Kwantitatieve analyse van echografiebeelden
van de premature hersenen

Quantitative Analysis of Ultrasound Images of the Preterm Brain

Ewout Vansteenkiste

Promotor: prof. dr. ir. W. Philips
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen

Vakgroep Telecommunicatie en Informatieverwerking
Voorzitter: prof. dr. ir. H. Bruneel
Faculteit Ingenieurswetenschappen
Academiejaar 2006 - 2007

UNIVERSITEIT
GENT

ISBN 978-90-8578-138-7
NUR 954
Wettelijk depot: D/2007/10.500/12

Dankwoord

Naar het schijnt ligt de waarde van woorden niet in hun aantal, en zou daarom ook een dankwoord kort en bondig moeten zijn. Dat dit waar is voor alles wat hierna komt daar kan ik mee leven, maar nu even niet.

In 2001 kwam ik voor het eerst TELIN binnen en ging ik slijtage van stofzuigers onderzoeken, of hoe praktisch beeldverwerking wel niet zijn kan. Stofzuigers werden algauw satellietbeelden, satellietbeelden werden beurskoersen, beurskoersen werden echografiebeelden van witte hersen-stof en de cirkel was rond. Dit om te zeggen dat de hoofdbrok van mijn onderzoek uiteindelijk in dit werk is terecht gekomen maar de vrijheid er altijd is geweest om zelf op zoek te gaan naar inspirerende onderwerpen (let wel: met die stofzuiger had ik niets te maken).

De atmosfeer waarin dit mogelijk was, en is, is maar aan de visie van 1 iemand te danken en dat is Prof. Wilfried Philips. Zijn begeleiding, vertrouwen en steun zijn de basis van mijn werk en al zal hij wel minder plezier beleefd hebben aan het meermaals moeten herlezen van dit boek, was niets ooit te veel gevraagd. Ik bedank hem dan ook van ganser harte voor alle grote en kleine dingen die hij voor mij deed en doet. In dezelfde zucht bedank ik Dr. Paul Govaert voor de zin die hij aan dit onderzoek gaf. Zijn gedrevenheid werkt aanstekelijk en met een druk artsenleven als het zijne toch nog een verschil proberen maken, maakt ook van hem eerder uitzondering dan regel. Zijn deur, zowel thuis als in Sofia, stond/staat altijd open voor een babbel over hele kleine en grote kindjes.

In de laatste vijf jaar heb ik redelijk wat verschillende bureaus op de TELIN-gang van binnen gezien om uiteindelijk toch mooi, samen met nieuw jong geweld, in het grootste van allemaal terecht te komen. Met plezier heb ik dan ook mijn laatste residentie aan "Annette Inc." verhuurd want zonder haar/hun zorgen zat ik waarschijnlijk nu nog ergens op een vliegveld in transit vast. Mijn oorarts dankt ook Philippe en Davy voor de blijvende oorschade door de cd-spelers die hier mooi op alle pc's geïnstalleerd zijn. Ikzelf bedank hen om Sam meermaals te reanimeren en al mijn medische goestjes te installeren.

Bedankt ook aan alle TELIN-collega's die ik ken en gekend heb. In het bijzonder Gjenna voor zijn steun bij het echo-onderzoek, Aleksandra en Filip voor de filterexperimenten, Stefaan en Alessandro voor hun hulp bij de implementatie van de psycho-visuele experimenten en de lay-out van dit werk. Mijn grote vriend doctor manager Vladimir dank ik voor zijn steun, slivovic en oosterse taxiritten. Jef en Bruno wil ik bedanken voor de mooie samenwerking. In hun begeleiding leidden ook zij mij naar nieuwe problemen en oplossingen. Ik hoop dat het jullie goed gaat in al wat je doet.

Alle dokters en professoren die hun tijd wilden nemen om de (medische) beelden te scoren, van feedback te voorzien of te filteren wil ik uitdrukkelijk bedanken. Hier zijn ze allemaal: Dr. Maarten Lequin, Dr. Nikk Conneman, Dr. Jeroen Dudink, Dr. Alex Korsten, Dr. Renate Swarte en Dr. Rene Kornelisse (EMC/Sofia kinderziekenhuis, Rotterdam), Dr. Kris Decoene en Dr. Alexandra Zecic (UZ, Gent), Prof. Linda De Vries and Dr. Floris Groenendael (Wilhelmina kinderziekenhuis/UZ, Utrecht), Dr. Gerda van Wezel-Meijler (UZ, Leiden), Dr. Liem en Dr. Mullaart (UZ, Nijmegen), Dr. Frances Cowan (Imperial College School of Medicine, Hammersmith Hospital, London), Prof. Terrie Inder (Royal Women's and Children's Hospital, Melbourne), Dr. Andrea Mewes and Dr. Goetz Welsch (Bringham and Women's Hospital, Boston), Dr. Alessandro Foi en Prof. Karen Egazarian (University of Tampere, Tampere), Prof. Javier Portilla (Universidad de Granada, Granada) en Prof. Ivan Selesnick (Polytechnic University, New York).

Prof. Jean-Bernard Martens bedank ik voor het ter beschikking stellen van zijn XGms code en voor het delen van zijn expertise rond psychovisuele testen. Dietrich dank ik voor zijn code rond de vaag-logische similariteitsmaten. Prof. David Vyncke en Bieke dank ik voor hun statistische telefoongesprekken. Coert dank ik voor de 3D-echografiesequenties, Sebastian voor de echografiebeelden van de halsslagader, Tom voor het echofantoom, Rino voor de textielbeelden, Youri en Benoit voor hun steun in de eerste jaren. Het Instituut voor de aanmoediging van innovatie door Wetenschap en Technologie in Vlaanderen bedank ik voor de financiële ondersteuning van dit werk.

Wijze mannen vertelden me ooit dat je eerste grote conferentie belangrijk en bepalend is voor je onderzoek, 5 minuten later stonden we aan het begin van een memorabele nacht die begon in the Sidney Opera House. Nu begrijp ik ongeveer wat zij toen bedoelden. De vier windstreken, Abake, Pau, ET, Peetse, Chris, Maribel, Jef, Lucas, Listopad, Gianluca, Greg, Vasilis, Monica en Stew bedank ik dan ook om via hun collegiale vriendschap-op-afstand mijn horizon te verruimen.

Ne specht voor de Cinemobiel en de Dimangeurs om er altijd en overal te zijn,

ik heb alleen mijn hart te geven. Taco, Griet, Wannes en al mijn vrienden, bedankt voor wie jullie zijn.

En dan zijn we er bijna. Mijn ouders, kleine zus en mama van Mieke dank ik om thuis te kunnen komen, te leren wat echt belangrijk is en me te maken tot wie ik ben.

En hiermee is de naam gevallen van zij die alles bij mekaar houdt. Mieke is mijn rots in de branding, mijn alles. Binnenkort heeft ze twee kindjes om voor te zorgen waarvan het ene hier plechtig belooft geen lange nachten meer te zullen werken en het andere hoop ik zal beloven ons geen al te lange nachten wakker te houden. Het zijn lastige maanden geweest en met woorden kom ik er toch niet uit om te zeggen wat het betekent dat jij naast mij staat, maar nu komen andere tijden, samen. Dit is voor jullie twee.

Takk.

Ewout aka Eddie
januari 2007

Samenvatting

Het bepalen van de graad van hersenschade is een complex gebeuren dat afhankelijk is van meerdere aspecten van de geneeskunde. Klinisch relevante informatie omtrent de hersenen komt op een onregelmatige en ongestructureerde manier tot uiting en data moet vaak gegroepeerd worden vooraleer er een betekenisvolle diagnostische waarde aan kan worden toegekend en een gerichte behandeling kan worden gestart. Deze thesis behandelt 1 specifiek aspect, namelijk de momenteel meest gebruikte beeldmodaliteit in de detectie en diagnose van premature hersenschade, zijnde *echografie*.

Echografiebeelden komen tot stand door ultrasone geluidsgolfjes (met frequenties van 2 tot 40 MHz) door het huidoppervlak te sturen en hun reflecties op zachte weefsels en organen om te zetten in een beeld. Er zijn heel wat voordelen aan deze manier van scannen. Echografietoestellen variëren in grootte van een aktentas tot iets groter dan een computer wat hen uiterst geschikt maakt om ze tot aan het bed van de patiënt te brengen. Echografie is een niet-invasieve techniek aangezien de probe van het echografietoestel, die de geluidsgolven uitstuurt en capteert, op de huid wordt geplaatst. Verder worden er in echografie geen contrastvloeistoffen noch radioactieve substanties gebruikt wat de techniek zeer veilig maakt. Daarenboven is de prijs van een echografietoestel veel lager dan die van andere courante scanners zoals bijvoorbeeld de magnetische resonantie scanner. Tenslotte is echografie een real-time beeldvormingsmodaliteit, wat betekent dat de arts de beelden in ware tijd ziet tijdens het scannen. Dit komt een snellere diagnose tegemoet. Uit al deze argumenten is het dan ook duidelijk dat hoemeer echografie kan worden ingeschakeld in de dagelijkse klinische praktijk, hoe beter voor zowel de patiënt als arts.

Het grootste nadeel aan echografie en de eraan gekoppelde diagnose daarentegen is echter de befaamde lage beeldkwaliteit van de beelden. Echografiebeelden zijn opgebouwd uit *spikkel*. Dit zijn korrelachtige witte vlekjes die ontstaan uit het interferentiepatroon van de gereflecteerde geluidsgolven. Spikkel bevat zowel klinisch relevante als irrelevante informatie. Pathologische weefselstructuur bijvoorbeeld manifesteert zich in de echografiebeelden vaak als een relevant spikkeltextuurpatroon. Backscattering daarentegen is het

gevolg van reflecties op ongeordende microstructuren of inhomogeniteiten aanwezig in alle weefsels, zowel gezond als ziek, en wordt vaak gezien als irrelevante informatie. In het geval van minder uitgesproken ziektebeelden of pathologieën is het vaak niet evident de relevante van de irrelevante informatie te onderscheiden. Dit bemoeilijkt een diagnose die alleen gebaseerd is op een visuele beeldinterpretatie, ook wel *kwalitatieve* diagnose genoemd.

De detectie van Periventriculaire Leukomalacie, ook wel witte-stofziekte genaamd, in echografiebeelden van premature baby's met een zeer laag geboortegewicht ($< 1500g$) is een typisch voorbeeld waar een kwalitatieve diagnose niet volstaat. Door een tekort aan zuurstof rond de geboorte sterft een deel van de witte hersenstof af of ontwikkelt die zich onvoldoende bij deze zeer vroeg geboren baby's. Gezien de kwetsbare aard van de patiëntjes is echografie de meest praktische beeldvormingstechniek om deze pathologie te onderzoeken. Daarentegen is het premature brein echter nog in volle ontwikkeling waardoor het vaak moeilijk is de subtiele verschillen in witte stof te detecteren enkel gebaseerd op een kwalitatieve echografiebeeldinspectie.

Daarom vertrouwen artsen dezer dagen meer en meer op zogenaamde computergebaseerde diagnosetechnieken in het geval van moeilijk te onderscheiden pathologieën. Deze technieken voorzien de arts van computeralgoritmen die specifieke informatie uit de beelden haalt via objectieve *metingen*. De diagnose gebaseerd op deze metingen wordt ook een *kwantitatieve* diagnose genoemd.

We vermeldden eerder dat pathologische weefselstructuur zich in de echografiebeelden manifesteert als een spikkeltextuurpatroon. In Periventriculaire Leukomalacie manifesteert pathologische witte stof zich in een helder textuurpatroon dat *flaring* wordt genaamd. In bepaalde varianten van de pathologie is deze flaring moeilijk met het blote oog te onderscheiden. Zodoende was een van de doelstellingen van deze thesis om een classificatiealgoritme te ontwikkelen dat mathematische textuurparameters extraheert uit het grijswaardenpatroon van de textuurgebieden om deze vervolgens als pathologisch of niet-pathologisch (flaring of geen flaring) te labelen, op een automatische manier en met een precisie die de visuele kwalitatieve diagnose overstijgt.

Naast de weefselkarakterisatie hebben artsen nood aan een betrouwbare manier om de flaringoppervlakte in de beelden af te kunnen bakenen. Het is immers zo dat flaring, zij het in mindere mate, ook voorkomt in niet-pathologische gevallen. Echter, in tegenstelling tot de pathologische gevallen waar de flaring met tijd uitbreidt, verdwijnt niet-pathologische flaring vaak na een aantal dagen. Het volgen van de evolutie van de flaring is dan ook essentieel voor een correcte diagnose. Het feit dat flaring in bepaalde gevallen moeilijk te onderscheiden is met het blote oog bemoeilijkt echter opnieuw een accurate manuele aflijning of segmentatie. Zodoende ontwikkelden we ook een

segmentatiealgoritme om de randen van de gebieden met flaring af te lijnen alsook de oppervlakte van die gebieden te bepalen.

Dit algoritme combineert de textuurparameters die we gebruikten voor de classificatie van pathologische witte stof en morfologische operatoren en presteert beter dan de gangbare kwantitatieve methode gebaseerd op actieve contouren. Verder zijn onze segmentatieresultaten ook vergelijkbaar met gouden standaard segmentaties. Gebaseerd op de principes van dit algoritme ontwikkelden we tevens een driedimensionaal hersenventrikelsegmentatiealgoritme en een segmentatiealgoritme voor de halsslagader, beide voor echografiebeelden.

Een volgend algoritme ontwikkeld in het kader van dit doctoraat behelst de multi-modale validatie van klinische data. Meer en meer combineren artsen complementaire klinische informatie afkomstig van verschillende modaliteiten in hun diagnose. Dit is ook het geval bij medische beeldanalyse waar beelden van verschillende modaliteiten vergeleken worden om na te gaan hoe een bepaalde pathologie zich manifesteert over de beelden heen. In het geval van Periventriculaire Leukomalacie gebeurt deze *meta-analyse* of *cross-validatie* van echografiebeelden vooral ten opzichte van magnetische resonantiebeelden, vaak beschouwd als de gouden standaard voor de analyse van witte-stofziekte (bij termen en kinderen op latere leeftijd). Om echter na te kunnen gaan hoe de pathologie zich manifesteert over de verschillende modaliteiten is de localisatie van de flaring in beide beelden alsook het aligneren van de beelden, zijnde de beeldregistratie, van cruciaal belang. We ontwikkelden dan ook een interactief beeldregistratiealgoritme dat tweedimensionale echografiebeelden aligneert met hun overeenkomstig driedimensionaal volume van magnetische resonantiebeelden.

Door na registratie de flaringgebieden in beide modaliteiten te vergelijken merkten we op dat, in bepaalde gevallen, flaring wel zichtbaar en aflijnbaar was op echografiebeelden waar dit niet het geval was op de magnetische resonantiebeelden. Hiermee toonden we voor het eerst een vermoeden aan dat artsen al een tijd deelden, namelijk dat alhoewel magnetische resonantie ontegensprekelijk de gouden standaard is voor witte-stofziekte bij termen en kinderen op latere leeftijd, echografiebeeldvorming van belangrijke diagnostische waarde is voor de zeer vroege detectie van de pathologie.

Zoals gezegd bevat spikkel zowel relevante als irrelevante klinische informatie. Verschillende onderzoekers stelden dan ook voor om via filtering irrelevante spikkel in de beelden te onderdrukken en klinische relevante spikkel te benadrukken. Hiervoor gebruikten wij een eigen ontwikkelde spikkelfilter in zowel ons driedimensionaal hersenventrikel segmentatie- als multimodale registratiealgoritme. Het effect van dergelijke filteroperaties op de totale echografiebeeldkwaliteit werd echter nog niet aangetoond. Meer bepaald is het niet duidelijk of het filteren van irrelevante spikkel ook de be-

trouwbaarheid, de snelheid of het gemak van de kwalitatieve diagnose verhoogt.

In een psychovisueel experiment waaraan 7 artsen deelnamen onderzochten we dit effect, gebaseerd op een multidimensionale schalingstechniek. Tegen onze verwachtingen in bleek de spikkelfilter een negatief effect te hebben op de diagnostische kwaliteit van de echografiebeelden. Alhoewel onze filtertechniek de spikkel in homogene gebieden onderdrukt en de randen van relevante anatomische kenmerken accentueert ervaren de artsen dit als onnatuurlijk en als een verlies aan structurele informatie.

Alhoewel dit laatste resultaat eerder negatief mag lijken leert het ons toch twee dingen. Vooreerst dat we de performantie van spikkelfilters moeten nagaan eerder door hoe ze kwantitatieve metingen, zoals segmentaties of registraties, beïnvloeden dan na te gaan hoe ze de beeldkwaliteit beïnvloeden. Daarnaast, bleek uit ons experiment dat artsen *gewend zijn* om naar alle spikkel in het beeld te kijken, ook als dit betekent dat het hun diagnose vaak hindert.

Samenvattend leverde dit onderzoek als belangrijkste resultaten 3 klinisch gevalideerde algoritmen en 1 psychovisueel experiment op om artsen bij te staan in een snelle, vroege, computergebaseerde kwantitatieve analyse van de premature hersenen:

1. Een classificatie-algoritme om pathologische witte stof te onderscheiden van niet-pathologische (flaring van niet-flaring), gebaseerd op spikkeltextuurkenmerken [Vansteenkiste et al., 2007a].

2. Een segmentatiealgoritme, gebaseerd op spikkeltextuurkenmerken gecombineerd met morfologische operatoren, om flaringgebieden af te lijnen en hun oppervlakte te schatten [Vansteenkiste et al., 2007b].

3. Een registratiealgoritme om tweedimensionale echografiebeelden te aligneren met hun overeenkomstige driedimensionaal magnetische resonantievolume [Vansteenkiste et al., 2006f].

4. Een uitbreiding van bestaande spikkelonderdrukkingsfilter en een psychovisueel experiment dat het effect van deze spikkelfilter op de kwalitatieve echografiebeeldkwaliteit uitdrukt [Pizurica et al., 2006].

Daarnaast leverde dit onderzoek ook volgende resultaten op, voortvloeiend uit bovenvermelde algoritmen:

1. Een uitbreiding van het flaringsegmentatiealgoritme naar een driedimensionaal hersenventrikelsegmentatiealgoritme, alsook naar een algoritme om de halsslagader te segmenteren in echografiebeelden.

2. Een uitbreiding van het flaringsegmentatiealgoritme naar een segmentatiealgoritme voor textiel dat een textuurpatroon vertonen gelijkaardig aan echografiespikkel [Morent et al., 2006].

3. Een tweede psychovisueel experiment dat de kwaliteit van 7 verschillende ruisonderdrukkingsalgoritmen test in niet-medische beelden [Vansteenkiste et al., 2006c], [Vansteenkiste et al., 2006b].

Dit proefschrift resulteerde in 6 A1-publicaties [Vansteenkiste et al., 2006f], [Vansteenkiste et al., 2006c], [Morent et al., 2006], [Pizurica et al., 2006], [Vansteenkiste et al., 2006b], [Vansteenkiste et al., 2007a]. Een ander tijdschriftartikel [Vansteenkiste et al., 2007b] is momenteel nog in review. 14 papers werden gepubliceerd in internationale conferenties [Vansteenkiste et al., 2003a], [Huysmans et al., 2004a], [Zlokolica et al., 2006], [Vansteenkiste et al., 2005a], [Vansteenkiste et al., 2006e], [Conneman et al., 2005], [Huysmans et al., 2004b], [Vansteenkiste et al., 2005c],[Vansteenkiste et al., 2006a], [Vansteenkiste et al., 2004b], [Vansteenkiste et al., 2004a], [Vansteenkiste et al., 2005b], [Vandemeulebroucke et al., 2006] en 4 in nationale conferenties en symposia [Vansteenkiste et al., 2002], [Vansteenkiste et al., 2003b], [Vansteenkiste et al., 2006d], [Govaert et al., 2006].

Summary

To determine the degree of brain injury we have to take into account many aspects of biomedicine. Clinical information on the brain usually surfaces in an erratic manner. Therefore, the relevant data must be grouped and filtered before meaningful diagnostic probabilities can emerge and treatment can be targeted to them. This thesis deals with the currently most widely trusted tool for diagnosing preterm brain injury, i.e., *Ultrasound imaging*.

Ultrasound images are formed from the reflections of ultrasonic (at frequencies of 2 to 40 MHz) wave pulses off different tissues and organs. There are numerous advantages to this imaging modality. Ultrasound equipment varies from the size of a briefcase to just over the size of a personal computer, which allows placement at bedside. The ultrasound probe, producing and receiving the ultrasound waves, does not penetrate the body and uses no contrast agents or radio-active substances. Also, ultrasound machinery is far less expensive than other medical imaging modalities as, e.g., Magnetic Resonance scanners. Finally, ultrasound imaging is performed in real-time, allowing the physician to see the images while scanning and perform a diagnosis interactively. Consequently, the more ultrasound imaging can be used in clinical diagnosis, instead of other imaging modalities, the better for both patient and physician.

The main drawback of ultrasound imaging is the poor image quality. Ultrasound images consist of *speckle*, i.e., granular white dots resulting from the interference pattern of the ultrasound wave reflections. Speckle is considered to contain a mixture of both relevant clinical information, e.g., pathological tissue *structure* results in a speckle *texture* pattern, and irrelevant clinical information, e.g., backscattering or reflections off *random* inhomogeneities or microstructures present in both pathological and non-pathological tissue. In the case of less pronounced pathologies, the difficulty to distinguish relevant from irrelevant information hampers an accurate diagnosis solely based on visual image inspection, also called a *qualitative* diagnosis.

The detection of Periventricular Leukomalacia or White Matter Damage, in ultrasound brain images of very low birth weight ($< 1500g$) preterm infants is a typical example where a qualitative diagnosis fails. Due to a lack of oxygen

around birth, some of the white matter in the preterm brain dies or does not develop sufficiently. Given its benefits, ultrasound imaging is the most suited imaging modality to investigate this pathology. However, since the preterm brain is still in full development the dectection of gradual changes in white matter through a qualitative ultrasound image inspection is complex.

Therefore, physicians nowadays rely more and more on what is called a computer-aided diagnosis (CAD). As the name itself suggests, CAD complements the qualitative diagnosis by providing computer algorithms or tools that extract important clinical information from the images through *measurements*. A diagnosis based on these image measurements is also called a *quantitative* diagnosis.

For example, as mentioned, the altered structure of pathological tissue results in an ultrasound speckle texture pattern. In the case of Periventricular Leukomalacia altered white brain matter results in a bright speckle pattern referred to as *flaring*. In certain variants of the pathology, it is difficult to detect this flaring with the unaided eye. As such, one of the goals of this thesis was to develop a classification algorithm that extracts mathematical texture features from the grey value texture pattern in the ultrasound images in order to label pathological tissue automatically and more accurately than by visual inspection.

Besides this structural white brain matter characterization, physicians state the need for the segmentation of flaring regions. More specifically, they are interested in the area of the flaring regions as it is not unusual for flaring to appear in non-pathological cases as well. However, non-pathological flaring usually disappears after a few days whereas pathological flaring spreads. Therefore, the evolution of the flaring area over time is critical information for the follow-up or *staging* of the pathology. So, in this thesis we also developed an interactive algorithm to delineate the flaring boundaries and estimate the flaring areas.

This algorithm combines both the textural information used in the quantitative tissue characterization and mathematical morphology operations, is shown to outperform an existing method based on active contours and is comparable to ground truth segmentations. Based on the principles of this segmentation algorithm we also presented a three-dimensional preterm brain ventricle segmentation algorithm and a carotid artery segmentation algorithm.

A third CAD algorithm presented in this thesis concerns the multimodal validation of clinical data. More and more, physicians are combining complementary clinical data to obtain a more accurate diagnosis. This is also the case in medical imaging where images from different scanning modalities are fused in order to compare how pathologies are manifested. In

the case of Periventricular Leukomalacia, cross-validation or meta-analysis is usually performed on Magnetic Resonance images which are considered the golden standard at term and later age. To compare the way the pathology manifests itself in the images, we need to both localize flaring exactly in the two modalities and align the images of both modalities. As performing these tasks manually is often very time-consuming, we developed a semi-automatic algorithm to register two-dimensional ultrasound images and their corresponding three-dimensional Magnetic Resonance image volumes.

By subsequently fusing our flaring segmentation results and the segmentations on the registered Magnetic Resonance image we noticed that, in certain cases, flaring is detected in the ultrasound images where it is not in the Magnetic Resonance images. This is the first published evidence of a conjecture by some physicians, namely that, although Magnetic Resonance imaging is indisputably the golden standard in Periventricular Leukomalacia at term and on later ages, ultrasound imaging is an important diagnostic tool in the early detection of White Matter Damage.

In the beginning, we mentioned that the ultrasound speckle pattern contains a mixture of relevant and irrelevant clinical information. As such, many researchers have proposed to suppress irrelevant speckle and enhance relevant speckle by filtering the images. We also applied this speckle-reduction approach in both our quantitative three-dimensional brain ventricle segmentation algorithm and multimodal registration algorithm, yet its effect on the overall perceived ultrasound image quality has not been tested. Moreover, it is unclear if a speckle-reduced image results in more reliable and useful information for a qualitative diagnosis.

Consequently, we presented a psycho-visual experiment in which 7 expert physicians took part. Based on a multidimensional scaling framework we investigated the image quality of our own speckle-reduction filter. Surprisingly, the outcome of this experiment showed that, contrary to preprocessing for quantitative image processing, our speckle-reduction filter is unsuited for diagnostic purposes. While our filter aims at enhancing anatomical edges of anatomical features and reducing speckle in homogeneous regions, it seems to also degrade the structural quality of the images and result in images perceived as unnatural.

Although this last result might appear predominantly negative, it teaches us two lessons. First of all, if the power of speckle-reduction filters is to be measured it should be done through their effect on, e.g., segmentation or registration results, rather than on how they improve the overall ultrasound image quality. The second lesson, that we should be careful in defining relevant and irrelevant information in a *qualitative* diagnosis since physicians are *accustomed* to look at all information present in the image, even if it hampers their diagnosis.

In conclusion, this research has resulted in four main contributions: three CAD algorithms, all of which have been clinically validated, and one psycho-visual experiment to assist physicians in a more objective, quantitative analysis of preterm US brain images:

1. A classification algorithm to characterize pathological flaring based on texture features [Vansteenkiste et al., 2007a].

2. A flaring segmentation algorithm combining texture information and morphological operators to delineate flaring boundaries and estimate flaring areas [Vansteenkiste et al., 2007b].

3. A registration algorithm to align two-dimensional ultrasound images and three-dimensional Magnetic Resonance volumes [Vansteenkiste et al., 2006f].

4. A modification of an existing speckle-reduction filter and a psycho-visual experiment to assess the effect of speckle-reduction on the diagnostic ultrasound image quality [Pizurica et al., 2006].

Besides these, this research also resulted in the following algorithms related to the ones above:

1. An extension of the flaring segmentation algorithm to both a three-dimensional brain ventricle segmentation algorithm and a carotid artery segmentation algorithm.

2. An extension of the flaring segmentation algorithm for textiles that show a texture pattern similar to speckle [Morent et al., 2006].

3. A second psycho-visual experiment to investigate the quality of 7 state-of-the-art noise-reduction filters on non-medical images [Vansteenkiste et al., 2006c], [Vansteenkiste et al., 2006b].

This work resulted in 6 A1-publications [Vansteenkiste et al., 2006f], [Vansteenkiste et al., 2006c], [Pizurica et al., 2006], [Morent et al., 2006], [Vansteenkiste et al., 2006b], [Vansteenkiste et al., 2007a]. One A1-paper is currently still in review [Vansteenkiste et al., 2007b]. 14 papers were published in international conferences [Vansteenkiste et al., 2003a], [Huysmans et al., 2004a], [Zlokolica et al., 2006], [Vansteenkiste et al., 2005a], [Vansteenkiste et al., 2006e],[Huysmans et al., 2004b], [Conneman et al., 2005], [Vansteenkiste et al., 2005c], [Vansteenkiste et al., 2006a], [Vansteenkiste et al., 2004b], [Vansteenkiste et al., 2004a], [Vansteenkiste et al., 2005b], [Vandemeulebroucke et al., 2006] and 4 papers were presented in national conferences and symposia [Vansteenkiste et al., 2002], [Vansteenkiste et al., 2003b], [Vansteenkiste et al., 2006d], [Govaert et al., 2006].

Quantitative Analysis of Ultrasound Images of the Preterm Brain

Contents

Symbols and acronyms used in this thesis

Symbols

x	Scalar value
$\mathbf{x}$	Vector
I	Greyscale or binary image
$M \times N$	Image size (rows and columns of pixels)
$f(x, y)$	Grey value of pixel at position (x,y)
f	Frequency of a sound wave
f_i	Function (i referring to its name)
v	Sound of speed in a material
λ	Wavelength of a sound wave
ρ	Density of a material / Opening by Reconstruction
Z	Acoustic impedance of a material
R	Reflection coefficient
T	Transmition coefficient
$\mathbf{T}$	Rigid transformation
G	Number of grey values
F	(discrete) Fourier transform
α	Attenuation coefficient
δ	Kronecker Delta / Conditional Dilation
θ	Angle
μ	Mean value
σ	Standard Deviation

Acronyms

US	Ultrasound
MRI	Magnetic Resonance Imaging
PVL	Periventricular Leukomalacia
VLBW	Very Low Birth Weight

SONAR	SOund Navigation and Ranging
CAD	Compter-Aided Diagnosis
ROI	Region Of Interest
1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
RF	Radio-Frequency
PET	Positron Emission Tomography
SPECT	Single Photon Emission Computed Tomography
MEG	Magneto-Encephalo-Gram
EMG	Electro-Myo-Graphy
DTI	Diffusion Tensor Imaging
fMRI	functional Magnetic Resonance Imaging
SAR	Synthetic Aperture Radar
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
PSNR	Peak Signal to Noise Ratio
SVM	Support Vector Machines
kNN	k-Nearest Neighbor
MAP	Maximum A posterio Probability
LDA	Linear Discriminant Analysis
FLD	Fisher's Linear Discriminant
PCA	Principal Componant Analysis
MV	Majority Voting
CT	Computed Tomography
ICC	Intra-class Correlation Coefficient
SI	Similarity Index
EF	Error Fraction
OF	Overlap Fraction
WI	William Index
3D-DFT	3D Discrete Fourier Transform
SA-DCT	Spatial Adaptive Discrete Cosine Transform
SPERRIL	Steerable Pyramid-based Estimation and Regularization of RIchardson-Lucy
BLS-GSM	Bayesian Least Squares - Gaussian Scale Mixtures
SROC	Spearman Rank Order Correlation
GenLik	Generalized Likelihood

Chapter 1

Introduction

I wanted to start my thesis by defining its topic in a few coherent sentences. So, as I gazed at an empty page for about a quarter of an hour, thinking about how to formulate them, I came up with "a fruit tree". It is not that apple or pear trees have anything to do with the actual content of this work, but to me it seemed like a suitable analogy, see Fig. 1.1.

The roots of this work are planted in *quantitative image analysis*. This implies that images are described in terms of numerical measurements. From these roots, a trunk of *medical ultrasound (US)* image analysis springs. The quantitative analysis of a brain pathology, called Periventricular Leukomalacia or White Matter Damage, feeds the rest of the work.

The trunk splits into several branches. One branch covers the classification of pathological white brain matter in preterm US images, based on tissue texture characteristics. A second branch covers the segmentation of pathological white brain matter areas in the preterm US images. A third branch covers the alignment or registration of preterm US images and Magnetic Resonance Images (MRI). A fourth and final branch describes the perceptual quality assessment of filtered US images. All these branches are heavily related by the information they exchange, and together they form the tree's crown.

Finally, some of the branches split even further and bear fruit mostly situated in the medical field, but not exclusively. The classification branch results in an automated US tissue classification algorithm. The segmentation branch splits into 3 smaller US twigs resulting in:

- An interactive pathological tissue delineation algorithm.
- A cerebral ventricle segmentation algorithm.
- A carotid artery segmentation algorithm.

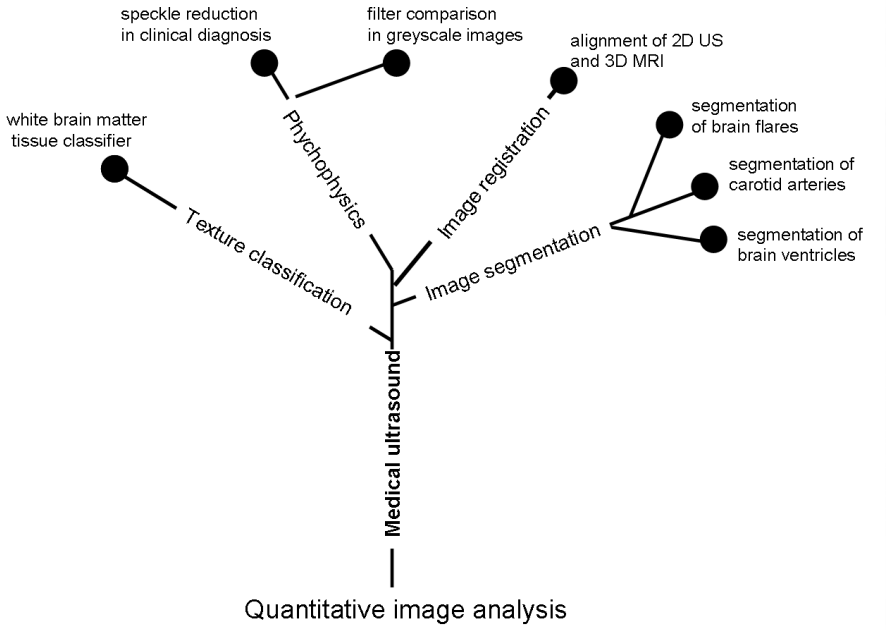


Figure 1.1: The structure of this thesis.

The registration branch results in an interactive two-dimensional (2D) US to three-dimensional (3D) MRI brain registration algorithm. The image quality branch produces both a quality assessment of filtered US images and noise-reduction filters in real-world greyscale images.

In the rest of this introduction, I will restrict myself to an initiation to the roots, in Section 1.1, and the trunk of the tree, in Sections 1.2 and 1.3. The structure of the thesis is described in Section 1.4. Finally, all novelties and contributions it produces are presented in Section 1.5.
Note that from here on I will continue the text in the *we*-form since my tree was nourished by the inputs of many colleagues.

1.1 Quantitative image analysis

Quantitative analysis comprises the *numerical* study, i.e., based on mathematical measurements, of data (images or image structures in our case). As such, it is the opposite of *qualitative* analysis, where the perceived quality of a sample, structure or image is expressed *in words*.

A simple example can clarify this. Image analysis has been part of standard medical diagnosis for many years now. Suppose that two physicians examine

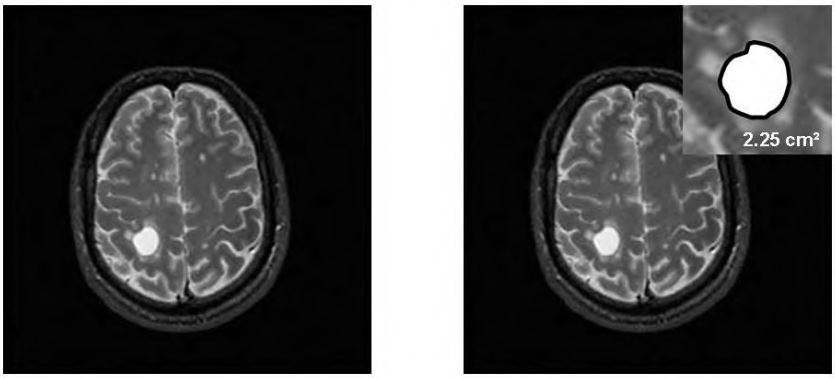


Figure 1.2: Left: brain tumor in a MR image visible as a bright white spot. Right: quantitative measurement of the area of the tumor by some computer algorithm.

a Magnetic Resonance (MR) brain image containing a tumor. Using their expertise they are able to describe (in words) how the tumor manifests itself in the image, e.g., as a "bright white spot" in a specific part of the image, as in Fig. 1.2 (left). Although both experts should normally detect the same object as a tumor, if asked about the size of the tumor, solely based on visual inspection, one expert might score it as being "big" where the other might score it as "average". Exactly this description (of medical images) in terms of the visual properties perceived is called a *qualitative* analysis.
Suppose now that the same experts score the same image but actually measure the size of the tumor using some computer algorithm, as in Fig. 1.2 (right). If the algorithm is constructed and applied well, both experts should reach a (close) consensus on the size, diameter, shape of the tumor. This description in terms of numerical information is called a *quantitative* analysis.

It needs no debate that in the case of less pronounced pathologies in low-quality images, quantitative analysis is the key to an accurate diagnosis. Where in the past, due to technical limitations, physicians were often restricted to a mere qualitative visual diagnosis, the advent of better image processing algorithms and computers with increasing computational power has nowadays allowed for more and more quantitative analysis.

In most computer-aided quantitative medical image analysis tasks, the main research question is: "How can we measure the maximal amount of diagnostically relevant information from a data set, preferably as accurately, quickly and expressed as concisely as possible?"

In this work specifically, this is translated into the following research questions:

1. "How can we characterize the structure of pathological and non-

pathological white brain matter, by as few texture features as possible?"

2. "How can we construct algorithms to segment pathological brain matter and align segmented regions in different imaging modalities?"
3. "How do we measure whether our algorithms are statistically significant?"
4. "How to measure perceived image quality and how does image quality relate to diagnostical quality?"

1.2 Analysis of medical ultrasound images

1.2.1 Ultrasound history

Although little good comes from warfare, it usually - although hardly ever for the right reasons - speeds up technological progress. Medical ultrasound is one of the many techniques initially developed for military war purposes.

At the end of World War I, Paul Langevin, a French professor in physics, developed a SONAR (SOund Navigation And Ranging) tool to locate German submarines, yet too late to be operationally useful [Glor, 2004]. His SONAR mechanism emitted ultrasonic waves that reflected off submarines and the captured echoes were used to image them. One of the first applications of the SONAR in the post-war era was the detection of icebergs, allowing ships to navigate safely through the Arctic.

World War II was the trigger for clinical ultrasound applications. An Austrian physician, Dr. Karl Theodore Dussik, found new ways to examine airmen with head and spinal cord injuries using ultrasound. He provided both the first experimental medical US machine as well as the first published work in medical ultrasonics [Dussik, 1942].
Shortly after World War II, Japanese researchers built the first A-mode (Amplitude-mode) US equipment that processed one-dimensional (1D) signals. Soon after, they introduced B-mode (Brightness-mode) equipment that produces 2D images. Their initial applications were restricted to the detection of gallstones, breast masses and tumors.

It took until the mid sixties before commercially available systems allowed a wide adoption of US diagnosis in the United States and Europe. At that time, pioneers in the United States also contributed extensive innovations to the field. Researchers learned how to use US to detect potential cancers and visualize tumors in living subjects and excited tissue.

Its applications in obstetrics however, are probably still the best known. The first fetal head measurements, to assess the size and growth of the fetus, date from the late fifties. Dr. Stuart Campbell's pioneering work on this fetal cephalometry led to the definitive methods for the study of fetal growth. As the technical quality of the scans improved, it soon became possible to study

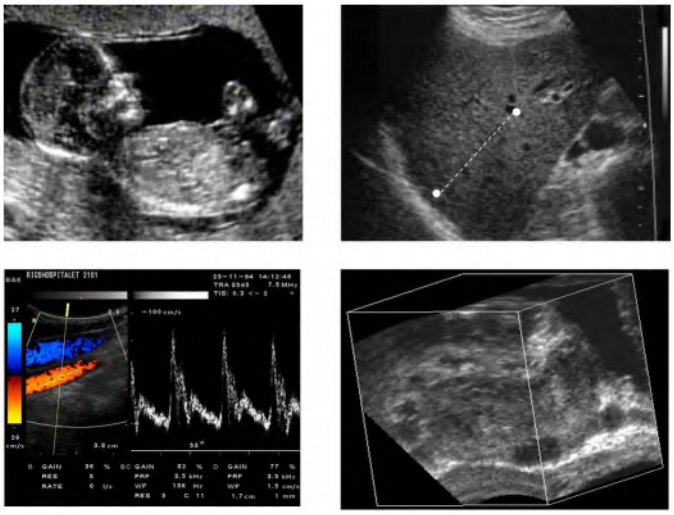


Figure 1.3: Upper left: a B-mode fetal US image. Upper Right: a B-mode liver images with a distance measurement. Lower left: Doppler image of the blood flow in the carotid artery (left-hand color picture), together with speed in the blood vessel as a function of time for two heart beats (right half of the picture). Lower Right: 3D US volumetric representation of the prostate.

pregnancy from start to finish and diagnose its many complications, such as multiple pregnancy, fetal abnormality and placenta praevia.

The recent advent of color Doppler and 3D US enables physicians to even measure blood flow in the images and create volumetric representations. Fig. 1.3 shows some classical B-mode, Doppler and 3D US examples.

The highly-portable, low-cost, almost completely harmless[1] and real-time imaging made US the second-most popular technique in medical imaging (the first still being X-ray) nowadays, with 25% of all medical imaging procedures involving US.

1.2.2 Difficulties of quantitative ultrasound analysis

Although US imaging is common medical practice, it still mainly complements the other modalities, as X-ray and MRI, rather than being an independent diagnostic instrument. In most applications, US image analysis is restricted to qualitative inspection and relatively simple quantitative measurements, e.g., length measurements based on manually selected image markers or features, see Fig. 1.3 (upper right). This is due to the "poor" image quality of the

[1]Up to now, the possible hazard of ultrasound imaging hasn't been proven, apart from possible effects due to heat generation. Tissue absorbs the ultrasound energy increasing its temperature. Continued elevated temperatures above 41 degrees (Celsius) can damage tissue.

US images. More precisely, the US image consists of small bright dots called *speckle*. Speckle results from the interference pattern of the ultrasound waves reflected off the tissues and organs. Although this speckle pattern contains a lot of valuable clinical information, part of it is often considered as (disturbing) noise when it comes to quantitative image analysis.

Related to this speckle, a common criticism on quantitative US analysis is the following: where other modalities, such as MRI, display the structural tissue composition, in US images the interfaces between tissues with different acoustic impedances are shown. Consequently, the tissue structure measured in US images is not the real physical tissue structure.
This criticism may be justified when it comes to describing absolute tissue characteristics, histological examinations, or comparing the same tissue over different modalities, but it does not prevent us from investigating relative organ or tissue differences within or amongst the US images.
For example, if we can measure that pathological and non-pathological tissue have different texture characteristics in US images or that the areas over which pathological and non-pathological tissue stretch out in the US images differ, we are able to perform a quantitative tissue comparison without an absolute tissue description.

However, it is true that this quantitative US image analysis is not straightforward. Quantitative image processing usually starts from an exact definition of what we want to measure, segment or calculate. In US image analysis, this information is usually only provided through qualitative visual image inspections by the physicians, since we lack the structural resemblance to almost all other imaging modalities. Due to the poor image quality, this information is often subjective in the case of less pronounced pathologies. This hampers the exact definition of what is called *ground truth* or *golden standard* information and is an issue we have to keep in mind when constructing and validating quantitative techniques.

A last pitfall for quantitative analysis is the imaging physics. As mentioned, US images are constructed from radio-frequency signals that are transformed by an US machine into grey values. The exact signal processing algorithms involved in the image formation are often kept secret by the manufacturers and differ from manufacturer to manufacturer.
In addition to this manufacturer dependency, multiple machine settings, which directly influence the grey values of the displayed image, can be tuned during image acquisition. These settings either have to be compensated for, by so-called compensation algorithms, or have to be bypassed by working directly on the raw radio-frequency data. Here again, most manufacturers do not provide any access to this data.

1.2.3 Benefits of quantitative ultrasound analysis

Despite the difficulties of quantitative US image analysis, its practical and diagnostic benefits are clear. From a practical point of view, as already stated, US imaging is a fast, low-cost, safe and bedside imaging modality. US machines vary from the size of a briefcase to just over the size of a personal computer (see Fig. 1.4) which allows them to be placed at bedside. US is non-invasive and needs no contrast agents or radio-active substances. Thereupon, US machinery is far less expensive than, e.g., MRI scanners, and imaging is performed in real-time, allowing the physician to diagnose while scanning. Consequently, the more clinical diagnosis can be based on US imaging, the better for both patient and physician.

From a diagnostic point of view, often subjective, visual image interpretations can be replaced by more objective quantitative measurements. The development of tissue classification, segmentation or registration algorithms provides the physician with supplementary, reliable, diagnostic information. Besides that, if these algorithms operate automatically or semi-interactively they can speed up otherwise time-consuming tasks, leaving the physicians more space for other work.

1.3 Periventricular Leukomalacia: clinical goals of the thesis

In this thesis, all issues discussed in Section 1.2 converge in the study of a preterm brain pathology. The recent increase in survival rate of Very Low Birth Weight (VLBW) infants ($\leq 1500g$) has caused increasing neurological morbidity [Davis, 1997]. Commonly, two different types of early brain damage, part of the spectrum of disorders in preterm white brain matter, are known to end in neurologic deficit.

The first consists of haemorrhage in the germinal matrix and lateral ventricle. The extensive types (with ventricular inundation or parenchymal venous infarction) are associated with a higher mortality rate and impaired further brain development.

The second disorder, which is the focus of this work, is called Periventricular Leukomalacia[2] (PVL). PVL is characterized by deep white matter lesions adjacent to the lateral ventricles. With a prevalence of 5-15% among infants at a postconceptional age of 32 weeks, PVL is one of the best predictors of cerebral palsy (complete or partial paralysis) in surviving preterm infants [Paneth et al., 1994].

