hoge aantal fout positieve detecties. Het aantal fout positieve detecties zou kunnen verminderen wanneer CAD programma's, net als radiologen, gebruik zouden maken van informatie uit meerdere opnamen zoals opnamen uit verschillende richtingen, opnamen van de linker en de rechter borst en opnamen verkregen op verschillende tijdstippen. In dit proefschrift onderzoeken we of het gebruik van voorgaande screeningsmammogrammen een positief effect heeft op de performance van een CAD systeem. We verwachten dat het gebruik van temporele informatie zal leiden tot een verbetering van zowel de detectie als de interpretatie van tumoren om de volgende redenen. Ten eerste kan het vergelijken van opeenvolgende mammogrammen kleine en subtiele afwijkingen aan het licht brengen die anders over het hoofd gezien zouden zijn. Ten tweede kan een verdachte regio beter beoordeeld worden wanneer deze vergeleken wordt met dezelfde regio op het voorgaande mammogram. In Hoofdstuk 4 bestuderen we welke veranderingen in de tijd optreden bij maligne tumoren. In deze studie zien we dat gemiddeld genomen maligne tumoren groter worden en dat het contrast toeneemt tussen twee opeenvolgende screenings. Een aanzienlijk deel van de tumoren echter verandert niet in grootte of wordt zelfs kleiner. Nadere inspectie laat zien dat we deze lesies in verschillende categorien kunnen indelen: architectuurverstoringen die weliswaar kleiner maar ook meer compact worden, lesies gelocaliseerd op de rand van het mammogram en lesies die echt kleiner worden. Hieruit kunnen we concluderen dat er veel verschil is in het temporele gedrag van maligne tumoren.

De rest van dit proefschrift wijden we aan de ontwikkeling en evaluatie van een temporeel CAD programma. Het temporele programma bestaat uit drie onderdelen: globale registratie, regionale registratie en extractie van temporele kenmerken. Hoofdstuk 5 presenteert een nieuwe methode voor regionale registratie met als doel corresponderende lesies op huidige en voorgaande mammogrammen aan elkaar te koppelen. De methode begint met een gesegmenteerde regio op het huidige mammogram. Op basis van de globale registratie maken we dan een eerste schatting van de locatie op het voorgaande mammogram waar deze lesie waarschijnlijk ontstaan is. Vervolgens definieren we rondom deze initiele schatting een zoekgebied. Op iedere locatie in dit zoekgebied rekenen we drie registratiematen uit: grijswaardecorrelatie tussen de regio op het huidige beeld en een kandidaat region op het voorgaande beeld, de *mass likelihood* van de locatie op het voorgaande beeld en de afstand tot de initiele schatting. Op basis van deze maten selecteren we de beste locatie. Daarna bepaalt ons nieuwe segmentatie algorithme de

contour van de geselecteerd regio op het voorgaande beeld. Tenslotte bepalen we twee soorten temporele kenmerken: verschil kenmerken en gelijkenis kenmerken. Verschil kenmerken meten de verandering in tumorkenmerken tussen de regio op het voorgaande beeld en de regio op het huidige beeld. Gelijkenis kenmerken meten of twee regio's er ongeveer hetzelfde uit zien.

In Hoofdstuk 6 gebruiken we het temporele CAD programma voor de detectie van tumoren. Als regionale registratie methode gebruiken we een simpele versie van de methode die beschreven is in Hoofdstuk 5. Als temporele kenmerken gebruiken we alleen verschil kenmerken. FROC (free response operating characteristic) analyse laat een kleine verbetering zien wanneer het CAD programma gebruik maakt van temporele kenmerken.

In Hoofdstuk 7 evalueren we het temporele CAD programma om lesies te classificeren als benigne of maligne. Het temporele CAD programma gebruikt eerst de regionale registratie methode uit Hoofdstuk 5 om iedere lesie op het huidige mammogram te linken aan een regio op het voorgaande mammogram. Daarna worden beide temporele kenmerken uitgerekenend: verschil kenmerken en gelijkenis kenmerken. Uit deze studie blijkt dat de classificatie performance verbetert door het gebruik van temporele kenmerken.