PVL comes in a focal as well as a diffuse variant. The focal, most severe, variant is characterized by tissue degeneration, the formation of cysts and dilatation

[2]Periventricular = around the ventricles, Leukomalacia = softening of the white matter.

Figure 1.4: Left: a physician next to the US machine. Right: a VLBW infant with the US coupling gel on the fontanelle where the image is captured.

of the lateral ventricles. This variant is also called cystic PVL (cPVL).
The diffuse, milder and more common, variant comes without cysts but is characterized by a more uniform spreading of the damaged tissue in the periventricular and even subcortical region. When echodense zones, i.e. zones with equal echogenicity, turn isodense, i.e. differing in echogenicity, without cyst formation or ventricular dilatation within the first few weeks of life, the phenomenon is described as a transient *flare* in (sonographic) literature. In later stages, these flares lead to mild ventriculomegaly (enlarged ventricles) and a decrease in white matter. This variant can also be called gliotic PVL (gPVL).

The early and accurate identification of PVL is important for counseling parents and targeting high risk neonates to appropriate rehabilitation. Cranial US is the common neuro-imaging procedure in the study of the premature infant [De Vries et al., 1993] since the anterior fontanelle of the preterm provides a convenient sonographic window, allowing non-invasive imaging in the deep midline regions of the brain, see Fig. 1.4. Physicians nowadays base their diagnosis on a visual inspection of these cranial US images.

The more pronounced cystic form of PVL is well visible on US images, see the middle image of Fig. 1.5. The milder, more frequent, gliotic variant is more difficult to distinguish with US, see the right picture of Fig. 1.5, and although cellular histopathology causes flaring, it is not abnormal for unaffected newborns to develop slight physiological flaring as well. This is partly related to normal white matter maturation (premyelination).

The detection of PVL in the first days of life has been a research focus for many years at the neonatology department of the Sophia Children's hospital in the Erasmus Medical Centre, Rotterdam, The Netherlands. At some point,

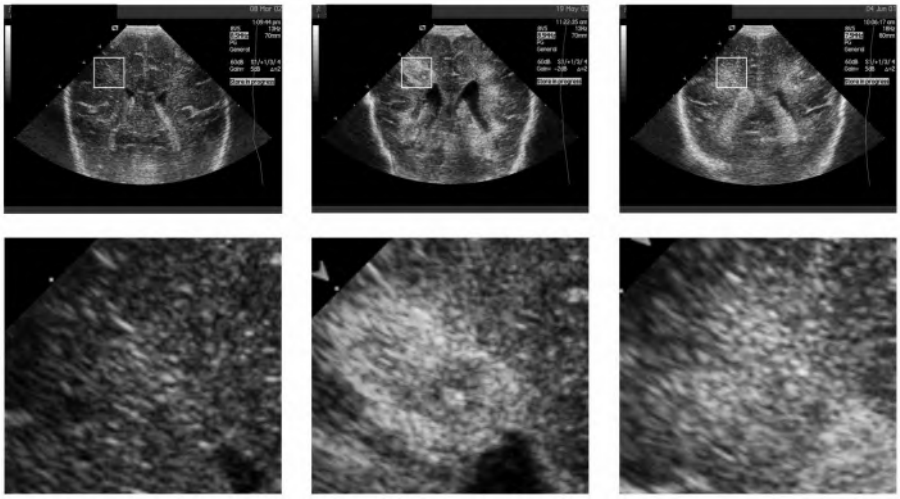


Figure 1.5: Left: a normal brain and an enlargement of the periventricular zone. Middle: brain with cystic PVL (focal variant) and an enlarged zone. Right: a brain affected by gPVL (diffuse variant) an enlargement of part of the area of interest (flaring area).

physicians there asked to investigate how quantitative PVL US image analysis could assist in a more objective US diagnosis, since both literature and their own experience pointed out that visual qualitative image scoring alone was insufficient to describe the subtle differences and gradual changes in image structure in the gPVL cases.
In literature, typical sensitivity scores for visual gPVL recognition lay around 70%, which is fairly low for accurate medical diagnosis [De Vries et al., 1993, Miller et al., 2003]. To illustrate the difficulty of visual scoring, Fig. 1.6 shows a gPVL pathological image on the left compared to a non-pathological image on the right. Clearly, describing the structural difference of the periventricular regions in these images, apart from a slight difference in intensity, is far from trivial. As such, subtle changes in structure have to be *measured* in a different way.

We believe a *quantitative characterization* of pathological and non-pathological white brain matter based on the US speckle texture patterns would allow physicians to characterize PVL in a fast and more objective way. The development of a classification algorithm that does exactly that, is the **first goal** of this work, and the initial step in a computer-aided diagnosis (CAD).

Next to tissue characterization, physicians state the clear need for flare segmentation. Primarily, the area of the flares is of interest. Since it is not unusual for (slight) flaring to appear in non-pathological white matter, the evolution of the flaring area over time is critical information for the follow-up of the pathology, also called *staging*. A persistence of the flares beyond 7-14 days (after

Figure 1.6: Left: a gPVL pathological brain, Right: a non-pathological brain. Regions of interest for flaring are delineated by ellipses.

birth) may be considered abnormal and indicative of damage. Therefore, a subclassification of flaring is suggested into flaring of brief duration (1-6 days) intermediate (7-13 days) and prolonged (14 days or more).
Secondly, yet of indirect interest, the delineation of the flare boundaries can be used to train new physicians. Since visual US diagnosis demands quite some experience, an automated accurate delineation of flares can assist the trainees in their understanding of the pathology.
Thirdly, for meta-analysis and to cross-validate PVL characterization on other modalities, such as MRI, the exact localization of the pathology in both modalities is crucial.
Consequently, a **second goal** of this work is to propose an interactive algorithm *delineating* pathological regions and determining their *area*. Characteristic artifacts such as speckle and a low signal-to-noise ratio often complicate this segmentation task. As such, the relatively low US image quality calls for both task-specific constraints and priors.

Validation of CAD algorithms w.r.t. to ground truth or golden standard information is indispensable. Although the sensitivity of, e.g., an automated tissue characterization algorithm might outperform that of visual inspection, it will never be flawless. Physicians using these algorithms have to be aware of both the limitations and accuracy of the algorithms. As mentioned before, due to the low image quality, this is particularly difficult in US imaging.
Therefore, a **third goal** of this work is the statistical, clinical and multimodal cross-validation of our quantitative US results. First of all, experiments with a panel of 14 physicians are used for both the *construction* of golden standard

information and the statistical and clinical validation of our classification and segmentation algorithms.
Furthermore, since at term and at later ages MRI is the common imaging modality for PVL research, we cross-validate our US segmentation results to preterm MR image data. For this purpose, we developed a semi-automatic algorithm to align 2D US images and 3D MRI volumes. This allows the comparison of structural resemblances and differences between pathological tissue on the registered multimodal images.

Finally, the effect of speckle-reduction on the overall US image quality is often questioned in the medical world. The diagnostic relevance of speckle-reduction has not yet been shown nor studied quantitatively for US images. A **fourth goal** is to quantify the effect of our own speckle-reduction filter on the diagnostic quality by means of a psycho-visual experiment involving 7 physicians.

Note that since at the Sophia Children's hospital preterms are routinely imaged by multiple physicians, we were provided with both the necessary number of images[3] and the indispensable clinical feedback for all these quantitative investigations.

1.4 Organization of the thesis

Now that the roots and trunk of the tree are described, this section presents an overview of the content of the various Chapters of the thesis.

Chapter 2: Medical ultrasound imaging. In this Chapter the fundamentals of (medical) US imaging are presented. The origin of speckle is discussed as well as how speckle reflects tissue properties. Also, the influence of machine settings on quantitative image analysis is commented on.

Chapter 3: Texture-based pattern recognition. In this Chapter we investigate how texture features can be used to classify pathological white brain matter. We investigate 7 first- and second-order texture features as well as 3 different statistical classifiers. Multiple features are combined in a statistically significant way and majority voting is used to link the results of multiple classifiers. This results in an algorithm that distinguishes pathological from non-pathological white brain matter in selected regions of interest with a sensitivity that outperforms the existing qualitative methods significantly.

Chapter 4: Ultrasound segmentation. This Chapter builds on the previous one and combines both tissue texture information and medical prior knowledge on relevant anatomical brain features to determine flaring areas. Mathematical morphology operations are applied to delineate flare boundaries.
Our technique is compared to both ground truth information, obtained from

[3]Verbal parental consent was given for all images used in this thesis. Parents were informed that US scanning is part of the medical routine. This was sufficient since US imaging does not need approval by the Dutch medical ethical committee.

averaged manual expert delineations, and to an existing method on active contours.
We show that our interactive algorithm segments pathological tissue with an accuracy that is comparable to the ground truth information, outperforms the existing method, and can be used as a stand-alone predictor for PVL.
Finally, two extensions of our method are presented based on mathematical morphology operations. A first extension is a 3D brain ventricle segmentation algorithm using a sequence of reconstructed 2D US images. A second extension is a carotid artery US segmentation algorithm for the study of atherosclerosis.

Chapter 5: Multimodal image registration. In this Chapter a cross-validation of the US segmentation algorithm of Chapter 4 is presented. First of all, 2D US images are interactively aligned or registered with their corresponding 3D MRI volumes to facilitate the comparison of pathological brain regions in both modalities. The registration algorithm, developed in close collaboration with a master thesis student [Vandemeulebroucke et al., 2005], consists of a B-spline interpolation of the MRI volumes and a combined speckle prefiltering of the US images. A mutual information metric is optimized using a regular step gradient descent criterion. An interactive initialization procedure is needed to assure good convergence. Our registration results are compared to manual expert registrations in a virtual environment called the I-SPACE.

Chapter 6: Psychophysical experiments on image quality. In this Chapter two psycho-visual experiments on image quality are presented. In these experiments, people score perceived image quality and their results are compared quantitatively based on a multidimensional scaling framework. A first experiment is aimed at assessing the quality of 7 state-of-the-art noise-reduction filters in real-world greyscale images. A second experiment is aimed at measuring the effect of our own speckle-reduction filter on the diagnostic quality of US images.

Chapter 7: Conclusions. This final chapter bundles the global conclusions of the thesis and points out in which direction further related research might proceed.

1.5 Novelties and contributions

The fruit, novelties and contributions, of this thesis can be divided in two baskets: (i) those related to US imaging, (ii) those originating from US research but resulting in applications in other fields. The first basket contains:

1. A multi-feature multi-classifier algorithm to determine if selected US white matter brain regions are pathological (for PVL) or not [Vansteenkiste et al., 2007a]. This algorithm outperforms all of the existing qualitative methods in both accuracy and sensitivity.

2. A user-interactive, clinically validated, flaring area estimation and delineation algorithm [Vansteenkiste et al., 2007b]. This algorithm outperforms an existing method on active contours and is currently used as a stand-alone PVL predictor in clinical practice.

3. A 3D brain ventricle segmentation algorithm and a faster alternative to an existing carotid artery segmentation, both based on mathematical morphology.

4. A first 2D US to 3D MRI brain registration scheme to semi-automatically align flares in both modalities [Vansteenkiste et al., 2006f].

5. A modification of the GenLik filter to US images [Pizurica et al., 2006] and a first psycho-visual US experiment to assess the effect of speckle-reduction on a qualitative diagnosis.

Some of the concepts used or developed for US images, can also be used in other application fields. The morphological segmentation algorithms can be extended to textile fabrics with a granular texture pattern similar to speckle. Also, the multidimensional scaling framework can be used to assess the quality of other, i.e., non-speckle related, noise-reduction algorithms. Since image denoising is one of the major research activities in our group, we performed a second psycho-visual experiment on state-of-the-art noise-reduction filters for non-medical images. Consequently, the second basket of results contains the following contributions:

1. An extension of the morphological segmentation algorithms to measure the influence of plasma treatment on the wicking behavior of textiles [Morent et al., 2006].

2. A psycho-visual assessment of perceived image quality of 7 state-of-the-art noise-reduction filters [Vansteenkiste et al., 2006c] and the investigation of fuzzy similarity measures as an alternative to existing instrumental similarity measures [Vansteenkiste et al., 2006b].

In terms of publications, so far this work resulted in 6 A1-publications [Vansteenkiste et al., 2006f], [Vansteenkiste et al., 2006c], [Vansteenkiste et al., 2006b], [Vansteenkiste et al., 2007a], [Morent et al., 2006], [Pizurica et al., 2006], one other A1-paper [Vansteenkiste et al., 2007b] is currently still in review.
Next to that, 14 publications were published in the proceedings of international peer-reviewed conferences, [Vansteenkiste et al., 2003a], [Huysmans et al., 2004a], [Vansteenkiste et al., 2005a],
[Vansteenkiste et al., 2006e], [Zlokolica et al., 2006], [Conneman et al., 2005], [Vansteenkiste et al., 2005c],[Huysmans et al., 2004b],
[Vansteenkiste et al., 2006a], [Vansteenkiste et al., 2004b],
[Vansteenkiste et al., 2004a], [Vansteenkiste et al., 2005b],

[Vandemeulebroucke et al., 2006] and 4 publications and abstracts in national conferences [Vansteenkiste et al., 2002], [Vansteenkiste et al., 2003b], [Vansteenkiste et al., 2006d], [Govaert et al., 2006].

Chapter 2

Medical ultrasound imaging

This Chapter briefly overviews the US imaging physics. Since US images are used throughout each of the following Chapters, it is necessary to understand how they are formed and what they consist of. This is the subject of Section 2.1. Subsequently, in Section 2.2 we discuss the characteristics of speckle, the building block of all US images, and show how it relates to tissue properties. Finally, the influence of US machine settings on the displayed images is discussed in Section 2.3.

2.1 The physics of ultrasound

As mentioned in Chapter 1, the US imaging process shows a great resemblance to radar principles where *information* is gathered from *wave* propagation.

In radar imaging, information is typically surface-related. Sea bottoms, urban or agrarian areas are imaged from the surface reflection or backscatter of typically long wavelength pulses. For example, RADARSAT, a Canadian advanced earth observation satellite project uses 5.6 cm wavelength sound waves to monitor environmental change and support resource sustainment. These "longer" wavelengths are used since the sound waves have to propagate long distances trough clouded air, tropical showers or troubled water.

In US imaging, on the contrary, the information is typically related to organ tissue structures located near the skin surface and the sound pulses are of very short wavelengths (below 1 cm), resulting in high frequency (typically 2 to 40 MHz) waves[1].

[1] As a reference, the human ear is sensitive up to 20 kHz, which explains the *ultra*sound terminology.

Figure 2.1: Left: an US machine. Right: an US transducer or probe.

To form an US image, the tissue or organ of interest is insonified using a specific transducer. Subsequently, the sound reflections, generated through tissue interaction during wave propagation, are registered by the US machine in function of time. The time elapsed between the emission of the pulse and reception of its echo is a function of the depth of the reflecting tissue or structures. The intensity of the echoing waves corresponds to the acoustic properties of tissue interfaces. Ultimately, the registered reflections are converted into a digital image. Fig. 2.1, shows both a typical US machine (left) and a transducer emitting and receiving the sound pulses (right).

In the following subsections, we discuss the major ultrasound imaging components. We start with the physics and constitution of the transducer in Subsection 2.1.1. The transducer generates pulses that are subject to reflection. How this reflection is characterized is discussed in Subsection 2.1.2. As any other radio pulse, ultrasound waves are also attenuated as they penetrate the tissue. This attenuation is discussed in Subsection 2.1.3. Upon reception of the reflections, the system's axial resolution is decisive for the image formation. This is discussed in Subsection 2.1.4. Finally, in Subsection 2.1.5 the conversion of the received echoes along a single scan line is described, introducing the concept of speckle, and extended to a sequence of scan lines resulting in the real 2D US image.

2.1.1 The transducer

The *transducer* or *probe* serves both as the transmitter and receiver of the sound pulses and is placed on the body surface. As such, the transducer can be considered both the mouth and ear of the US machine. Basically, it consists of piezo-electric crystals that convert electrical energy into mechanical energy.

The piezo-electric crystals contain electrical dipoles, oriented in the certain direction. Compression of the crystal changes the orientation of the dipoles,

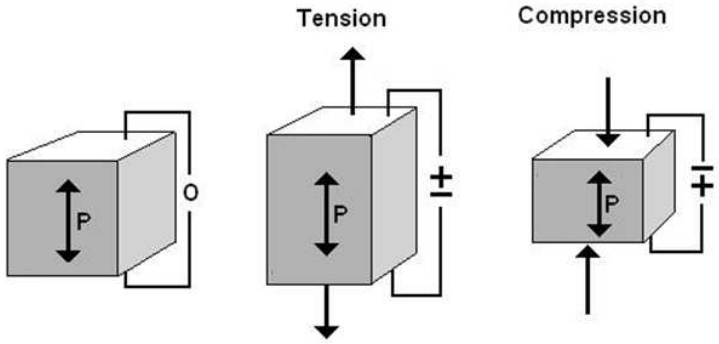


Figure 2.2: The Piezo-electric effect. Left: a piezo-electric crystal with no stress applied. Middle: tension on or expansion of the crystal inducing a potential difference between both sides of the material. Right: compression of the crystal inducing inverted polarization.

inducing a potential difference between both sides of the material. Tension of the piezo-electric material results in an potential difference of reversed polarity. As such, the material will act as a receiver and produce an alternating voltage when subjected to an acoustic pressure wave. This is called the *direct piezo-electric* effect, see Fig. 2.2.
Conversely, applying an electric field across the material results in either compression or expansion, depending on the polarity of the field, which results in the production of a mechanical sound wave. This is called the *inverse piezo-electric* effect.

There are two basic ways of inducing the piezo-electric material to vibrate. One way is to apply a sinusoidal electric potential across the material. Another way is to apply a short electrical pulse across the material, which causes it to vibrate at the resonant frequency. The latter is like hitting a bell with a hammer which will cause the bell to resonate at a certain frequency. The smaller the bell the higher the resulting frequency and vice versa.
In the case of a piezo-electric slab the resonant frequency is determined by the thickness of the slab, being equal to half of the wavelength of the sound within the piezo-electric material [Hughes, 2001].

For example: suppose we want to build a transducer that emits US waves at 8 MHz, and we know the velocity of sound in the piezo-electric material v is about 4000 m/s. The wavelength corresponding to 8 MHz then is

$$\lambda = \frac{v}{f} = \frac{4 \times 10^3}{8 \times 10^6} = 5 \times 10^{-4}\ m = 0.5\ mm. \tag{2.1}$$

So, in order to emit at the selected frequency we need a slice of piezo-electric material that is 0.25 mm thick.

A sound pulse emitted by the piezo-electric material typically consists of 3 to 4 cycles (with wavelength λ) and takes less than 1 μs to produce. During the time a pulse travels through the body, the transducer listens to the reflected waves instead of immediately emitting another pulse. The duration between two pulses is usually chosen between 250 μs and 500 μs. It is important to note that during insonation the *duty* factor of the transducer, i.e., the time the transducer actively transmits sounds, is typically less than 1% of the total time. This means about 99% of the time is spent listening for reflections.

2.1.2 Reflection of the sound waves

We are all familiar with echoes. When shouting your name from a mountain top, you often here the echo down in the valley. This echoing differs when you are inside of a cave, a room full of soft furnishing or concrete walls as carpets seem to absorb sound whereas concrete walls reflect almost all sound. It is the constitution of these materials that is responsible for the amount of reflection.

Exactly the same holds for wave propagation through the body. Sound wave reflection occurs at interfaces of tissues with a different *acoustic impedance*, and on small inhomogeneities with the size of approximately the wavelength (like cell conglomerate and small blood vessels), called *scatterers*. The acoustic impedance of a material Z is defined as the product of its density ρ and the velocity of sound v in the material

$$Z = \rho v. \tag{2.2}$$

The bigger the acoustic impedance of a material, the less sound will penetrate it. The percentage of sound energy reflected at an interface of two materials with different impedances Z_1 and Z_2 is called the *reflection coefficient* R and is defined as follows[2]:

$$R = \Big(\frac{Z_2 - Z_1}{Z_2 + Z_1}\Big)^2. \tag{2.4}$$

The percentage of sound energy transmitted is called the *transmission coefficient* $T = 1 - R$. The bigger the difference in acoustic impedance at the interface of two tissues, the stronger the reflection of an incident sound wave will be. In Table 2.1, an overview of the acoustic impedance of various tissues is presented. There, the particularly low value for air is noticeable. Sound is

[2] In general, when the incidence angle of the incoming wave is not equal to zero, equation 2.4 is given as follows

$$R = \Big(\frac{Z_2 \cos\theta_i - Z_1 \cos\theta_r}{Z_2 \cos\theta_i + Z_1 \cos\theta_r}\Big)^2 \tag{2.3}$$

where θ_i and θ_r are the incidence and refraction angles respectively.

Table 2.1: This table shows the density, velocity of sound, acoustic impedance Z and attenuation coefficient α (at 1 MHz) for different materials.

Material	**Density** ($10^3\ kg\ m^{-3}$)	**Velocity** ($m\ s^{-1}$)	$\boldsymbol{Z}$ ($10^3\ kg\ m^{-2}\ s^{-1}$)	α ($dB\ cm^{-1}$)
Air	0.0012	330	0.0004	1.38
Water	1	1430	1.43	0.0025
Soft tissue	1.1	1540	1.69	0.5-1.0
Liver	1.05	1570	1.65	1.1
Fat	0.95	1450	1.38	0.6
Bone	1.91	4080	7.8	10.0
Aluminum	2.7	6420	17	0.021

known to reflect almost entirely at a tissue-air interface. The reflection coefficient R_{at} of a tissue-air transition is:

$$R_{at} = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1}\right)^2 = \left(\frac{1.63 - 0.0004}{0.0004 + 1.63}\right)^2 = 0.99. \tag{2.5}$$

This means 99% of the sound energy is reflected and only 1% is transmitted. Because of this strong reflection, physicians always put some coupling gel on the transducer surface during investigation to avoid air between the probe and the skin. This coupling gel provides an acoustic pathway between the transducer and the skin. For a maximum acoustic transmission the coupling gel must have an acoustic impedance between that of the transducer surface (probe to coupling gel interface) and the skin (coupling gel to skin interface) and is usually chosen between 1.45 and 1.56. This results in typical transmission coefficients of around 99.9% [Sonotech, 2003].

2.1.3 Attenuation of the sound waves

As with most radiation processes the intensity or amplitude of a pulse is attenuated exponentially as it passes through a medium, i.e., a fraction of the energy is removed per unit length of travel. When an US pulse passes through a medium its amplitude A is exponentially attenuated according to the equation

$$A = A_0 e^{-\alpha x}, \tag{2.6}$$

where A_0 is the initial amplitude, α is the attenuation coefficient and x is the traveled distance. The attenuation coefficient α varies with the frequency f of the pulse and is approximately proportional to f^2 for water and $f^{1.2}$ for soft tissue. Table 2.1 also shows the attenuation coefficients of some tissues and materials.

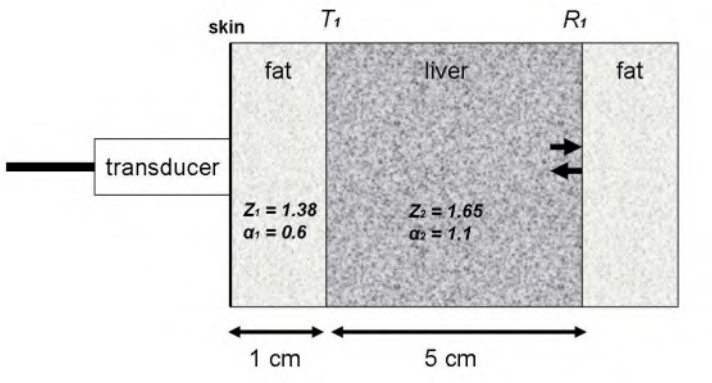


Figure 2.3: Example of an US pulse reflection.

Now that reflection and attenuation are defined we are able to calculate the amplitude of an US pulse returning from inside the body. Assume the simplified[3] model in Fig. 2.3. Pulses travel through a (1 cm) fat layer followed by a (5 cm) layer of liver. The sound reflects off the far liver-fat interface and travels back to the transducer. The amplitude A of the returned pulse is then calculated as:

$$\begin{aligned} A &= A_0 e^{-\alpha_1 x_1} T_1 e^{-\alpha_2 x_2} R_1 e^{-\alpha_2 x_2} T_1 e^{-\alpha_1 x_1} \\ &= A_0 T_1^2 R_1 e^{-2(\alpha_1 x_1 + \alpha_2 x_2)}. \end{aligned} \tag{2.7}$$

The initial amplitude A_0 is first attenuated by the fat (α_1). Part of the pulse is transmitted by the tissue interface[4] (T_1). Subsequently, the pulse is attenuated by the liver itself (α_2). A reflected part (R_1) is again attenuated by the liver (α_2) and transmitted at the interface (T_1) before being attenuated a last time by the fat layer (α_1). Based on the numbers in Table 2.1 we get $A/A_0 = 3.47 \times 10^{-8} \simeq -75$ dB. This means the attenuation of the fat and liver results in an amplitude drop of about 75 dB.

2.1.4 Axial resolution

Apart from reflection and attenuation, there is one more aspect we need to consider before moving on to the actual imaging, namely the reception precision of the transducer also called the *axial resolution*. This resolution is directly proportional to the sound frequency.

[3] In reality multiple layers are passed and reflections off the skin-gel interface are also neglected for simplicity reasons.

[4] A part is also reflected but we will also neglect this for simplicity reasons.

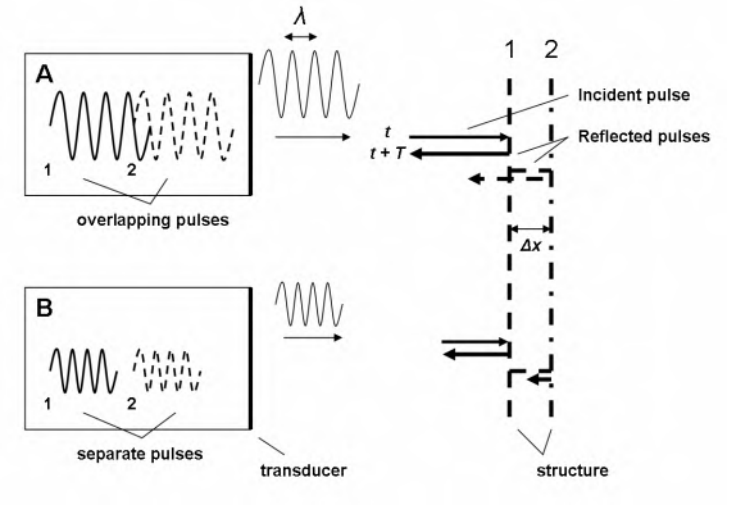


Figure 2.4: Schematic diagram depicting the relation between frequency and axial spatial resolution.

Suppose we have a transducer producing pulses at two different frequencies, "low" and "high", see Fig. 2.4. The low frequency pulse is called pulse **A**, the high frequency pulse, pulse **B**. Both pulses contain 4 cycles of wavelength λ (for the moment we do not worry about the value of λ). At the first interface **1** the energy of both pulses, emitted at the same time, is partly reflected. The rest travels through to the second interface **2** where the same happens. The pulses reflected off interface **2** travel back through interface **1**. The return pulses are also partly reflected at interface **1**, but we need not worry about that for the purpose of this discussion.
If we know the time it has taken the pulses to return to the transducer and consider the speed of sound being constant, we can calculate the distance traveled by each pulse and draw a plot representing the distance from the probe. If we also know the initial amplitude, we can even calculate the received amplitude, as was done in equation (2.7).

As can be seen from the diagram in Fig. 2.4, on arrival in the transducer the reflections of tissue **1** and **2** will be closer together in the case of pulse **A** than in the case of pulse **B**.
If the reflections overlap very strongly they will even appear as one. In that case, we will be unable to tell if the sound has reflected off two interfaces or not. So the question will be, for any given frequency, what is the smallest separation between two interfaces we can resolve?

From our diagram, we see that if the tail of the first reflecting pulse has not left interface **1** before the head of the second reflecting pulse started passing through interface **1**, the two pulses will overlap. This comes down to the

distance between interface **1** and **2** (Δx) being at least half the pulse length. When we poor this into an equation, we get

$$\Delta x = \frac{n\lambda}{2} = \frac{nv}{2f}, \tag{2.8}$$

where n is the number of periods in an ultrasound pulse, v is the velocity of sound in a medium and f is the frequency. This means that we can improve the *axial resolution* by either decreasing the number of periods in a pulse or by increasing the frequency.

Consider again the case of an 8 MHz transducer. Also assume that the speed of sound in human tissue is constant and equal to about 1540 m/s. When we have a 4 period pulse, then the axial resolution is

$$\Delta x = \frac{nv}{2f} = \frac{4 \times 1540 \times 10^2}{2 \times 8 \times 10^6} = 0.4\ mm \tag{2.9}$$

meaning that structures more than 0.4 mm apart are distinguishable at this frequency.

Note however, that in Section 2.1.3 we stated that the attenuation of the sound in human tissue increases with the frequency ($f^{1.2}$). This leads to a smaller penetration depth of the sound waves. As a result, the frequency choice is a trade-off between the necessary penetration depth and the desired axial resolution.

In medical US imaging, for different organs, depending on their shape and position in the body, different transducers with different frequencies are chosen. Typical frequencies are between 2 and 40 MHz. For example, in the neonatal brain it is most common to use a 7.5 to 8.5 MHz transducer. From what we computed, an 8 MHz transducer, fitted for neonatal US will have a piezo-electric slab thickness of 0.25 mm and an axial resolution of 0.4 mm.

2.1.5 Ultrasound image formation

Now that we defined all elements influencing the sound waves, we are ready to explain how the actual US *image* is constructed.

2.1.5.1 The Radio-Frequency signal

Consider the simplest case of a transducer consisting of one piezo-electric crystal insonating a homogeneous medium containing 4 pointlike scatterers, all of which lay within the axial resolution Δx.[5] When insonified, these scatter-

[5]This is a normal assumption since in practice we can not both penetrate the tissue as deeply as we want and assure an optimal axial resolution at the same time.

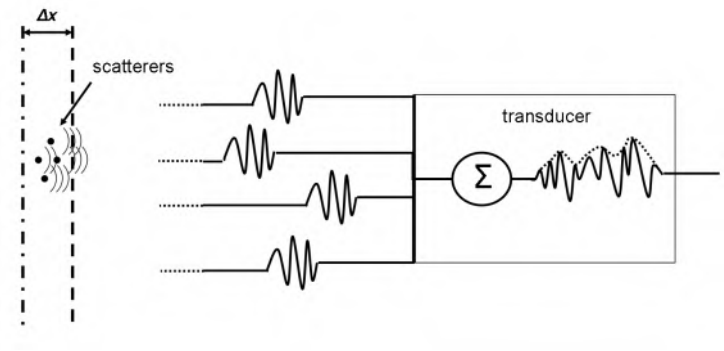


Figure 2.5: Interference of backscattered sound waves from 4 scatterers within the critical axial resolution Δx (Source: [Thijssen and Oosterveld, 1990]).

ers usually yield spherical waves that will arrive at the transducer at slightly different times, see. Fig. 2.5.

Upon reception, the transducer will transform the echoes in an electrical *Radio-Frequency signal* (RF-signal) that is the algebraic sum of the instantaneous sound pressures of the backscattered waves, see again Fig. 2.5. This algebraic sum is in fact the interference pattern of the four different scatterers and the dashed line in the figure represents the peaked demodulated echogram. This peaked interference pattern is called *speckle*.
Neither the number nor the amplitude of the peaks is directly related to the number of scatterers in the tissue. They rather reflect the distribution of the scatterers within the resolution cell (Δx).

2.1.5.2 Beam forming and 2D probes

Initially, US technology was based on this single rotating piezo-electric element, often called the lighthouse variant of US imaging. The resulting 1D-linescan is also called the A-mode (Amplitude-mode) echo. Nowadays, transducers are composed of a *linear* or *curvilinear* array of piezoelectric elements, see Fig 2.6 a) and b). The field of view of a linear array is small and limited to the footprint of the probe. A curvilinear array has a curved surface, creating a field in the depth that is wider than the footprint of the probe however at the cost of a reduced *lateral resolution*, i.e., the resolution perpendicular to the line-scans.

The array of piezo-electric elements is excited in a certain sequence in order to obtain *beam forming*. This can be done in two different ways. In the *sequential* method, the array elements shoot beams one after the other. In the *phased-array* method, probes excite *all* transducer elements of the transducer array, creating a wavefront. Usually, small delays are inserted between pulses in neighboring elements so that all pulses arrive in phase at a specific point.

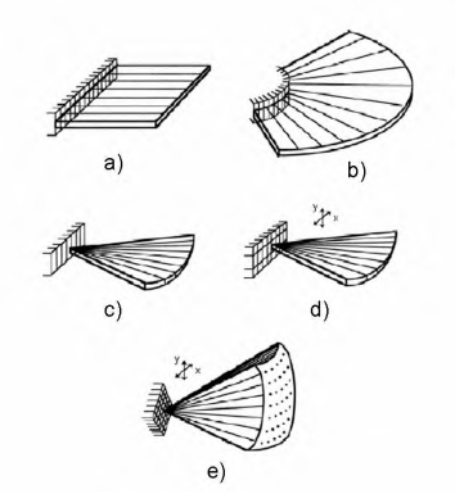


Figure 2.6: Arrangement of commercially available array transducers. a) a linear sequential array. b) a curvilinear array. c) a linear phased array. d) a 1.5D array which is phased along the X-axis and focussed along the Y-axis. e) a 2D phased array is shown for 3D US imaging. (Source: [Mischi, 2004]).

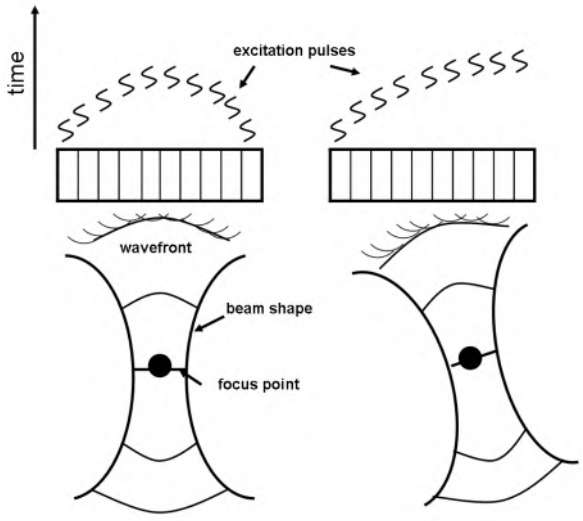


Figure 2.7: Left: a small delay profile of the set of induced pulses creates a wavefront that is focussed at one specific point. Right: A different delay profile not only focuses the wavefront but also steers it in a given direction.

This point is called the *focal point*. All sound waves converge there resulting in a maximal amount of reflection information, i.e., both the lateral and axial resolution are optimal there. Furthermore, by varying the delays, we can also steer the beam in any desired direction. Fig. 2.7 shows different excitation patterns of a phased-array probe, resulting in different beam form focusing and steering.
Because of its beam forming possibilities a very small phased-array transducer can image a large area in the far field. That is why these are the transducers of choice in applications like cardiac imaging where one has to deal with the small spaces between the ribs through which the much larger heart needs to be imaged. Fig. 2.6 c), d) and e) shows some examples of phased-array probes.

In a similar way as in the linescan, the transducer now sums all received pulses into an RF-signal containing the plane's speckle pattern. The result of these reflections is called the B-mode (Brightness-mode) echogram.

2.1.5.3 The actual ultrasound image

To be able to convert the RF-signals into an image, several transformations in the US machinery are needed. The block diagram in Fig. 2.8 shows the entire track of the RF-signal from transducer to the image on screen.

The first RF-signal transformations are the overall gain compensations. These include the *Gain* and *Time Gain Compensation* which serve as amplifications of the signal. Subsequently the signal is, as mentioned earlier, demodulated before it goes on to the most important transformation, the log-compression. This compression is needed to reduce the dynamic range of the input signal to match the (lower) dynamic range of the display device and the human visual system. The typical dynamic range of the input signal is in the order of 50-70 dB, whereas a typical display would has a range of the order of 20-30 dB.

Finally, a scan converter transforms the compressed signal into the appropriate (4 to 8 bit) greyscale image. Fig. 2.9 and Fig. 2.10 show two examples of a conical-shaped image and rectangular shaped US image. We notice that the constructive and destructive interference in the *RF-speckle* pattern is transformed in a granular bright dot pattern called *image speckle*.

2.2 Specific speckle characteristics

By now, it is clear that speckle resulting from the RF-signals is the building block of the US image. As such, its nature has been a major subject of investigation. In this Section, we briefly review the characterization of speckle in the RF-signal, Subsection 2.2.1. Following this, in Subsection 2.2.2 we explain how speckle patterns in the US images reflect actual tissue characteristics and how they can be used in speckle-reduction filters.

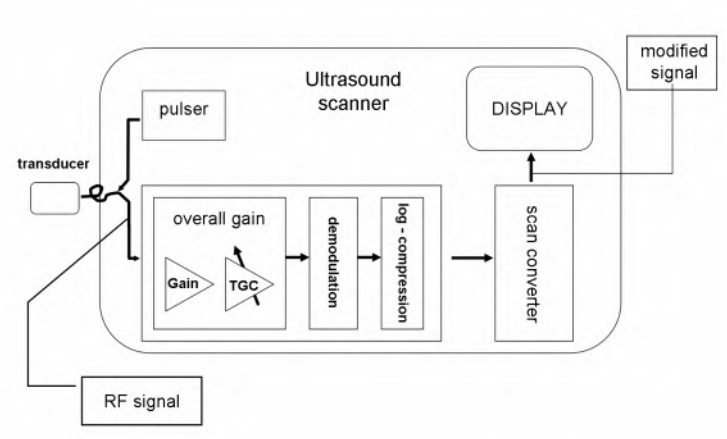


Figure 2.8: Block diagram of the processing steps of the RF-signal inside the US scanner (Source: [Stippel, 2004]).

Figure 2.9: A conical US images of the preterm brain, captured with a 8.5 MHz curvilinear probe.

Figure 2.10: A linear US image captured with a 13 MHz linear probe.

2.2.1 Speckle in the Radio-Frequency signal

From the way speckle is defined, it should be clear that when a fixed, rigid object is scanned twice under exactly the same conditions, we obtain exactly the same interference patterns, hence exactly the same speckle pattern. However, in practice it is almost impossible to acquire exactly the same image twice since the scatterers in human tissue are continuously in motion. Consequently, researchers will rather study the *distribution* of the speckle peaks, which reveals info on the density of the scatterers, rather than characterize speckle/scatterers individually.

A distinction is typically made between *fully developed* speckle, *partially developed* speckle and *low scatter density* speckle *with structural components*. We do not go into detail on the exact meaning of these terms, yet intuitively they are related to the distribution and density of the scatterers they result from. One example: speckle that is formed from uniformly distributed backscattered amplitudes reflecting of spatially uniformly distributed high density scatterers (> 10 per resolution cell), is called "fully developed" speckle.

Different mathematical models that characterize the various first-order statistics (amplitude distributions) of the RF-signals, as function of the different scatterer distributions in a resolution cell Δx, are proposed throughout literature:

- The first-order statistics of the RF-signal for fully developed speckle follow a Rayleigh distribution [Wagner et al., 1983].
- The first-order statistics of the RF-signal for partially developed speckle follow a Rice distribution [Jakeman and Pusey, 1976].

- The first-order statistics of the RF-signal for low scatterer density with structural components follows a homodyned K-distribution [Dutt, 1995].

We refer to the cited papers as well as to [Stippel, 2004] for a broader overview and the exact technical details on these distributions.

First-order RF-signal statistics are commonly used to investigate the statistical properties of backscattering [Wang and Shung, 1997]. By distinguishing different distributions in the RF-signal we can distinguish structured scatterers from non-structured scatterers.
A direct application is the characterization of pathological tissues in medical diagnosis, where different tissues consisting of different scatterer distributions yield different RF-characteristics [Ishii et al., 2003, Anuja et al., 2002, Jeremias et al., 1999]. Besides that, first-order RF-signal information is also used in algorithms that compensate for machine settings [Xiao et al., 2002].

2.2.2 Speckle in the ultrasound images

Apart from first-order characteristics extracted from the RF-signal, second-order speckle statistics like the size and spatial correlations of speckle are usually extracted from the US images.

In [Thijssen and Oosterveld, 1990] it is shown that the speckle size depends strongly on the frequency (higher-frequency results in better spatial resolution), and the geometry of the employed transducer (linear or curvilinear). Furthermore, they show that the attenuation by the insonated tissue yields a depth-dependent increase of mainly the lateral speckle size, in addition to the depth-dependency caused by the beam formation (usually focussed on one spot and then diverging).

The main application of image speckle characterization is speckle-reduction. Some well-known filters that either use the first- or second-order image speckle characteristics are:

- The classical filters of [Lee, 1980] and [Frost et al., 1982] based on first-order image statistics.
- Techniques by [Gupta et al., 2004, Achim et al., 2001] and [Sattar et al., 1997] reduce speckle in a multi-resolution wavelet-based approach.
- In [Stippel, 2004] several morphological, textural and single-speckle identifying filters are presented.

Again, the cited papers contain all the technical details. Later on, we will also present our own modification of an existing multi-resolution wavelet-based approach called the GenLik filter [Pizurica et al., 2003].

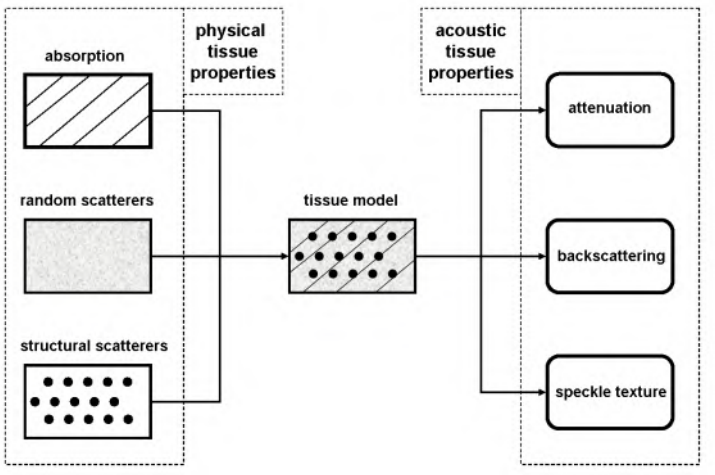


Figure 2.11: The usual tissue model in ultrasound imaging (Source: [Thijssen and Oosterveld, 1990]).

The speckle pattern in the US images can also be used to characterize tissue properties. The most prevalent model in literature to explain the effect that occurs when human tissue is insonified is explained in Fig. 2.11. This figure shows the correlation between the physical tissue properties and their corresponding acoustic properties. Tissue is modeled as a sound *absorbing* medium containing microstructures that scatter the sound. These scatterers are both *random and structured inhomogeneities* approximately equal to the wavelength of the sound.

The corresponding US equivalents for absorption, random inhomogeneities and structural scatterers are the following: as described in section 2.1.3 the sound absorption by the tissue results in the attenuation of the sound beam. Random inhomogeneities result in diffuse reflection and backscattering. Most importantly however, the presence of structural scatterers results in *speckle texture patterns*. This is the key to quantitative US tissue characterization.

Although the US image does not depict the real tissue characteristics, the tissue structure is visible in the speckle texture. Note also that according to [Wagner et al., 1983] the backscattering caused by random inhomogeneities is also considered as valuable information in medical tissue characterization. Therefore, speckle-filtering of any kind is discouraged in US tissue characterization.

What [Thijssen and Oosterveld, 1990] also show is that for fully developed speckle the tissue characteristics are exclusively reflected in the first-order characteristics, not in the speckle size. If the density of scattering sites within the tissue is relatively low, the speckle characteristics are dependent on the second-order statistics and tissue characteriza-

tion is feasible using mathematical methods that analyze these. Numerous different tissues such as the liver [Kadah et al., 1996, Sun et al., 1996], the prostate [Basset et al., 1993, Schmitz et al., 1994, Huynen et al., 1994] and carotid plaques [Christodoulou et al., 2003] have already been described based on these second-order speckle texture characteristics.

In the Chapter that follows, we quantify how pathological (PVL) white brain matter can be distinguished from normal white brain matter, exactly by investigating an exhaustive set of first- and second-order characteristics of the manifested speckle pattern. We overcome the depth-dependency of the speckle characteristics by always comparing tissues from the same regions in the US image. Besides that, we do not apply any form of speckle-reduction, maintaining the optimal amount of (clinical) information in the image. Also, we consistently use the same transducer at a fixed frequency and (focus) depth during acquisition, so we do not have to bother about their effects either.

However, during scanning physicians alter typical machine settings to obtain an optimal image. So, what remains before we can move on to the actual tissue characterization, is to describe the effect of these machine settings on the speckle formation.

2.3 Influence of the US machine settings

As we moved from the RF-signal analysis to the *transformed* greyscale image analysis, the influence of the signal processing in the US machine becomes significant. Up to now, we only briefly explained the role of the US machinery, namely in Section 2.1.5 where we discussed the transformation of the RF-signal into an actual greyscale image.

Machine settings are routinely altered by the medical experts to obtain an image of optimal diagnostic quality. While these settings produce images of an optimal visual quality, their variation influences the image greyvalues directly, disturbing most quantitative image analysis.

First of all, there is the *Gain* option by which the physician amplifies the overall RF-signal. Due to attenuation, we know that as sound waves move through the tissue they loose energy. As such, the expert has the option to amplify the reflected, attenuated signal.
Secondly, reflected signals returning from deeper in the tissue, travel a longer way and hence are more attenuated. In order to obtain a uniform image the level of amplification can be depth-adaptive. Therefore, commercial US machines provide a *Time Gain Compensation (TGC)* option to manually adjust the gain per depth segment or region. Mostly, the Gain is tuned by one single button or slider, whereas the TGC is tuned by setting multiple sliders. As an example of this amplification, Fig. 2.12 shows two images of the same phantom with different Gain settings.

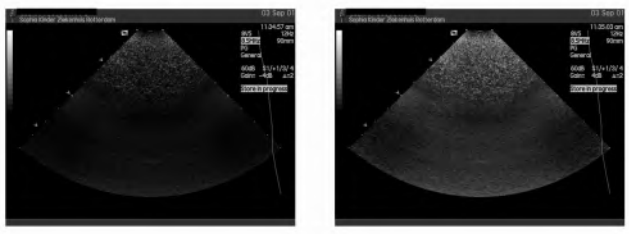


Figure 2.12: Left: a phantom image captured at a Gain setting of -4 dB. Right: the same phantom captured at a Gain setting of +4 dB.

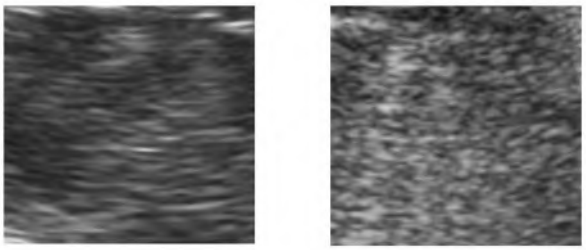


Figure 2.13: US brain images captured with two different US machines.

In order to compare US images quantitatively, e.g., by looking at the greyscale speckle texture patterns, we either have to assure all images are captured with the same settings or have to compensate for these settings, creating a *normalized* image. Compensation algorithms that construct such normalized images have already been proposed in the past [Xiao et al., 2002, Simaeys et al., 2000]. These algorithms incorporate assumptions or models on the way the US machine (trans)forms the images.

Although these compensation algorithms fit their purpose well, in most cases they are heuristic as most manufacturers, for obvious economical reasons, don't provide the exact signal transformations they apply. Apart from the obvious transformations in Fig. 2.1, speckle suppression techniques, higher-order harmonics, interpolation, frequency compounding or even digital image postprocessing are applied.

Since each manufacturer provides unique machinery this also results in machine-dependent speckle. Fig. 2.13 shows an US brain region captured with two different machines. We see that although the same kind of tissue is imaged, speckle patterns differ.

Chapter 3

Texture-based pattern recognition

The surface of most natural objects shows a certain (ir)regular, periodic or stochastic pattern referred to as visual texture. In image processing, texture serves as a strong object characteristic in many pattern recognition problems. In this Chapter we show quantitatively how texture information can be applied to tissue characterization in medical US diagnosis. This results in a multi-feature multi-classifier algorithm for the detection of pathological PVL white brain matter.

3.1 Introduction

In image processing, pattern recognition is defined as the *automatic identification* of objects or patterns in an image based on *characteristic features* such as shape, length, color, outlines, or texture. To illustrate this, we start with a simple example.

Suppose company X produces tuna fish and salmon salad. A conveyor belt is installed, on which both the salmon and tuna fish enter the factory. To speed up the production process, the selection of the tuna fish and salmon is done by cameras that visually inspect the fish that pass on the belt. Every time a fish passes under a camera an image is taken and the length of the fish, size of the head and color of the scales are measured using image processing algorithms. Depending on these features, the fish are then automatically sorted in 2 different baskets for salad processing.

Obviously, the manufacturer will want to classify as many fish as possible, as fast as possible. Besides that, it is clear, considering the price they pay,

consumers usually will not complain if there is salmon in their tuna fish salad, whereas if there is too much tuna in their salmon salad they might. As such, it is very important for the camera system to function without (too many) errors and if incorrect classifications occur they should be in favor of the tuna fish salad for the obvious economical priors.

To achieve all this, the system has to control multiple parameters. First of all, the captured images have to be of an acceptable quality (color and resolution in this case) in order to measure in an optimal way. Secondly, possibly multiple fish have to be detected/segmented in each image. Thirdly, the (few) features measured from the segmented fish have to be representative for both salmon and tuna fish. A good system training on a large set of salmons and tuna fish has to assure this. Finally, a set of rules is needed to classify the fish.

From this example, we can now derive the three major building blocks pattern recognition typically consists of:

1. **Data-acquisition and preprocessing.** A representative set of images or objects is crucial to any pattern recognition problem. We have to assure the data represents all image or object categories we want to recognize by a sufficient number of well-chosen examples.
   Often a preprocessing step is needed to condition the data correctly. This preprocessing can range from the simple definition or delineation of regions of interest in an image to a more advanced removal of acquisition noise and compensation for machine settings.

2. **Feature extraction and selection.** Features are numerical measurements that simplify the object's characteristics and, if selected properly, speed up the recognition process. They are considered the signatures of the object or image by which we can distinguish them from other objects. Extracting the object or image features accurately, but also selecting the most relevant ones amongst them, are crucial for the recognition outcome.

3. **Classification or decision algorithm.** Automatically grouping or clustering the features so that different object or image classes can be formed is the task of a classification of decision algorithm. Often within a probabilistic framework taking into account prior probabilities, these algorithms label an object or image according to the class or group its features are most likely to belong too.

The way these building blocks are organized depends highly on the application. In this Chapter, we focus on an application of texture-based classification. More precisely, the quantitative analysis of pathological white brain matter texture in preterm US images. This application is very challenging due to the presence of speckle that both degrades the visual image quality, hindering the expert's diagnosis, and contains the relevant information on tissue structures, as discussed in the previous Chapter.

Apart from speckle, another challenge is the possible (medical) consequence of missing pathological samples. By this we mean that in certain medical applications where a follow-up or staging of a pathology over a longer period (possibly followed by a later treatment) rather than an actual immediate treatment is important, the performance of a pattern recognition system is not only measured by the total number of misclassified samples (pathological and non-pathological) but also by the number of pathological samples that are misclassified.

When we consider a non-pathological sample initially as pathological, follow-up might reveal the sample is non-pathological after all and no harm is done. However, if we consider a pathological sample initially as non-pathological, we risk not monitoring the sample which might have serious implications, such as later treatments becoming ineffective or no longer possible. Since the main purpose of investigating white matter damage in the preterm brain is to be able to diagnose as early as possible in view of the right long-term guidance (i.e., non-invasive physical or mental stimulation), this comment is important.

The outline of this Chapter is as follows: since our pattern recognition is built around texture analysis, in Section 3.2 we start with an introduction to image texture. We define/describe texture as well as different mathematical texture descriptors. Subsequently, Section 3.3 introduces the basic concepts of classification. In Section 3.4 we present our own work on classification. We combine 7 texture feature extraction methods and 3 different classifiers in a multi-feature multi-classifier US brain tissue algorithm that outperforms the state of the art. The overall conclusions of the Chapter and hints for future improvements are presented in Section 3.5.

3.2 Texture

Artificial objects or images are typically characterized by uniform color or greyscale regions, e.g., the simple cartoon picture in Fig. 3.1 (left). Images of real objects, on the contrary, hardly ever contain regions of uniform intensities, e.g., the wooden surface in Fig. 3.1 (right) contains intensity variations that result in an irregular pattern referred to as *visual texture*.
These texture patterns are the result of either physical surface properties, such as roughness and oriented strands, or reflectance differences such as surface colors [Jain and Tuceryan, 1998].

3.2.1 Definition or Description?

Although people are good at visually detecting image texture it seems non-trivial to define it formally. This is demonstrated by the number of different texture definitions attempted by vision researchers. Some examples of a catalogue of texture definitions compiled in [Coggins, 1982] are presented here:

Figure 3.1: Left: an artificial (simple) cartoon image showing only uniform regions, i.e., without texture. Right: a natural scene containing tree bark texture.

- "We may regard texture as what constitutes a macroscopic region. Its structure is simply attributed to the repetitive patterns in which elements or primitives are arranged according to a placement rule." [Tamura et al., 1978]
- "A region in an image has a constant texture if a set of local statistics or other local properties of the picture function are constant, slowly varying, or approximately periodic." [Sklansky, 1978]
- "The image texture we consider is non-figurative and cellular... An image texture is described by the number of types of its primitives and the spatial organization or layout of its primitives... A fundamental characteristic of texture: it can not be analyzed without a frame of reference of the primitive being stated or implied. For any smooth gray-tone surface, there exists a scale such that when the surface is examined, it has no texture. Then as the scale increases, it takes on a fine texture and then a coarse texture." [Haralick, 1979]
- "Texture is an apparently paradoxical notion. On the one hand, it is commonly used in the early processing of visual information, especially for practical classification purposes. On the other hand, no one has succeeded in producing a commonly accepted definition of texture. The resolution of this paradox, we feel, will depend on a richer, more developed model for early visual information processing, a central aspect of which will be representational systems at many different levels of abstraction. These levels will most probably include actual intensities at the bottom and will process through edge and orientation descriptors to surface, and perhaps volumetric descriptors. Given these multi-level structures, it seems clear that they should be included in a definition of, and in a computation of, texture descriptors."[Zucker and Kant, 1981]
- "The notion of texture appears to depend upon three ingredients: (i) some local order is repeated over a region which is large in comparison to

Figure 3.2: Left: the periodic texture visible in a floor pattern. Each vertical strip of floor tiles can be considered as a texel. Middle: the irregular or chaotic texture of grass where it is non-trivial to define texels. At best, each individual blade of grass could be considered as one texel. Right: the stochastic texture of a cloud pattern where texel definition is even impossible.

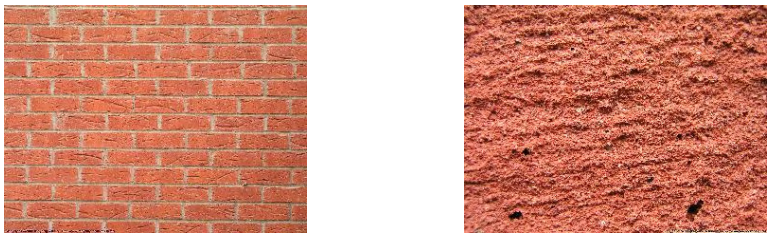


Figure 3.3: Left: periodic texture pattern of a brick wall viewed at on a coarse scale. Right: irregular texture of one brick looked at on a very fine scale.

> the order's size, (ii) the order consists in the non-random arrangement of the elementary parts, and (iii) the parts are roughly uniform entities having approximately the same dimensions everywhere within the textured region."[Hawkins, 1969]

Although we notice that texture definitions clearly differ in nature (perceptually motivated by some, whereas others hold a more application-driven approach), two components are consistently present throughout most: the *primitive texture element* and the *scale-dependency*.

The primitive texture element, called *texel*, is considered the basic texture building block. It is the ordering or arrangement of these texels that leads to the texture patterns. For regular or periodic textures it is usually fairly easy to detect texels, whereas in chaotic, irregular or stochastic textures this is far less trivial or even impossible, see Fig. 3.2.

The scale dependency is related to the texels. When observing images at different scales, different textural structures are revealed. For example, when observing the brick wall in Fig. 3.3 from a far-away distance, i.e., on a coarse scale, we notice the regular texture pattern of the bricks (texels). However, when we zoom in on one brick of the wall, i.e., look at a very fine scale, we notice the totally different irregular pattern of the brick itself. Consequently,

selecting the appropriate texture scale depends on the desired result.

Based on the same two characteristic components, we opt for a texture *description* rather than a texture *definition*. We describe image texture in terms of the following set of properties:

1. Texture is a local image characteristic, meaning it is measured over a certain area or neighborhood of pixels rather than on individual pixels.
2. Texture changes according to the scale it is looked at. The scale selected depends on both the texture structure and the desired result.
3. Texture can only be characterized or perceived well when the number of texels present in the image or image region is large enough. When few texels are present, these constitute a group of countable objects rather than texture.
4. Texture regions are characterized by a specific distribution or correlation of pixel values. Thus methods accounting for neighborhood correlations are effective texture analysis tools.

The first 3 properties are directly related to the primitive texture element and scale-dependency of the texture as discussed above. The fourth property is related to the conversion of image texture into numerical features, as we will discuss next.

3.2.2 Texture descriptors

In actual pattern recognition applications, we need practical ways of handling image texture. First of all, perceived texture *qualities* are identified. Some of the most common perceived qualities are the uniformity, entropy, contrast, periodicity, homogeneity, linearity and directionality of a texture. Then, these qualities are transformed into numerical measures called *texture features*.
As the fourth property of our texture description suggests, texture qualities are reflected by the spatial distribution of the pixel values. Methods that quantify these distributions are called *texture descriptors*.

In what follows, we present a short overview of important state-of-the-art texture descriptors.

3.2.2.1 First-order descriptors

Features derived from the image grey value histogram describe the *first-order* statistics of a texture. Typical examples are the mean, median and standard deviation of the pixel values. Since these features do not incorporate any information on the spatial distribution of the pixel values, they are often used in combination with descriptors that do so, called *second-order* descriptors. All of the following descriptors are second-order descriptors.

1	1	2	2
1	1	2	2
1	3	3	3
3	3	4	4

$$P_{1,0} = \begin{bmatrix} 2 & 2 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 3.4: Left: a 4×4 image where numbers represent grey values. Right: the corresponding co-occurrence matrix for distance $d = 1$ and angle $\theta = 0$.

3.2.2.2 Statistical methods

1. ***Co-occurrence matrices.*** Haralick's co-occurrence matrix [Haralick et al., 1976] is one of the oldest texture descriptors. The method originates from Remote Sensing and is based on a matrix representation of the 2D histogram of grey values of pixel pairs located at a specific relative position, expressed by a pixel distance d and orientation θ.
   Denote by I an $M \times N$ pixel image containing G different grey values and by $\triangle x = d\cos\theta$ and $\triangle y = d\sin\theta$. Furthermore, the grey value of a pixel (m, n) is denoted by $f(m, n)$. The entry $P_{d,\theta}(i, j)$ on position (i, j) in the $G \times G$ co-occurrence matrix is then calculated as:

   $$P_{d,\theta}(i, j) = \#\{(m, n) | f(m, n) = i \wedge f(m + \triangle x, n + \triangle y) = j\}, \quad (3.1)$$

   where # represents the cardinality of a set. As such, the entry $P_{d,\theta}(i, j)$ is the cardinality of the set of image pixels (m, n) for which both $f(m, n) = i$ and $f(m + \triangle x, n + \triangle y) = j$. As an example, consider a 4×4 pixel image containing 4 different greyvalues, see Fig. 3.4 (left). The co-occurrence matrix for $d = 1$ and $\theta = 0$ is then computed by scanning all pairs of neighboring pixels in the horizontal direction, and increasing the entry (i, j) in the co-occurrence matrix by one for each pair of respective grey values i and j, see Fig. 3.4 (right).

   This co-occurrence matrix is easy to compute and straightforward to derive texture features from, as we will show later on. The downside is that for any given combination of d and θ, the co-occurrence matrix only expresses a small fraction of the texture's second-order statistics and computing matrices for a range of different d and θ is often time-consuming.

2. ***Sum and Difference histograms.*** Unser [Unser, 1986] developed a technique to extract features from the histograms of sums and differences of grey values of pairs of pixels separated by a distance d in a direction θ, as an alternative to the co-occurrence matrix.
   Denote again by I a $M \times N$ pixel image containing G different grey

3	1	2	2
1	1	2	2
1	3	3	3
3	3	4	4

$$P_0 = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Figure 3.5: Left: a 4×4 image where numbers represent the grey values. Right: Run Length matrix for an angle $\theta = 0$.

values. Let (m, n) denote a pixel and $\triangle x = d\cos\theta$ and $\triangle y = d\sin\theta$. Then the pixel sums and differences are calculated as $s_{m,n} = f(m, n) + f(m + \triangle x, n + \triangle y)$ and differences $d_{m,n} = f(m, n) - f(m + \triangle x, n + \triangle y)$, with $f(m, n)$ the grey value of pixel (m, n). The sums $s_{m,n}$ take on values in the interval $[0, 2G]$, the differences in the interval $[-G, G]$. The sum histogram $S_{d,\theta}$ is then defined as $S_{d,\theta}(i) = h_s(i)/R$ with $h_s(i) = \#\{(m, n) \in I | s_{m,n} = i\}$ and $R = \sum_{i=0}^{2G} i$. The difference histogram $D_{d,\theta}$ is computed in a similar way using the differences $d_{m,n}$. Again, as we will show later on, multiple texture features can be computed from these histograms.

3. ***Run Length matrices.*** Another texture descriptor related to the co-occurrence matrix is the Run Length matrix. A *run* is a set of connected pixels of constant intensity on a straight line of a given orientation θ. For a specific θ, the Run Length matrix P_θ is obtained by counting the number of runs of a given length for each grey level. The matrix entry $P_\theta(g, r)$ then denotes the number of runs of grey value g and size r.
An example of a Run Length matrix for $\theta = 0$ is shown in Fig. 3.5 for a 4×4 pixel image, again containing 4 different grey values. The rows in this matrix correspond to different grey values, the columns to different run lengths. We obtain a 4×3 matrix since there are 4 different grey values and the maximum run length in the horizontal direction is of size 2. Note that r starts from the value 0. Again, multiple features can be extracted from the Run Length matrix, as we will show later on. Run Length matrices are very suited for textures that show some regularity or periodicity since, depending on the texel size, typically more long or small Run Lengths will be prominent.

   Before computing the Run Length matrix, grey levels are usually quantized to ensure sufficiently long runs. There are two straightforward ways to alter the number of grey levels. Suppose we want to reduce the number of grey levels G in an image I of size $M \times N$ pixel to G'. We can map each of the initial grey values g to the interval $[0, G' - 1]$ as follows:

$$f_{eq}(g) = \left\lfloor \frac{G'}{MN - h(0)} \sum_{i=1}^{g} h(i) \right\rfloor, \tag{3.2}$$

where $\lfloor x \rfloor$ is the first integer smaller than or equal to x and $h(.)$ is the original grey value histogram. This transform spreads the histogram uniformly, and is also called *histogram equalization*. A second method is by *scaling* the grey values so that the first grey value is set to 0 and the rest gradually move up in the $[0, G' - 1]$ interval as follows:

$$f_{sc}(g) = \text{round}\left[(G' - 1) \times \frac{g - G_{min}}{G_{max} - G_{min}}\right], \tag{3.3}$$

where G_{min} and G_{max} represent the minimum and maximum grey values of the original image.

4. ***Autocorrelation function.*** This function represents how well a shifted version of the texture matches itself. As many textures are periodic or repetitive by nature, the shape of the autocorrelation function is used to assess both texture regularity and the texture coarseness/fineness. Suppose again I is an $M \times N$ pixel image, the autocorrelation function $f_{aut}(x, y)$ at pixel position (x, y) is then defined as:

$$f_{aut}(x, y) = \frac{MN}{(M - x)(N - y)} \sum_{i=1}^{M-x} \sum_{j=1}^{N-y} \frac{I(i, j)I(i + x, i + y)}{I^2(i, j)}. \tag{3.4}$$

If the texture is coarse, i.e., changes gradually, the autocorrelation function will drop off slowly. If the texture is fine, i.e., changes rapidly, the autocorrelation function will drop off very rapidly. For regular or repetitive textures, the autocorrelation function drops once we move away from a texel and rises once we approach the next texel, creating peaks and valleys.

3.2.2.3 Geometrical methods

Geometrical methods relate directly to the ordering of the primitive elements or texels in the image, rather than to pixel neighborhoods. Once these texels are extracted there are two major geometrical approaches to analyze texture. Either the properties of the texels, such as their shape or size, or the geometrical placement and inter-texel relationships, such as their distribution, are used as texture descriptors. We present one example:

1. ***Voronoi tessellation.*** Once the texels have been extracted the Voronoi tessellation associated with their centroids reflects local spatial neighborhoods in terms of the shapes of the Voronoi polygons.
Suppose the texel centers are represented by a set of points T (in a 2D Euclidean space). If we consider two points $\mathbf{x}$ and $\mathbf{y}$, the bisector of the line joining $\mathbf{x}$ and $\mathbf{y}$ is the locus of points equidistant to both $\mathbf{x}$ and $\mathbf{y}$ that divides the plane in halves. Denote by H_x^y the half plane that contains

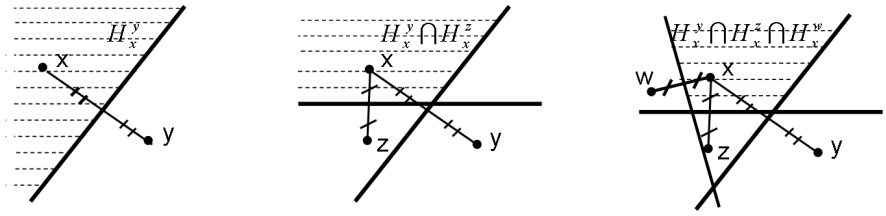


Figure 3.6: Left: two texel centroids, **x** and **y**, are shown together with their bisector and the half plane H_x^y. Middle: intersection of the half plane H_x^y and H_x^z for an extra texel **z**. Right: result of the half plane intersections when a fourth texel **w** is added.

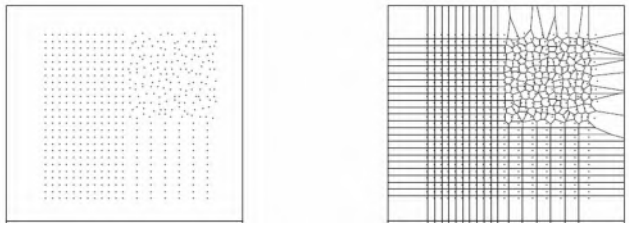


Figure 3.7: Left: a texture structure represented as a set of texel center points. Right: the corresponding Voronoi tessellation (Source: [Jain and Tuceryan, 1998]).

the points closer to **x** than to **y**, as shown in Fig. 3.6 (left). For any given point **x** such half plane is obtained for all its surrounding **y**.
The intersection $\cap H_x^y$ of all these half planes defines a polygonal region consisting of points closer to **x** than to any other point. This is shown in Fig. 3.6 (middle and right) for 2 extra points. Such a region is called the *Voronoi polygon* associated with the texel.
The set of complete polygons is called the *Voronoi Diagram* of T. The voronoi diagram together with the incomplete polygons in the convex hull define a *Voronoi tessellation* of the entire plane, see Fig. 3.7.
Typical texture features are the distance of the texel centroid to its voronoi centroid, the size and shape of the voronoi polygon and orientation of the polygon's longest diagonal.

3.2.2.4 Filter Domain methods

The large degree of structure and/or regularity present in most textures usually results in a lot of edge information. It has been shown that the human visual system relies a lot on this edge-frequency information for pattern recognition tasks. For example, a zebra will be recognized, e.g., in a herd of horses,

predominantly because of its high contrast striped pattern. Edge-frequency information is mostly obtained through image filtering.

1. ***Spatial Domain filters.*** Spatial domain filters are probably the most straightforward way to capture texture edge properties. Usually, texture edge-densities are measured using Sobel, Robert and Laplacian operators [Theodoridis and Koutroumbas, 1999].
   Amelung [Amelung, 1995] derives features from the grey value and gradient histograms. He defines 2 histograms, one for the X-component and one for the Y-component of the grey value gradient in each image pixel. These components are estimated by convolving the image with a Sobel filter kernel in each direction. For the horizontal component of the gradient, the Sobel kernel is defined as:

$$\frac{1}{4}\begin{bmatrix}1\\2\\1\end{bmatrix}. \tag{3.5}$$

   The horizontal component of the gradient images are convolved with the following kernel:

$$\frac{1}{2}\begin{bmatrix}-1 & 0 & 1\end{bmatrix}. \tag{3.6}$$

   Laws [Laws, 1980] also defined a set of texture features by first convolving the image with small filter kernels, and then combining statistics of the filter responses. The 2D convolution kernels he proposes for texture discrimination are generated by convolving the following set of five 1D convolution kernels:

$$\begin{aligned}
L &= [\ 1 & 4 & 6 & 4 & 1\ ]\\
E &= [-1 & -2 & 0 & 2 & 1\ ]\\
S &= [-1 & 0 & 2 & 0 & -1\ ]\\
W &= [-1 & 2 & 0 & -2 & 1\ ]\\
R &= [\ 1 & -4 & 6 & -4 & 1\ ]
\end{aligned}$$

   where L performs local averaging, E is an edge detector, S detects spots and the W and R vectors act as wave detectors. The derived texture features of both Amelung's and Laws' descriptors are again presented later on.

2. ***Fourier Domain analysis.*** Using the Fourier transform we move away from the image domain and obtain an image representation of in the frequency domain. Consider an $M \times N$ pixel image I in the spatial

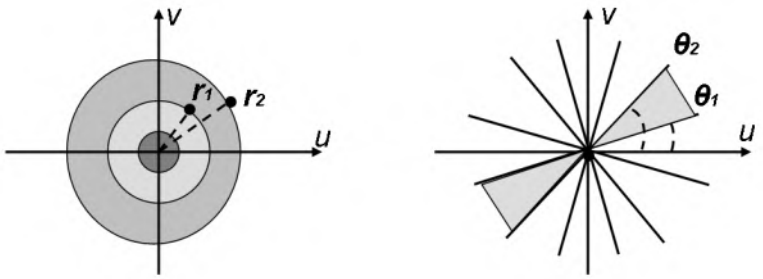


Figure 3.8: Left: division of the Frequency domain in spectral bands. Right: division of the Frequency domain in oriented wedges.

domain. The discrete Fourier transform $F(u, v)$ is then defined as:

$$F(u,v) = \frac{1}{MN} \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} I(x,y) e^{-j2\pi(\frac{ux}{M} + \frac{vy}{N})}, \tag{3.7}$$

where j denotes the imaginary number. Furthermore, define $\|F(u,v)\|^2$ as the spectral energy density, where $\|.\|$ denotes the modulus of an imaginary number.
In Fourier texture analysis the frequency domain is divided into rings (for frequency discrimination) and wedges (for orientation discrimination), as shown in Fig. 3.8. The total energy in each of these rings or along the wedges is then used as texture features.
Denote by $r = \sqrt{u^2 + v^2}$ and $\theta = \arctan(v/u)$, the texture energy in a band f_{r_1,r_2} or along a wedge f_{θ_1,θ_2} are then defined as:

$$f_{r_1,r_2} = \int_0^{2\pi} \int_{r_1}^{r_2} \|F(u,v)\|^2 dr d\theta \tag{3.8}$$

and

$$f_{\theta_1,\theta_2} = \int_{\theta_1}^{\theta_2} \int_{-\infty}^{\infty} \|F(u,v)\|^2 dr d\theta. \tag{3.9}$$

3. ***Gabor Filters.*** The disadvantage of the Fourier transform is that it provides the frequency spectrum of the entire image. However, often a spatially localized frequency analysis is preferred, e.g., when we want to characterize or segment more than one texture in the same image.

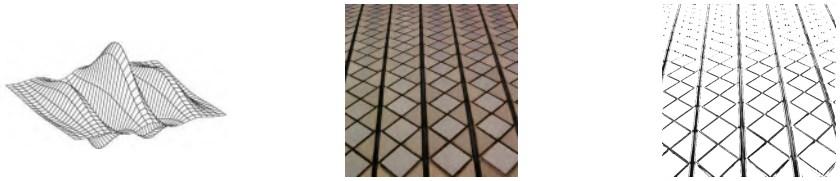


Figure 3.9: Left: frequency characteristics of a 2D Gabor filter. Middle: tile floor texture image. Right: Magnitude of the Gabor transform response of the texture image.

For this purpose, Gabor filters, having a *finite* filter support, see Fig. 3.9 (left), are very well suited. A 2D Gabor function consist of a sinusoidal plane wave for a certain frequency and orientation modulated by a Gaussian envelope and has the following impulse response:

$$w_{\lambda,\phi,\theta,\sigma}(x,y) = e^{-\frac{1}{2\sigma^2}[x'^2+y'^2]}\cos(2\pi\frac{x'}{\lambda}+\phi), \tag{3.10}$$

with $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, λ and ϕ the frequency and phase of the sinusoidal wave and σ the width of the Gaussian envelope. The image $I(x,y)$ is first convolved with the Gabor function and subsequently the resulting in the response $m(x,y)$ is computed as:

$$m(x,y) = w_{\lambda,\phi,\theta,\sigma}(x,y) * I(x,y) = \sum_k \sum_l I(x+k, y+l) w_{\lambda,\phi,\theta,\sigma}(k,l). \tag{3.11}$$

Then, the (texture) energy density defined as

$$\sum_{x=0}^{M}\sum_{y=0}^{N} \|m(x,y)\|^2, \tag{3.12}$$

where $\|.\|$ again denotes the modulus, is a typical texture feature. By tuning the function parameters, we can compute these features over various orientations and window sizes.
There are two common approaches to tune the Gabor filter parameters. The first, and most elaborate one, is to use a large filter bank of selected Gabor filters with different predetermined parameters $(\lambda, \theta, \sigma, \phi)$ to analyze the filter responses. A second approach is to design a specific set of Gabor filters. An example of a filter-band approach will be presented later on.

3.3 Classification

Extracting texture features is only half of the work. We also need to evaluate their discriminating power in order to use them in a specific pattern recognition problem. As such, in this section, the fundamentals of building a reliable texture feature classifier are discussed.

Basically, a texture feature classifier works as follows: given a texture pattern or sample $\mathbf{x}$ characterized by a vector of d specific texture features $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, the sample is assigned to one of m possible classes $C_1, C_2, \ldots, C_m$ in such a way that a predefined classification error criterion is optimized. These classes can either be disjunct or overlapping. In the case of disjunct classes, a perfect classification is possible. In the case of class overlap, usually classification errors will occur since there is no clear boundary separating the classes. The size of the feature vector and the error criterion used depend usually on both the application and size of the data set.

This section is structured as follows: Subsection 3.3.1 presents an overview of different classification strategies. In Subsection 3.3.2, we describe one specific class of classifiers, namely statistical classifiers. How a classifier is trained and what the role of the curse of dimensionality is, is described in Subsection 3.3.3. Related to this, feature selection is described in Subsection 3.3.4. Finally, different ways to calculate the classifier's accuracy are presented in Subsection 3.3.5.

3.3.1 Classification strategies

In pattern recognition usually four main categories of classifiers are distinguished, based on the heuristics used.

1. ***Template matching.*** This is the oldest and most straightforward approach where a sample is compared to a prototype or template of the object of recognition to measure its resemblance. This approach stands or falls with the feasibility to construct a good template. As such, it is mostly used in applications of low-complexity.
   For example, in office automation and digital library applications template matching is used to retrieve the most similar match for any input document image in a prestored document image data set [Peng et al., 2003]. A template then consists of the shapes of the different paragraphs or blocks a text document consists of. In most applications however, templates are difficult to construct, or simply do not exist, and more sophisticated approaches are needed.

2. ***Statistical matching.*** In this approach patterns are described in terms of random variables. Pattern recognition and/or classification then comes down to statistically modeling the feature vector describing the object or

pattern. For example, probabilities that express how likely a feature vector is to belong to a certain class can be computed for all classes based on some statistical rule, such as Bayes rule (see later). A common classification rule is then to assign the sample (feature vector) to the class it has the highest probability/likelihood to belong to. Statistical matching is by far the most popular and widely spread approach in pattern recognition.

3. ***Syntactical matching.*** In this approach a complex pattern or object is divided into simpler subpatterns or subobjects. Pattern classification then depends on defining the relations between the different subpatterns or subobjects, using a set of *grammatical* rules and a recognition system often referred to as an *automaton*.
   For example, if we want to differentiate a human being from an animal in a surveillance application we can look for the limbs, body and head in the image and based on prior rules of how they interact, e.g., humans walk on 2 legs whereas most animals on 4 paws, decide whether the object is human or not. Often statistical and syntactical matching are combined leading to so-called *belief networks*.

4. ***Neural networks.*** In this approach, related to statistical matching, statistical rules are used to describe patterns or objects. What is different here is that the rules have a learning procedure related to the way biological learning is done in the human brain.
   A Neural Network consists of a number of highly interconnected processing elements, called *nodes* or *neurons*, that are tied together with weighted connections, called *links*. Learning involves adjustments of the link weights and occurs through training, e.g., by exposure to a specific set of input/output patterns. One application field where Neural Networks are commonly used is Optical Character Recognition. There, a character or letter is recognized/classified by a network that is trained on different instances of the character by multiple persons [Avi-Itzhak et al., 1995].
   Although Neural Networks adapt themselves to the data in the "most natural way", their main drawbacks are complexity and a black box behavior. By this we mean that to get good classification results usually multiple connected layers of multiple neurons are needed which makes it difficult to track what exactly goes on in the network.

For our application at hand, i.e., the white matter tissue texture classification in US images, the lack of real ground truth information on the texture pattern prevents us from creating reliable templates. A syntactical approach could be useful if we would consider individual speckle in the US textures as subpatterns. However, we describe the relation of these subpatterns in the texture feature extracted so there is no need to do this a second time in our classifier.
Mostly because of their black box behavior, (complex) Neural Networks are also excluded for our application. This leaves us with statistical matching as the main weapon for our classification task. Consequently, we now present a detailed overview of this particular class.

3.3.2 Statistical Matching

As we mentioned, in statistical matching pattern recognition and/or classification comes down to statistically modeling the feature vector describing the object or pattern. As statistical matching comprises a broad range of classifiers we consider three subdivisions.

A first subdivision is made between *supervised* and *unsupervised* classifiers. In supervised classification a number of feature vectors with known class labels is available. This allows us to train the classifier, meaning we can extract some class-related statistics, based on *prior* information.
In unsupervised classification, on the contrary, none of the feature vectors are labeled a priori. This means the classifier itself has to group those vectors most likely to belong to the same class, in order to train itself. Unsupervised sample grouping is also referred to as *clustering*. Clearly, unsupervised clustering/classification is far more difficult than supervised classification, for sure in cases where there is class overlap or ambiguity within the data set.

A second subdivision is made between *parametric* and *non-parametric* classifiers. Parametric statistical matching is based on assumptions of the distribution of the samples within a class.
The advantage of a parametric approach is that it lowers the complexity of a classifier drastically by reducing the number of degrees of freedom. Assume the samples within a certain class to be, e.g., normally distributed. This implies that we can characterize this class by estimating two simple distributional parameters, the mean and standard deviation (or covariance matrix in the multivariate case). The downside of this approach is that in order to make reliable estimates on these distributional parameters, we need a sufficient amount of data. Consequently, when only few samples of a particular class are available, parameters might be over- or underestimated leading to unreliable class distributions.

Non-parametric models on the contrary, impose no prior knowledge on the distribution of the data set but inspect, e.g., the shape of the feature histograms to describe the structure or density of a data set in a more empirical way. Non-parametric models are primarily used in unsupervised clustering since there the lack of prior knowledge is often complete.

A third subdivision is made between *linear* and *non-linear* classifiers. A linear classifier is a classifier that bases its decision on a linear function of its input features. For a two-class classification problem and multiple features, we can describe the operation of a linear classifier as splitting a high-dimensional input space with a hyper plane: all points on one side of the hyper plane are classified into one class, while the others are classified in the other class. Because linear classifiers are often of low-complexity they are commonly used in classification problems where processing time is important. Well-known examples of linear classifiers are Minimum Distance classifiers, Linear Discriminant classifiers (LDC), and Support Vector Machines (SVM).

A non-linear classifier, as the name suggest, bases its decision on a non-linear function of the inputs. These classifiers are often more complex and less time-efficient. One of simplest non-linear classifiers is the Bayesian Maximum A Posteriori (MAP) probability classifier, where *quadratic* decision surfaces (or hyperquadrics) partition the feature space.

Given these subdivisions, we now present some examples of supervised linear and non-linear statistical parametric and non-parametric classifiers, that will later be applied to our US classification problem. We do not need unsupervised methods since we have labeled training data at our disposal.

1. ***Minimum Distance Classifiers***. These techniques group and/or classify the samples based on distance measurements in the feature space. In most cases, these distances are Euclidean distances, yet other distances such as Minkowski norms, Mahalanobis and frequency-sensitive distances are also applied [De Backer, 2002],[Theodoridis and Koutroumbas, 1999]. The most common minimum distance classifier is the *k-means* classifier. In *supervised k-means* classification first the center of gravity for each class is determined based on a set of training samples.
   Suppose we have k classes $C_1, C_2, \ldots, C_k$ and $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in}$ are n samples belonging to class C_i, the center of gravity m_{C_i} is then computed as

$$m_{C_i} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_{ij}. \tag{3.13}$$

   This is done for all k classes. Subsequently, for each new sample the distance to all class centers is calculated and the sample is assigned to the class with the smallest distance to its mean. Once the sample is assigned to a class, the center of this particular class is recalculated, adding this new sample, before a next sample is presented. In *unsupervised k-means* clustering the k centers of gravity are chosen randomly and the same procedure is repeated. The centers of gravity can also be computed using weighting coefficients for each sample, in which case we speak of *fuzzy k-means*.

2. ***k-Nearest Neighbor (kNN) classifier.*** This is a supervised non-parametric classifier where a sample or feature vector is assigned to predominant class amongst its k *nearest neighbors*. We use the kNN classifier in the following way. Suppose again we have a two class problem C_1 and C_2. For a sample $\mathbf{x}$ and a surrounding set $N = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ of n nearest neighboring samples, we define $M_{C_i}(\mathbf{x})$ as

$$M_{C_i}(\mathbf{x}) = \sum_{\mathbf{y}_j \in N \cap C_i} \frac{1}{d(\mathbf{x}, \mathbf{y}_j)} \tag{3.14}$$

with $d(\mathbf{x}, \mathbf{y}_j)$ the Euclidean distance between sample $\mathbf{x}$ and sample $\mathbf{y}_j$ in N that belongs to the class C_i. The reliability for $\mathbf{x}$ if classified as a member of C_i is then calculated as

$$rel_{C_i}(\mathbf{x}) = \frac{M_{C_i}(\mathbf{x})}{M_{C_1}(\mathbf{x}) + M_{C_2}(\mathbf{x})}, \tag{3.15}$$

and the sample $\mathbf{x}$ is assigned to the class (C_1 or C_2) with the highest reliability. This method is also called the *k Nearest Neighbor Density Estimation*.

3. ***Bayesian Maximum a Posteriori (MAP) classifier.*** This non-linear statistical classifier is based on Bayes' rule and assigns each sample to the class with the maximum a posteriori probability.
Suppose again we have an m-class problem $C_1, C_2, \ldots, C_m$. The distribution of the samples $\mathbf{x}$ in each of the classes $C_i,\ i = 1 \ldots m$ is denoted by $p(\mathbf{x}|C_i)$. This probability is also called the class-conditional probability. Furthermore, the probability for each of the classes $C_i,\ i = 1 \ldots m$ to occur is denoted by $p(C_i)$ and is called the prior probability. According to Bayes rule: given any (non-labeled) sample represented by a feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ the a posteriori probability $p(C_i|\mathbf{x})$ expresses the probability that the sample belongs to class C_i and is calculated as:

$$p(C_i|\mathbf{x}) = \frac{p(C_i)p(\mathbf{x}|C_i)}{\sum_{j=1}^{m} p(C_j)p(\mathbf{x}|C_j)}. \tag{3.16}$$

The Bayesian MAP rule assigns the sample $\mathbf{x}$ to the class with the highest a posteriori probability over all classes, i.e., to the class C_i with

$$p(C_i|\mathbf{x}) > p(C_j|\mathbf{x}), \quad \forall j \neq i. \tag{3.17}$$

In most cases the class conditional probability density functions are assumed to be normal. In that case $p(\mathbf{x}|C_i)$ has the following functional form:

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{d/2}|Q_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T Q_i^{-1}(\mathbf{x}-\mu_i)} \tag{3.18}$$

where μ_i and $Q_i = E[(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T]$ denote the mean and covariance matrix of class C_i respectively and $|.|$ denotes the matrix determinant. Prior probabilities $p(C_i)$ are usually either chosen to be equal for all classes, i.e.,

$$p(C_i) = \frac{1}{m},\ i \in \{1, 2, ..., m\} \tag{3.19}$$

or are estimated from the data set, i.e.,

$$p(C_i) = \frac{n_i}{N},\ i \in \{1, 2, ..., m\} \tag{3.20}$$

where N denotes the total number of samples in the entire (training) data set and n_i denote the number of samples belonging to C_i.

There is also still a third possibility where we assign the prior probabilities ourselves. Referring to our initial example, prior probabilities can be imposed to steer the classification process. Remember that we mentioned that if classification errors were to occur, they should be in favor of the tuna fish salad. By assigning a lower prior probability for salmon than for tuna fish, we lower the chances of a fish being classified as salmon, allowing only the fish with really specific salmon characteristics to be classified as salmon.

4. ***Fisher's Linear Discriminant classifier.*** This classifier is built around the principle of Linear Discriminant Analysis (LDA) where the linear combination of features is sought that best separates two (or more) classes of objects or events.
Suppose we have a total of N samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ and $C_i,\ i = 1 \ldots m$ classes each containing n_i samples. Furthermore, a sample $\mathbf{x}$ is characterized by its feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and a linear combination of features is denoted by $\mathbf{wx} = w_1x_1 + w_2x_2 + \ldots w_dx_d$. The purpose of LDA is to determine the $\mathbf{w} = (w_1, w_2, \ldots, w_d)$ that optimizes the following criterion:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}, \tag{3.21}$$

where S_b stands for the *between-class* scatter matrix and S_w stands for the *within-class* scatter matrix, defined as

$$S_b = \sum_{i=1}^{m} n_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{3.22}$$

$$S_w = \sum_{i=1}^{m} \sum_{\mathbf{x}_j \in C_i} (\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T \tag{3.23}$$

with

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \tag{3.24}$$

and

$$\mu_i = \frac{1}{n_i} \sum_{\mathbf{x}_i \in C_i} \mathbf{x}_i. \tag{3.25}$$

Intuitively, optimizing this criterion corresponds to transforming the data so that the distance between different class means becomes as large as possible, while the variance within each class stays as low possible. This is exactly what we want because the gap between the classes is then expected to be big. Once the optimal $\mathbf{w}$ is found, the transformed class-means $\mathbf{w}\mu_i$ are computed for all classes and a new sample $\mathbf{y}$ is, after transformation $\mathbf{wy}$, classified to the class with the smallest distance to its mean, just as in the minimum distance classifier described earlier.

3.3.3 Training and the curse of dimensionality

In practice, classification is always performed on both a *finite* data set and a *finite* feature set. In the introductory example on the salmon and tuna fish, it is practically infeasible to measure *all* physical properties of *every* salmon and *every* tuna-fish on earth. As such, finite sets of salmon and of tuna fish, sufficiently large to be representative for the species, and a finite number of features, believed to characterize the differences between the species, are selected. Based on this data, classifiers are *trained* and *tested* in such a way that they can be used for any future data set.

In the case where sufficiently large data sets are available, usually a division is made between independent *training* and *test* sets. The training set is used to determine/estimate the classifier parameters, e.g., the class-distribution parameters in a parametric statistical classifier. The test set is used to measure the accuracy of the trained classifier, e.g., the percentage of samples classified incorrectly. Normally, the ratio between the number of training and test samples is around 70/30 down to 50/50.

A second way of training a classifier is based on the *leave-one-out principle*. The classifier is then trained on all but one samples of the data set. This is done for all samples individually using each excluded sample as a test sample. The classifier accuracy is then defined by counting how many times the single test sample is classified correctly.

The ability of a classifier to deliver a good performance on a set of data samples, not used in training, is called the *generalization* ability of the classifier. It needs no debate that the goal of each pattern classification problem is to design a classifier that is as general as possible.

The main factor influencing the generalization ability is the ratio of the number of features used relative to the number of training samples. When a classification of high-dimensional feature vectors is sought, the so-called *curse of dimensionality* comes into play.

The curse of dimensionality is inherent to the sparseness of high-dimensional spaces, implying that, in the absence of simplifying assumptions, the number of training data needed to get reasonably low variance estimators on different parameters is really high.
If for example, class densities or distributions are completely known, an increase in the number of features will not necessarily result in an increase of the probability of misclassification. However, it has often been observed in practice that adding features may actually degrade the performance of a classifier if the number of training samples used to design the classifier is relatively small compared to the number of features [De Backer, 2002, Bins, 2000].

A simple explanation for this phenomenon is as follows: as most statistical classifiers use parametric estimations to find the probability densities of the different classes, they estimate the unknown parameters. For a fixed sample size, as the number of features is increased, the reliability of the parameter estimate decreases. Consequently, the performance of the resulting classifiers may degrade with an increase in the number of features.

A common approach to overcome the curse of dimensionality, is to assume that most of the discriminative information in the data actually lies in a low-dimensional space rather than a high-dimensional space. The dimension of this optimal subspace (optimal in terms of classification accuracy) is called the *intrinsic* dimension of a problem.

Finding the intrinsic dimension of a problem is probably the hardest issue in classification. However, many people have proposed different rules or bounds for the selection of the number of features, given the number of training samples. Some well-known approaches are the Vapnik-Chervonenkis and Structural Risk Minimization [Vapnik, 1998, Theodoridis and Koutroumbas, 1999] criterion. To fully exploit these boundaries however would lead us too far and is of out the scope of this work.
Therefore, a simplification applied to US breast tissue recognition presented in [Finette et al., 1983] is used. There it is shown that, for statistical classifiers in US pattern recognition, when N is the number of training samples used and l is the number of features, a ratio $\frac{N}{l} \geq 20$ assures good generalization properties.