Tenslotte beschrijven we in Hoofdstuk 8 een studie die we uitgevoerd hebben om te bepalen welk effect een temporeel CAD programma kan hebben op de diagnostische accuraatheid van radiologen. In totaal deden zes radiologen mee met deze studie. Iedere radioloog beoordeelde 198 cases, waarvan er 99 een benigne en 99 een maligne lesie bevatten. Het temporele CAD systeem beoordeelde dezelfde cases. We vergeleken de volgende situaties: individuele beoordeling door één radioloog, onafhankelijke combinatie van de beoordelingen van twee radiologen en onafhankelijke combinatie van de resultaten van het temporele CAD systeem en een radioloog. De resultaten laten zien dat de interpretatie van tumoren verbetert voor zowel onafhankelijke combinatie van de beoordelingen van twee radiologen als voor onafhankelijke combinatie van de beoordeling van een radioloog en het CAD programma. De verbetering verkregen door het gebruik van het CAD programma was het grootste. Hieruit blijkt dat een temporeel CAD systeem nuttig kan zijn om radiologen te helpen met de beoordeling van lesies. Er is meer onderzoek nodig om te bekijken op welke manier een CAD systeem het beste in de praktijk gebruikt kan worden.

Dankwoord

De afgelopen vier jaar heb ik met veel plezier gewerkt in de groep van Nico Karssemeijer aan het ontwikkelen van computer gestuurde technieken voor de detectie van borstkanker. Dit werk stelde mij in staat mijn interesses uit verschillende richtingen te combineren: natuurwetenschappen, informatica en geneeskunde. Mijn copromotor Nico heeft mij daarin uitstekend begeleid. Enerzijds heb ik veel van hem geleerd door zijn kennis en ervaring, anderzijds heeft hij mij de ruimte heeft gegeven om zelf onderzoek te doen en eigen ideeen in te brengen. Nico, bedankt voor de fijne tijd die ik in Nijmegen heb gehad.

Mijn promotor Stan Gielen wil ik bedanken voor zijn begeleiding bij het voltooien van dit proefschrift. De promotiecommissie dank ik voor het lezen en beoordelen van dit proefschrift.

Door het voortijdig overlijden van Jan Hendriks heeft hij helaas de voltooiing van dit boekje niet kunnen meemaken. Als radioloog beschikte hij over zeer veel ervaring op het gebied van borstkankerscreening. Daarnaast was hij altijd zeer betrokken en enthousiast over de projecten van de CAD groep.

Mijn kamergenoten Celia en Saskia wil ik bedanken voor de fijne samenwerking en voor de gezelligheid en goede sfeer op onze kamer. Daarnaast zorgden mijn collegae van de CAD groep en het LRCB voor vele afwisselende gesprekken en interessante discussies. Ook voor inhoudelijke vragen kon ik altijd bij iemand van de groep terecht.

In het bijzonder wil ik de fietsclub ‘Bergh in het zadel’ noemen die gezorgd heeft voor financiele sponsoring van dit project. I also thank Deborah Lee Soltesz for creating the beautiful cover of this thesis (Soltesz 2006).

Lieve Nicky, mijn dierbaarste en mijn grote liefde. Dank je voor je onvoorwaardelijke steun en dat je me zo gelukkig maakt. Eigenlijk had ik dit boekwerk willen vullen met een enumeratie van ’Lieve Nicky’. Dat zou zonder twijfel veel interessantere leeskost hebben opgeleverd dan het huidige.

Curriculum Vitae

Sheila Timp was born on March 29th, 1973 in The Hague, the Netherlands. After completing grammar school she entered medical school in 1991 at the University of Utrecht and obtained her medical degree in 1999. Then, in 2000, she joined the chair of Signals and Systems of the Faculty of Electrical Engineering, University of Twente, Enschede to work on a project for real-time visualisation of three dimensional medical data. In September 2001 she started a Ph.D. project under supervision of dr. N. Karssemeijer at the Radiology Department of the University Medical Centre St. Radboud Nijmegen. Her research focused on the use of temporal information to improve computer aided detection and diagnosis programmes for mammography.

January 2006 she started a radiology residency at the Department of Radiology, University Medical Centre Groningen.

She is married to Nicky van Foreest and has four sons: Jorden (1999), Lucas (2001), Pieter (2002), and Tristan (2005).