3.3.4 Feature selection

To avert the curse of dimensionality, we have to select among our features. This can be done in different ways.

The most straightforward way to select features is to just test all possible combinations of all features in a given set, up to a certain dimension. This is called *exhaustive* searching. Although this approach theoretically assures that the optimal combination will be found, its computation time increases exponentially with the number of features selected. In practice, this makes it often impossible to actually find the optimal combination within acceptable time limits.

Another approach is to investigate the discriminative power of all single features based on some *figure of merit*. In the case of a two-class problem, a simple figure of merit for the discriminatory value of each individual texture feature (or a combination of features) is the *Mahalanobis distance* between the classes. Suppose we have a two-class problem C_1 and C_2, and a training set of labeled samples. The Mahalanobis distance δ_{mah} for a feature x_i is defined as follows:

$$\delta_{mah}(x_i) = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{3.26}$$

where μ_1 and μ_2 are the mean values and σ_1 and σ_2 are the standard deviations of the feature over the two classes in the training set. This distance characterizes the inter-cluster distance, taking into account the variance of the clusters. The best discriminating features are considered to be the ones with the greatest δ_{mah}, i.e., the greatest inter-cluster distance. Usually, features are ranked according to this distance and all features above a certain threshold are selected.

A third common feature selection procedure is called *sequential searching*. In sequential *forward* searching we start from one feature, e.g., the one with the highest individual discriminating power in terms of the Mahalanobis distance. Sequentially, a second feature is added so that this couple of features again reaches a maximum in Mahalanobis distance (over all couples). Then, a third feature is added to this couple so that Mahalanobis distance is optimized over all triplets. This procedure is iteratively repeated until either a desired maximum number of features or a certain stopping criterion is reached.

In sequential *backward* searching, we start from the entire feature set and exclude one feature at each iteration. Suppose we have n features, then in a first step all combinations of $n-1$ features are tested, e.g., based on the Mahalanobis distance. The remaining feature for the set with the highest Mahalanobis distance is then omitted. Then the same procedure is repeated for all sets of $n-2$ features out of the $n-1$ features.

A last and very popular class of feature selection techniques is *combining* existing features into new features. The most widely known approach is *Principal Component Analysis* (PCA), where high-dimensional feature vectors are reduced to lower-dimensional ones based on an eigenvalue decomposition. Technically speaking, PCA is a linear transformation that reduces the high-dimensional data (feature vectors) to a new lower-dimensional coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. By selecting dimensions according these coordinates we reduce feature space.

Although PCA is usually more computation-efficient than exhaustive, sequential of figure of merit-based searching, its drawback is that by combining the

individual features into new features, often the notion of what the new, reduced feature(s) actually expresses is lost. Therefore, when is it important to maintain the ability of describing the characteristic or quality each feature expresses, we usually do not opt for PCA.

3.3.5 Classifier accuracy

Once a presumed optimal classifier is trained with the right amount of data and selected features, its accuracy is measured based on an error criterion. The most common way to determine the *overall accuracy* is by calculating the percentage of correctly classified samples.

However, apart from the overall accuracy, two other measures are important to judge a classifier's performance: the *sensitivity* and the *specificity*. To be able to explain those concepts we introduce the concepts of a *false-negative* and *false-positive*.
Let us return to our initial example for the last time. As we know, we are mostly concerned about how tuna fish is classified. As such, a tuna fish classified as tuna fish will be called a true-positive. A salmon classified as salmon will be called a true-negative. Furthermore, a salmon classified as tuna fish is called a false-positive and a tuna fish classified as salmon a false-negative. Based on our (economical) priors, we allow for false-positives more than for false-negatives.

The sensitivity and specificity are ratios that express the frequency of false-positives or false-negatives. The specificity expresses the proportion of true-negative samples and the samples actually classified as negative and is defined as:

$$specificity = \frac{\#true - negatives}{\#true - negatives + \#false - positives}. \tag{3.27}$$

The sensitivity expresses the proportion of true-positive samples of all positively classified samples and is defined as:

$$sensitivity = \frac{\#true - positives}{\#true - positives + \#false - negatives} \tag{3.28}$$

where # expresses the cardinality. In terms of the fish, a sensitivity of 100% means the system recognizes all tuna fish as tuna fish, whereas a specificity of 100% means the system recognizes all salmon as salmon. So, in our fish classifier what we want is a good overall accuracy with a sensitivity that is as high as possible. This again shows that in real-world applications, sensitivity and specificity scores are often considered as important as the overall accuracy.

Now that all aspects involved in a texture-based pattern recognition algorithm are described, we start our actual study.

3.4 Tissue texture classification in preterm ultrasound images

In Chapter 1, Section 1.3, we described the characterization of Periventricular Leukomalacia (PVL) in preterm US brain images as the driving force to our research. We mentioned that the qualitative description of pathological white brain matter is not straightforward and often leads to a subjective diagnosis. As such, a quantitative computer-aided diagnosis (CAD) in the form of a (more objective) tissue classifier is desired to improve both the diagnostical accuracy and finesse.

In Chapter 2, Section 2.2.2, the key to CAD in US was presented as we showed that tissue *structure* is represented as speckle *texture*. This means that if there are structural differences in pathological and non-pathological white matter, we should be able to pick these up as differences in image texture.

Consequently, in this section we extract quantitative texture information from PVL US images and show how this results in a tissue classification algorithm that can be used to confirm or refute visual qualitative scoring.

The outline of this section is as follows: Subsection 3.4.1 addresses the state of the art in PVL tissue characterization and pinpoints where and how improvements can be made. In Subsection 3.4.2, the experimental setup is discussed, i.e., what our data set consists of, how our US images are acquired and how machine-settings are handled. Subsection 3.4.3 describes 7 texture feature sets extracted from the images. In Subsection 3.4.4, 3 different classifiers are compared and a majority voting procedure is presented to combine them. In Subsection 3.4.5, we describe the results on the individual texture features and classifiers, as well as on the combination of multiple features sets and multiple classifiers. Finally, in Subsection 3.4.6, we present the discussion of our results.

3.4.1 State of the art in PVL tissue characterization

In previous work, due the a higher structural image quality, pathological white brain matter is commonly studied on MRI images at term and later ages [Counsell et al., 2003, Skranes et al., 1998, Schouman-Claeys et al., 1993, Keeney et al., 1991]. Since preterms are often ventilated, obtaining the same quality of MRI information from non-sedated preterms in the first days of life is not trivial and serial imaging is even more difficult.
Contrary to MRI, as mentioned in the previous Chapters, US imaging is well-suited for preterms since it can be performed at the bed and in a rapid and cost-efficient manner. Comparative studies of PVL in MRI and US are reported in [Childs et al., 2001, De Vries et al., 1993, Sie et al., 2000, Miller et al., 2003, Inder et al., 2003, Maalouf et al., 2001]. Finally, a majority of studies have presented results on the characterization of preterm PVL in US images alone [Pierrat et al., 2000, Babcock and Ball, 1983, Kuban, 2001,

Costello et al., 1988, Dammann and Leviton, 1997, De Vries et al., 1988, DiPietro et al., 1986, Fawer et al., 1985, Hope et al., 1988, Horsch et al., 2005, Laub and Ingrisch, 1986, Tamisari et al., 1986, Townsend et al., 1999, Levene et al., 1983].

All of these papers rely on the qualitative interpretation of structural brain changes. For example, the lateral ventricle volume is scored as small, intermediate or large; the echogenicity of periventricular white matter is scored increased or decreased compared to reference structures such as the choroid plexus. Combinations of this categorical scoring then result in final conclusions on sensitivity and specificity.

Although this qualitative strategy is appropriate in the case of well-pronounced pathologies, such as cPVL (cystic PVL) [Pierrat et al., 2000, De Vries et al., 1992, Townsend et al., 1999], it is less suited for more gradual and subtle image changes, as in gPVL (gliotic PVL). As a result, the studies on gPVL report sensitivity scores that usually do not exceed 70% [Inder et al., 2003, Kuban, 2001].

The difficulty of visual scoring was already demonstrated in Fig. 1.6 of Chapter 1. There, we showed that it is not straightforward to perceive structural differences in the periventricular regions with the unaided eye, apart from a potential difference in echogenicity. We believe *quantitative* and observer-independent, textural descriptions of the periventricular tissue should succeed in better distinguishing pathological from non-pathological white brain matter.

The advantage of numerical texture features over qualitative descriptions is that they are more sensitive to gradual differences in structure. For example, physicians might classify tissue as having a high, medium or low contrast when they score the images on a coarse scale, potentially missing valuable structural information, whereas texture contrast features will describe a more continuous contrast transition in a group of pixels, i.e., on a much finer scale.
This relates to the discussion in Section 3.2.1 on the scale of a texture. What we basically try to do is investigate if when we compute texture features on a fine scale, i.e., in a local pixel neighborhood, rather than visually inspecting the entire images, we pick up more subtle structural differences in the characterization of PVL.

In medical pattern recognition literature, texture descriptors have already been reported successful in difficult US classification problems. The most common applications are US tissue characterization of the liver [Kadah et al., 1996, Sun et al., 1996], the prostate [Basset et al., 1993, Schmitz et al., 1994, Huynen et al., 1994] and atherosclerotic carotid plaques [Christodoulou et al., 2003]. While there are differences between the structure of neonatal brain, liver and prostate, they are all soft tissues that have in common that a disease process disrupts cell function and tissue architecture resulting in altered US speckle patterns [Hope et al., 2004]. This implies that parameters that quantify aspects of the brain tissue structure should also be able to provide a higher-level description of white matter damage. However,

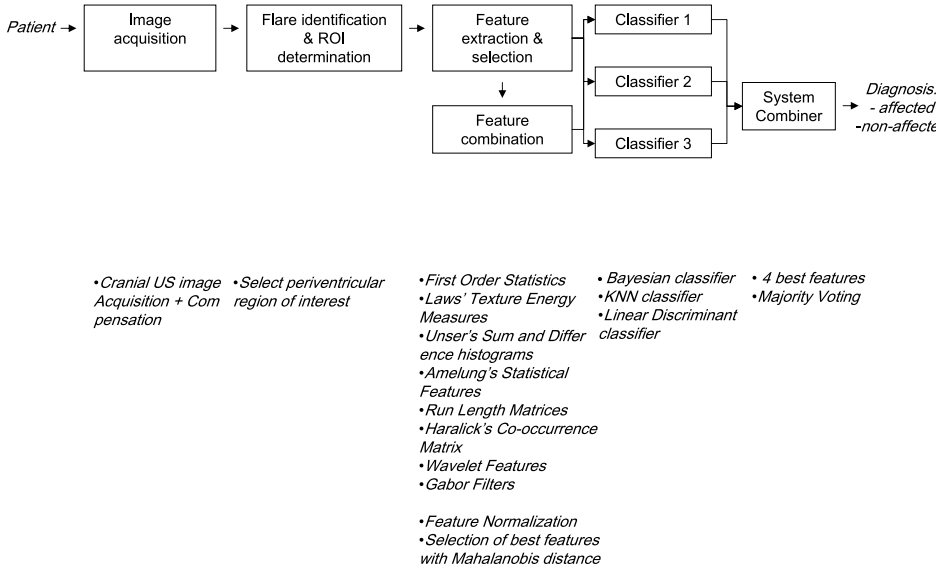


Figure 3.10: Flowchart of the multi-feature multi-classifier algorithm presented.

up till now, only few PVL studies actually incorporate texture structure.

A texture-based characterization of brain tissue using the co-occurrence matrix was attempted on *simulated* US line-scans in [Valckx and Thijssen, 1997]. In [Mullaart et al., 1999], real US images from 39 *non-pathological* terms and preterms were investigated. In the latter, the main focus was on distinguishing white from grey brain matter based on first-order features and the co-occurrence matrix, not on characterizing pathological PVL tissue. In [Stippel, 2004] first-order and co-occurrence features were computed on a data set of 58 PVL images (16 pathological and 42 non-pathological). No results on specificity, sensitivity nor overall accuracy are presented, only a figure of merit for the individual texture features and two threshold values for the most discriminative feature: the grey mean value and contrast. Finally, in [Hope et al., 2004] Gabor features were calculated along the axial direction of the US, on a set of 18 preterm images (12 pathological and 6 non-pathological). Measurements there were performed an digitized analogue US scans and a two-tailed t-test was used to account for statistical significance. Again no results on sensitivity are presented, nor did they account for the influence of machine settings.

As such, this study is the first to use multiple higher-order texture feature sets and multiple statistical classifiers on a representative data set of real preterm US brain images for the purpose of classifying PVL quantitatively with statis-

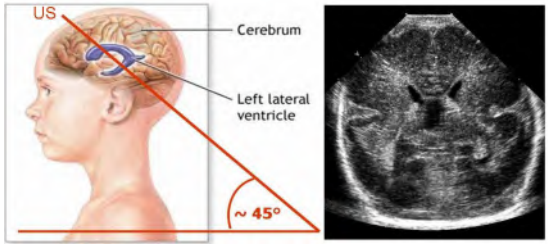


Figure 3.11: US images were acquired under a scan angle of about 45 degrees (to the coronal plane) and in such a way that both the left and right lateral ventricle atria and choroid plexus are visible on screen.

tical significance. A flowchart of our method is presented in Fig. 3.10.

3.4.2 Experimental setup

The study involved 140 coronal US brain samples from equally many preterms with a postconceptional age up to 32 weeks.[1] 50% of the images displayed PVL and 50% were unaffected according to whether subsequent sonographic evolution or acute MRI demonstrated unequivocal changes in white matter. The images were captured in the first week after birth at the Sophia Children's hospital, Erasmus Medical Centre Rotterdam, The Netherlands, always by the same physician using an Acuson Sequoia 512 US machine and a hand-free curvilinear phased-array 8.5 MHz probe.

The position of the freehand probe was not fixed nor recorded, since those operations demand for tracking devices (attached to the probe). Not only are these devices expensive, due to their relatively big size they often hinder the physician in acquiring the optimal image. Also, most tracking devices need to be recalibrated for each new scan, which makes them very impractical in serial scanning. However, because of the specified protocol all selected images were acquired under a scan angle of about 45 degrees (to the coronal plane) and in such a way that both the left and right lateral ventricle atria and choroid plexus are visible on screen, see Fig.3.11. The captured image size is 768 x 576 pixels of 0.1 mm x 0.1 mm in actual size, see Fig. 3.12.

When capturing an US image the physician was allowed to select various scanner settings to optimize the image on display. In our case these included the Power, controlling the amplitude of the emitted waves, the Gain and Time Gain Compensation (TGC) using different levels of amplification for signals returning from different depths. We found that in our data set the Gain varied in the [-11,0] dB interval and the Power was set fixed to 60 dB. The TGC

[1]Images were acquired between 2000 and 2004 and, as far as we know, this is the largest labeled PVL databases for quantitative neonatal US brain image research.

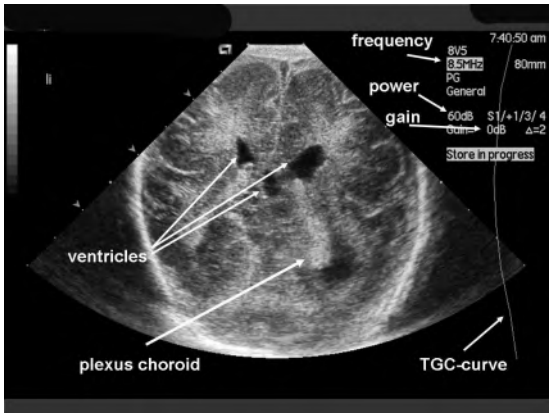


Figure 3.12: Typical example of a coronal US image captured under the medical protocol described. Machine settings are also visible in the image on display.

settings could not be measured directly since the US machine didn't allow to read out the numerical values.

As described in Chapter 2, Section 2.3, machine settings differ from patient to patient, directly influence the greyvalues displayed and hence complicate a quantitative image analysis. The common approach to overcome this problem is normalizing the images [Xiao et al., 2002, Simaeys et al., 2000, Christodoulou et al., 2003]. In our study, all images are normalized based on the algorithm of [Simaeys et al., 2000] developed earlier at our department. This algorithm was adapted to our image data set in the Master thesis of [Wydooghe et al., 2004].

3.4.3 Texture feature extraction

We use square regions of interest (ROI) as training areas to compute the texture features from. On the one hand, the size of these ROI has to be large enough to contain enough texels. On the other hand, if the ROI is chosen too large the computed texture parameters will represent a mixture of the texture of interest and other non-desired regions of interest (cortex or ventricles) and the diagnostic value will decrease.

The location of the regions of interest was based on prior indications of the physicians. We set up an experiment where we asked 5 different physicians to segment pathological tissue in 10 different US images. Fig. 3.13 shows the delineation result for 5 different experts on the same image. The outcome of this experiment supports the thesis that scoring on visual inspection alone is indeed subjective. As we can see, there is no real uniformity in the delineation of the pathological tissue. However, from the experiment we noticed that in

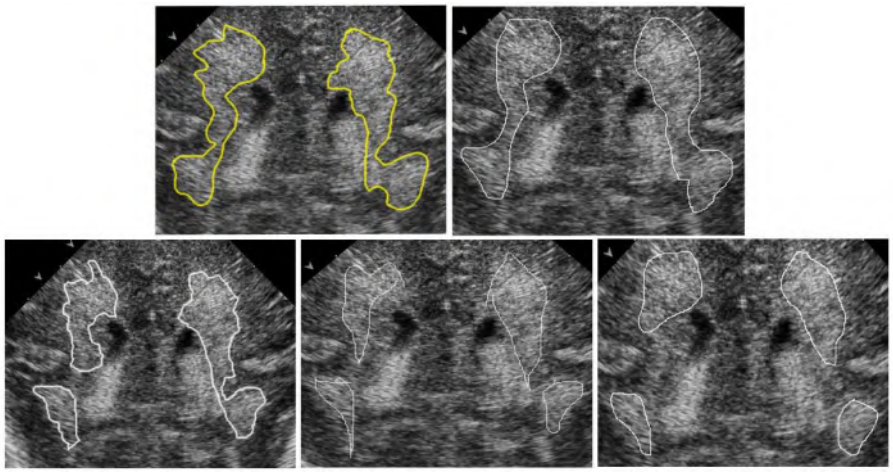


Figure 3.13: Manual delineations of pathological regions (flaring) in the same image by 5 different physicians.

all images, all experts agree on the upper periventricular region. Therefore, we confine our ROI to these zones. Two typical examples of our squared training regions are depicted together with an expert delineation in Fig. 3.14.

In Section 3.2.2, we presented a number of possible texture features extractors. In this research we used 7 sets of texture descriptors: First-order descriptors, Co-occurrence matrices, Run Length matrices, Sum and Difference Histograms, Texture energy measures, Gradient Histograms and Gabor filters, presenting us with both first- and second-order, edge-frequency and spatial correlation information on the speckle structure.

We now explain how we calculate the features exactly and what texture qualities the features express. In what follows, denote by I a ROI of size $N \times N$ pixels and again by $f(x, y)$ the grey value at pixel position (x, y). Finally, the total number of grey values in the ROI is denoted by G.

1. ***First-order statistics.*** The following standard grey value statistics: the Mean Grey Value (MGV), Standard Deviation, Skewness, Kurtosis and Signal to Noise Ratio are computed over the ROI defined as in table 3.1.

   The first parameter defines the mean intensity value of all pixels and reflects the echogenicity of the region. The standard deviation expresses the spreading of the grey values around the mean intensity. The skewness is a measure for the symmetry of the distribution/histogram of grey values. A distribution has a positive skew (right-skewed) if the right (higher grey value) tail of the histogram is longer than the left tail and a negative skew (left-skewed) if the left (lower grey value) tail is longer than the right tail.

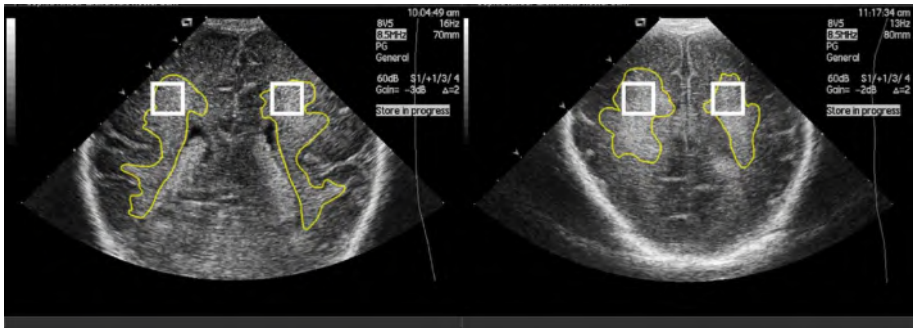


Figure 3.14: The manual delineations of the flaring zones by the physician that captured the US images (yellow), inside the delineations of the ROI the texture features are computed (white).

Table 3.1: First-order parameters.

Parameter	Definition
Mean Grey Value	$\mu = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} f(x_i, y_j)$
Standard Deviation	$\sigma^2 = \frac{1}{N^2-1}\sum_{i=1}^{N}\sum_{j=1}^{N}(f(x_i, y_j) - \mu)^2$
Skewness	$\frac{1}{(N^2-1)\sigma^5}\sum_{i=1}^{N}\sum_{j=1}^{N}(f(x_i, y_j) - \mu)^3$
Kurtosis	$\frac{1}{(N^2-1)\sigma^6}\sum_{i=1}^{N}\sum_{j=1}^{N}(f(x_i, y_j) - \mu)^4$
Signal to Noise ratio	$\frac{\mu}{\sigma}$

Finally, the kurtosis measures the peakedness of the distribution of grey values where higher kurtosis means more of the variance is due to rare but extreme grey value differences as opposed to frequent average-size differences.

2. ***Co-occurrence features.*** Although co-occurrence matrices were already investigated in previous PVL research, we tested more co-occurrence parameters than in previous studies. We calculated the co-occurrence matrix for distance parameters $d = 1, \ldots, 10$ and angles $\theta \in \{\frac{k\pi}{4} | k = 1 \ldots 4\}$ as well as the average over all four angles. A visualization of the co-occurrence, computed as in equation (3.1) for $d = 1$ and $\theta = 0$ is shown in Fig. 3.15.

   From the co-occurrence matrix 6 parameters are calculated. Denote by $P(i, j)$ the co-occurrence matrix entry at column i and row j, then the Angular Second Moment (ASM), Entropy, Contrast, Inverse Difference Moment, Correlation, Homogeneity, Signal to Noise Ratio, are defined as

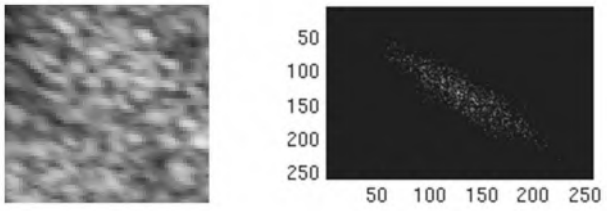


Figure 3.15: Left: a pathological flaring ROI. Right: a visualization of the corresponding co-occurrence matrix for $d = 1$ and $\theta = 0$ degrees. The whiter the entries in the matrix are, the higher their value is.

Table 3.2: Parameters derived from the co-occurrence matrix.

Parameter	Definition
Angular Second Moment	$\sum_{i=1}^{G}\sum_{j=1}^{G}(P(i,j))^2$
Entropy	$-\sum_{i=1}^{G}\sum_{j=1}^{G}P(i,j)\log(P(i,j))$
Contrast	$\sum_{i=1}^{G}\sum_{j=1}^{G}(i-j)^2P(i,j)$
Inverse Difference Moment	$\sum_{i=1}^{G}\sum_{j=1}^{G}\frac{1}{1+(i-j)^2}P(i,j)$
Correlation	$\frac{1}{\sigma_x\sigma_y}\sum_{i=1}^{G}\sum_{j=1}^{G}(ijP(i,j)-\mu_x\mu_y)$
Homogeneity	$\sum_{i=1}^{G}\sum_{j=1}^{G}(i-\mu)^2P(i,j)$

in table 3.2.

Note that for the fifth parameter in this table, i.e., Correlation, μ_x, μ_y, σ_x and σ_y are respectively the grey mean values and standard deviations of $P_x(i)$ and $P_y(j)$, which are defined as:

$$P_x(i) = \sum_{j=1}^{G} P(i,j), \qquad (3.29)$$
$$P_y(j) = \sum_{i=1}^{G} P(i,j).$$

The ASM reflects the uniformity of an image. A non-uniform image has predominantly different inter-pixel grey value transitions, i.e., many small entries in the co-occurrence matrix, resulting in a low ASM value. A uniform image has a small number of large entries in the matrix, increasing

the ASM. Entropy is a measure for the (dis)order in the image and is related to ASM.
As the name suggests, Contrast expresses the contrast in the image. The contrast is reflected by the number of matrix entries away from the diagonal compared to entries on the diagonal, since typically larger pixel transitions amount to the image contrast. In the Inverse Difference Moment on the contrary, the entries on the diagonal are favored more than those away from it. Finally, Correlation expresses the linear dependency of neighboring pixels.

3. ***Sum and Difference histogram features.*** Sum and difference Histograms $S_{d,\theta}$ and $D_{d,\theta}$ were computed using the same combinations of d and θ as used in the co-occurrence matrix. Four features were extracted from the Sum and Difference histograms: Histogram Mean, Histogram Angular Second Moment, Histogram Contrast and Histogram Entropy as defined in Table 3.3.

Table 3.3: Parameters derived from the Sum and Difference Histograms shown for the Sum histogram S.

Parameter	Definition
Histogram Mean	$\mu = \sum_i iS(i)$
Angular Second Moment	$\sum_i S(i)^2$
Contrast	$\sum_i (i-\mu)^2 S(i)$
Entropy	$-\sum_i S(i)\log S(i)$

4. ***Run Length Features.*** Run length matrices were computed for angles $\theta \in \{\frac{k\pi}{4} | k = 1 \dots 4\}$ on images reduced to 8 grey values through histogram equalization, see Fig. 3.16. For each angle, eleven features are extracted from the run length matrix. Denote by R the maximum Run Length in the ROI and by n the total number of runs in the image, computed as

$$n = \sum_{g=0}^{G-1} \sum_{r=1}^{R} P(g,r). \tag{3.30}$$

With these notations, the Short Run Emphasis, Long Run Emphasis, Grey Level Distribution, Run Length Distribution, Run Percentage, Low Grey Level Emphasis (LGLE), High Grey Level Emphasis (HGLE), Long Run High Grey Level Emphasis (LRHLGE), Long Run Low Grey Level Emphasis (LRLGLE), Short Run High Grey Level Emphasis (SRHGLE)

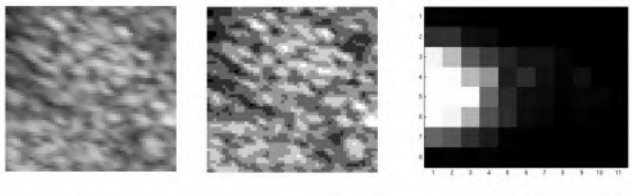


Figure 3.16: Left: the pathological flaring ROI. Middle: reduction of the ROI to 8 grey values. Right: Run Length matrix for $d = 0$ and $\theta = 0$. The whiter the color, the higher the value of the matrix entry.

and Short Run Low Grey Level Emphasis (SRLGLE) are defined as in Table 3.4.

The Short Run Length Emphasis is high for low run lengths, for the Long Run Length Emphasis this is the opposite. The Grey Level Distribution feature will be high when the image consists of few grey values, or if some grey value is dominant. The more the grey values are spread the smaller this feature will be. The same reasoning holds for the Run Length Distribution but now this feature will be high if some run length(s) is (are) dominant, if the run lengths are spread this feature will be smaller.
The meaning of LGLE is analogous to the Short Run Length Emphasis with the difference that small grey values are favored instead of small Run Lengths. The HGLE will penalize small grey values and favor big ones. The last 4 features combine both the Run Lengths and the grey values. As such, the LRHGLE is a measure for the long Run Lengths of high grey value, and similar reasonings hold for the last 3 features.

5. *Gradient histogram features.* The histograms of the X-component and Y-component of the gradients are computed from the ROI filtered by the Sobel-operators, as in equations (3.5) and (3.6), see Fig. 3.17. From the gradient histograms six features are computed: Histogram Mean, Histogram Variance, Third Moment, Fourth Moment, Angular Second Moment and Entropy, defined as in table 3.5. These features hold information on the distribution of the horizontal and vertical (speckle) edges in the US images.

6. ***Texture energy measures.*** We combine the five 1D convolution kernels in equation (3.7), into twenty-five 2D convolution kernels. The result of filtering a ROI by each of these features is shown in Fig. 3.18. Suppose I is our region of interest and t_{ij} is the response of the 2D convolution kernel obtained from the 1D filters L_i and L_j. Then the filter response $g_{ij}(x, y)$ is determined as

$$g_{ij}(x, y) = t_{ij}(x, y) * I(x, y). \tag{3.31}$$

Table 3.4: Parameters derived from the Run Length Matrix.

Parameter	Definition
Short Run Length Emphasis	$\frac{1}{n}\sum_{g=0}^{G-1}\sum_{r=1}^{R}\frac{P(g,r)}{r^2}$
Long Run Length Emphasis	$\frac{1}{n}\sum_{g=0}^{G-1}\sum_{r=1}^{R}r^2P(g,r)$
Grey Level Distribution	$\frac{1}{n}\sum_{g=0}^{G-1}\left(\sum_{r=1}^{R}P(g,r)\right)^2$
Run Length Distribution	$\frac{1}{n}\sum_{r=1}^{R}\left(\sum_{g=0}^{G-1}P(g,r)\right)^2$
Run Percentage	$\frac{n}{N\times N}$
LGLE	$\frac{1}{n}\sum_{g=0}^{G-1}\sum_{r=1}^{R}\frac{P(g,r)}{g^2}$
HGLE	$\frac{1}{n}\sum_{g=0}^{G-1}\sum_{r=1}^{R}g^2P(g,r)$
LRHGLE	$\frac{1}{n}\sum_{g=0}^{G-1}\sum_{r=1}^{R}g^2r^2P(g,r)$
LRLGLE	$\frac{1}{n}\sum_{g=0}^{G-1}\sum_{r=1}^{R}\frac{r^2}{g^2}P(g,r)$
SRHGLE	$\frac{1}{n}\sum_{g=0}^{G-1}\sum_{r=1}^{R}\frac{g^2}{r^2}P(g,r)$
SRLGLE	$\frac{1}{n}\sum_{g=0}^{G-1}\sum_{r=1}^{R}\frac{1}{g^2r^2}P(g,r)$

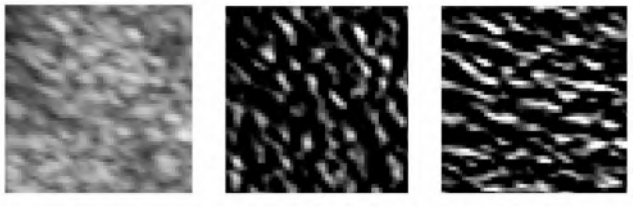


Figure 3.17: Left: the original ROI. Middle: gradient image after filtering with the horizontal Sobel operator. Right: gradient images after filtering with the vertical Sobel operator.

Table 3.5: Parameters derived from the gradient Histograms h_X (similar for h_Y).

Parameter	Definition
Histogram Mean	$f_1 = \frac{1}{N^2}\sum_{g=1}^{G} h_X(g) \cdot g$
Histogram Variance	$f_2 = \frac{1}{N^2}\sum_{g=1}^{G} h_X(g) \cdot (g - f_1)^2$
Third Moment	$f_3 = \frac{1}{N^2}\sum_{g=1}^{G} h_X(g) \cdot (g - f_1)^3$
Fourth Moment	$f_4 = \frac{1}{N^2}\sum_{g=1}^{G} h_X(g) \cdot (g - f_1)^4$
Angular second moment	$f_5 = \frac{1}{N^4}\sum_{g=1}^{G} (h_X(g))^2$
Entropy	$f_6 = \frac{-1}{\log(\frac{1}{G})}\sum_{g=1}^{G} h_X(g) \cdot \log(\frac{h_X(g)}{N})$

This filter output is used to extract initial texture features l_{ij} defined as:

$$l_{ij} = \sum_{x=0}^{N}\sum_{y=0}^{N} g_{ij}(x, y) \tag{3.32}$$

and this for all 2D convolution kernels. Then the parameters l_{ij} are added to l_{ji} since one kernel is the transposed of the other. This reduces the number of features to 15 (10 combinations of the different parameters, and 5 combinations of the same parameter). Finally, the parameter l_{11} is used to normalize all others which brings the total down to 14 features.

7. ***Gabor Energy Features.*** We use a Gabor filter-bank approach to compute Gabor texture features for different wavelengths and orientations. Referring to equation (3.10) we choose $\lambda \in \{2, 4, 8\}$ and $\theta \in \{0, 45, 90, 135\}$, $\phi = 0$ and $\sigma = 0.65\lambda$ to create symmetrical kernels under four different orientations and three different spatial frequencies, see Fig. 3.19. These parameters were selected since this filter-bank was previously found successful for US tissue classification [Yiqiang and Dinggang, 2003]. From each impulse response the energy feature defined in equation (3.12) is computed over the filter responses, resulting in 12 texture energy features.

3.4.4 Classifiers

As argued in Section 3.3.1, we apply statistical classifiers to our US problem. Furthermore, we classify samples (ROI) as pathological or not pathological, i.e., we consider a two-class problem. A Maximum A Posteriori probability (MAP) Bayesian classifier, k Nearest Neighbor (kNN) classifier and Fisher Linear

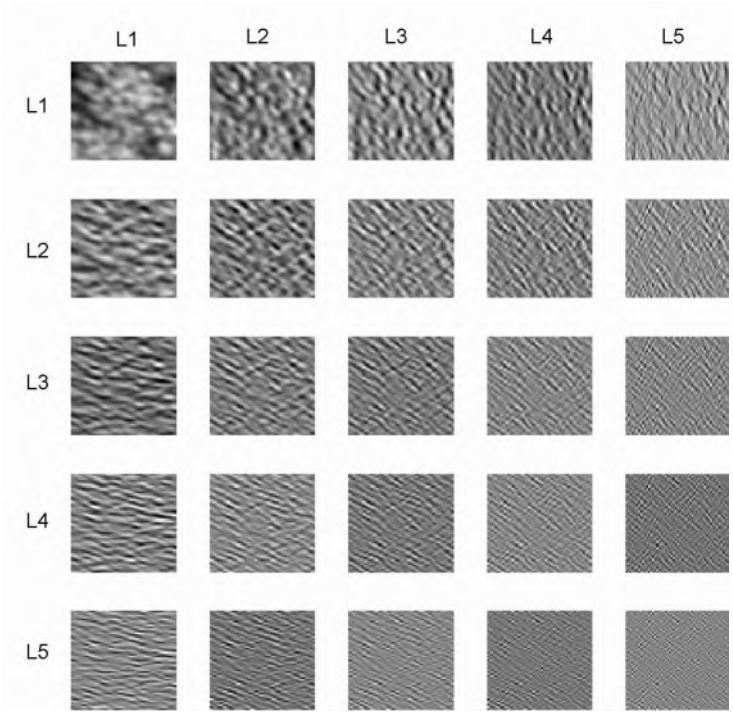


Figure 3.18: The response of filtering a ROI with the 25 different kernels. Consider this figure as a matrix of images where the entry on row i and column j is the response of the ROI after filtering with 2D convolution kernel l_{ij}, Where $L_1 = [1, 4, 6, 4, 1]$, $L_2 = [-1, -2, 0, 2, 1]$, $L_3 = [-1, 0, 2, 0, -1]$, $L4 = [-1, 2, 0, -2, 1]$, $L5 = [1, -4, 6, -4, 1]$.

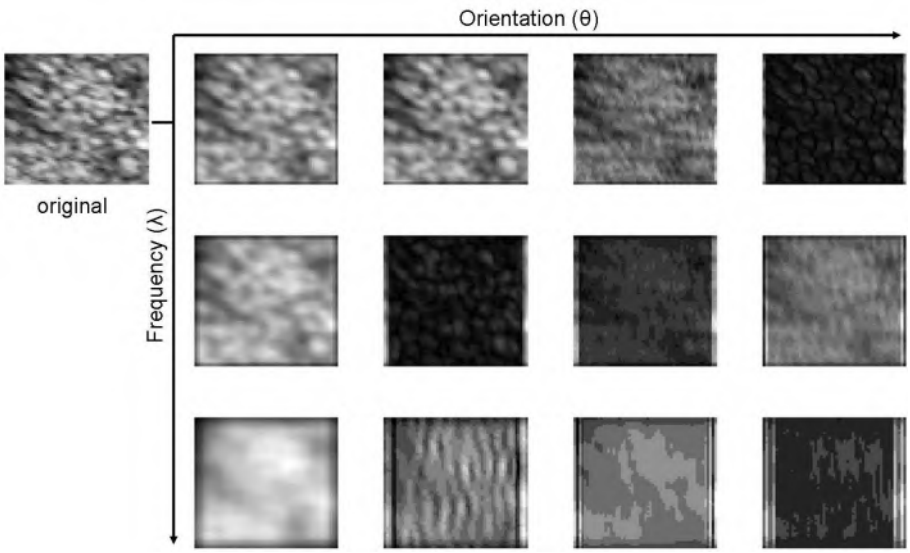


Figure 3.19: Gabor filter response for an US ROI and the $\lambda \in \{2, 4, 8\}$, $\theta \in \{0, 45, 90, 135\}$ filter-bank.

Discriminant (FLD) based classifier, all as described earlier, are compared. Note that comments can always be made on any choice of classifiers and one will never be able to prove there exists no better approach (unless perfect classification accuracies are reached). However, the choice of our classifiers was motivated as follows.

The non-linear parametric Bayesian MAP classifier was chosen because we expect class overlap, meaning that a simple linear classifier will not suffice. Besides that, the MAP classifier has a low-complexity, is fast, and easy to train. If desired, it also allows to incorporate prior knowledge, by tuning the prior probabilities. Normal multi-variate class-conditional distributions were considered with means and covariance matrices calculated from the training data.
The kNN classifier was added because, contrary to the MAP classifier, it uses only the local neighborhood information of a sample to classify it. For samples at the borders of two classes, where the a posteriori probabilities might converge, kNN incorporates the alternative information on the distribution of only the neighboring (training) samples. Neighborhood sizes of $k = 1, \ldots, 15$ samples were tested. Finally, Fisher's Linear Discriminant Classifier was added to test if linear combinations of multiple features could make the problem linearly separable.

Prior to classification, all features were normalized by subtracting their mean value and dividing by their standard deviation $n(x_i) = \frac{(x_i - \mu_i)}{\sigma_i}$ where μ_i is the

mean value and σ_i is the standard deviation of feature x_i (over the training set). This is necessary since our classifiers are based on Euclidean distances.

The leave-one-out principle was used to test the overall accuracy. To test if we do not overtrain our classifier in this way, a bootstrap technique was used to compute the standard deviation on the overall accuracy. Bootstrapping in our case means that classification is repeated over 10 sets of 120 (60 pathological + 60 non-pathological) randomly selected samples from our data set of 140 samples.

Since the amount of texture features l is high compared to the number of training samples N we follow the $\frac{N}{l} \geq 20$ rule [Finette et al., 1983] to overcome problems with the curse of dimensionality. Therefore, we select texture features in the following way. For the classification of each texture feature set individually, an exhaustive search up to 4 features is performed. To test the effect of combining features from different sets, we select the 15 most discriminative features based on the Mahalanobis distance figure of merit, and then reduced this set by eliminating highly correlated features.

Finally, in the case of difficult pattern recognition problems, the combination of the outputs of multiple classifiers can improve the overall classification performance. Combining classifiers increases the probability that the errors of one individual classifier are compensated by correct results of the others [Christodoulou et al., 2003, Bins, 2000].

First, the Bayesian MAP, kNN and Linear Discriminant classifiers are studied for individual performance. After that, the results of the best classifiers are combined using *Majority Voting (MV)*. In majority voting a sample is assigned to the class the majority of classifiers assigns it to. In our case, that is the class on which at least two classifiers agree.

3.4.5 Experimental results

3.4.5.1 Texture feature comparison

Table 3.6 (first seven rows, first three columns) shows the classification results for the seven different feature sets and for each of the three different individual classifiers. The numbers represent the classification accuracy (in %) with corresponding standard deviations over the bootstrap sets. Overall, the Gradient Histogram features perform best (91.8%), followed by the Co-occurrence (90.8%), First-order (90%), Sum and Difference Histogram (86.7%). The Run Length features (83%), Gabor and texture energy features (71%) perform worst.

Subsequently, out of the seven feature sets we select the subset of features with the best individual discriminative power. As a performance measure we use the inter-class Mahalanobis distance, see equation (3.26). Table 3.7 shows the 14 most discriminative texture features (we cut off at a Mahalanobis distance of 0.5 since we noticed a considerable drop in distance over our feature set at that value). We see all of them can be found in four of the seven feature sets.

Table 3.6: Comparison of the classification accuracy results of seven different feature sets and three different classifiers. The table shows the classification accuracies as well as the standard deviations over the bootstrap sets. The combined classifier results based on a majority voting are also shown. The bottom row shows the results for the set of four best individual features.

	MAP	**KNN**	**FLD**	**Majority Voting**
First-order	87.6 ± 1.8	87.3 ± 0.9	86.2 ± 1.4	90 ± 0.4
Texture energy	69 ± 1.1	71 ± 2.1	67 ± 1.3	69 ± 1.7
Run Length	74.3 ± 1.0	82.5 ± 1.5	74 ± 1.4	83 ± 1.7
Sum and Diff.	84 ± 2.3	84.2 ± 2.0	81.3 ± 0.4	86.7 ± 1.0
Gradient hist.	89 ± 0.9	88 ± 0.6	90.5 ± 0.5	91.8 ± 0.4
Co-occ. matrix	88.2 ± 0.8	88.4 ± 1.4	86.9 ± 1.4	90.8 ± 0.9
Gabor filters	80.3 ± 0.7	82.2 ± 1.2	82.5 ± 0.9	84 ± 0.8
4 best feat.	90.5 ± 0.5	90 ± 1.1	88.6 ± 0.7	92.5 ± 0.4

By using all 14 features at once, we are sure to loose our generalization properties, due to the curse of dimensionality. On the other hand, not all features contain independent information. As such, to further reduce the feature set, we calculate the correlation matrix for those 14 features. The lower the value of the matrix entry on column i and row j, the less the feature in column i is (linearly) correlated to the feature on row j.

It is now expected that when we add a feature to a classifier that provides the same kind of information the classifier already contains, the overall performance of the classifier will not improve drastically. As such we start with the feature in table 3.7 with the highest mahalanobis distance and exclude all features that have a correlation coefficient to this feature that is above 0.5, based the correlation matrix shown in Fig. 3.20. From this reduced set of features, we then take the feature with the second-highest mahalanobis distance and repeat the same procedure.

Note that this procedure is somehow similar to what PCA-based methods would do. As mentioned in Section 3.3.4, PCA transforms features in such a way that they become uncorrelated while retaining the maximal variance within the data set. In our approach, this maximal variance is captured by the Mahalanobis distance and the uncorrelation by choosing the least correlated features. The advantage of our method however is that we do not need to transform or recombine our features, allowing us to interpret the exact qualities they express.

Using this feature selection procedure, we end up with 4 parameters: the Angular Second Moment, Contrast, Skewness and mean of the Y-gradient histogram. When we combine these features into a new feature set, we see that we can further improve the classification accuracy, see Table 3.6 (row 9, first 3 columns).

Concerning the angles and distances used in the Co-occurrence, Sum and His-

Table 3.7: The 14 best texture features computed from the 140 (70 affected, 70 non-affected) Ultrasound brain images, ranked according to their interclass (Mahalanobis) distance.

	δ_{mah}	rank
Features	$\frac{\lvert m_1-m_2\rvert}{\sqrt{\sigma_1^2+\sigma_2^2}}$	#
First-Order features		
Skewness	1.17	1
Kurtosis	0.69	4
Sum and Difference features		
mean of Sum Histogram	0.52	13
Gradient Histogram features		
mean of Y-gradient Histogram	1.2	2
variance of the X-gradient Histogram	0.64	8
variance of the Y-gradient Histogram	0.82	3
3th-order moment of Y-gradient Histogram	0.60	9
4th-order moment of grey value histogram	0.54	11
Angular Second Moment	0.79	4
Entropy	0.68	7
Co-occurrence Matrix features		
Homogeneity	0.50	14
Entropy	0.58	10
Contrast	0.74	5
Angular Second Moment	0.52	12

togram and Run Length features, $d = 1$ and $\theta = 0$ performed optimally. We also investigated the optimal size of the selected ROI. Fig. 3.21 shows the optimal classification results for different window sizes and all classifiers, where we used the set of 4 optimal features to determine the classification accuracy. We notice the 55x55 pixel ROI always resulted in the lowest classification errors. As such, we found the optimal tissue texture scale for our purposes.

Finally, as a last experiment, we computed what happens if we do not compensate for machine settings but compute the texture features on the uncompensated images. Surprisingly, we found that the classification results differ only by 1.5% from the ones obtained on the compensated images.

3.4.5.2 Classifier comparison

Table 3.6 (column 1 to 3) shows the results over all individual classifiers. The percentage of correctly classified samples and its standard deviation over the 10 bootstrap sets is shown. On most of the feature sets, the Bayesian MAP and kNN classifiers (for $k = 8$, see Fig. 3.22) perform better than the Linear Discriminant Classifier. In other words, creating new features by linear combinations of existing ones does not improve classification accuracy of make the

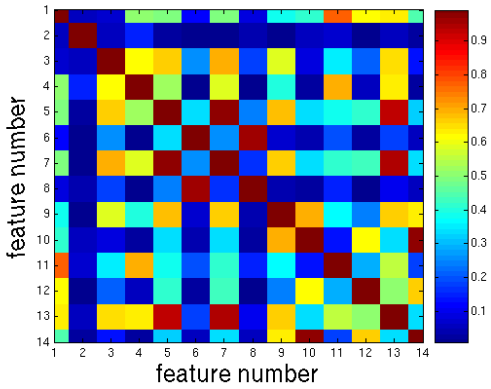


Figure 3.20: Linear correlation matrix where the rows and columns represent the 14 best individual features. Colors indicate the correlations strength.

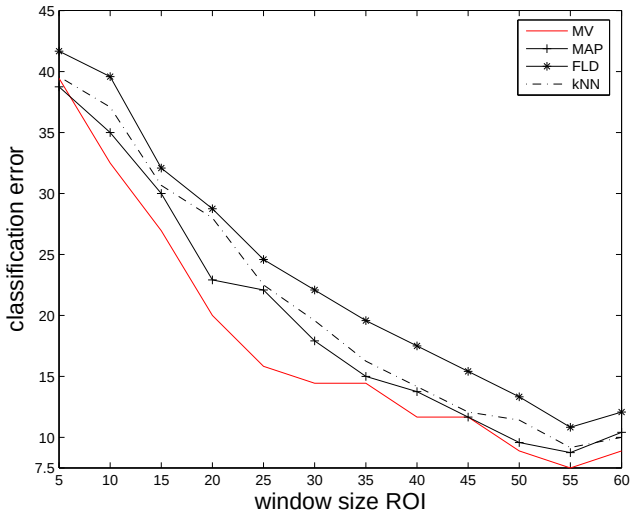


Figure 3.21: the classification errors for all classifiers (Y-axis) versus the window size of the regions of interest (X-axis).

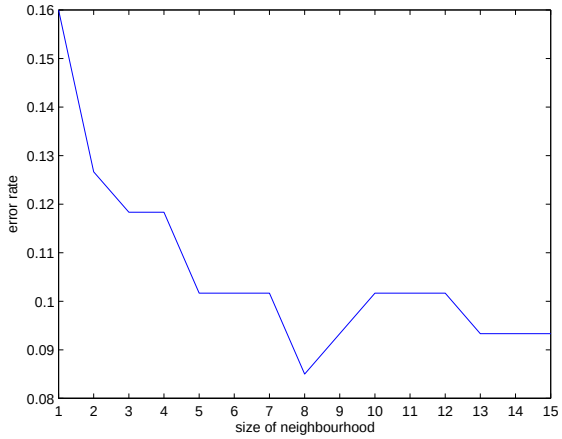


Figure 3.22: The classification error (Y-axis) for the kNN classifier for different neighborhood sizes k (X-axis).

problem linearly separable.
The last column of Table 3.6 show the results for the combination of classifiers. We see that by combining all classifiers, using Majority Voting, we improve the classifier accuracy.
Finally, by combining both multiple feature sets and multiple classifiers we achieve the highest classification result of 92.5%.

To test the generalization ability of our system we bootstrapped over 10 sets of 120 samples. We see that all standard deviations over the bootstrap sets range from 0.2% up to 2.31% with an overall average of 1.31%. As such, we conclude that our classifiers are accurate up to 1.3%. Furthermore, the sensivity and specificity for our majority voting over the reduced set of for 4 features were computed. We obtained scores for sensitivity of 88% and for specificity of 94%.

3.4.6 Discussion

Seven different texture feature sets and three different statistical classifiers were compared for the quantitative characterization of pathological and non-pathological white matter tissue. This resulted in a combination of 3 different classifiers and 4 texture features to classify pathological periventricular white matter with an accuracy of 92.5% and an sensitivity of 88%.

Although we were aware that not all of features extracted contain independent texture information, all features were included to test their discriminative power. Amelung's Gradient Histogram features, the Co-occurrence features and First-order statistics appeared the most performing feature sets.

The power of the First-order statistics is related to the qualitative description of PVL which is based on differences in echogenicity. Co-occurrence features, also investigated in previous PVL studies [Stippel, 2004], have here been proven discriminative on a more representative data set and with statistical significance. Gradient histograms show that, although there is real dominant orientation in the images, there is discriminative information in the horizontal speckle-edge distribution.

By selecting features from the different sets based on the Mahalanobis distance and a minimal correlation criterion we retain the 4 texture features that result in the optimal classification results. The advantage of reducing our feature set without actually combining features into new ones is that, based on the qualities the individual features express, we can translate the classification results of the new feature set in to a more objective qualitative description of pathological and non-pathological tissue. In general, the affected areas are denser in high grey values (positive Skewness), show a higher contrast (Contrast) and are less homogeneous (Angular Second Moment), and show a higher distribution in horizontal speckle edges (Y-gradient histogram) than the non-pathological regions.

A possible explanation for the relation between this feature combination on the one hand and the tissue type on the other is the following. Non-pathological tissue creates a certain backscatter energy, resulting in a specific structural speckle pattern with a particular grey value distribution. What happens in the tissue is that because of an increase of the cicatrix tissue in the areas affected by PVL, there will be more strong scatterers. This will result in brighter peaks in the image which causes a higher contrast. The more affected tissue becomes the more strong scatterers there are, so the higher the contrast and the better speckle edges can be detected. The accompaning increase in higher grey values explains the skew in the intensity histogram. Finally, as there is no reason to assume these stronger scatterers appear in a regular or periodical way, this will result in a less homogenous speckle pattern, hence the difference in Angular Second Moment.

From Table 3.6 we can conclude that combining classifiers (although not spectacularly) improves the overall classification results, whereas linear combinations of features do outperform the combination of existing ones. The optimal accuracy is 92.5%, with a sensitivity of 88% and a specificity of 94%. Compared to the classification results based on the visual inspection [De Vries et al., 1993, Miller et al., 2003, Inder et al., 2003], where optimal sensitivity scores never exceeded 70%, we outperform all existing methods by about 18%.

Concerning the classifier accuracies, all standard deviations on the bootstrap sets range from 0.2% up to 2.31% in percent with an average of 1.31%, which means that if our classifier is tested on any new data set, accuracies should differ no more than 1.31%.

We are aware that our classifiers were tested and trained on images from one

particular US machine, yet for the moment so is any other texture descriptor technique in this field. Given the different machine specificities each manufacturer uses, we run the risk of overtuning our method to the characteristics of that particular machine.

Finally, what is not completely clear to us yet is the effect of the compensation algorithm. In an experiment where we omitted compensation, classification results didn't differ more than 1.5% from the results on the compensated images. This could either mean that the (combination) of features are invariant to US machine settings or that the machine settings are within a range that does not alter the classification results significantly.

Given the definition of our 4 best features the invariance to the settings is very unlikely. None of the features are invariant to multiplication of the grey values with a certain scalar (effect of the Gain), to name one thing, let alone that we know what effect the real machine settings and transformations exactly have on the grey values. As such, the second explanation is more likely to be the correct one. Although the values of the texture features may be altered by the machine settings, this may happen in a similar way for both pathological and non-pathological tissue which would leave the discriminative power of the feature unaltered. Yet, in order to test that assumption we need to have both pathological and non-pathological samples scanned with different settings and know how these settings influence the grey values exactly. However, for ethical reasons, preterms are not scanned more often than necessary, so this kind of information was not incorporated in the study.

Performing an equivalent experiment on a hardware phantom could have been a possible way to overcome this. However, the material of our hardware phantoms at hand is unlikely to correspond well to the real brain tissue. Consequently, invariance of classifiers to a range of machine settings according to the phantom would not necessarily imply invariance according to real tissue.

3.5 Overall conclusions and hints for future work

The aim of this chapter was to prove that a quantitative description of pathological white matter in VLBW infants results in a more objective diagnosis than the current qualitative image inspection. A comparative study on 7 different texture feature sets and 3 classifiers resulted in a computer-aided diagnosis (CAD) pattern recognition algorithm that quantifies the texture differences between pathological and non-pathological tissue. We showed how to construct an algorithm that classifies periventricular tissues with an accuracy above 90% and a sensitivity of 88%, outperforming all current qualitative descriptions. Our classifier has a low-complexity enabling real-time classification, and shows good generalization properties.

Note that our sensitivity score of 88% means we still classify 12% of the pathological images incorrectly. Further improvements could be made in the following directions. More complex classifiers, such as Neural Networks, Self-Organizing Maps and other more complex non-linear classifiers could be tested to (possibly) get improved classification rates, yet at the risk of overtraining and slowing down the classification. Another approach is to expand our existing classifier.

A new classifier, specially trained for wrongly classified samples of our current classifier(s) could improve the overall accuracy as well as the sensitivity of our system even more. This however implies that we need to obtain enough samples to train this new classifier. A second way of further tuning our existing classifier is by altering the prior probabilities so that the sensitivity of our system improves further. However, there is a good reason why we did not follow that line.

In Chapter 1, Section 1.3, we showed that flaring is not confined uniquely to pathological images. Amongst the difficulties of scoring PVL inherent to the poor US quality, non-pathological tissue might also show some slight flaring. This means that in quantifying the texture properties of flaring, it is difficult to define the transition zone between pathological and non-pathological tissue as an exact border or cut-off point. Consequently, limitations in overall accuracy and sensitivity will always exist for any CAD classifier.

Therefore, instead of trying to optimize the sensitivity by tuning classifier parameters, we try to further improve our results by measuring the area over which pathological tissue stretches. As pathological flaring usually spreads where non-pathological flaring disappears, flaring areas should bring along the extra discriminative information needed to obtain a better CAD description. This is the main focus of the next Chapter.

Chapter 4

Ultrasound segmentation

In the previous Chapter we described pathological white brain matter based on US image texture characteristics. Although this proved extremely useful for the detection of visually less perceptible PVL, the emphasis in clinical PVL diagnosis is on the area estimate of the pathological regions, i.e., the *flares*. In this Chapter we propose a new flare segmentation algorithm based on both binary mathematical morphology operations and prior knowledge on the texture properties of pathological tissue. Subsequently, since apart from altered white matter tissue structure brain ventricle enlargements are also directly related to PVL, we extend our method to a 3D brain ventricle segmentation algorithm. Finally, a non-PVL related morphological segmentation algorithm is proposed to speed up the segmentation of the carotid artery for the diagnosis of vascular diseases.

4.1 Introduction

Segmentation is one of the oldest tasks in image processing and can be defined as the subdivision of an image into objects, patterns or areas with similar characteristics such as texture, shape or color.
Conceptually, there are two main approaches to image segmentation. Either we look for uniform regions in an image, in which case we speak of *region-based* segmentation or we look for the boundaries between regions with different characteristics, in which case we speak of *edge-based* segmentation.

The choice for one approach or the other is usually based on both the segmentation *task* we want to fulfill and the *image content*. Let us first explain the content-dependency. In applications where images consist of multiple texture regions, typically a region-based approach will be used since it is often not

Figure 4.1: Left: an artificial mixture of different texture regions. Middle: a cut-out of an agrarian IKONOS high-resolution satellite image. Right: samples of unicellular model organisms used in the study of fundamental metabolic processes.

trivial to determine the exact texture boundaries due to possible false edges in the texture structures, see Fig. 4.1 (left). Another example where region-based segmentation is appropriate is in high-resolution satellite images where agrarian regions, acres or woods are characterized in terms of their homogeneity in color or texture, see Fig. 4.1 (middle). An edge-based approach is applied when image structures or objects show the same color or greyscale characteristics as their surroundings, or as the background, and differ mostly in shape or morphology. In biology and microscopy typical examples are found in, e.g., microscopic cell segmentation, see Fig. 4.1 (right).

Apart from this content-driven selection, the segmentation task is equally decisive. Segmentation is either a goal on itself or is used as a preprocessing step in a more general framework. If segmentation is the final goal region-based methods will be applied when, e.g., we want to quantify the area or volume of certain objects such as the size of a tumor in medical imaging. Edge-based methods will be applied to characterize the shape or contours of an object, e.g., the roughness of coastal lines in geographical applications.

Examples of segmentation as a preprocessing step in a more general framework are found in tracking or surveillance applications where region- and/or edge-based segmentation is applied to track the contours of possible intruders. Another example is image registration where, e.g., similar image structures are first segmented in two images and subsequently the optimal geometrical transformations that align both images according to these (segmented) structures are computed.

By now, the diversity in segmentation approaches should be clear. Therefore, depending on the application and the field, different choices have to be made. Turning to the PVL characterization in medical US images: what result do we want and which choices do we make?

Concerning the US brain *image content*, we notice speckle patterns resulting in image texture on the one hand, and black cavities (regions where no reflection occurs, such as ventricles) on the other hand. To quantify tissue structures a region-based approach will be most suited since the exact boundaries between different speckle regions are usually unclear. Additionally, as we show later on,

the granular nature of speckle often hinders edge-based approaches.
In the case of the black cavities, the boundaries are often more visible so edge-based approaches are possible. Nevertheless, here also region-based approaches prove to be equally suited.

Concerning the US segmentation *tasks*, also different choices are made. As mentioned in Chapter 1, both flaring areas and ventricle enlargements are indicative for PVL. So, quantifying the flaring areas and ventricle volumes are goals on itself.
In addition, 3D ventricle segmentations can be used in a segmentation-based registration approach to align multi-modal image volumes, as we show in Chapter 5, and flaring delineations can be used to train new physicians. These are but two applications where segmentation is applied in a larger framework.

Multiple techniques have already been presented for the practical implementation of region- and edge-based segmentation both in US and other fields. However, throughout this Chapter we focus on *mathematical morphology* operations, *thresholding* and *texture classification* as the basis techniques of our region-based US segmentation approaches.

Since mathematical morphology is introduced here, we start with a description of its fundamentals in Section 4.2. Subsequently, we move on to our medical US applications where we present a region-based segmentation of PVL flaring areas in Section 4.3. The 3D segmentation of preterm brain ventricles is presented in Section 4.4. The segmentation of the carotid artery is presented in Section 4.5. Finally, the overall conclusion of this Chapter and hints for future work are presented in Section 4.6.

4.2 Mathematical morphology

4.2.1 Introduction

Mathematical morphology is a framework for image processing introduced by [Matheron, 1975, Serra, 1982, Serra, 1988]. Its fundamentals are quite elaborate, therefore we restrict ourselves to explaining those necessary for the understanding of this Chapter. Mathematical morphology is based on *set theory* and is used as a tool to investigate the geometrical properties of binary as well as greyscale images [Ledda, 2006]. Morphological operations can simplify objects and eliminate irrelevant image structures, while preserving the object's essential shape characteristics. These properties make them very suited for segmentation purposes.

We restrict ourselves to binary morphology, meaning the operators are defined on binary images. In a binary image, all pixels either have a value 1 (white) or 0 (black). As morphological operators are defined using set theory, we represent

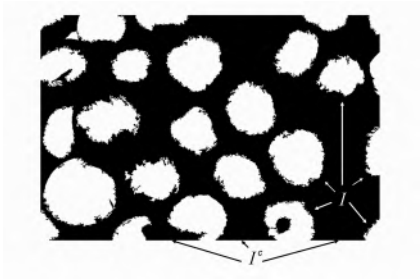


Figure 4.2: Example of a binary image with some white (foreground) objects ($\in I$) as well as black background regions ($\in I^c$).

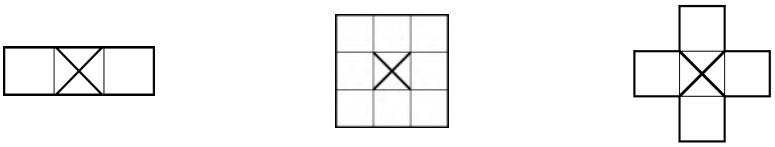


Figure 4.3: Three different symmetrical structuring elements are shown: a line, a square and a cross. Each square represent a white pixel. The crosses indicate the origins of the structuring elements and are centered in all three cases.

a binary image in terms of mathematical sets I and I^c:

$$\begin{aligned} I &= \{\mathbf{r} | f(\mathbf{r}) = 1\} \\ I^c &= \{\mathbf{r} | f(\mathbf{r}) = 0\}. \end{aligned} \tag{4.1}$$

where $f(\mathbf{r})$ is the binary function that assigns values 0 or 1 to each pixel $\mathbf{r}$ in the image. A set representing a binary image can consist of several objects. If we consider connected white pixels in a binary image as objects, by definition I is the set containing all image objects and I^c, i.e., the set complement of I, is the set containing the image background, see Fig. 4.2.

Binary morphological operators are defined in terms of *unions* ($\cup$), *intersections* ($\cap$), and set *differences* ($\setminus$) between a (fixed) binary image and a (moving) *structuring element*.

A structuring element can also be seen as a binary image (object), albeit a very small one in terms of the number of pixels it contains. Fig. 4.3 shows three different structuring elements: a line segment, a square and a cross. We show later that the shape of the element is chosen in function of the task to perform. Note also that, contrary to the examples shown, structuring elements do not necessarily need to be symmetrical. This structuring element is the basis of all morphological operations since the binary image is altered by comparing it to shifted versions of the structuring element (in a local neighborhood).

The cross inside the structuring element denotes its origin. This is the reference

position used to move the structuring element around in the binary image. Again, the choice for the origin position is free and it does not necessarily have to be centered or even be inside the structuring element. However, in what follows, we only discuss symmetrical structuring elements with a centered origin.

Before defining the basic morphological operations we introduce one more definition. Since structuring elements are moving over the image, a *translation* of a structuring element B over a vector $\mathbf{r}$ is defined as:

$$\begin{aligned} T_{\mathbf{r}}(B) &= \{\mathbf{b} | \mathbf{b} - \mathbf{r} \in B\} \\ &= \{\mathbf{b} + \mathbf{r} | \mathbf{b} \in B\}. \end{aligned} \tag{4.2}$$

4.2.2 Morphological operators

We are now ready to describe the morphological operations applied in this Chapter. The binary dilation, described in Subsection 4.2.2.1, and the binary erosion, described in Subsection 4.2.2.2, are the basic morphological operations. Based on those, we define the binary closing in Subsection 4.2.2.3, the binary opening in Subsection 4.2.2.4, the binary gradient in Subsection 4.2.2.5 and the binary opening by reconstruction in Subsection 4.2.2.6.

4.2.2.1 Binary dilation

The *binary dilation* combines (the objects of) a binary image I and a structuring element B and is defined either using the vector addition of the set elements, also called the Minkowski addition, or in terms of union set operations:

$$I \oplus B = \{\mathbf{r} | \mathbf{r} = \mathbf{i} + \mathbf{b} \ , \ \ \mathbf{i} \in I \wedge \mathbf{b} \in B\} \tag{4.3}$$

$$= \bigcup_{\mathbf{b} \in B} T_{\mathbf{b}}(I). \tag{4.4}$$

Fig. 4.4 shows the effect of a morphological dilation. In practice, the origin of the structuring element is sequentially translated to every object pixel, adding at every position all pixels that are part of the union of the structuring element and the binary image to the dilated image. Consequently, a dilation enlarges the images and has the properties of filling holes and smoothing contour lines.

4.2.2.2 Binary erosion

The *binary erosion* again combines a binary image I and structuring element B and is defined either again using the Minkowski addition, or in terms of

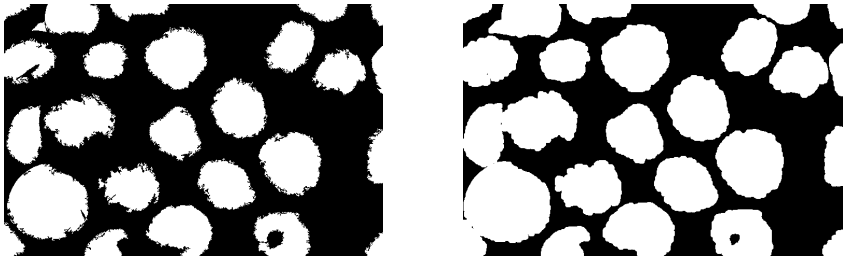


Figure 4.4: Left: original binary image. Right: dilation by a square structuring element of size 5×5 pixels.

intersection set operations:

$$I \ominus B = \{\mathbf{r} | \mathbf{r} + \mathbf{b} \in I \ , \forall \mathbf{b} \in B\} \tag{4.5}$$

$$= \bigcap_{\mathbf{b} \in B} T_{-\mathbf{b}}(I). \tag{4.6}$$

Fig. 4.5 shows the effect of a morphological erosion. In practice, the origin of the structuring element is again translated to every object pixel, one at a time. If every pixel of the translated structuring element is also part of the object, the pixel at the origin of the structuring element is retained in the eroded image. As such, an erosion shrinks an image and has the properties of disconnecting images linked by a small lesion. Also, small objects may be eroded completely.

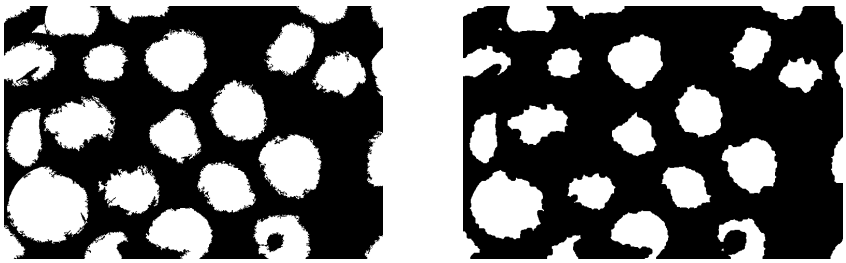


Figure 4.5: Left: original binary image. Right: erosion by a square structuring element of size 5×5 pixels.

4.2.2.3 Binary closing

The dilation and erosion are the building blocks for all other morphological operations. Combining operations can be done in several ways, yet the most straightforward approach is sequentially applying one basic operator after the other onto the image. In that way, the *morphological closing* of a binary image

I by a structuring element B is defined as a dilation followed by an erosion:

$$I \bullet B = (I \oplus B) \ominus B. \tag{4.7}$$

Note that the same structuring element B is used both for the dilation and the erosion. Fig. 4.6 shows the effect of a morphological closing. The swelling of the objects, related to the dilation, is partially canceled by the shrinking of the erosion. The dilation-related smoothing of the contours is also partially undone by the erosion. Consequently, a closing operation smoothes the objects and fills holes, without overall enlarging them. On the downside, once small holes in an object are completely filled by the dilation step, it is impossible for the erosion to make them reappear.

Figure 4.6: Left: original binary image. Right: closing by a square structuring element of size 5×5.

4.2.2.4 Binary opening

Contrary to the morphology closing, the *morphological opening* of a binary image I by a structuring element B is defined as an erosion followed by a dilation:

$$I \circ B = (I \ominus B) \oplus B. \tag{4.8}$$

Note that again the same structuring element B is used both for the dilation and the erosion. Fig. 4.7 shows the effect of a morphological closing. Here the dilation will compensate for the shrinking effect of the erosion, yet objects that have been eroded completely can not be recovered by the dilation.

4.2.2.5 Morphological gradient

The *morphological gradient* operation is defined as the set difference of a dilation and an erosion:

$$G_B(I) = (I \oplus B) \setminus (I \ominus B). \tag{4.9}$$

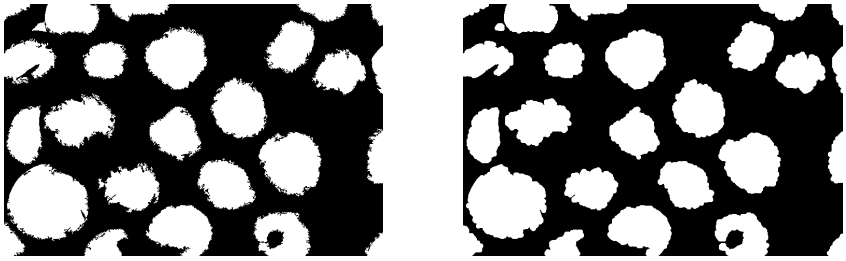


Figure 4.7: Left: original binary image. Right: opening by a square structuring element of size 5×5 pixels.

Note that in the equation again the same structuring element B is used for both the dilation and the erosion. If desired however different structuring elements can be used. Fig. 4.8 shows the effect of a morphological gradient.
As could be expected, the gradient is very suited for boundary detection. Since the dilation enlarges image objects and the erosion shrinks them, the set difference will contain the object's boundary information. We show later on the size of the structuring element determines the thickness of the extracted boundary.

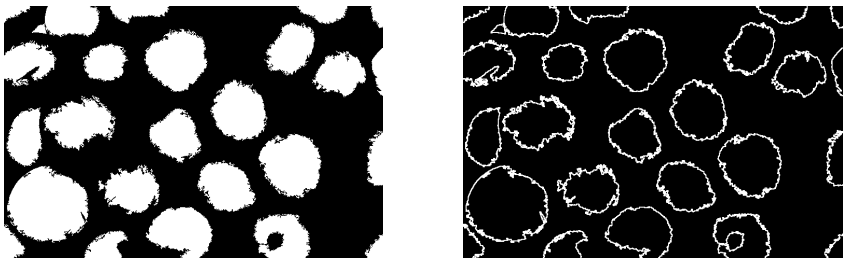


Figure 4.8: Left: original binary image. Right: morphological gradient by a square structuring element of size 5×5 pixels.

4.2.2.6 Opening by reconstruction.

The last morphological operator we need to define is an opening by reconstruction. For this purpose, we introduce the concept of a *conditional dilation*. We already showed that a normal dilation enlarges objects. If we would repeat this dilation process iteratively, objects will keep on enlarging until they fill up the entire background. The conditional dilation restricts this swelling by a *mask* element M, which defines the boundaries of the image growth and is chosen to be larger than an object or region of objects. A dilation of an image I by a structuring element B, conditioned to a mask M and is defined as:

$$\delta_B(I|M) = (I \oplus B) \cap M. \tag{4.10}$$

Note that the image I is this case is also referred to as the *marker image*. Next, we define the *conditional dilation of size n* as:

$$\delta_B^n(I|M) = \underbrace{\delta\delta\ldots\delta}_{n\text{times}}(I|M), \tag{4.11}$$

which is a concatenation of n conditional dilations. A *morphological reconstruction* of an image I by a structuring element B and mask M is then defined as:

$$\rho_B(I|M) = \bigcap_{n\geq 1}\{\delta_B^n(I|M)\}. \tag{4.12}$$

In words, this is a sequence of conditional dilations until idempotency, meaning until the output result no longer changes.
The *morphological opening by reconstruction* is defined as:

$$\rho_C(I \circ B|I). \tag{4.13}$$

Since the opening $I \circ B$ is smaller than I we can use it as a marker image and the (larger) input image I as the mask image. The structuring element C is then used to reconstruct I from the opening $I \circ B$. Fig. 4.9 shows the effect of a morphological opening by reconstruction. Starting from a smaller, incomplete opened image, determined by the size of structuring element B, image objects are iteratively *reconstructed*, by the structuring element C, yet bounded by the initial image I.

Concretely, image structures that were big enough to remain after the initial opening are reconstructed to their original shape, while smaller objects, that vanished due to the opening, are filtered out of the final image.

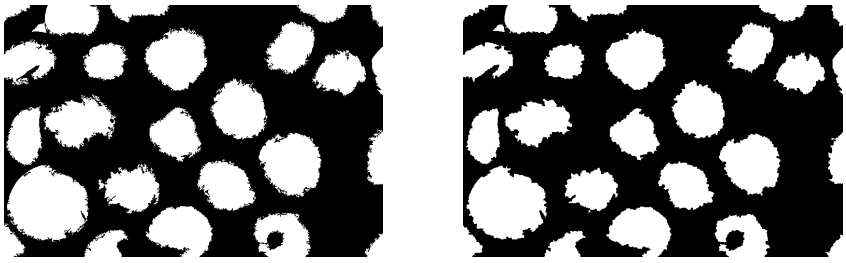


Figure 4.9: Left: original binary image. Right: morphological opening by reconstruction by square structuring elements B and C of size 5×5 pixels.

Combinations of all these morphological operations are applied throughout the different segmentation algorithms we propose. The main advantage of most operators is that by definition they have a low-complexity. Another benefit is that operators can be combined almost endlessly, and are easy to control

through the structuring element. This makes them suited for very specific segmentation tasks.

4.3 Segmentation of PVL flaring areas

We remember from the discussion on the visual inspection of PVL that the more pronounced cystic form of PVL is well detected on the US images. The milder, more frequently occurring, gliotic variant is more difficult to detect on US images and although cellular histopathology causes flaring, it is not abnormal for unaffected newborns to develop slight physiological flaring as well, partly related to normal white matter maturation (premyelination).

Consequently, apart from characterizing pathological from non-pathological white matter, the evolution of flaring over time is critical information for the follow-up of the pathology also called *staging*. A persistence of the flares beyond 7-14 days may be considered abnormal and indicative of damage. Therefore, a subclassification of flaring is suggested into flaring of brief duration (1-6 days) intermediate (7-13 days) and prolonged (14 or more). Apart from staging, for meta-analysis and to cross-validate PVL characterization on other modalities, such as MRI, the exact localization of the pathology in both modalities is crucial.

Nowadays, the only criterion for flaring area estimation is comparing the echodensity of the periventricular regions and the choroid plexus, an anatomical feature located in the lateral brain ventricles [Laub and Ingrisch, 1986, Levene et al., 1983, Tamisari et al., 1986, Townsend et al., 1999]. Flaring is usually considered when the periventricular regions appear more echodense, i.e., brighter in most studies, than the plexus. However, this criterion alone is not always sufficient for an accurate area segmentation since defining this difference in echodensity solely based on visual inspection is subjective.

Consequently, we present an interactive flaring area segmentation algorithm physicians can use in a computer-aided diagnosis. We build on the tissue texture classification of the previous chapter and combine it with a quantification of the choroid plexus criterion, thus incorporating medically accepted prior information. Finally, mathematical morphology operations are used to describe the actual flaring boundaries.

In the following Subsections, we start by addressing the state of the art in US segmentation in Subsection 4.3.1. Subsection 4.3.2 concerns our experimental set-up. Subsection 4.3.3 describes the texture-based area estimation, followed by Subsection 4.3.4 on the morphological segmentation. Subsection 4.3.5 presents the experimental results and Subsection 4.3.6 their clinical as well as statistical validation. We end with a small discussion in Subsection 4.3.7.

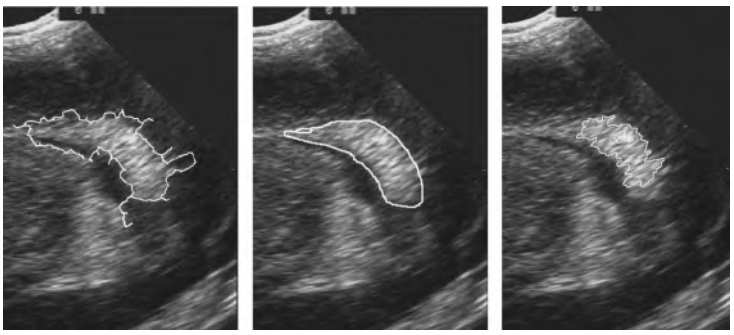


Figure 4.10: Left: active contour flaring segmentation by the technique of [Stippel et al., 2001], without a speckle-reduction preprocessing. Middle: ground truth manual expert delineation. Right: flaring area segmentation using the watershed-based technique of [Stippel, 2004] where an unstable merging procedure results in an undersegmentation.

4.3.1 State of the art in ultrasound segmentation

Just recently, an extensive survey on US segmentation techniques was published [Noble and Boukerroui, 2006]. This is the first comprehensive review of US segmentation methodology in a broad sense and comprises the main techniques used in cardiology (two-, three- and four-dimensional endocardial border, myocardium and epicardium detection), breast cancer, prostate, gall bladder and liver tumor detection, vascular diseases, obstetrics and gynecology. However, we highlight only the particular techniques that are also relevant to neonatal brain images, i.e., deformable models, watersheds and texture-based approaches.

4.3.1.1 Deformable models

The most popular edge-based segmentation approach is to use deformable models or active contours, also called *snakes*. A snake is a closed curve $\mathbf{x}(s) = [x(s), y(s)],\ s \in [0, 1]$ in the spatial domain, drawn around or inside an object of interest. These curves deform under the influence of force fields [Zu and Prince, 1997], optimizing an energy functional

$$E = \int_0^1 \frac{1}{2}(\alpha|\mathbf{x}'(s)|^2 + \beta|\mathbf{x}''(s)|^2) \ + \ E_{ext}(\mathbf{x}(s))ds. \tag{4.14}$$

This functional contains grey value related potential functions, *potential forces* E_{ext} such as the magnitude of the image gradient $E_{ext}(x, y) = -|\nabla(I(x, y))|^2$, combined with *elasticity* and *bending* forces, represented by $\alpha|\mathbf{x}'(s)|^2$ and $\beta|\mathbf{x}''(s)|^2$ respectively, which pull the snake together or bend it to the object's edges.

In US images, where undoubtedly speckle is the bottleneck for segmentation, often a preprocessing speckle-reduction step is needed since contours tend to get stuck on isolated speckle disturbing the force fields, see Fig. 4.10 (left). Recent work in US speckle-reduction include wavelet-based methods [Gupta et al., 2004, Achim et al., 2001], anisotropic diffusion methods [Yu, 2002, Abdel-Elmoniem et al., 2002] and others [Evans and Nixon, 1996, Karaman et al., 1995, Avianto and Ito, 2001]. The risk of speckle-reduction however is that, together with speckle, possibly valuable clinical information is filtered out.

In [Stippel et al., 2001] multiple speckle-reduction filters as well as an active contour-based technique for the delineation of flaring areas are presented. In [Tauber et al., 2004] a B-spline active contour is used on fetal echocardiographic images. Finally, [Noble and Boukerroui, 2006] present numerous other examples of active contour techniques in non-fetal related US problems, e.g., in [Jendoubi et al., 2004] an improved active contour model is used for the segmentation of the prostate and in [Chen et al., 1995] active contours are used for breast tumor segmentation.

4.3.1.2 Watersheds

Another approach is *watershed* segmentation, where first an image is partitioned into small segments based on local gradient information. Subsequently, segments are grouped into the final segmentation based on merging criteria such as a texture- or intensity-based similarity [Hernandez and Barner, 2005] or on more advanced graph matching [Haris et al., 1998b, Haris et al., 1998a] or level set algorithms [Seongjai and L., 2005, Zhu and Tian, 2003].

Consider a greyscale image as a topographical map, as shown in Fig. 4.11 (upper right), with *valleys* and *peaks* created by the grey values. If we would flood this map and build artificial dams each time when two basins of water merge, we obtain a so-called watershed image, see Fig. 4.11 (lower left and right). Unsurprisingly, due to the US image speckle, the resulting watershed image is highly oversegmented. Therefore, the watershed transform is usually computed either over a speckle-reduced version of the image or over a (thresholded) gradient image resulting in a less oversegmented image.

In [Abdel-Dayem et al., 2005] a watershed-based technique for carotid artery US images is presented. A watershed-based method for sonographic breast tumor detection is proposed in [Huang and Chen, 2004]. Again, more non-fetal related examples are found in [Noble and Boukerroui, 2006]. In [Stippel, 2004] the only existing watershed-based method for flaring is presented.

In the latter, an initially oversegmented watershed image is merged based on the texture correspondence of the individual watershed segments. However, in US images the watershed segments often correspond to individual speckle dots (texels) which are too small to derive accurate texture features from, see our texture description in the previous Chapter. More concretely, this

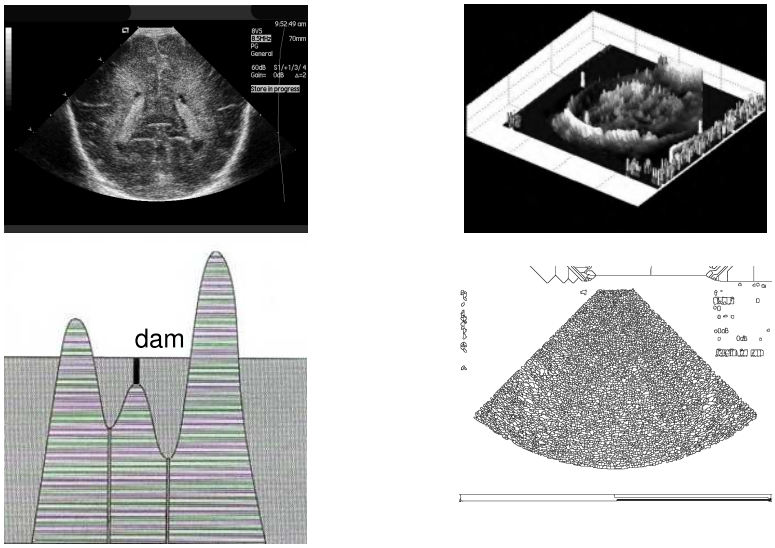


Figure 4.11: Upper left: an original greyscale US image. Upper right: a topographical representation of an US image. Lower left: the merging of two valleys is shown, where an artificial dam is constructed. Lower right: the resulting watershed image where all dams are shown.

means that the end result after merging often depends highly on the texture characteristics of the starting segment from which the watershed segments are merged. If by chance this segment has texture characteristics differing too much from the majority of the segments, the merging procedure might stop prematurely. Fig. 4.10 (right) shows a typical undersegmention due to a failing merging procedure.

4.3.1.3 Texture-based approaches

Finally, local *texture* characteristics of the speckle pattern can be used to distinguish between different tissues. Some examples are again presented in [Noble and Boukerroui, 2006]. In [Mohamed et al., 2003] and [Richard and Keen, 1996] Gabor and Laws' texture energy features are used for the segmentation of the prostate. In [Wu et al., 1989] sum and difference texture features are used for the segmentation of abdominal US images. Since in the previous Chapter we showed that none of these features performs well to classify flaring, we can not use these methods.

The segmentation technique we present is based on our previous texture characterization of pathological white matter, with the features that did show to perform well, and is the first to incorporate the texture structure and arrangements of flaring speckle *groups*, rather than the *individual* speckle statistics, in US brain images. In this way, we avoid the instability in the watershed-based

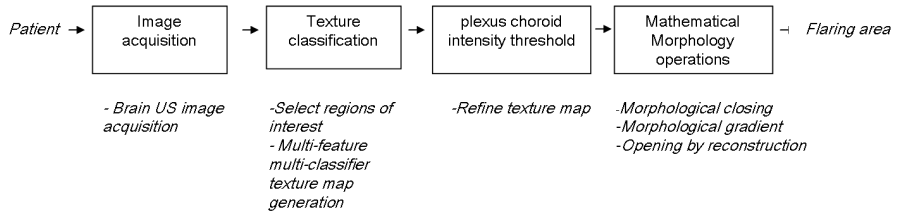


Figure 4.12: flowchart of the segmentation technique.

technique of [Stippel, 2004], and reduce the hinder of isolated noisy speckle in the active contour technique of [Stippel et al., 2001].

Another novelty is that in our technique we quantify prior medical knowledge on the choroid plexus criterion, used by the physicians in their visual image interpretation. Finally, contrary to most of the existing techniques, our algorithm is validated both statistically as in clinical practice. A flowchart of the algorithm is shown in Fig. 4.12.

4.3.2 Experimental setup

A total of 98 coronal US brain images (56 pathological, 42 non-pathological) were used in this study, obtained from as many VLBWs at a postconceptional age of 32 weeks. All images were captured at the Sophia Children's Hospital, Erasmus Medical Centre Rotterdam, the Netherlands, using a freehand curvilinear 8.5 MHz probe. For efficient operation, again the position of the hand-free US probe was neither fixed nor recorded. Again, multiple saggital and coronal sections were scanned per patient. However, because of a specified medical protocol only the images captured under a scan angle of about 45 degrees (to the coronal plane) with both the lateral ventricle atria and choroid plexus visible were retained for investigation. The captured image size is again 768 x 576 pixels of 0.1 mm x 0.1 mm in actual size.

4.3.3 Texture segmentation map

In Chapter 3, Section 3.4, we characterized the global texture properties of pathological white brain matter quantitatively. Texture features were extracted from specific flaring regions and used to build a multi-feature, multi-classifier algorithm. From a square region of interest (ROI) of 55×55 pixels 4 texture features were extracted: the skewness (f_1), the co-occurrence matrix contrast (f_2) and angular second moment (f_3) and the histogram mean of the Y-component

of the gradient (f_4) are computed (since they resulted in the optimal classification of pathological and non-pathological regions).

Suppose I denotes the selected 55×55 pixel ROI and μ and σ represent the mean grey value and standard deviation in this ROI. Additionally, $P_{d,\theta}$ denotes the grey level co-occurrence matrix (with distance $d = 1$ and angle $\theta = 0$ degrees) computed over I and defined as in equation (3.1). Furthermore, h_Y denotes the histogram (with n bins) of the Y-component of the grey value gradient of the ROI. The texture feature vector $\mathbf{x} = (f_1, f_2, f_3, f_4)$ is then computed as:

$$f_1 = \frac{1}{(55^2-1)\sigma^5}\sum_{i=1}^{55}\sum_{j=1}^{55}(I(i,j)-\mu)^3 \tag{4.15}$$

$$f_2 = \sum_{i=1}^{g}\sum_{j=1}^{g}(i-j)^2 P_{d,\theta}(i,j) \tag{4.16}$$

$$f_3 = \sum_{i=1}^{g}\sum_{j=1}^{g}(P_{d,\theta}(i,j))^2 \tag{4.17}$$

$$f_4 = \frac{1}{55^2}\sum_{i=1}^{n} h_Y(i)\cdot i \tag{4.18}$$

$$\tag{4.19}$$

This feature vector is presented to three classifiers and the output is combined in a majority voting scheme.

This procedure was initially developed to label a specific periventricular ROI. Now, we apply our classifier on a pixel to pixel basis over the entire image, constructing an initial flare area *segmentation map*. Instead of selecting a ROI at a specific place in the image, the classifier is applied in a square 55×55 neighborhood of each pixel. If the pixel, i.e., its neighborhood, is classified as pathological (flaring) it is assigned the label 1, otherwise it is assigned the label 0 (non-flaring).

Figure 4.13: Left: the original unpreprocessed ultrasound image with choroid plexus, skull, sinus and ground truth flaring regions annotated. Middle: the texture-map as obtained through our multi-classifier system. Right: the overlap of the original image and texture-based segmentation map.

Fig. 4.13 (middle) shows an example of the binary segmentation map. We notice that indeed few noisy or isolated speckle are included in the map since the textural *neighborhood* is taken into account for each pixel. On the downside, as is clear from Fig. 4.13 (right), which shows the US image part related to the segmentation map, the map also contains large non-flaring regions. Some lower parts of the skull as well as regions close to the fontanelle and sinus are included in the map.
These brain structures are either very echolucent and show a high contrast or non-homogeneous texture pattern. As such, they share the texture characteristics of flaring and it is not surprising that they are included in the map. However, they are irrelevant to white matter damage and can easily be removed since they are located far from the actual flaring areas. We will show later on how this is done exactly.

A drawback of the ROI-based classification approach however is that for the pixels at the borders of the flaring regions, both flaring and non-flaring tissue are mixed, resulting in less reliable texture features and possible inaccurate flaring border regions (slight oversegmentation in our case, as can be seen in Fig. 4.13 when comparing flaring regions). Therefore, we refine our texture map using an extra, first-order criterion.

Fig. 4.13 (right) also illustrates the choroid plexus, a brain area saturated with blood located mainly inside the lateral ventricles, is included in the segmentation map. As mentioned in the introduction, physicians believe there is a relation between the echodensity of the plexus and flaring. In visual PVL diagnosis, flaring regions are detected by comparing the echodensity of the plexus to that of the periventricular region, where echodensity is usually translated in terms of grey value intensity. Since the grey value intensity is a first-order statistic, can be measured at pixel level, and as such does not depend on neighborhood information, we try to quantify this relation and use it as extra information to refine the borders of our segmentation map.

We measured the plexus-flaring brightness ratio ourselves based on the following experiment. From 70 clearly affected images we calculated the brightness ratio of manually selected choroid plexus and flaring areas. Fig. 4.14 shows a plot of the mean intensity of the choroid plexus (X-axis) versus the mean intensity of flaring (Y-axis). Through linear regression, we find that

$$Y = 0.0084 + 0.83X. \tag{4.20}$$

This shows that there is indeed a (linear) relation between the intensity of flaring and plexus and that the intensity ratio $flaring/plexus \approx 0.83$. Note however also that where physicians thought flaring to be brighter than plexus, in fact it is the opposite[1]. This result is confirmed by quantitative findings presented in [Stippel, 2004] where a ratio of 0.85 was proposed (albeit on a subset

[1]This does not necessarily mean that the criterion physicians use in their visual diagnosis is wrong. It points out that echodensity is not necessarily translated into intensity alone and that it is indeed subjective to say what is brighter or darker solely based on visual inspection.

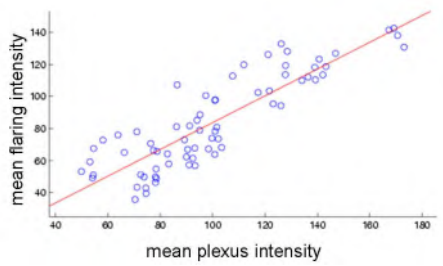


Figure 4.14: The mean intensity values of plexus (X-axis) are plotted versus the mean grey values of flaring regions (Y-axis) for 70 samples where flaring was clearly detectable. The line in the graph show the linear regression Y = 0.0084 + 0.83X

Figure 4.15: Left: the texture-map as obtained through our multi-classifier system. Middle: the refined map after the plexus intensity thresholding. Right: binarization of this refined map or in other words the thresholded initial binary segmentation map.

of only 20 subjects). Besides that, [Stippel, 2004] also presented a ratio of 0.74 for non-flaring. As such, we will incorporate our new *quantified* information in our segmentation map as follows.

After manually, i.e. interactively, delineating the choroid plexus in the image all pixels classified as flaring in our segmentation map and with a grey value below 0.83 times the mean value of this selected plexus are excluded from the segmentation map. Fig. 4.15 shows the initial US segmentation map (left) and the refined, thresholded, segmentation map (middle). On the right the refined segmentation map is also binarized (we will use that image for the morphological segmentation to be explained later). We notice the thresholding step thins out the initial flaring regions retaining only those pixels that contain *both* flaring-texture characteristics *and* an intensity-based correspondence to the plexus.
In other words, we used a new quantification of the plexus-flaring relation, yet not on the entire image but on those pixels that through our initial segmentation map are expected to be good flaring candidates.

Note that in the refined segmentation map, parts of the plexus still remain. This is to be expected since pixels in these regions share both the structural

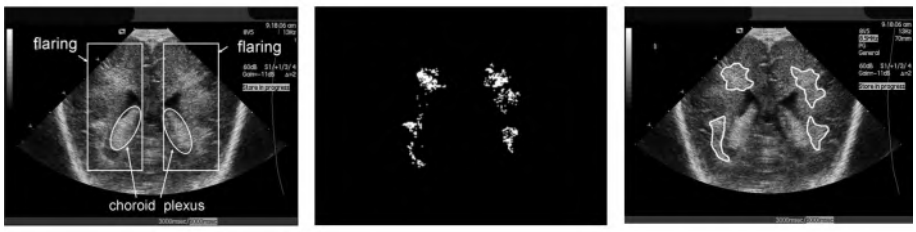


Figure 4.16: Left: the selection of the flaring and plexus bounding boxes for the left and right flaring areas. Middle: the refined texture-based segmentation map. Right: ground truth flaring delineation.

and intensity characteristics of the flaring area and thus are indistinguishable for the algorithm. However, this is no problem since by manually selecting the plexus ROI, we know which pixels belong to the plexus and have to be excluded from the refined segmentation map.

To conclude, in practice the complete texture-based characterization is executed as follows. Initially, the user (physician) is asked to draw two bounding boxes. A first one to define a certain area of interest and exclude possible skull and sinus from the map and a second one containing the choroid plexus. Once this is done, the algorithm automatically determines the (binary) refined texture segmentation map and excludes the irrelevant structures based on the bounding boxes. Fig. 4.16 show a bounding box selection and the resulting left and right flaring area as well as a manual expert flaring delineation.

Note that the size of the first bounding box is not critical for the quality of the segmentation algorithm as long as the skull and sinus are not included. The selection of the plexus bounding box however is more critical. Quantitative results on this box selection will be presented later on.

4.3.4 Morphological area delineation

From Fig. 4.16 and Fig. 4.17 (b) we notice the resulting flaring area still contains *holes* around the border areas, whereas medical experts usually delineate the flaring area by a continuous line. To determine the final (smoother) flaring area boundary we use binary mathematical morphology operators. As mentioned, morphological operations simplify image data and eliminate irrelevant objects while preserving the objects' essential shape characteristics.

We know from Section 4.2 that closing acts like a dilation (it fills holes), but does not enlarge the objects, so the original object size is retained. The morphological *gradient* operation extracts the object's edges and is useful to determine the flaring boundaries.

By morphologically closing the preprocessed image with a disc of radius 6 pixels we fill the flaring holes, as shown in Fig. 4.17 (c). The shape of the structuring

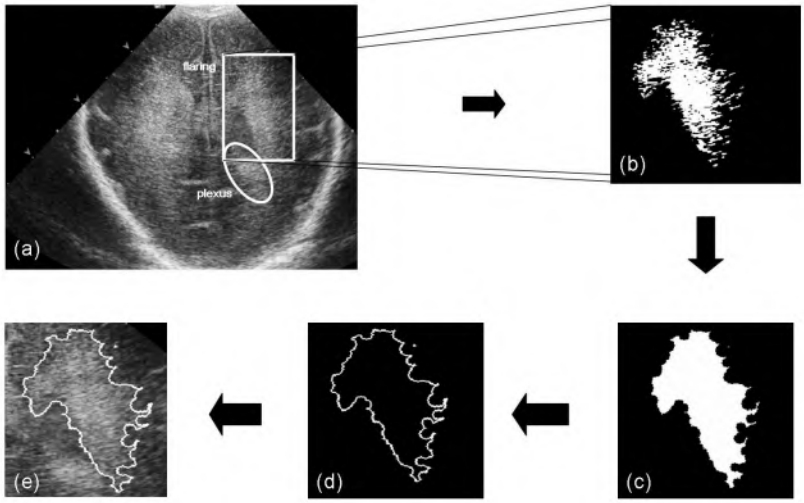


Figure 4.17: (a): the original unpreprocessed US image where the choroid plexus is manually delineated as well as the bounding box where flaring is expected. (b): the texture-based refined segmentation map. (c): the result after morphological closing with a spherical structuring element. (d): the result after the morphological gradient operation. (e): the final contour projected back onto the original image.

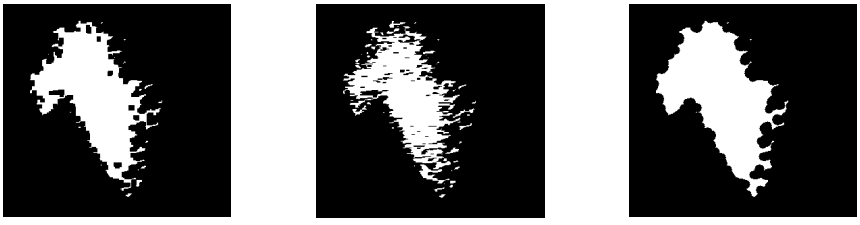


Figure 4.18: Left: the original image (Fig. 4.17 (b)) closed with a square of size 6 pixels. Middle: result of closing with a horizontal line segment of length 6 pixels. Right: result of closing with a disk of radius 6.

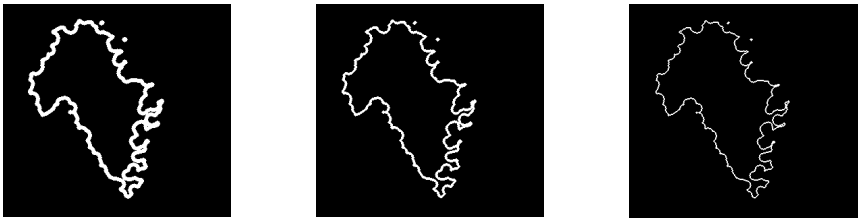


Figure 4.19: Left: gradient operation with a disk of radius 3. Middle: gradient operation with a disk of radius 2. Right: gradient operation with a disk of radius 1.

element is usually chosen depending on the shape of the object to be processed or holes to be filled. In general, if the object or holes are predominantly horizontally or vertically oriented, predominantly vertically and horizontally shaped structuring elements are used. If the objects or holes show no predominant direction, a spherical structuring element is chosen. We selected a spherical structure element since it leads to the most natural results, i.e., smoothest boundaries, see Fig. 4.18. The radius size was set sufficiently large as to close all holes in the boundary neighborhood.

Extracting the morphological gradient of this closed image, using a disc of radius 1 pixel, results in the flare boundary. The radius size is set so as to obtain a thin contour, see Fig. 4.19. The result of the gradient operation is shown in Fig. 4.17 (d).

It is possible that little isolated islands remain after the gradient operation. To overcome this island forming, optionally, we perform an *opening by reconstruction* postprocessing by a disc of radius 3 for both masking and reconstruction.

4.3.5 Visual results

In this section we present some visual results of the segmentation technique. In the next session we validate these experimental results in a quantitative way.

First, we discuss the influence of the plexus bounding box. Although the medical protocol was set so that the plexus is present in the US images, in 4 of the 98 images this was not the case, see Fig. 4.20. In those cases the plexus thresholding is impossible and no measurements, apart from the initial texture map, can be made unless we possess another image of the same patient containing the plexus.

Fig. 4.21 and 4.22 show typical segmentation results for pathological and non-pathological images. Although flaring is present in both pathological and non-pathological images, we notice a clear difference in the area over which it stretches out.

Finally, Fig. 4.23 and 4.24 show flaring area results of our method (middle) compared to a manual expert delineation (left) and an existing active contour

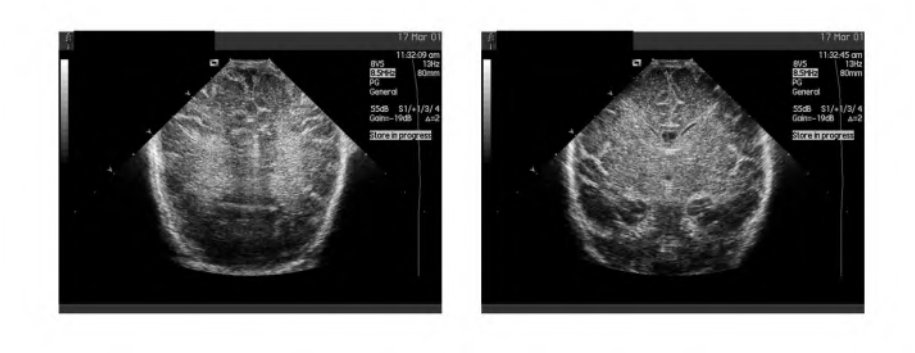


Figure 4.20: Two images where the choroid plexus is not visible.

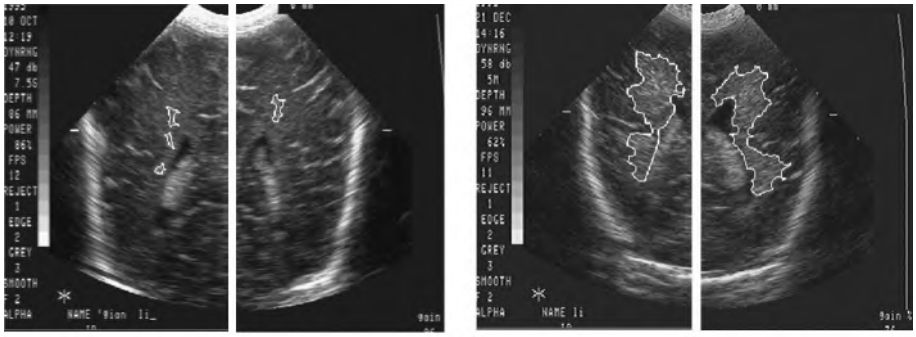


Figure 4.21: Flaring area for a non-pathological case (two leftmost pictures), and a pathological case (two rightmost pictures).

method (with speckle-reduction preprocessing) [Stippel et al., 2001] (right). As can be seen on both figures, our method is less sensitive to isolated speckle surrounding the flares which makes the contour follow the boundary more rigidly.

4.3.6 Validation

To verify the performance of our method we conducted multiple tests and experiments. In Subsection 4.3.6.1, we validate the inter-observer variability and clinical significance of our method based on hypothesis testing and an experiment with 2 physicians. The flaring area accuracy and the comparison of our technique to [Stippel et al., 2001], based on golden standard information derived from an experiment involving 12 physicians, are described in Subsections 4.3.6.2 and 4.3.6.3.

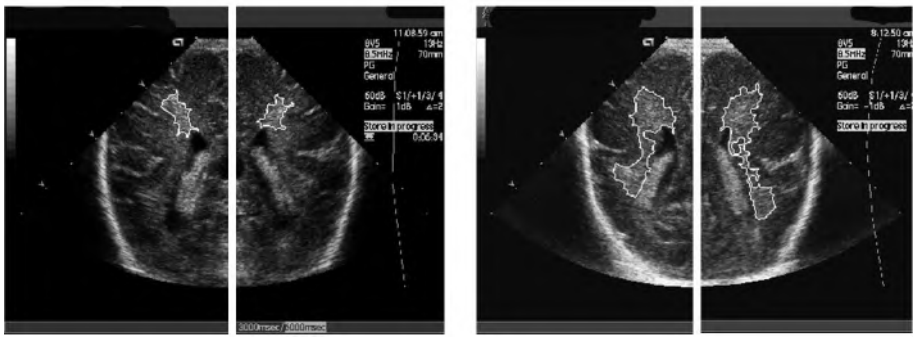


Figure 4.22: Flaring area for a non-pathological case (two leftmost pictures), and a pathological case (two rightmost pictures).

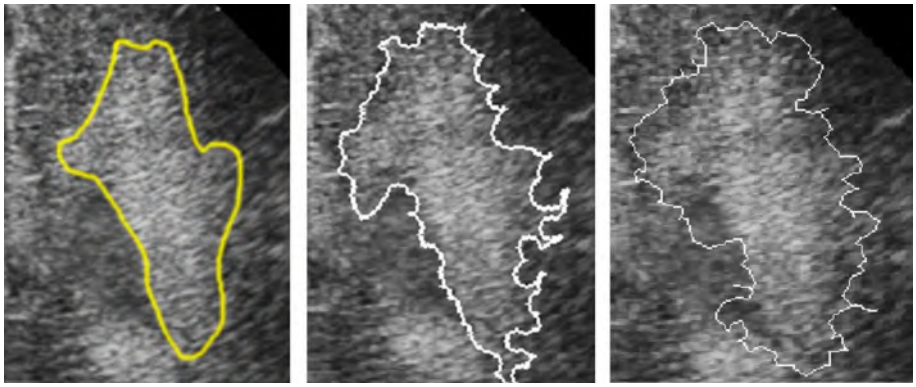


Figure 4.23: Left: the delineation of physician. Middle: segmentation result with our method. Right: an existing active contours method.

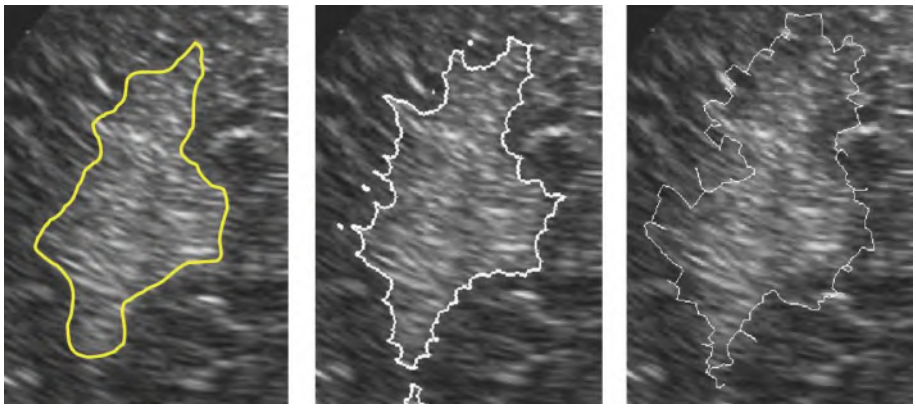


Figure 4.24: Left: the delineation of a physician. Middle: our method. Right: an existing active contours method.

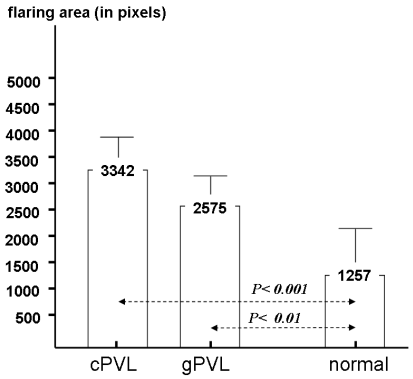


Figure 4.25: The plot of the mean (computed over left and right ventricle) flaring area estimates obtained by 2 physicians using our technique. The two-tailed P-value using the Mann-Whitney U test, are found to be ≤ 0.001 when the group of cPVL samples is compared to the normal group and ≤ 0.01 when the group of gPVL samples is compared to the normal group, based on our measurements.

4.3.6.1 Clinical validation of the method

Two physicians were asked to segment the left and right flaring areas in 64 US images, using our method. The images displayed coronal sections through the atrium, using an Acuson Sequoia scanhead, this time at identical Gain and Depth settings. All US images were of preterms who had first week reference MRI or where lesions became cystic based on which they were priorly classified as follows: normal MRI (43 images), gliotic PVL on MRI (11 images), and extensive cystic PVL on MRI (10 images). The physicians who took part in this experiment had no access to this prior classification.

In Fig. 4.25, the scoring grade is shown on the X-axis versus the segmented flaring area (in pixels) on the Y-axis (mean flaring area per class + 1 standard deviation). We compared the normal group to the pathological ones, based on the Mann-Whitney U test [Mann and Whitney, 1947]. This test expresses whether measurements belong to same population or not, i.e., if measurements for pathological and non-pathological areas differ significantly or not. We found that the two-tailed P-value is < 0.001 for the difference between normal and cPVL. Comparing the normal group to the gPVL one, we get a P-value of < 0.01. Both values are considered extremely significant. In other words, the size of the segmented flaring area can serve as a predictor for pathological and non-pathological PVL in the clinical diagnosis.

We now use the class means in Fig 4.25, i.e., a cPVL-area equals 3342 pixels, a gPVL area 2575 pixels and normal area 1257 pixels, to reclassify the data set used in Chapter 3 as follows. We determine the flaring area for all samples using our method and classify each sample according to this area (using a minimum-

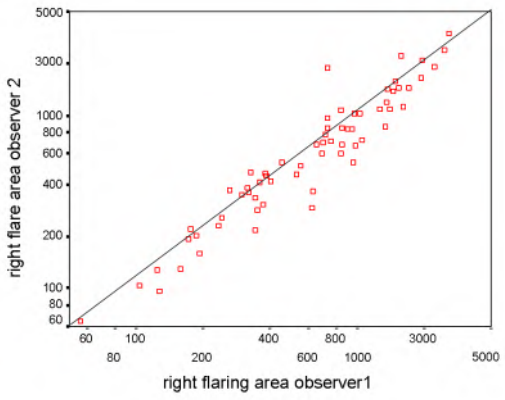


Figure 4.26: The right flaring areas are plotted for the physicians (observer 1 on the X-axis, observer 2 on the Y-axis) segmenting the images independently.

Table 4.1: The minimum, maximum and mean difference (in units of 1000 pixels) between the flaring areas, scored by two clinicians independently, as well as the standard deviation are tabulated over the test set of $N = 64$ images.

	N	***Min***	***Max***	μ	σ
Difference left flaring	64	-0.17	0.85	0.0618	0.13771
Difference right flaring	64	-0.33	0.41	0.0166	0.11367

distance classifier) and the representative class-areas we just determined. In doing so, we obtain a sensitivity of 98% (pathological versus non-pathological).

In Fig. 4.26, the flaring area derived by the first observer is plotted against the area derived by the second observer. If both physicians had exactly the same score the plot in Fig. 4.26 would be a perfect line since there would be no variance in the scoring of observer 1 and 2. All of the variance in the experiment would be due to the different images used in the experiment. We can obtain a measure of the degree of relationship between both observers by asking what proportion of the total variance in the data set (subject-related + image-related) is image-related. This is exactly what the intraclass correlation (ICC) expresses. In our case we get a ICC of 0.95 meaning only 5% of the variance is due to the difference in scoring of the observers. The explanation for (small) inter-rater differences is that physicians sometimes tend to overestimate the initial flaring bounding boxes incorporating non-flaring related features as parts of the skull, or do not segment the plexus completely in which case parts of it can remain in the final segmentation.

We also tested the inter-observer variability of our technique by computing

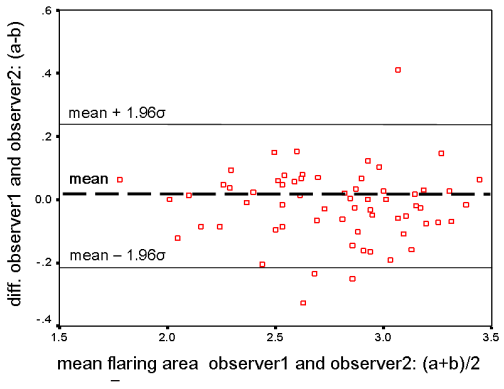


Figure 4.27: The Bland and Altmann plot comparing the flaring area scored by two clinicians. On the X-axis the mean flaring area (for right flaring) of observer 1 and observer 2 is plotted (times 1000 pixels). On the Y-axis, the difference in flaring area for observer1 and observer2 is plotted (again times 1000 pixels). The dashed line corresponds to the mean difference, the two other lines in the plot to the mean ± 1.96 times the standard deviation on the mean, i.e. the 95% confidence interval.

the Bland and Altman plot [Bland and Altman, 1986] for both observers, see Fig. 4.27. In this graphical method, the mean (in units of 1000 pixels) of the area estimates of both observers is plotted against the difference (in units of 1000 pixels) between their estimates, for all 64 images. Since both observers use our technique, a low inter-observer variability of our technique should correspond to a mean difference around 0.
Table 4.1 shows this mean difference and its standard deviation for both the left and right flaring area where we notice that indeed the mean difference of only 0.0618 and 0.0166 ($\times$ 1000) pixels. Furthermore, the mean ± 1.69 times the standard deviation are considered (the 95%) confidence intervals so that from Fig. 4.27 we notice only 4 outliers.

A final remark on the flaring areas is that in the medical literature flaring is typically considered symmetrical [Govaert and De Vries, 1995]. Flaring is assumed to be equally spread around both ventricles. We are able to prove this quantitatively using our technique. In Fig. 4.28 we plot the mean flaring area (of both observers) for left hemisphere flaring on the (X-axis) against the mean area for the right hemisphere on the (Y-axis). Through regression we get that

$$Y = 0.52 + 1.00X \tag{4.21}$$

or that in general the area-ratio $left/right \approx 1$. In other words, flaring can indeed be considered symmetrical.

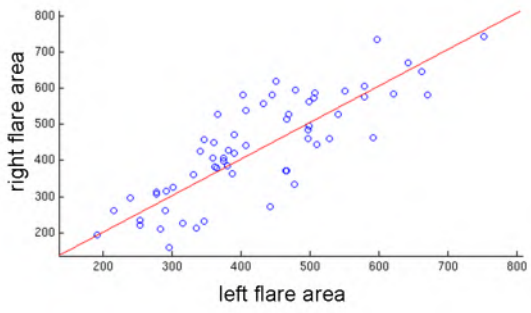


Figure 4.28: The mean pixel area of both observers (in units of 50 pixels) for the left hemisphere (on the X-axis) and right hemisphere (on the right-axis)

4.3.6.2 Panel experiment on flaring area accuracy

A drawback for the validation of most US segmentation algorithms is the difficulty to obtain ground truth information. Consequently, when there is no real ground truth information, such as a hardware phantom or other imaging modalities, available, we *construct* our ground truth in order to draw conclusions on the accuracy of our technique. Commonly, this is done based on the manual delineations by different experts considered as our benchmarks.

In a second experiment involving multiple experts, we collected a set of 8 images displaying pathological flaring and sent them to 12 physicians from different institutions, asking them to delineate the flaring areas manually. Fig. 4.29 shows the result for one of the images and 6 expert delineations. We see that, as with the former experiment on the ROI for texture classification, experts do not always agree.

Based on these results, we can either choose to retain only the intersection all experts agree on, or the union, or something in between. Retaining the common intersection favors the smallest individual delineation too much, retaining the union favors the largest individual delineation. As such, we choose the intermediate way. We calculated an average expert flaring segmentation and considered this as our ground truth. This average delineation is based on first establishing a one-to-one correspondence between the points constituting two or more curves. This initial single-point correspondence between the multiple delineations is then iteratively optimized until one stable average delineation is reached, see Fig. 4.30.

Given a set of m delineations $D_i,\ i = 1 \ldots m$, we determine the average delineation Y using the following iterative procedure. In the first iteration, we define n equidistant points $\mathbf{x}_{1j},\ j = 1 \ldots n$ on D_1 and choose one starting point $\mathbf{x}_{11}$ on D_1. We then look for a point closest to $\mathbf{x}_{11}$ on each other delineation $D_i,\ i = 2 \ldots m$. These points are denoted by $\mathbf{x}_{i1}, i = 2 \ldots m$ respectively. Then,

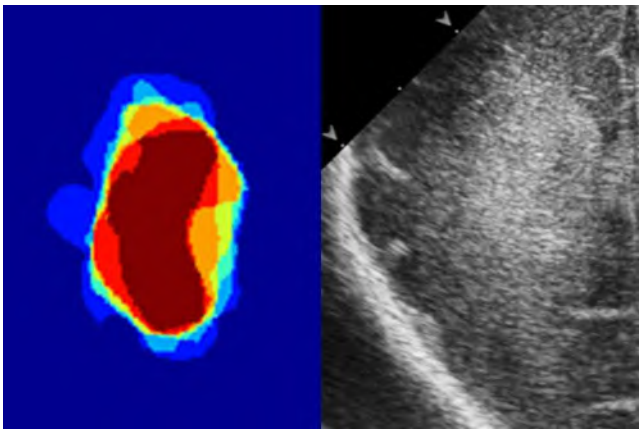


Figure 4.29: Left: the overlap of 6 different expert delineations for the left flaring area. Right: one of the 8 test images sent to the physicians.

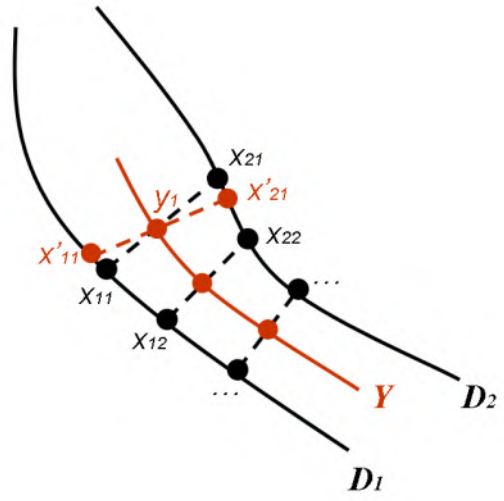


Figure 4.30: In black two initial contours D_1 and D_2 with 3 equidistant points. the full red line represents the average delineation after one iteration. The dashed red line represents the normal to the average contour and determines the new starting points for the next iteration.

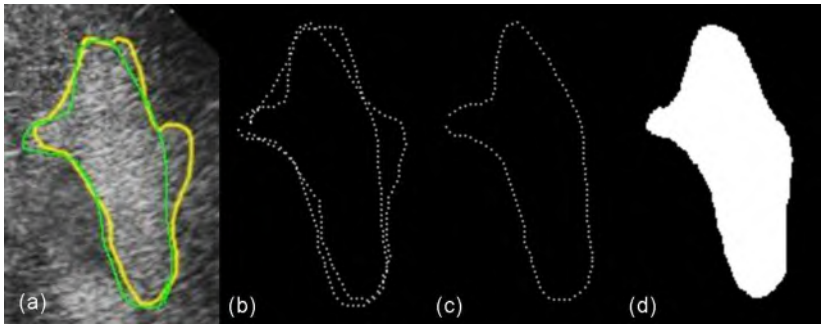


Figure 4.31: (a) 2 overlapping expert delineations, (b) 100 equidistant points chosen on both delineations, (c) the average delineation after 4 iterations, based on the correspondence between the points (d) the resulting average overlap.

starting from these points we define $n-1$ equidistant points on $D_i,\ i = 2 \dots m$, clockwise.
The correspondence of these points is established sequentially, i.e., the point $\mathbf{x}_{12}$ on delineation D_1 corresponds to points $\mathbf{x}_{i2},\ i = 2 \dots m$ on delineations $D_i,\ i = 2 \dots m$ respectively. Once this is done, the centroid of n corresponding points

$$y_i = \frac{1}{m} \sum_{j=1}^{m} x_{ij} \tag{4.22}$$

for each $i = 1, 2, \dots, n$ is defined as one of the points of the average delineation. Next, a normal to the average delineation is drawn from each point on Y and the intersection of this normal for each of the M input curves is determined. These intersections form a new set of corresponding points. Then the process is repeated on this new set of corresponding points and iterated until the average curve no longer changes.

Fig. 4.31 (a) shows 2 overlapping expert delineations, (b) shows 100 equidistant points chosen on both delineations, (c) the average delineation based on the correspondence between the points for the two delineations and 4 iterations and (d) the resulting average overlap of both initial delineations. Once our ground truth is determined we can define similarities between the delineation of the flaring area of our technique compared to the ground truth.

As illustrated in Fig. 4.32 we use the following quantities to compare the results:

- The intersection of the delineation by our segmentation (SEG) and the ground truth (GT) is the set of correctly classified pixels (overlap).
- The intersection of $\overline{GT}$, the complement of GT, and SEG corresponds to the false positives (extra).

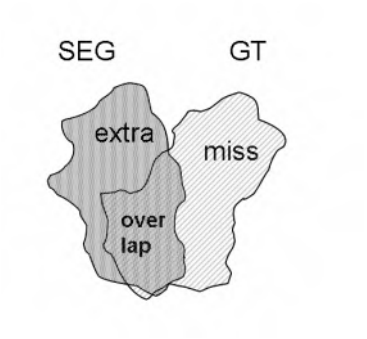


Figure 4.32: Comparison between the semi-automatic SEG and ground truth GT segmentations.

- The intersection of GT and $\overline{SEG}$ corresponds to the false negatives (missed flaring).

Based on this we define the similarity index (SI), the overlap fraction (OF) and the extra fraction (EF) as:

$$SI = \frac{2\|GT \cap SEG\|}{\|GT\| + \|SEG\|}, \tag{4.23}$$

$$OF = \frac{\|GT \cap SEG\|}{\|GT\|}, \tag{4.24}$$

$$EF = \frac{\|\overline{GT} \cap SEG\|}{\|GT\|}, \tag{4.25}$$

where $\|GT\|$ and $\|SEG\|$ denote the (pixel) area of the ground truth and our delineation respectively.
The SI is a measure for the correctly segmented area relative to the total area of the flaring in both the ground truth as well as the area of our technique, and ranges between 0 (worst) and 1 (best). The OF measures the correctly segmented area relative only to the flaring area in the ground truth image and again ranges between 0 (worst) and 1 (best). The EF measures the area that is incorrectly segmented as flaring and ranges between 0 (best) and 1 (worst). Results for the 8 test images can be found in Table 4.2.

We see that on average we obtain a SI of 0.889 and OF of 0.858. These numbers are significant since as recommended by [Zijdenbos et al., 1994] a good overlap occurs for coefficients above 0.700. The average EF is 0.060 which means very

Table 4.2: Accuracy of our technique compared to the panel of experts.

	SI	**OF**	**EF**
image 1	0.90	0.90	0.14
image 2	0.91	0.91	0.015
image 3	0.88	0.83	0.031
image 4	0.85	0.82	0.049
image 5	0.94	0.91	0.17
image 6	0.89	0.85	0.028
image 7	0.85	0.80	0.030
image 8	0.89	0.85	0.014
mean	**0.88**	**0.85**	**0.060**

few pixels are detected where there is no actual flaring.

Finally, to compare our segmentation to all manual expert delineations instead of an averaged golden standard we used a modified version of the Williams Index (WI) which computes the ratio between the average computer-to-observer agreement and the average inter-observer agreement [Chalana and Ki, 1997]. If a panel of $n+1$ observers, numbered 0 to n, segment s images, this statistic aims to compare observer 0 with the reference group of n observers. In our case, observers 1 to n are the physician's delineations and 0 is our computer delineation.

We first define the proportion of agreement $P_{j,j'}$ for a pair of observers (j, j') as the reciprocal of the disagreement between the delineations:

$$P_{j,j'} = \frac{1}{Q_{j,j'}}. \tag{4.26}$$

This disagreement $Q_{j,j'}$ can be defined in several ways. A common approach in contour comparison is to choose $Q_{j,j'}$ as

$$Q_{j,j'} = \frac{1}{s}\sum_{i=1}^{s} e(D_{ij}, D_{ij'}), \tag{4.27}$$

where $e(.,.)$ is a distance measure and D_{ij} denotes the delineation of image i by subject j. This distance is defined as follows. Suppose again delineations D_1, D_2 are defined as sets of m points $D_1 = \{\mathbf{x}_{1i},\ i = 1 \dots m\}$ and $D_2 = \{\mathbf{x}_{2j},\ j = 1 \dots m\}$. The distance $e(D_1, D_2)$ is then defined as the average difference between corresponding points $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ computed as:

$$e(D_1, D_2) = \frac{1}{m}\sum_{i=1}^{m} \|\mathbf{x}_{i1} - \mathbf{x}_{i2}\|, \tag{4.28}$$

with $\|.,.\|$ the Euclidean distance. Now that we defined the proportion of agreement, the average agreement between observer 0 and the reference group is computed as

$$P_0 = \frac{1}{n}\sum_{j=1}^{n} P_{0,j} \tag{4.29}$$

and the average level of agreement between the n observers by

$$P_n = \frac{2}{n(n-1)}\sum_{j}\sum_{j':j'\neq j} P_{j,j'}. \tag{4.30}$$

Once we compute all of these, the WI for observer 0 compared to n observers is defined as:

$$WI = \frac{P_0}{P_n}. \tag{4.31}$$

We can define a confidence interval (CI) for this index based on a leave-one-out criterion. Denote by X_i the set of delineations with delineation i excluded. We construct n such data sets, one for each expert observer, which lead to n estimates of the WI, denoted by WI'_i, $i = 1, \ldots, n$. The estimate of the standard error in the computation of the WI is then presented in as:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}[WI'_i - WI'_a]^2 \tag{4.32}$$

where

$$WI'_a = \frac{1}{n}\sum_{i=1}^{n} WI'_i. \tag{4.33}$$

As such, the 95 % CI for the estimate is $WI'_a \pm z_{0.95}\sigma$ where $z_{0.95} = 1.96$ is the 95th percentile of the standard normal distribution.

If the upper limit of the confidence interval of the WI is greater than 1, we can conclude that the measurement data are consistent with the hypothesis that our delineation agrees with the group of experts at least as well as the group members agrees with each other, i.e., our delineation is a reliable member of the group.

We used $m = 100$ equidistant points on each curve to compute the distance between two delineations. Table 4.3 shows the WI for the $s = 8$ test images where we notice all upper CI values are indeed above 1. Since all indexes are also above 1 we can even conclude there is even a better agreement between

Table 4.3: Williams Index for the panel of 12 physicians compared to our algorithm. The corresponding 95% confidence intervals are also shown.

	WI	**95% CI**
image 1	1.07	(1.01,1.13)
image 2	1.10	(1.04,1.16)
image 3	1.1	(1.05,1.15)
image 4	1.4	(1.30,1.51)
image 5	1.14	(1.05,1.23)
image 6	1.12	(1.06,1.18)
image 7	1.05	(1.00,1.11)
image 8	1.16	(1.11,1.21)

Table 4.4: Accuracy of the active contour technique compared to the panel of experts.

	SI	**OF**	**EF**
image 1	0.85	0.81	0.12
image 2	0.84	0.80	0.15
image 3	0.78	0.77	0.081
image 4	0.81	0.81	0.12
image 5	0.79	0.79	0.18
image 6	0.85	0.84	0.17
image 7	0.84	0.81	0.061
image 8	0.81	0.78	0.094
mean	**0.82**	**0.80**	**0.12**

our technique and the different expert delineations than between the individual expert delineations themselves.

4.3.6.3 Comparison to Active Contours

In Section 4.3.5 we already presented the visual difference between our segmentation and an existing segmentation algorithm based on active contours presented in [Stippel, 2004]. We will now compare the flaring areas of both techniques quantitatively by again calculating the SI, OF and EF measures for the active contour method and the constructed ground truth. The parameter settings used to obtain the active active contour delineations were $\delta = 0.2$, $\mu = 0.1$, $\alpha = 0.05$, $\beta = 0$, $\gamma = 1$ and $\kappa = 0.05$ using 80 iterations to compute the Gradient Vector Flow and 40 to iterate the snake. The code of the algorithm was provided to us by the author.

From Table 4.4 we notice our technique outperforms the active contours since as well the SI and OF are higher and the EF is lower. The difference in EF can be explained by the fact that the active contours tend to get stuck on bright speckle, which leads to more false positives.

4.3.7 Discussion

We presented a technique for the segmentation of flaring areas in US images. Our technique requires the interactive delineation of two bounding boxes. A first bounding box is used to exclude possibly disturbing features in the US images as the skull, and sinus. Although the definition of the ROI is free, and user-dependent, we noticed it results in a intra-class correlation of 95%, meaning the selection does not influence delineations significantly. We want to mention that it is utopic, when it comes to tissue segmentation, to try and segment fully automatically in a modality as difficult as US.

A second bounding box, containing the plexus is also selected manually. The presence of the plexus is crucial for the technique to perform well. We quantified the plexus-flaring relation in echogenicity and incorporated this to refine our texture segmentation map.

Morphological processing allows us to define the actual flaring boundary. Although we didn't really investigate the shape of the flaring boundary, the mere contour can be used to train new physicians in the detection of the pathology. The advantage of using morphological processing was that it is simple and easy to control. Structuring elements were fixed for all investigation although, if desired, the user can tune them.

Validation of our technique was one of the main topics of this study. Our first, and most important, experiment showed that our flaring segmentation can indeed be used as a reliable indicator for PVL. The Mann and Whithney tests proved extremely significant and based on the flaring area, we succeeded in further improving the sensitivity of detection in the data set of Chapter 3 to 98%.

In a second experiment, we constructed a golden standard from manual expert delineations. Based on our constructed ground truth we were able to show our method resembles the golden standard in term of overlap ratios and outperforms the state of the art method based on active contours. Besides that, the Williams Index showed our segmentation can be considered as a valid member of the class of individual expert segmentations. Yet, both the same index and the visual inspection of the expert delineations pointed out that there is indeed quite some difference amongst the manual delineations. Therefore, the main result to remember is the predictive value of our method. We showed our segmentation result is a statistically significant indicator for PVL and can *replace* the subjective manual delineation.

4.4 Segmentation of preterm brain ventricles

In Chapter 1, Section 1.3, we mentioned PVL is not only characterized by structural changes in the periventricular tissue. Dilations of the cerebral ventricles are considered indicative as well. Studies on the ventricle volume at term and later age are routine in MRI research. However, so far nobody has studied ventricle volumes quantitatively in US images. In this section, we propose a first low-complexity technique, developed in collaboration with the Biomedical Imaging Group of the Rotterdam Erasmus Medical Center, to measure the preterm ventricle volumes in 3D US representations of the preterm brain.

This section is structured as follows: in Subsection 4.4.1, we address 3D ultrasound since we work on volumetric US representations. This subsection describes both the acquisition and 3D reconstruction of sequences of 2D US images. Subsequently, we describe our 3D segmentation algorithm based on thresholding and mathematical morphology in Subsection 4.4.2. Some first results are presented in Subsection 4.4.3. We conclude with a discussion in Subsection 4.4.4.

4.4.1 3D ultrasound imaging

Over the last decade, the development of advanced transducers has resulted in volumetric US information. Based on the transducer movement, we can distinguish different approaches to obtain 3D US information.

In the *constrained* or motorized method the transducer is moved by a mechanical instrument during acquisition, see Fig. 4.33. Since in this approach the transducer movement is totally controlled and known, 3D reconstruction is possible. *Sensorless* systems, on the contrary, determine the transducer position by registering or aligning consecutive images to one another.

A third approach is based on 2D *phased-array* transducers which allow the sound beam to be steered in two directions and acquire real-time 3D volumes. Although this approach is less sensitive to body movements than the first two, its downside is that, up to now, it still results in relatively low-resolution volumes which often prevents it from being used for diagnostic purposes.

Finally, the cheapest yet most elaborate solution is attaching a (magnetic or optical) *position sensor* to a standard 2D transducer. When calibrated correctly, this results in a 3D freehand US system where specialized software allows for the right 3D reconstruction of a sequence of 2D US images.

A constrained approach is not feasible in our case since preterms are often ventilated and incubated. The sensorless approach is usually less accurate and 3D phased-array transducers were not available at the hospital. So, because of both cost-efficiency and practical reasons but also because of its high-resolution imaging, the freehand approach was chosen.

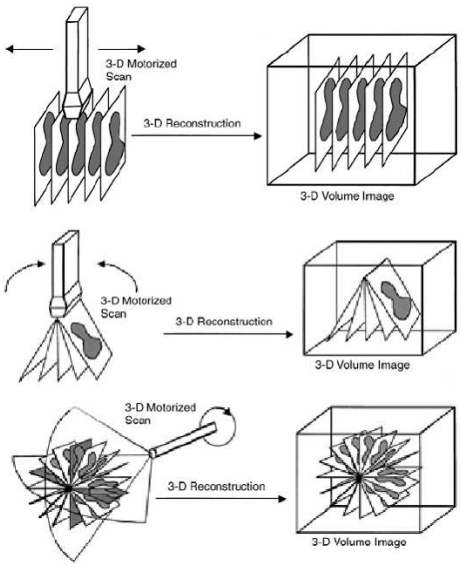


Figure 4.33: Mechanical solutions for 3D ultrasound scanning (Source: [Mischi, 2004]).

4.4.1.1 Freehand image acquisition

The 3D reconstruction of freehand 2D US images requires knowledge on the position of every 2D image captured. The position of a 2D plane in a 3D volume is defined by 6 parameters: 3 specifying the position of the transducer (x, y, z) and 3 specifying the orientation of the transducer (ϕ,θ,ψ).
To determine these parameters, at the Biomedical Imaging group of the Erasmus Medical Center, a 3D position sensor of a DC magnetic tracking system (AscensionTech Flock of Birds) was mounted on a curvilinear 8.5 MHz transducer of an Acuson Sequioa 512 US machine using a specially constructed socket, see Fig. 4.34 (left and middle). The US machine is connected to a notebook PC (Dell Latitude D810). The transducer is then placed on a fixed position of interest on the preterm's fontanelle and swept hence and forth to cover a conical brain segment. Specialized software, like StradWin[2], captures the image data and probe tracker position data simultaneously during scanning. In our application, images are acquired using StradWin 3.0 with the following settings:

- "Auto-correct from image data" turned off to keep the highest frame rate possible.

[2]Freely available at `http://mi.eng.cam.ac.uk/~rwp/stradwin`.

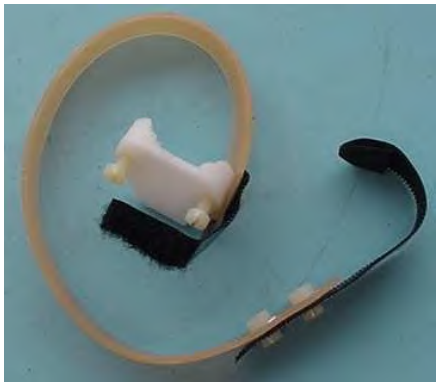



Figure 4.34: Left: Specially constructed position sensor socket. Middle: Acuson Sequoia 512 US machine.

- "Video rate" set to "Full speed" because of possible very small out of plane transducer movements.
- "Image format" set to greyscale and "Data format" set to B-scan.

4.4.1.2 Probe calibration

Using the freehand approach, the relation between the image (p), receiver (r), transmitter (t) and volume (c) coordinate systems is defined by:

$$\mathbf{x}_c = \mathbf{T}_{ct}\mathbf{T}_{tr}\mathbf{T}_{rp}\mathbf{x}_p \tag{4.34}$$

where $\mathbf{x}_p$ is a pixel location in the 2D acquired image, $\mathbf{x}_c$ is the resulting pixel location in the reconstructed volume and $\mathbf{T}_{ij}$ denotes a transformation from coordinate system i to coordinate system j. An illustration of the coordinate systems can be found in Fig. 4.35.
The most straightforward transformation is $\mathbf{T}_{tr}$ which is determined directly from the position sensor readings. In the context of reconstruction, $\mathbf{T}_{ct}$ is included largely as a matter of convenience. However, usually $\mathbf{T}_{ct}$ is omitted by aligning the reconstructed volume with the transmitter. This leaves us with $\mathbf{T}_{rp}$, which is determined through calibration.

Mercier at al. [Mercier et al., 2005] present an elaborate overview of different calibration methods together with their reported accuracy and precision. One of the easiest methods assuring an accurate calibration is the method using a single-wall phantom. This method is also described in [Prager et al., 1998b] and approximates $\mathbf{T}_{rp}$ by reconstructing the bottom of a water bath.

Because of its simplicity and ease of use, the water bath method is chosen to calibrate our 3D US configuration. A roughened perspex sheet (11 mm thick)

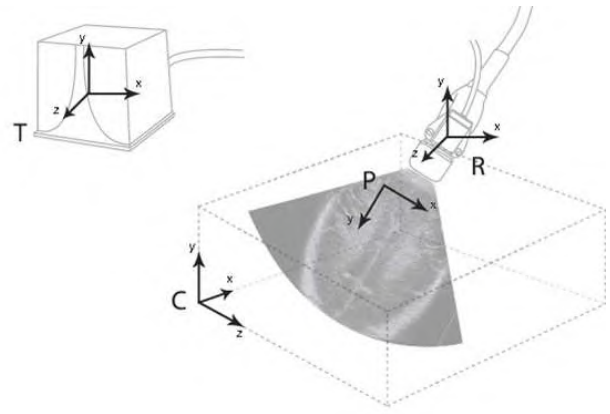


Figure 4.35: Different coordinate systems in 3D freehand ultrasound reconstruction. c stands for the reconstructed 3D US volume, p stands for the 2D US B-scan, r stands for the position sensor attached to the ultrasound transducer and t stands for the magnetic receiver.

is placed at the bottom of the bath to obtain better visual reflections in the US images. To get an accurate calibration, the velocity of sound in the water bath used must be equal to the considered velocity of sound in human tissue, and is again chosen equal to about 1540 m/s. Hereto, either a water temperature of 20°C and an alcohol percentage of approximately 10% or a pure water temperature of 50°C are reported as optimal configurations [Martin and Spinks, 2001]. We opted for the second configuration. The calibration procedure is then carried out exactly as described by Prager et al. in [Prager et al., 1998a], also using StradWin. Note that both the acquisition and calibration were done entirely by the Biomedical Imaging Group of the Rotterdam Erasmus Medical Center.

4.4.1.3 3D ultrasound reconstruction

Once the probe is calibrated and the images acquired, a rigid in-plane correction of the acquired images is performed. Due to small differences in probe pressure or patient movements during acquisition, slight out-of-plane distortions are possible. StradWin corrects the position of each individual 2D image plane through an intensity-based image registration, see Fig. 4.36.

Applying this procedure after acquisition has the same effect as applying it during acquisition, yet without lowering the video frame rate. The result of the combined rigid correction and acquisition, performed in StradWin, is shown in Fig. 4.37.

Following this correction, another package called StradX[3] is used to generate

[3]Freely available at `http://mi.eng.cam.ac.uk/~rwp/stradx`

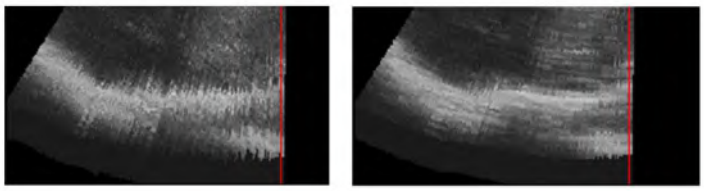


Figure 4.36: Left: a 2D sweep of US frames seen from the side, the serrated line at the bottom showing the out-of-plane distortion. Right: the same 2D sweep after rigid correction.

an isotropic 3D volume by linearly interpolating the reconstructed in-plane corrected sweep, see Fig. 4.38.

4.4.2 3D ventricle segmentation

Before explaining our ventricle segmentation technique three comments need to be made. First, we choose to segment slice by slice in the isotropic 3D volume rather than first segmenting all 2D images and then reconstructing them.

Secondly, the reconstructed 3D US volumes cover only certain sections of the brain. This often leads to partial ventricle information. Hence, the segmentation/reconstruction of the complete brain ventricle is not guaranteed for all volumes.[4]

Thirdly, since ventricles in essence are cavities filled with cerebrospinal fluid they have the nice property of being depicted as uniform black regions in the US sections, see Fig. 4.39 (a). However, the choroid plexus, producing the cerebrospinal fluid, is also located inside the ventricle and shows a completely different and more complex structure, see again Fig. 4.39 (a). The plexus has a speckle texture structure and is as such completely different from the black ventricle cavities, which makes the segmentation task more difficult. However, for diagnostic evaluation of the ventricles physicians are merely interested in the cavities, so in what follows we only focus of the segmentation of the cavities and not the plexus.

Given these comments, our technique works as follows. In a first step, we enhance the dark ventricle cavities by speckle reduction and histogram equalization. In a second step, a grey value thresholding combined with mathematical morphological operations determines the 2D ventricle areas. Once all 2D ventricle areas are gathered a final morphological smoothing step is applied to obtain a 3D ventricle volume.

[4]The main reason for these partial ventricles is that initially the sequences were not acquired for the purpose of ventricle segmentation but rather for tissue characterization through 3D registration to MRI. As such, during acquisition the attention was focused on the inclusion of the tissue of interest rather than the entire ventricles.

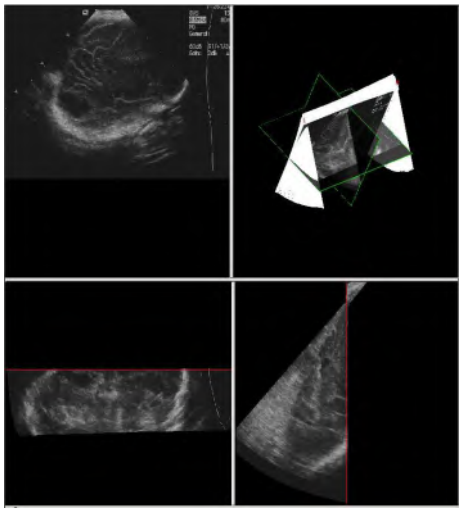


Figure 4.37: Example of the StradWin results. Upper left: a saggital 2D US B-mode image. Upper right: the sequence of acquired 2D slices. Lower left: A transversal reconstructed slice through the 3D US volume at the location of the transversal green plane in the upper right image. Lower right: a coronal reconstructed slice through the 3D ultrasound volume at the location of the coronal green plane in the upper right image.

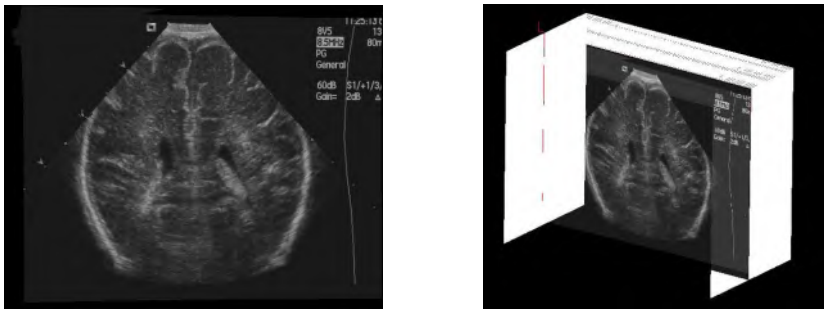


Figure 4.38: Left: reconstructed 2D US slice. Right: its place in the bounded isotropic 3D US volume created by StradX.

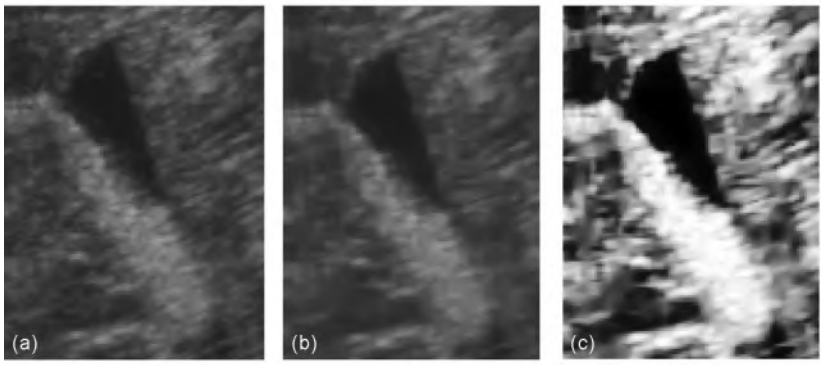


Figure 4.39: (a) a cut-out of the US image showing the ventricle cavity as a dark spot in the middle of the image. The plexus choroid is located right below it as a bright feature. (b) the denoised version of (a) using the GenLik filter with a threshold $T = 4$ (c) the histogram equalization of (b).

4.4.2.1 First step: 2D contrast enhancement

To enhance the ventricle cavities we reduce surrounding speckle using a modified version of an edge-preserving speckle-reduction filter, called the GenLik filter. The exact details of this filter will be explained in Chapter 6, Section 6.4.1, but this wavelet-based shrinkage filter basically removes speckle in the interior of the ventricle while preserving and enhancing its surrounding edges, see Fig. 4.39 (b). Next, we enhance the edge contrast further by equalizing the histogram of the filtered image using equation (3.2), see Fig. 4.39 (c).

4.4.2.2 Second step: 2D ventricle area

Now that we have enhanced the image we can consider different edge-detection operations to determine the ventricle boundaries. High-level edge-detection, e.g., based on active contours, usually demands some user-interaction which hinders fast and automated processing.
Therefore, we compute a simple threshold value instead and process the thresholded image using fast binary mathematical morphology operations. The enhanced US images are thresholded at a grey value of 30. All grey values above 30 are set to 0, all grey values below 30 are set to 1. This value was determined empirically by inspecting the effect of multiple thresholds on 10 ground truth test images taken at different positions along the reconstructed volume. The criterion for a good threshold was that the resulting ventricle area matched its ground truth (manual) ventricle area delineation as well as possible, in terms of the similarity index SI introduced in equation (4.23). On average, over the 10 test samples a grey value of 30 resulted in optimal SI. An example of this

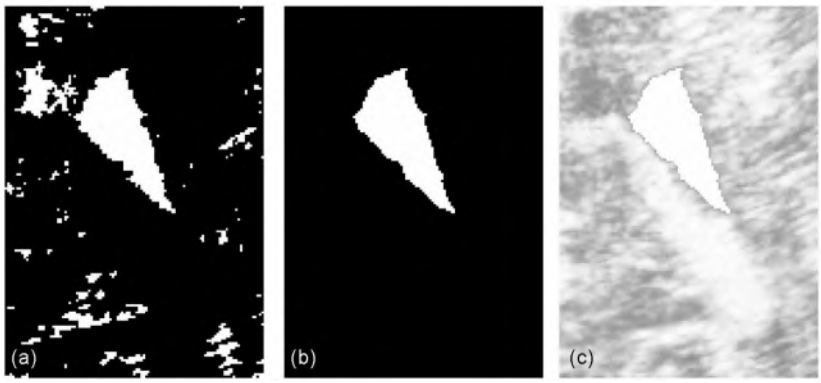


Figure 4.40: (a) the thresholded version of the histogram image. (b) the result after an opening by reconstruction followed by a closing. (c) the segmentation result backprojected on the histogram equalization.

thresholding is presented in Fig. 4.40 (a).

After that, remaining image objects not related to the ventricle are excluded from the threshold map again using mathematical morphology operations. By applying an opening by reconstruction with a spherical structuring element of radius 6, we sieve out the surrounding objects. Once these surrounding objects are removed we morphologically close the threshold map with a disc of radius 3 to obtain a smooth ventricle area. The combined result of the morphological processing is shown in Fig. 4.40 (b) and backprojected onto the enhanced image in (c).

4.4.2.3 Third step: 3D reconstruction

Steps one and two are repeated for each of the 2D slices in our reconstructed volume independently. Although consecutive ventricle areas will never differ a lot, due to the remaining speckle small discontinuities on the ventricle surface are unavoidable. To overcome these discontinuities we smooth the volume of consecutive 2D segmentations by morphologically closing it with a 3D ball of radius 3 pixels.

4.4.3 Experimental results

The left and right ventricles were segmented from reconstructed volumes of two preterm infants. Fig. 4.41 shows four different ventricle results. Note that, as mentioned in one of the comments in the former section, the US sequences do not necessarily contain the entire ventricles since the transducer's sweep is restricted to a specific part of the brain. Consequently, up to now, the

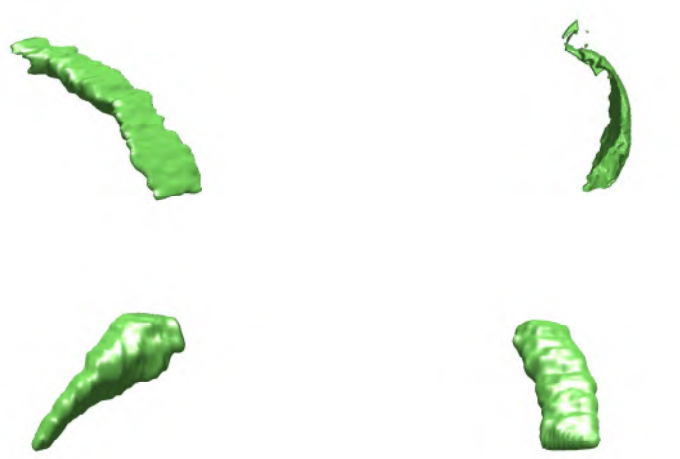


Figure 4.41: 4 different Examples of reconstructed 3D ventricle volumes.

segmentations we obtain are related to the parts of the ventricles that *are* present in the volumes.

Therefore, we cannot present any reliable quantitative results on the entire brain ventricle volumes yet. The only feedback we received is that the ventricle shapes have been scored as accurately as possible through a visual inspection by experts.

The entire segmentation process is fully automated and relatively fast. In StradWin and StradX, it takes about 120 seconds to reconstruct the US volume and about 100 seconds to segment the ventricle in Matlab (AMD 1.8 GHz processor).

4.4.4 Discussion

The aim of this small study was to present a low-complexity, automated algorithm to segment lateral ventricles in real preterm US volumes. First, a sequence of 2D US images was reconstructed into a 3D US volume based on a freehand US system using magnetic probe tracking information.

Setting up the US configuration takes about 30 minutes. This is due to the assembly of the different parts of the system (US machine, position tracking system, laptop) and the calibration procedure. If the system needs to be used in daily clinical practice this is too long and other options should be considered. Installation time could be decreased significantly by using a transducer with a permanently attached or internal position sensor. In that case the calibration procedure can be skipped.

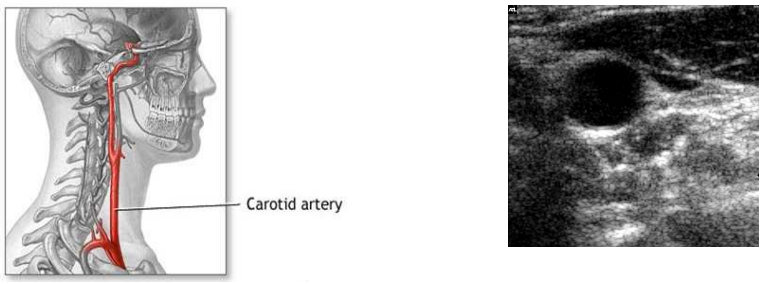


Figure 4.42: Left: the position of the carotid artery is highlighted. Right: transversal US slice where the carotid is visible as the big dark spherical region.

In the near future, the quality of 3D phased-array transducers is expected to improve sufficiently for 3D neonatal brain imaging. The total configuration could then be replaced by such a system which saves installation and calibration time.

Once the 3D volumes are acquired they are reconstructed using specialized (free) software. Subsequently, the combination of an image contrast enhancement, thresholding and morphological operations results in a slice by slice ventricle area delineation. These 2D areas are smoothed and merged into the final ventricle reconstruction.

Absolute, quantitative volume measurement are not yet possible and require both better 2D sequences and ground truth data such as, e.g., MRI volume measurements. However, preliminary visual inspection of the 3D ventricles by expert physicians shows our preliminary results are reliable.

4.5 Carotid artery segmentation

A final US segmentation application lies in a different medical field. The carotid artery is susceptible to a number of vascular diseases as there are: carotid dissection, the pathological condition where blood intrudes the intimi media and atherosclerosis, where thickening, hardening and loss of elasticity of the arterial wall results in impaired blood circulation. Less frequent but more severe is a carotid aneurism, a widening of the carotid lumen. Since the carotid is close to the body surface, US imaging is used frequently in diagnosing these different pathologies, see Fig. 4.42.

In this context, the Cardiovascular Mechanics and Biofluid Dynamics Research Unit of our university developed a 3D reconstruction method to perform flow measurements in a carotid bifurcation [Glor, 2004]. This reconstruction was based on a slice-to-slice 2D segmentation of the carotid, then combined into a

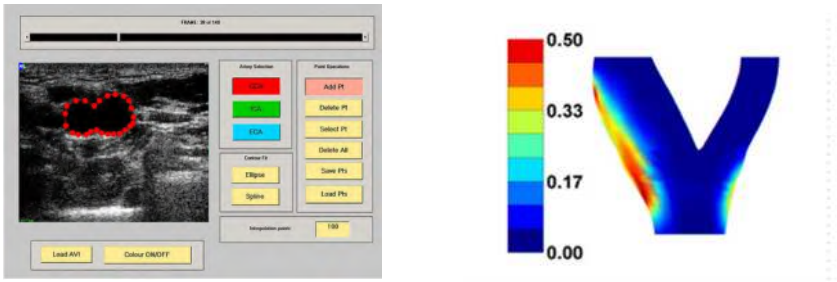


Figure 4.43: Left: the existing segmentation method where vessel contour points are annotated. Right: A 3D bifurcation reconstruction result where the color code corresponds to the Oscillatory Shear Index (OSI), a blood flow related measurement.

3D volume using both interpolation techniques and magnetic probe tracking information.

The bottleneck of this technique is the 2D slice-to-slice segmentation where multiple marker points on the vessel wall had to be selected manually. These points are used to fit a smooth cubic spline or ellipse, see Fig. 4.43. Since there are about 114 images per 3D scan and as each scan has to be annotated with 20 up to 30 points, this procedure is very time-consuming. As such, in this section we propose a first approach towards an automatic 2D segmentation again using morphological operations.

4.5.1 The proposed method

The objective in this application is again to segment a blood vessel rather than some organ tissue. Also, to save computation time, we do not preprocess or filter the images in advance.

The first operation we do apply is again thresholding of the image. We determined the threshold value in the same way as with the brain ventricle and found a grey value of 40 as optimal, see Fig. 4.44 (b). Consequently, we again proceed with an opening by reconstruction to clear the carotid further from its surrounding structures or cavities. The opening by reconstruction is done by spherical discs with radii of 3 pixels, see Fig. 4.44 (c). Finally, to fill the cavities in the borders of the vessel boundary we close the image using a spherical structuring element of the same radius, see Fig. 4.44 (d). A morphological gradient operation leads to the actual vessel boundary segmentation, see Fig. 4.44 (e) and (f). On overview of consecutive segmentations is found in Fig. 4.45.

4.5.2 Discussion

From Fig. 4.45, we notice that we succeed in segmenting the 2D vessel boundary relatively accurately, and this in a fully automatic way. Although the final

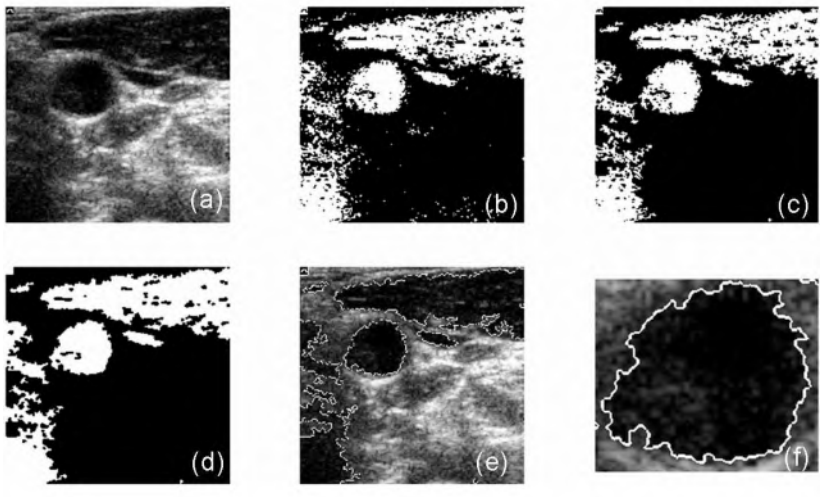


Figure 4.44: (a) the original 2D US slice, showing the carotid. (b) the thresholded version of image (a). (c) the thresholded image after opening by reconstruction. (d) the result of closing the reconstructed image. (e) the result after applying the gradient operation to the opened image. (f) cut-out of the segmented carotid.

boundary is still quite rough, due to the speckle in the vessel wall, a smoothing step, similar to the one for the brain ventricles, could be applied in the reconstruction step. In future, this segmentation method could even be incorporated in the existing reconstruction algorithm by sampling the contour down to 20 or 30 equidistant points as an initialization of the method. In that way, we both eliminate the user-interactivity and preserve the desired edge-smoothness. However, since we did neither have the reconstruction program nor the probe tracking information at our disposal, we were unable to perform a 3D reconstruction based on the segmentations ourselves yet.

4.6 Overall conclusions and hints for future work

In this Chapter we explained how mathematical morphology combined with texture information and image thresholding can be applied to US segmentation. Primarily, a segmentation algorithm was developed to determine to which extent flaring spreads out in Periventricular Leukomalacia. This algorithm started from the texture characterization of pathological tissue and incorporated quantified information on the plexus-flaring intensity ratio. Finally, morphological operations result in the contour delineation of the flaring regions.

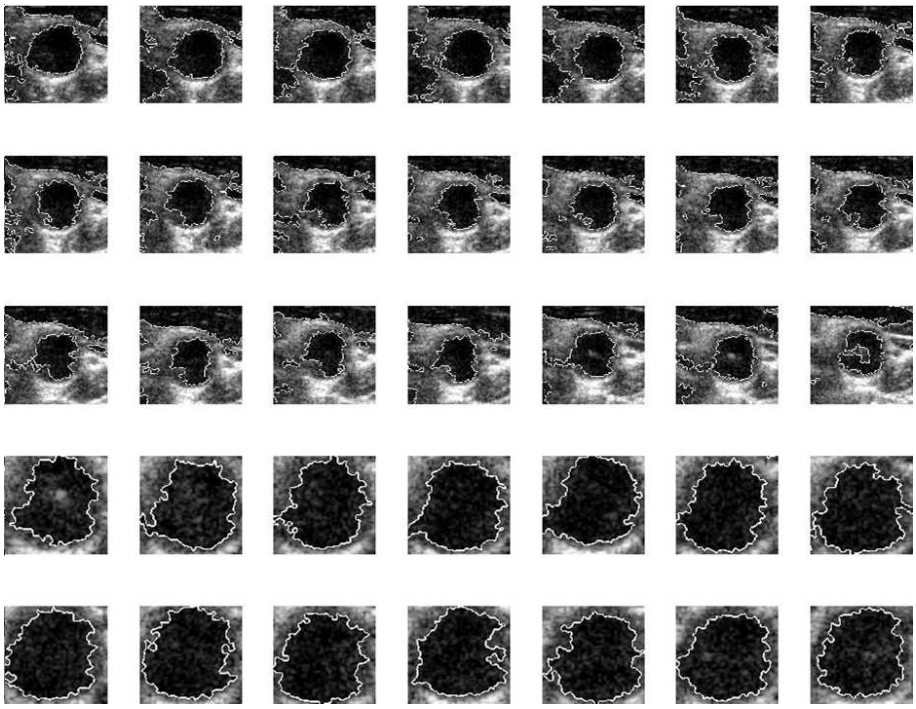


Figure 4.45: Segmentation results for consecutive 2D US slices. The top 3 rows who the carotid and its neighborhood. The bottom two rows zoom in on the carotid itself.

Using this method we are able to describe flaring to its full extent. Through experiments we showed that by segmenting the flaring regions, we succeed in an improved characterization of pathological white brain matter.

Besides that, we compared our method to ground truth information obtained by averaging manual expert delineations. Although, based on this information, we showed our method corresponds well to expert delineations and outperforms an existing active contour method, we mentioned that we have to be cautious in interpreting these results. Namely, we also showed that our delineation corresponds better to the group of individual expert delineations than the expert delineations correspond to one another. Although we did not really need the manual (ground truth) expert delineations to prove the diagnostic power of our method, it is worthwhile to tackle this validation problem from another point of view.

If golden standard information is unlikely to be found in US imaging, we can switch to other modalities and cross-validate with, e.g., MRI. This is exactly the main focus point of our next Chapter where we align 3D MRI and 2D US images in order to enable a cross-validation of our current algorithm.

Next to the segmentation of flaring regions, we also presented a morphology-based method for 3D brain ventricle segmentation. Starting from a sequence of 2D US slices acquired with probe-tracking, we reconstructed the parts of the ventricles present in the sequences. Quantitative measurements are not possible yet, since up to now we lack both image sequences that contain the complete ventricles and ground truth information. However, as we will also explain in the next Chapter, a future application of this segmentation is the registration of 3D US and 3D MRI data.

A preliminary result on the application of morphological segmentation to the carotid artery was also presented. However, due to the lack of probe-tracking information, a complete reconstruction of the carotid was not possible yet. As such, both a quantitative comparison to the existing (time-consuming) method and bloodflow simulations on our reconstructed model are the main applications for future research.

Finally, we want to note that the techniques presented in this Chapter have also resulted in an application in a non-medical application field. The transport of perspiration is an important contributor to the thermal comfort of fabrics worn next to the skin. Due to body activity, the wearer perspires and the cloth that is in contact to the skin wets. These moistured fabrics reduce the body heat and will eventually fatigue the wearer.

As such, cloth worn next to the skin should possess two important properties: Primarily, it must be able to evaporate the perspiration from the skin surface and secondly transfer the moisture to the atmosphere, also called wicking. Therefore, measuring wicking in fabrics is of topical interest. In collaboration with the Research Unit of Plasma Technology of our university a new wicking experiment was performed to measure wicking behavior and investigate the

influence of plasma treatment of the clothing on the wicking. A morphology-based segmentation algorithm succeeded in measuring the wicking area semi-automatically and was proven effective for determining the wicking rate of woven and non-woven fabrics [Morent et al., 2006].

Chapter 5

Multimodal image registration

In Chapter 3 we described the texture characteristics of pathological and non-pathological white brain matter. In Chapter 4 we combined this information with morphological operators to develop a new interactive flaring area segmentation algorithm. To validate this algorithm, we compared it to both manual expert delineations and an existing algorithm on active contours.
In this Chapter, we propose a multimodal 2D US to 3D MRI registration algorithm to semi-automatically align the segmented US flares and the corresponding regions in the MR images. This allows experts to further analyze the pathology and to cross-validate our segmentation results with the modality considered the golden standard in PVL outcome at term and later age.

The registration problem, i.e., the geometrical alignment of a 2D plane and its corresponding position in a 3D volume, is not trivial due to the high number of degrees of freedom involved. Additionally, the registration of US and MRI data is extra challenging due to the differences in imaging characteristics.

5.1 Introduction

The advent of new medical imaging modalities such as Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT), Diffusion Tensor Imaging (DTI), functional Magnetic Resonance Imaging (fMRI) and magnetoencephalograms (MEG), combined with an exponential increment in computer processing power has led to a growing interest in merging complementary information. The main rationale supporting this *multimodal* research is that the *fused* image often reveals extra diagnostic information.

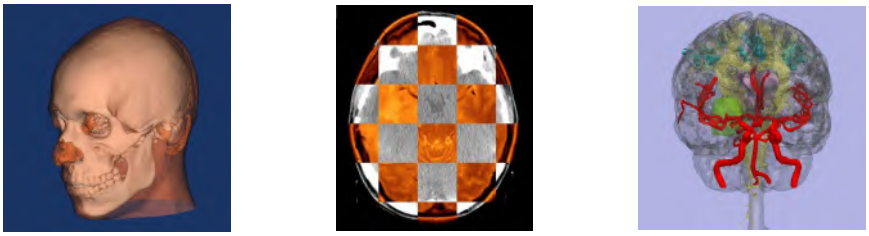


Figure 5.1: Left: 3D fusion of a skull and skin model derived from MRI images. Middle: fusion of CT and MR brain images. Right: fusion of 3D PET and MRI brain volumes.

Examples of multimodal applications are ubiquitous: in dentofacial registration 3D volumes of the dental cast and face of the same patient are merged to plan orthopedic surgery.
In radiotherapy planning Computed Tomography (CT) imaging is used almost exclusively. However, the combined use of MRI and CT is beneficial as the former is better suited to delineate tumor tissue while the latter is needed for accurate radiation dose computation.
In brain research, PET imaging reflects the functionalities of different parts of the cerebral neurosystem. If this information is fused with normal structural MRI, it provides us the combined functional behavior and structural constitution of pathological areas. Three visualizations of how these fused multimodal examples are usually represented are shown in Fig. 5.1.

Image registration, or finding the optimal geometrical transformation that aligns image structures, is the necessary preprocessing step to this image fusion. Because this is usually quite a complex procedure, registering images manually is often too time-consuming in clinical practice. Consequently, (semi)-automatic registration is one of the hottest topics in CAD nowadays.

A distinction is made between *rigid* and *non-rigid* image registration. A registration is called *rigid* when its geometrical transformations preserve distances, e.g., translations and rotations. Rigid registration is applied when no structural deformations are expected between the different modalities. Since the human brain is contained by a rigid skull, this is also the kind of registration we apply. In other applications, such as pre- to intra-operative registration where tissues or organs might deform due to pressure changes or when there is anatomical variability across the images to be registered, as in atlas-studies, rigid registration will not do the trick. In those cases, transformations that alter shapes and distances, such as scaling, are needed. Since we do not use non-rigid registrations we will not discuss them further but refer to [Maintz and Viergever, 1996] for a broad overview.

Apart from a rigid versus non-rigid subdivision, we can classify registration methods into six different groups according to the principles they are based on:

1. ***Feature-based*** methods. These methods use sets of corresponding points

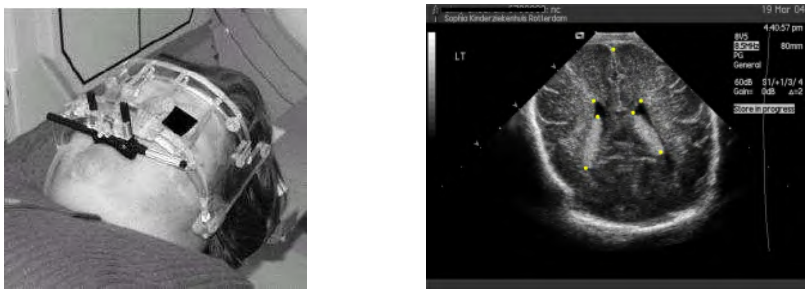


Figure 5.2: Left: example of extrinsic landmarks in the form of a mask placed on the face of a patient. Right: example of intrinsic landmarks, white dots, in the US brain as identified by a physician.

assigned to both images. These points can either be real anatomical features or artificial markers attached to the body prior to acquisition. These salient points are often called *landmarks* or *fiducial points*, see Fig. 5.2. The goal of these registration methods is to find the geometrical transformation that correctly aligns corresponding landmarks in both images. Typical measures for the displacements of the landmarks are based on average (Euclidean) distance errors.

2. ***Intensity-based*** methods. This is the most widely used approach in medical image registration. By comparing the pixel or voxel (the pixel analogous in a volume) values at corresponding positions in two images, a similarity between the images is determined based on a predefined criterion. Commonly used criteria are the minimal Mutual Information, which we will discuss later on, the Maximum log-Likelihood and minimal Kullback-Leibner distances. The reason why intensity-based methods are so popular is because they need little preprocessing as opposed to, e.g., the feature-based method where landmarks or fiducial points need to be determined manually.

3. ***Segmentation-based*** methods. These methods are related to the feature-based approach but differ in the fact that the set of landmarks or fiducial points is replaced by segmentations of anatomical features. Optimal transformation parameters then correspond to the optimal alignment of corresponding segmentations. The drawback of these methods is that the registration accuracy is limited to the segmentation accuracy.

4. ***Gradient-based*** methods. These relatively new methods use gradient images and normals to surfaces or curves defined in the images. The goal of these registration methods is to look for corresponding image surface normals or gradients, considering their amplitudes and orientations.

5. ***Hybrid*** methods. These methods combine the methods described above. Typically multivariate similarity-measures are constructed that incorpo-

rate, e.g., both intensity and gradient information or intensity and fiducial distance errors.

6. ***Non image-based registration*** methods. Is seems paradoxical that registration of (multimodal) images can be non image-based but this is possible if the imaging coordinate systems of both modalities are somehow calibrated to each other. This usually necessitates combining both imaging modalities in the same physical place and immobilizing the subject. US imaging is one of the few modalities where this is possible. Suppose that US images are acquired with a probe tracker during a CT or MRI scan. We can then register the US image and the CT or MRI volume coordinate system solely based on the probe tracker information, i.e., without using the information in the images.

Mostly, rigid intensity-based registration methods are presented that align 2D images or 3D volumes. The 2D to 2D case is often the easiest to handle since the number of degrees of freedom is restricted to a translation and a rotation in the plane. The 3D to 3D case is computationally more expensive since degrees of freedom are added due to extra translation and rotation parameters. However, due to the fact that we possess 3D data, there is usually a lot more image content at hand than in the 2D case to reliably find the alignment.

In 2D to 3D registration we have the same number of degrees of freedom as in the 3D to 3D case and at each positioning of the 2D image in the 3D volume there is only as little overlap information as in the 2D to 2D case. This complicates the 2D to 3D registration method significantly. Nevertheless, in some cases, due to physical limitations in time or space or due to the lack of appropriate imaging equipment, 2D to 3D registration is the only option.

It is clear that, due to similar image characteristics, registering images from the same modality, i.e., unimodal registration, is almost always easier than multimodal registration. The fact that the characteristic speckle in US images is not present in most other image modalities makes that few multimodal registration methods exist involving US as one of the modalities.

In [De Bruin et al., 2002, Amin et al., 2003] feature-based methods are used to bypass the speckle problem in US to MRI and CT registrations. Their approach however only works when landmarks are clearly detectable on both modalities. In the absence of that kind of information others [Shekhar and Zagrodsky, 2002, Chen and Abolmaesumi, 2005] use Mutual Information based methods either combined with a speckle suppressor or restricted to regions of interest in the images in a 2D to 2D registration.

We present a rigid intensity-based registration method for 2D US to 3D MRI preterm brain data to investigate if valuable information on PVL can be obtained from the simultaneous inspection of the US image and the corresponding MR image. We choose an intensity-based over a feature-based approach since, apart from the central ventricles, we lack good landmarks in the preterm brain. Working on these ventricles alone, in a segmentation- or feature-based

approach, is also no option in a 2D to 3D approach since we do not posses a 3D ventricle segmentation algorithm. However, if in future a MRI ventricle segmentation would become available, a segmentation- or feature-based approach using our 3D ventricle segmentation technique of Section 4.4.2 could become interesting.

The drawback of an intensity-based method however is that we have the granular US speckle pattern on the one hand and the more homogeneous MR image on the other. As such, simply comparing intensities at corresponding places will not work. To control the speckle influence, we filter the images again using a modified GenLik filter. Subsequently a Mutual Information metric, taking into account the distribution of grey values rather than the individual grey values, is optimized by a regular step gradient descent searching algorithm. As, to our knowledge, this is the first registration algorithm for 2D US to 3D MRI brain registration we can, for now, only compare our results to the manual registrations of medical experts. We do this in a novel way, i.e., by using a virtual environment called the I-SPACE.

The structure of this Chapter is as follows: in Section 5.2, we describe the experimental setup, i.e., the US and MRI acquisition procedure. Subsequently, we discuss US and MR image preprocessing in Section 5.3, followed by the explanation of the registration algorithm in Section 5.4. In Section 5.5, we describe our experimental results and in Section 5.6 we present the validation of our results. In Section 5.7, we compare flaring in US and MRI based on our registration algorithm. Finally, the overall conclusions and hints for future work are presented in Section 5.8.

5.2 Experimental setup

A total of 28 coronal 2D US brain images were analyzed, obtained from 28 preterms at a postconceptional age of 32 weeks. All images were captured in the first 3 days after birth at the Sophia Children's hospital, by one medical expert using the Acuson Sequoia 512 ultrasound machine and a hand-free 8.5 MHz curvilinear phased-array probe, and fixed machine settings. Similar as in Chapter 3, we obtained no probe tracking information but all scans were captured under an angle of approximately 45 degrees to the coronal plane so that both central ventricles and the flaring zones are clearly visible. The US images have a resolution of 768×576 pixels with a physical size of 0.16×0.16 mm, see Fig. 5.3 (left).

Besides that, 56 MRI volumes (two per patient) were acquired at day three using a General Electric 1.5T Sigma Infinity scanner: a T1 weighted volume containing 85 images at a 256×256 pixel resolution and a voxel size of $0.85 \times 0.85 \times 1$ mm and T2 weighted volume containing 20 images at a 256×256 pixel resolution at a voxel size of $0.70 \times 0.70 \times 4$ mm, see Fig. 5.3 (middle and right). The acquisition time for the MRI images ranged between 4 and 5 minutes and preterms were not immobilized nor sedated.

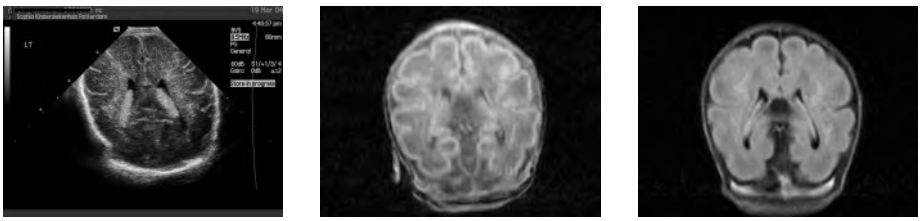


Figure 5.3: Left: a 2D US images used in the registration study. Middle: a coronal 2D intersection of a T1 MRI volume. Right: a coronal 2D intersection of the T2 MRI volume.

5.3 Image preprocessing

We are again restricted to the medical US and MRI acquisition protocols used in the hospital, which obviously influence the usability of the data for our purpose.

Regarding the US images, see Fig. 5.3 (left), we notice there is a lot of information in the image frame that is irrelevant for the registration algorithm. First of all, inherent to the probe used, the image has a conical shape. This implies that almost half of the image content is not subject-related and contains information as machine settings, scan date, etc.

Since this part of the image will confuse the intensity-based registration, we mask the image by defining a conical region of interest, as shown in the upper part of Fig. 5.4. However, the lower corners of this mask, outside of the skull, contain little valuable information since the head of the preterm is relatively small (6 to 8 cm in diameter) and the skull is too thick for US waves to penetrate. As such, we further adapt the mask by omitting also the corners, as shown in the middle part of Fig. 5.4. Eventually, since we are mostly interested in registering the periventricular white brain matter, we also exclude the skull from the mask to emphasize the correspondence of the inner structures of the brain, see lower part of Fig. 5.4.

Apart from the irrelevant information in the US images, there is the ever-present speckle. Speckle is both the main source of information and the main obstacle for a multimodal intensity-based registration method since its characteristics reflect both relevant and irrelevant tissue characteristics and differ from most other imaging modalities. Therefore, we reduce irrelevant speckle and smooth image features by filtering the US images with a modified version of the GenLik filter. The technical details of this filter will be explained in Section 6.4.1 of the next Chapter but basically the filter reduces speckle, based on a multi-resolution wavelet approach in a probabilistic framework, while maintaining the relevant features in the images.

The result of this US filtering is shown in Fig. 5.5 where it is also compared to a Gaussian low-pass filter and a grey mean value filter. We notice that

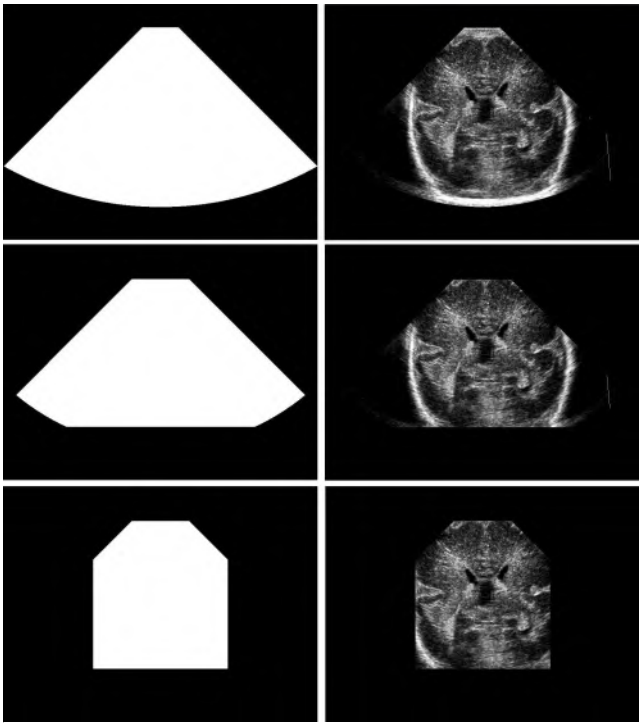


Figure 5.4: Masks for the US images in the registration method.

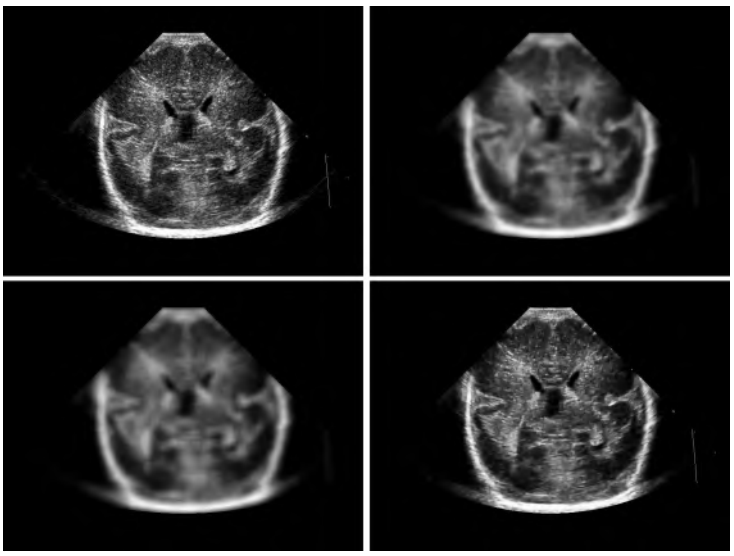


Figure 5.5: Upper left: the original US image. Upper Right: low-pass filtering with a Gaussian filter kernel. Lower Left: mean grey value filtering with a 15×15 pixel window. Lower Right: GenLik filter with a threshold $T = 4$.

the GenLik filter succeeds in reducing speckle while preserving more structural information (such as edges) than standard filters.

Concerning the MRI volumes, the influence of the acquisition protocol is much bigger. Obtaining high-quality MRI information from non-sedated preterms is not trivial. Artifacts due to patient movements are common, so the imaging time is often limited to 4 up to 5 minutes per scan. In our case, in the T2 volumes this leads to volumes with voxels of $0.70 \times 0.70 \times 4$ mm in size. As such, the resolution in the Z-direction is about 6 times the resolution in the X- and Y-direction[1]. To illustrate this, Fig. 5.6 shows a 3D volume rendering of T2 MRI data. The effect of a low Z-resolution is even more emphasized when we look at an intersection along the Z-direction, see Fig. 5.7 left.

Consequently, we upsample the MRI volume, through interpolation, to obtain a smoother image, i.e., with a higher sampling density in the Z-direction. Note that in doing so, we do not incorporate new patient information in the volume but rather create information at intermediate positions based on the scan information already present. We choose to interpolate the MRI volumes using B-splines since their finite support and smoothing qualities [Unser, 1999] make them very suited for image interpolation.

B-splines are piecewise polynomials of a certain degree or order n and defined

[1]In what follows, we consider the X- and Y-direction to be the axes in the coronal plane.

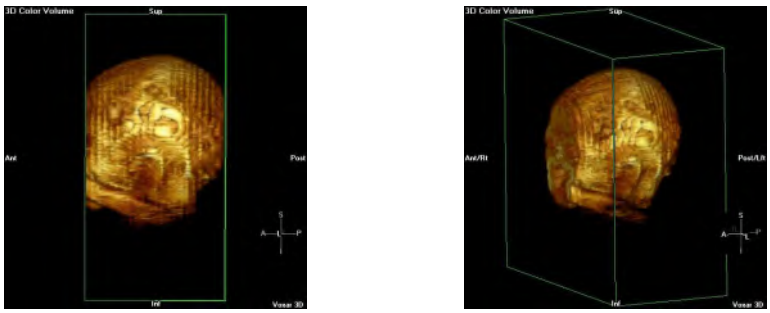


Figure 5.6: Two volume renderings of a 3D MRI data set where the poor image resolution in the Z-direction is clearly visible.

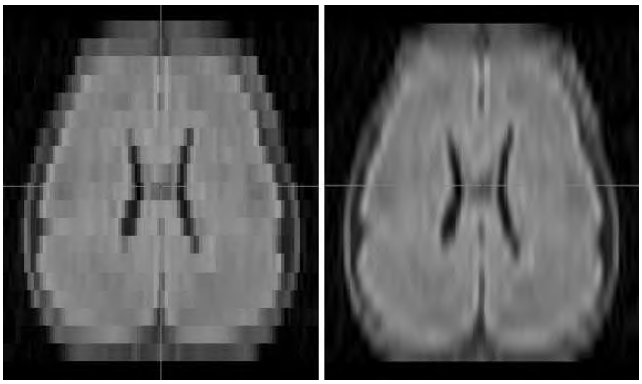


Figure 5.7: Left: an original MR image along the Z-direction. Right: the B-spline interpolated MR image.

as:

$$\beta^n(x) = \frac{1}{n!}\sum_{k=0}^{n+1}(-1)^k\binom{n+1}{k}\left(x-k+\frac{n+1}{2}\right)_+^n, \tag{5.1}$$

where $n! = 1\times 2\times\ldots\times n-1\times n$, $\binom{a}{b}$ is a binomial coefficient[2] and the function $(x)_+^n$ is defined as:

$$(x)_+^n = \begin{cases} x^n, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{5.2}$$

[2] a binomial coefficient is defined as $\binom{a}{b} = \frac{a!}{(a-b)!b!}$

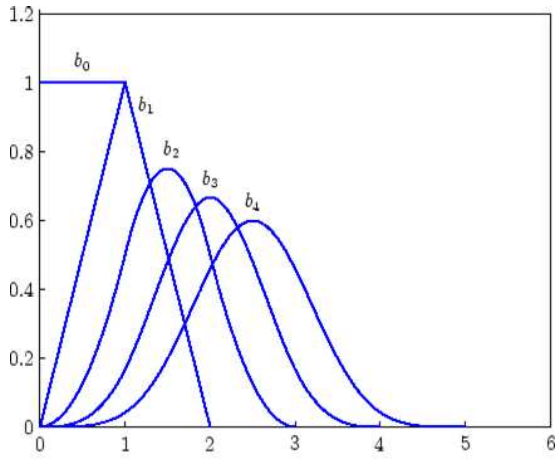


Figure 5.8: B-splines of degree 0 to 4.

It is also possible to define B-splines in recursive way [Unser et al., 1993] as:

$$\beta^n(x) = \frac{(\frac{n+1}{2} + x)\beta^{n-1}(x + \frac{1}{2}) + (\frac{n+1}{2} - x)\beta^{n-1}(x - \frac{1}{2})}{n}. \tag{5.3}$$

Fig. 5.8 shows the first five (degree 0 to 4) B-splines. To interpolate, the input signal is preprocessed with a recursive prefilter $(b_n)^{-1}$ that samples the signal, followed by a convolution with the B-spline function. Suppose I is the original 1D input signal, the B-spline interpolation I' of degree n is then defined as:

$$I' = \sum_{m=-\infty}^{m=\infty} ((b_n)^{-1} * I)(m)\beta^n(x - m), \tag{5.4}$$

where $*$ denote the convolution operation. Depending on the degree of the B-spline, different types of interpolation are defined. The 0-degree spline β^0 is known as the kernel of *nearest-neighbor interpolation*. The first-order B-spline β^1 is called the *triangular function* which is the kernel of linear interpolation. The B-splines for $n = 2$ and $n = 3$ are called the *quadratic* and *cubic* B-splines, respectively.

These 1D kernels can be easily extended to higher dimensions by separable filtering the image (or volume) with B-splines along each of the dimensions [Unser, 1999]. Usually, for registration purposes a trade-off has to be made between the degree of upsampling and the size of the interpolated image. Convolving our MRI volume with B-splines of degree 4 resulted in isotropic voxels of $0.5 \times 0.5 \times 0.5$ mm and upsampled MRI volumes of about 30 MB. Fig. 5.7 (right), shows the upsampling effect on one of the MRI slices. We notice a smoother and more natural looking image.

5.4 Proposed registration algorithm

As mentioned in the introduction, registration amounts to finding the geometrical transformation that optimally aligns images. This alignment is performed in a given coordinate system, commonly an Euclidean space. Since the US and MR images/volumes have a different (pixel) resolution we can not just compare pixel to pixel and therefore reason in terms of the physical spatial coordinates of each pixel in that space. For example, suppose we have an image of physical dimensions 21×12 cm and a resolution of 7×6 pixels. The spacing between pixel coordinates will then be 3 cm in one direction and 2 cm in the other, see Fig. 5.9. As such, images are from now on represented by the grid of coordinates of their pixel centers.

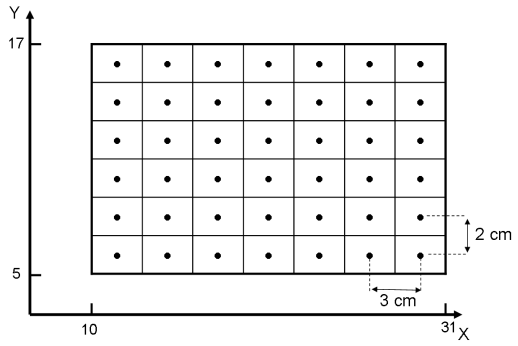


Figure 5.9: Coordinate spacing of the input images.

Once the images/volumes are represented in our coordinate space, the 2D to 3D registration can start. First of all, we have to decide which image will be fixed in our space coordinate system, and which one will be rotated and translated. We fix the MRI volume and align the US scan with it since less coordinate points have to be transformed in the case of the 2D US image, which speeds up the registration process.

From now on we will use the notation $V = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$ for the set of coordinates of the pixels of the US image before rotation and translation and $U = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m\}$ for the set of coordinates of the voxels of the reference MRI volume (both in a 3D Euclidean space). Furthermore, $\mathbf{r}'_i = R\mathbf{r}_i + \mathbf{t}$ represents the coordinate point $\mathbf{r}_i$ after a rigid transformation, with R a rotation matrix and $\mathbf{t}$ a translation vector to be defined later on. Finally, $f_V(\mathbf{r}_i)$ denotes the grey value assigned to a point $\mathbf{r}_i$ in the US image and $f_U(\mathbf{s}_j)$ the grey value of a coordinate point $\mathbf{s}_j$ in the reference MRI volume.

For each transformation of the US image, a similarity between the grey values (of the coordinate points) of the moving image and fixed volume is calculated.

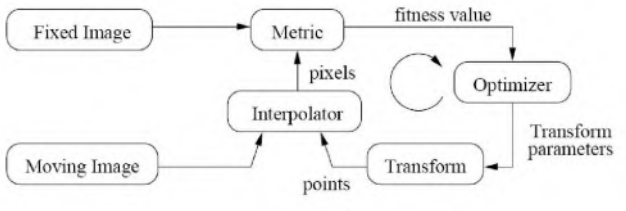


Figure 5.10: Schematic overview of the registration.

Based on this "metric" value, an optimizer then proposes new transform parameters and the cycle is repeated until the optimal transform is found or a stopping criterion is reached. Fig. 5.10 illustrates our registration scheme.

Suppose $f_m(f_U(\mathbf{r}'_i), f_V(\mathbf{r}_i))$ is the metric function of the grey values of coordinate points in U and V, under transformation $\mathbf{r}'_i = R\mathbf{r}_i + \mathbf{t}$, that reaches an optimum when both images are aligned optimally. Then, the registration goal is to find the transformation that optimizes this metric.

Note however that the transformation $\mathbf{r}'_i = R\mathbf{r}_i + \mathbf{t}$ does not always guarantee that $\mathbf{r}'_i$ will coincide exactly with a point $\mathbf{s}_j$ in U. Therefore, $f_U(\mathbf{r}'_i)$ is usually determined through interpolation of the grey values $f_U(\mathbf{s}_j)$ of the points $\mathbf{s}_j$ in U neighboring $\mathbf{r}'_i$.

In the following Subsections, we will detail the most important parts of this registration method: the rigid transform through which the US image is moved around in the coordinate space is presented in Subsection 5.4.1. The linear interpolation needed to compute $f_U(\mathbf{r}'_i)$ is presented in Subsection 5.4.2. The metric f_m we use is called Mutual Information and is based on the joint intensity histogram of two images. This is described in Subsection 5.4.3. Finally, a regular step gradient descent optimization algorithm is used to find the optimal value of the metric. This is discussed in Subsection 5.4.4. The whole registration procedure starts with an interactive initialization presented in Subsection 5.4.5.

5.4.1 Rigid transforms

Since we apply a rigid registration method, we only consider translations and rotations around the principal axes of our Euclidean space. If $\mathbf{r}_i = (r_{ix}, r_{iy}, r_{iz})$ again represents a point of V in the coordinate space, then the transformed coordinate point $\mathbf{r}'_i = (r'_{ix}, r'_{iy}, r'_{iz})$ is defined as rotations around the principle axes $R = R_z R_y R_x$ followed by a translation $\mathbf{t}$, i.e., $\mathbf{r}'_i = R\mathbf{r}_i + \mathbf{t}$ and computed

as

$$\begin{bmatrix} r'_{ix} \\ r'_{iy} \\ r'_{iz} \end{bmatrix} = R_z R_y R_x \begin{bmatrix} r_{ix} \\ r_{iy} \\ r_{iz} \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{5.5}$$

with

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix} \tag{5.6}$$

in the case of a rotation (of angle θ_x) around the X-axis,

$$R_y = \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \tag{5.7}$$

in the case of a rotation (of angle θ_y) around the Y-axis, and

$$R_z = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.8}$$

in the case of a rotation (of angle θ_z) around the Z-axis. Consequently, our transformation has 6 degrees of freedom, i.e., 3 rotation and 3 translation parameters. As such, from now on we will characterize a transformation by its parameter vector $\mathbf{p} = (\theta_x, \theta_y, \theta_z, t_x, t_y, t_z)$.

5.4.2 The linear interpolator

Depending on the transformation parameters, the moving US image, represented by its set of points V, takes on different positions in the coordinate space. However, a transformation does not guarantee that a transformed point $\mathbf{r}'_i = R\mathbf{r}_i + \mathbf{t}$ will coincide exactly with a point $\mathbf{s}_j$ of the reference volume U, as is shown in Fig. 5.11 for the 2D case. Since for our metric we need to define the grey values $f_U(\mathbf{r}'_i)$ of the reference image at those points we need to interpolate these from the grey values of the neighboring coordinate points $\mathbf{s}_j$ of the reference volume.

The choice of this interpolator influences both the accuracy of the results and the computation time of the registration method since interpolations are needed in every iteration. A trade-off will have to be made between computational complexity and accuracy. B-spline interpolation (of a high degree), as used in our preprocessing step, is rather too time-consuming so we choose a trilinear interpolation technique.

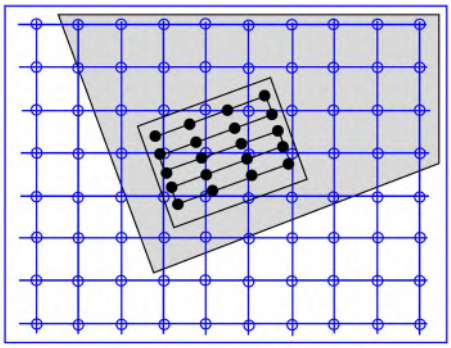


Figure 5.11: The overlap of the transformed image with the reference image is shown as the grey area (in the case of 2D images). It is clear that the coordinate points of the moving image, represented by the black coordinate points, and the reference image, characterized by the blue coordinate points, do not necessarily coincide.

Since we align a 2D US image to a 3D reference volume, we have eight possible neighboring grid points to interpolate from. As such, the intensity $f_U(\mathbf{r}'_i)$ at position $\mathbf{r}'_i = R\mathbf{r}_i + \mathbf{t}$ in 3D the reference volume is interpolated from its 8 closest grid points $\mathbf{s}_j$ in U.

Suppose we want to interpolate at a point with non-MRI grid coordinates (x, y, z), i.e., we want to determine $f_U(x, y, z)$. Define

$$\begin{aligned} x_d &= x - \lfloor x \rfloor \\ y_d &= y - \lfloor y \rfloor \\ z_d &= z - \lfloor z \rfloor \end{aligned} \tag{5.9}$$

with $\lfloor x \rfloor$ representing the largest MRI grid X-coordinate smaller than x. The interpolated grey value $f_U(x, y, z)$ is then computed as

$$\begin{aligned} f_U(x,y,z) = {} & f_U(\lfloor x \rfloor, \lfloor y \rfloor, \lfloor z \rfloor)(0.5 - x_d)(0.5 - y_d)(0.5 - z_d) + \\ & f_U(\lceil x \rceil, \lfloor y \rfloor, \lfloor z \rfloor)(0.5 - y_d)(0.5 - z_d) + \\ & f_U(\lfloor x \rfloor, \lceil y \rceil, \lfloor z \rfloor)(0.5 - x_d)y(0.5 - z_d) + \\ & f_U(\lfloor x \rfloor, \lfloor y \rfloor, \lceil z \rceil)(0.5 - x_d)(0.5 - y_d)z_d + \\ & f_U(\lceil x \rceil, \lfloor y \rfloor, \lceil z \rceil)x_d(0.5 - y_d)z_d + \\ & f_U(\lfloor x \rfloor, \lceil y \rceil, \lceil z \rceil)(0.5 - x_d)y_d z_d + \\ & f_U(\lceil x \rceil, \lceil y \rceil, \lfloor z \rfloor)x_d y_d(0.5 - z_d) + \\ & f_U(\lceil x \rceil, \lceil y \rceil, \lceil z \rceil)x_d y_d z_d, \end{aligned} \tag{5.10}$$

where $\lceil x \rceil$ represents the smallest MRI grid X-coordinate bigger than x and 0.5 corresponds to the spacing of the isotropic MRI volume.

5.4.3 The Mutual Information metric

The metric expresses the quality of a solution, based on the grey values of the coordinate points of the moving and reference image. This is often the most critical component in a registration scheme.

As mentioned before, in the US to MRI registration we are comparing modalities with quite different image (content) characteristics. Therefore, simply comparing the intensities of corresponding grid points in the moving and reference image and calculating, e.g., the sum of squared differences as a similarity-measure, is not a good solution. Comparing landmarks is also not feasible since experts find it difficult to detect good corresponding fiducial points. As explained earlier, ventricle segmentation comparison is also not possible. We will determine the quality of an alignment from the *joint intensity histogram* of the coordinate points in the images, using a measure called *Mutual Information*.

Let us first explain the joint intensity histogram. Denote by U and V again the set of coordinate points defined as earlier and the grey values of these coordinate points takes on values in the $[0, 255]$ interval, i.e., $0 < f_U(\mathbf{s}_j), f_V(\mathbf{r}_i) < 255$. The joint intensity histogram H then is represented as a matrix 256×256 matrix with entry $H(m, m)$ calculated as:

$$H(m, n) = \#\{(\mathbf{r}_i | f_U(\mathbf{r}'_i) = m \wedge f_V(\mathbf{r}_i) = n\}, \tag{5.11}$$

where # denotes the set cardinality.

Mutual Information is a similarity measure that originates from information theory. Suppose X_1 and X_2 are two random variables, then the mutual information $MI(X_1, X_2)$ measures how much information X_1 conveys about X_2 and is defined as:

$$MI(X_1, X_2) = \int_{X_1} \int_{X_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} dx_1 dx_2 \tag{5.12}$$

with $p(x_1, x_2)$ the joint probability density of X_1 and X_2 and $p(x_i)$ the marginal probability density of X_i.

We can translate this measure to image processing if we consider the grey values of our reference volume and moving image as multi-variate random variables and their joint distribution is presented by the joint intensity histogram. Furthermore, denote by $H_{\mathbf{p}}(m, n)$ the entry at position (m, n) of the joint intensity histogram of the reference volume and the moving image after a transformation with parameter vector $\mathbf{p}$ and denote by

$$H_{\mathbf{p}}(m) = \sum_{n=0}^{255} H_{\mathbf{p}}(m, n), \qquad H_{\mathbf{p}}(n) = \sum_{m=0}^{255} H_{\mathbf{p}}(m, n) \tag{5.13}$$

the marginal distributions after a transformation with parameter vector $\mathbf{p}$. Then, the Mutual Information f_m corresponding to a transformation with parameter vector $\mathbf{p}$ is defined as:

$$f_m(f_U(\mathbf{r}'_i), f_V(\mathbf{r}_i)) = \sum_{m,n=0}^{255} H_{\mathbf{p}}(m,n) \log \frac{H_{\mathbf{p}}(m,n)}{H_{\mathbf{p}}(m)H_{\mathbf{p}}(n)}. \tag{5.14}$$

The transformation that optimizes f_m assures that the grey values in the specific region of the reference volume are the best match for the grey values of the moving image. In our registration scheme, we use an implementation for the Mutual Information as proposed by [Mattes et al., 2001].

5.4.4 The gradient descent optimizer

The optimizer searches a locally optimal value of the metric, preferably as quickly as possible and as closely as possible to the global optimum. The most reliable way to find the transformation that maximizes the similarity criterion is to exhaustively search the entire parameter space. It is clear that this strategy is very time-consuming since we have 6 degrees of freedom in our transformation, i.e., $\mathbf{p} = (\theta_x, \theta_y, \theta_z, t_x, t_y, t_z)$. Therefore, optimization algorithms have been proposed that search the six-dimensional parameter space in specific ways.

The gradient descent algorithm is an optimization method which uses the gradient of the metric and the strategy of steepest descent. It is assumed that small steps in the direction of the gradient of the metric lead to points of higher similarity.

Formally, our Mutual Information metric $f_m(f_U(\mathbf{r}'_i), f_V(\mathbf{r}_i))$ will for now be denoted as $f^*_m(\mathbf{p})$ with $\mathbf{p} = (\theta_x, \theta_y, \theta_z, t_x, t_y, t_z)$ since the position of the moving image is determined completely by the transformation and the transformation is determined completely by its parameter vector. As such, it is in the parameter space that we are looking for the optimal solution. The gradient descent method interactively calculates:

$$\mathbf{p}_{j+1} = \mathbf{p}_j + l\nabla f^*_m(\mathbf{p}_j) \tag{5.15}$$

with ∇ the mathematical gradient operation with respect to the transformation parameters, i.e., $\nabla = \{\frac{\partial}{\partial\theta_x}, \frac{\partial}{\partial\theta_y}, \frac{\partial}{\partial\theta_z}, \frac{\partial}{\partial t_x}, \frac{\partial}{\partial t_y}, \frac{\partial}{\partial t_z}\}$, and l is called the step size. This iteration process is stopped once either a maximum number of iterations is reached or the difference between consequent iterations is below a certain threshold value. Fig. 5.12 shows the principle of the gradient descent method for a one-dimensional metric (where ∇ in this case is the derivative).

We notice from equation (5.15) that each iteration is determined both by the size of the gradient, which we can not control ourselves, and by the step size l, which we can control ourselves. A big step size results in an optimization that

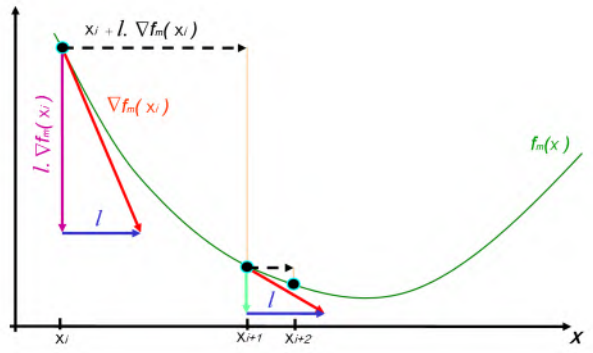


Figure 5.12: Visual representation of the evolution of the gradient descent optimization for a one-dimensional metric $F(\mathbf{x})$.

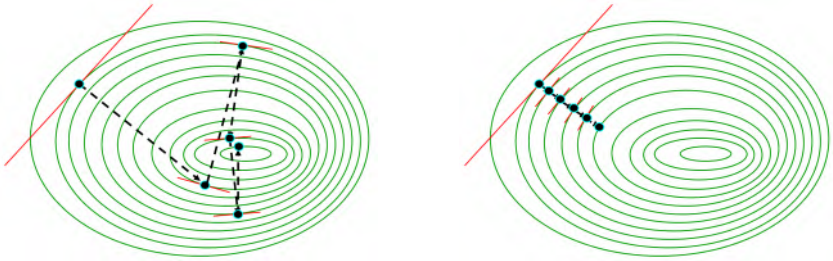


Figure 5.13: A 2D-metric is plotted in the form of ellipses of constant value of the metric, the optimum of the metric being located in the center of the smallest ellipse. Left: the Gradient Descent convergence with an overshooting large step size. Right: Gradient Descent convergence a step size that is too small.

easily overshoots the function optimum and a small step size will usually result in a long convergence time. This is shown in Fig. 5.13 for a 2D-metric surface.

In *normal* gradient descent methods a step size is determined initially and kept fixed throughout the iteration process. As such, the behavior of the optimizer depends on the gradient magnitude. In the *regular step* method we also choose an initial step l_1 but this step is decimated $l_2 = l_1/2$ each time the gradient changes direction, see Fig. 5.14. This change in direction corresponds to the scalar product or dot product of the current and next gradient (vectors) changing sign.

The idea behind this approach is that if a gradient changes direction, this indicates the method has passed an optimum. The division of the step length then prevents the method from overshooting, allowing for a more precise search

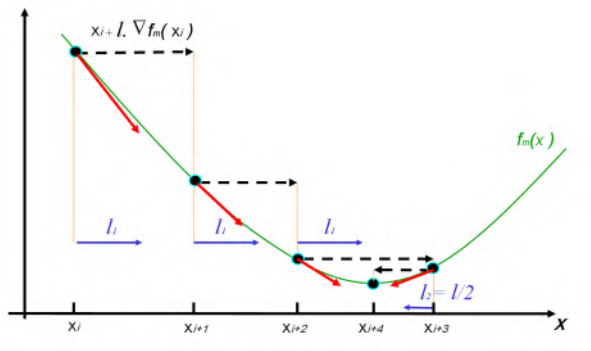


Figure 5.14: Visual representation of the regular step gradient descent optimization again in the case of a one-dimensional metric.

around the optimum. This division is continued until either a lower step size limit l_f is reached or until an initially set lower limit gradient size ∇f^*_{stop} is reached or until again a maximum amount of iterations is reached.

The regular step gradient descent method is an attractive optimization method because it is conceptually straightforward and often converges quickly. As with any other practical algorithm however, the major drawback of this method is that it converges to a local optimum and that the optimum it reaches depends on the initialization.

To counter a possible suboptimal convergence, we impose a strict initialization procedure and run the regular step gradient descent algorithm with multiple initial regular step sizes. [Ibanez et al., 2005] propose step sizes from up to 20 times the image spacing, i.e., the physical distance between image coordinate points, down to 0.5 times the spacing. We choose $l_1 = 10, 9, 8, 7, 6, 5, 4, 3, 1, 0.5$. The initialization, discussed in the next subsection, prevents the optimizer from wandering in regions that are completely out of interest, thus putting some constraints on the search space.

5.4.5 Initialization

Normally, the US image are assumed to be captured under a scan angle of approximately 45 degrees to the coronal plane. However, when we inspected the images ourselves, we noticed the best visual correspondence between the US image and the MRI volume was more likely to be found under a 35 degrees, see Fig. 5.15. This means that although the physician planned to scan at 45 degree angle, it is more likely he used a 35 degree angle. Note that without probe tracking it is difficult for a physician to tell exactly under which scan angle the images are captured, let alone to always maintain the exact same

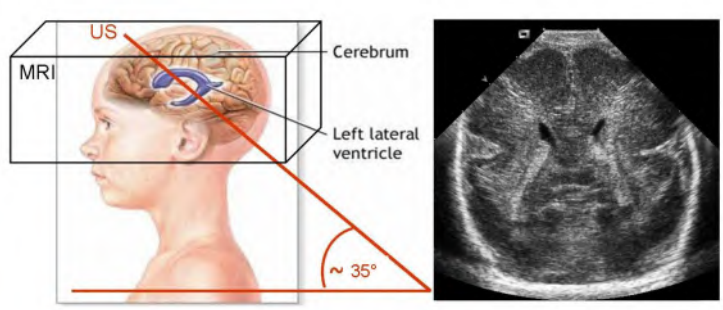


Figure 5.15: Through visual image inspection we noticed US images were more likely to be captured under a 35 degree angle than the suggested 45 degree angle.

scan angle for different patients. As such deviations are possible.

Knowing this, the first part of our initialization was to rotate the MRI reference volume of this 35 degree angle. Subsequently, we align the center of the head in the US volume with the center of the head in the MRI volume to assure a good overlap and position the origin or our coordinate system there. Finally, starting from that position the US image is manually moved through the rotated MRI volume until a visually acceptable initialization slice is found. The transformation vector $\mathbf{p}_1$ corresponding to this position is used as the initialization of our optimization, see Fig. 5.16. The selection of this optimal slice is the interactive part of the algorithm and needs to be done (by the physician) before the program starts.

5.5 Examples

In Fig. 5.17 and Fig. 5.18 the registration results are shown for T1 and T2 weighted MRI volumes of 8 different subjects. We notice the brain lobi in the T1 image are well registered, as are the ventricles. In the T2 images, is it more difficult to detect the brain lobi on the MR images, yet we also notice a good correspondence between ventricles. The T2 volumes also show sharper edges, are smoother, and generally contain less artifacts.

All results were obtained by running the algorithm with 10 different initial step sizes $l_1 = 10, 9, 8, 7, 6, 5, 4, 3, 1, 0.5$ and a stopping criterion of $l_f = 0.001$ with a maximum of 10000 iterations. We found that all registrations converged before the maximum number of iterations was reached and that in most cases the following step sizes $l_1 = 9, 5, 4, 3, 0.5$ resulted in the optimum of the metric. Convergence times differ from run to run and are usually between 2 to 3 minutes. In some cases however, a combination of a small step size and a flat metric surface result in long registration times, i.e., up to 10 minutes.

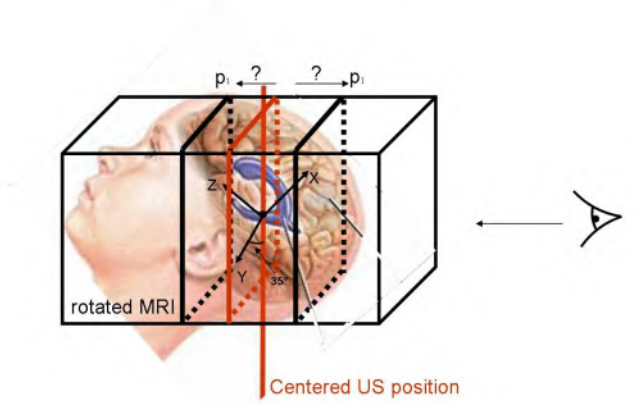


Figure 5.16: The MRI volume is rotated over a 35 degree angle (around the X-axis). The US image is automatically placed at the center of the MRI cube and the origin of the transformation space is moved there too. Subsequently, the physician selects the best initial visual correspondence by moving through the MRI volume. The registration then starts from the transformation parameter vector $\mathbf{p}_1$ corresponding to this slice.

We asked two physicians to visually score the results of the 56 registrations (28 T1 and 28 T2) performed on the data set at hand. 44 out of the 56 registrations, i.e., 79%, were considered good visual correspondences. The criterion for a good registration was that recognizable brain features such as the ventricles, lobi, skull and sinus were aligned well.

Both the image quality of the US scan and MRI volumes determine the registration success. If the US image lacks sufficient structural information, i.e., if it does not display the ventricles or brain lobi, it is difficult to find a good initialization. If we do not initialize in the neighborhood of the expected solution, the optimizer will converge to a wrong optimum. Concerning the MRI volumes mostly the artifacts due to the head movement disturb the metric and lead to inferior registrations. Fig. 5.19 show examples of both a bad US image, containing little structural information, and a bad MRI volume, containing ringing artifacts due to head movements.

5.6 Quantitative validation

Up to now, there are no other existing registration schemes for 3D MRI to 2D US brain registration. As such, to validate our technique quantitatively we compared it to manual expert registrations. For this purpose, we used a brand new technology that allows experts to register images in a virtual environment,

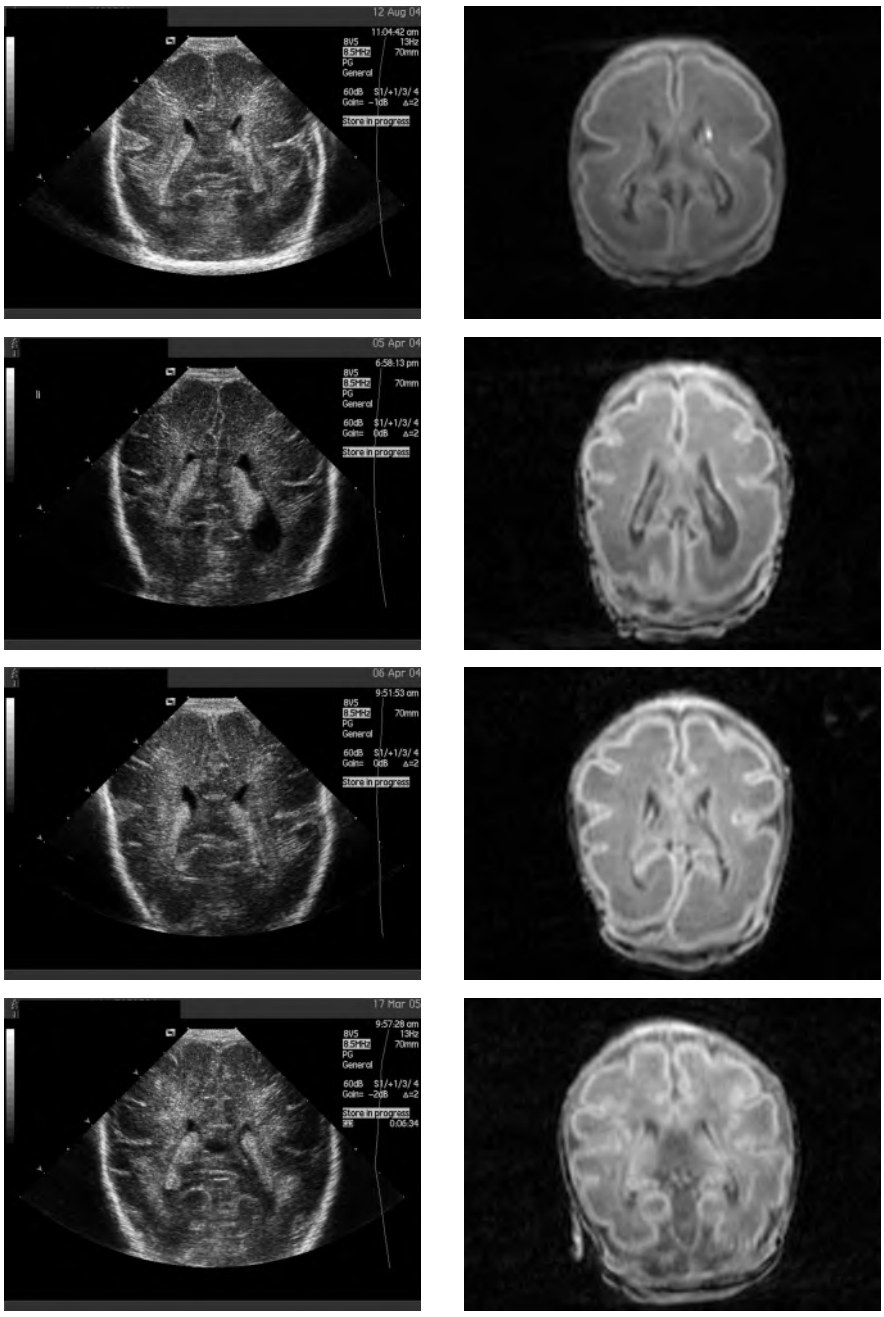


Figure 5.17: Registration results of the T1 MRI volumes.

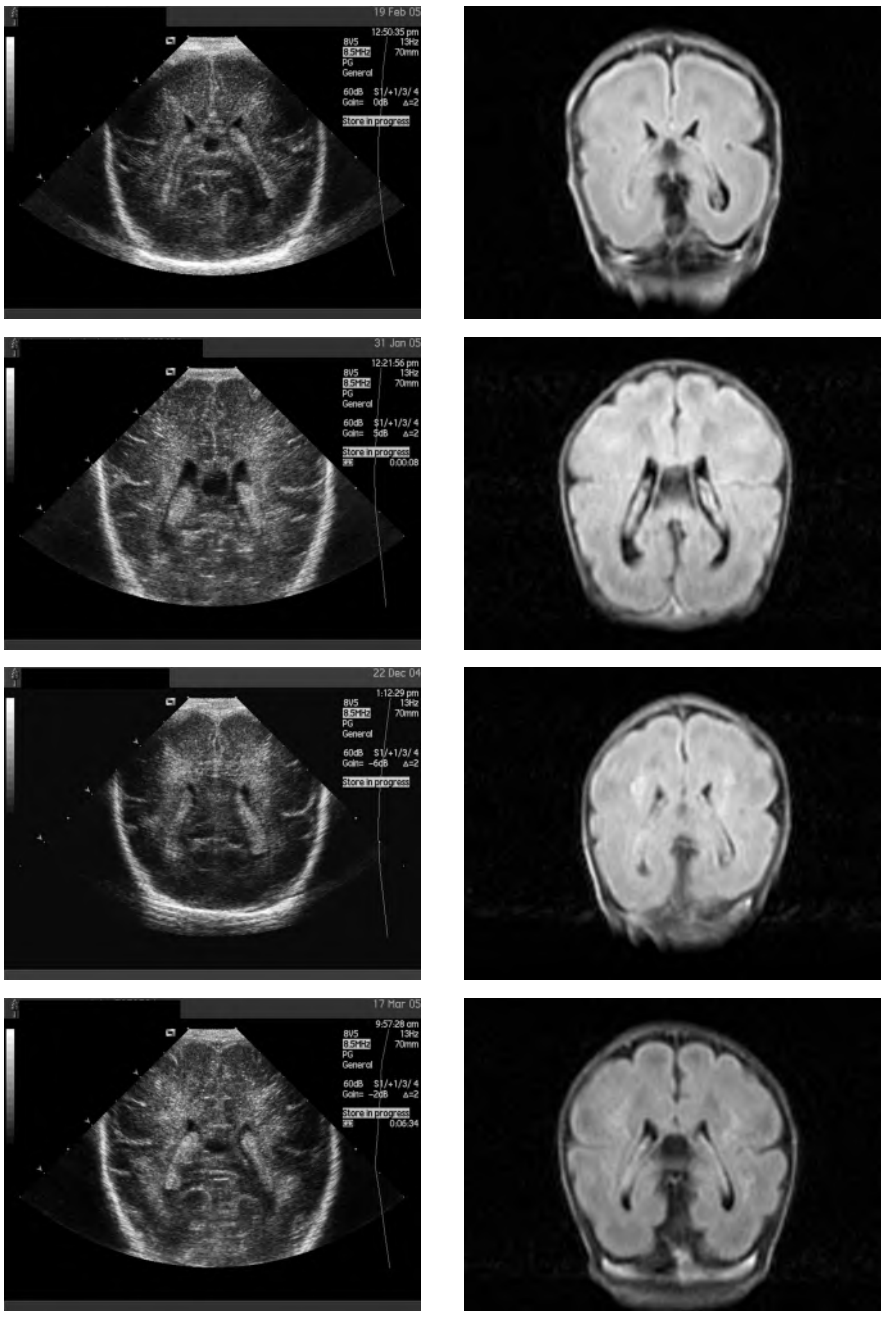


Figure 5.18: Registration results of the T2 MRI volumes.

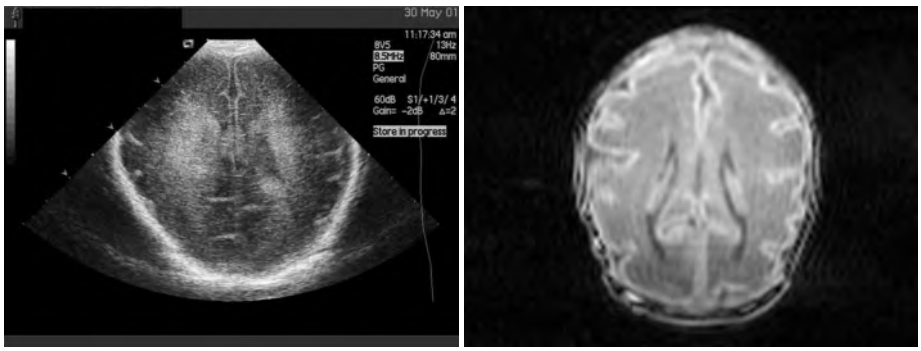


Figure 5.19: An US scan and MRI slice of the same patient that are both of a bad quality for registration. The US images lacks structure to initialize the image well. The MR image shows ringing artifacts due to head movements.

namely the BARCO I-Space installed at the Erasmus Medical Center. This is a four-walled CAVE-like virtual reality system, see Fig. 5.20, that originates from the Visualization Laboratory of the University of Chicago.

Figure 5.20: The I-space virtual reality cave used for the manual expert registrations.

In the I-Space, physicians are surrounded by computer-generated stereo images, projected on three walls by high-quality video projectors. Inspecting these images while wearing a lightweight pair of glasses with polarizing lenses allows the viewer to perceive depth images. As such, they perceive a real 3D representation of the images and are able to look at the images from various angles, see Fig. 5.21. A hand-held joystick also enables the physician to scroll through and even fuse images. The automatic adaptation of the images to the viewer's perspective is made possible by a wireless tracking system that sends the leading viewer's position and orientation to the computer generating the

Figure 5.21: Expert user in the I-space navigating with a joystick and wearing special glasses with polarizing lenses.

Table 5.1: Average differences between 23 T2 expert registrations and our semi-automatic registration.

	θ_x (rad)	θ_y	θ_z	$t_x(mm)$	t_y	t_z
$mean\ diff.$	0.0423	0.0342	0.0452	1.01	1.501	1.402

images.

To register the 2D US image and 3D MRI volume the interpolated MRI volume is projected into the I-SPACE. With the joystick the physician can then guide the US images through the MRI volume until the optimal correspondence is found. Although this procedure requires some training, it is a more sophisticated and interactive way of aligning images than procedures based on existing visualization packages as, e.g., Brainvoyager and Slice-O-matic, where the registration has to be done on a computer screen.
If we can extract the transformation parameters of the optimal manual registration by the physician we can calculate the translation and rotation differences to our method.

In total 23 T2 MRI volumes were manually registered in I-SPACE. Only T2 images were registered since the experts considered these as containing the optimal amount of clinically relevant information. Once physicians fixed their optimal registration, the rotation and variation parameters could be read from the upper left corner of the image, see Fig. 5.22. To compare the registration results quantitatively, the average differences for all registration parameters are calculated and presented in Table 5.1.

We notice that for the translation parameters our registration results are always

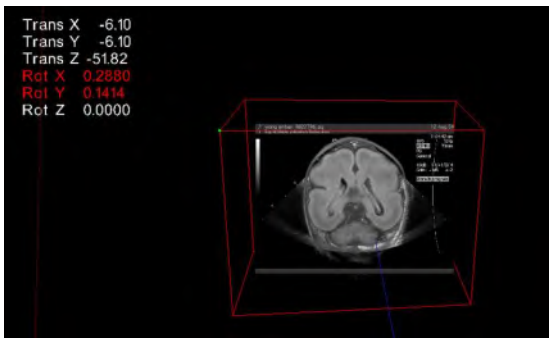


Figure 5.22: Visual result of manually registering the MRI volume to the 2D echogram.

very close to the manually registered image, usually within the range of 1 to 1.5 mm. For the rotational properties, we notice differences in the range of 0.034 up to 0.0423 radians (corresponding to angles of about 2.5 degrees) which is also close.

However, in some cases (4 out of 23 T2 MRI volumes) our registration fails, as shown in Fig. 5.23. Fig. 5.24 shows the surface of the Mutual Information metric in a 2D parameter plane for the upper two images of Fig. 5.23. The graphs show the translation in the XY-plane and the rotation around the X- and Y-axis. We notice that for the translation parameters the metric's surface shows a clear optimum, close to the optimum (manual registration) located at the origin.
For the rotations, the metric's surface is valley-shaped and multiple points can be considered as possibly optimal solutions.

Fig. 5.25 shows the deviations from the optimal result, i.e., the result of a manual expert registration, for the translation parameters and the rotation parameters. This figure shows that we indeed obtain good translation parameters but that the position of optimal metrical value differs significantly from the correct position for the rotation parameters, even up to 0.3 radians (17.9 degrees).

Concerning the reproducibility of our technique we compared the results of different users. 5 T2 MRI volumes were registered by two different users each using their own initialization. Tables 5.2 and 5.3 show the initial $\mathbf{p}_1$ and final $\mathbf{p}_f$ transform parameters for the different users. Table 5.4 shows the average difference over the final transform parameters over the 5 registrations. We notice from Tables 5.2 and 5.3 that the algorithm indeed converges to an optimum in the neighborhood of the initialization and from Table 5.4 that the inter-rater differences are very small.

In terms of time-efficiency, the interactive initialization of our algorithm takes about 1 minute. As mentioned in the previous section, the time per automatic

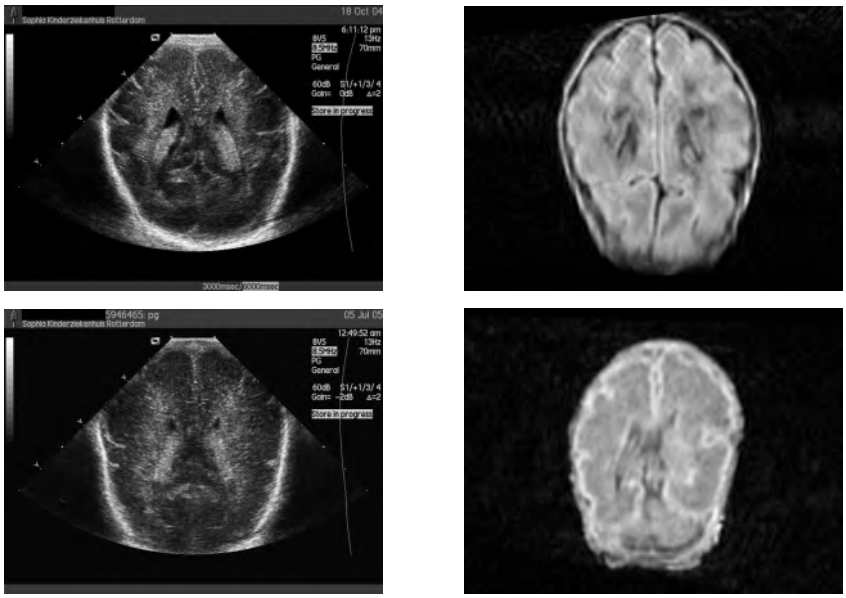


Figure 5.23: Two results of poor registration using our method. The lower two images correspond to a T2 volume registration, the upper two images correspond to a T1 volume registration.

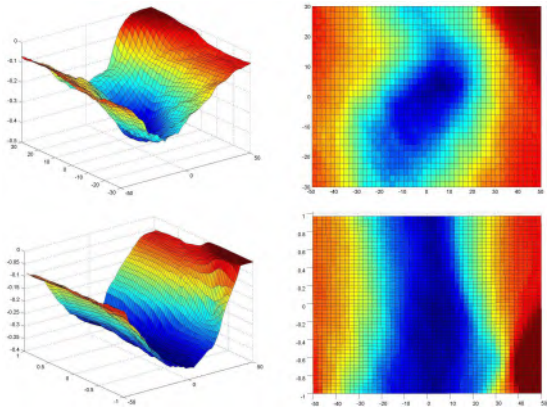


Figure 5.24: The surface of the Mutual Information metric and its corresponding 2D-projection, for a patient in the dataset where registration failed. The upper graph shows the translation in the XY-plane and the lower graph the rotation around the X- and Y-axis.

Table 5.2: Initializations and final registration results for the first user on 5 T2 volumes.

	θ_x (rad)	θ_y	θ_z	t_x (mm)	t_y	t_z
$\mathbf{p}_1$	-0.6	0	0	0	-9.3	-13.6
$\mathbf{p}_f$	-0.616	0.0872	0.023	-0.772	-3.71	-8.6
$\mathbf{p}_1$	-0.6	0	0	0	-10.3	-15.3
$\mathbf{p}_f$	-0.583	0.0511	-0.140	-1.73	-1.25	-16.0
$\mathbf{p}_1$	-0.6	0	0	0	-7.6	-11.1
$\mathbf{p}_f$	-0.567	0.0717	0.0798	1.28	-13.2	-16.4
$\mathbf{p}_1$	-0.6	0	0	0	-11.3	-16.5
$\mathbf{p}_f$	-0.639	-0.0748	0.0125	-1.54	-5.03	-19.7
$\mathbf{p}_1$	-0.6	0	0	0	-15.4	-14.5
$\mathbf{p}_f$	-0.645	0.0247	0.044	-3.5	-19.2	-9.3

Table 5.3: Initializations and final registration results for the second user on 5 T2 volumes.

	θ_x (rad)	θ_y	θ_z	t_x (mm)	t_y	t_z
$\mathbf{p}_1$	-0.6	0	0	0	-10.3	-14.6
$\mathbf{p}_f$	-0.614	0.0862	0.022	-0.761	-3.66	-8.5
$\mathbf{p}_1$	-0.6	0	0	0	-11.1	-16.1
$\mathbf{p}_f$	-0.585	0.509	-0.138	-1.72	-1.29	-16.1
$\mathbf{p}_1$	-0.6	0	0	0	-8.1	-13.1
$\mathbf{p}_f$	-0.566	0.0717	0.0798	1.24	-13.3	-16.4
$\mathbf{p}_1$	-0.6	0	0	0	-10.3	-14.5
$\mathbf{p}_f$	-0.638	-0.0750	0.0126	-1.56	-5.03	-19.7
$\mathbf{p}_1$	-0.6	0	0	0	-16.4	-13.5
$\mathbf{p}_f$	-0.645	0.0246	0.043	-3.4	-19.2	-9.4

Table 5.4: Average registration differences between the two users over our 5 T2 images.

	θ_x (rad)	θ_y	θ_z	$t_x(mm)$	t_y	t_z
$mean\ diff.$	0.0012	0.00084	0.0001	0.0184	0.038	0.04

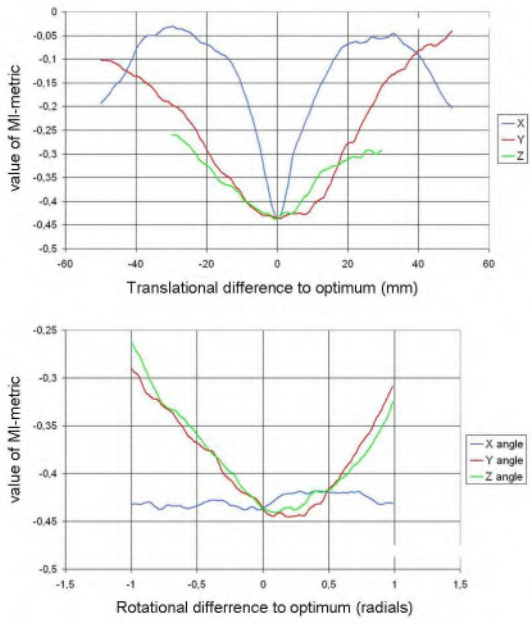


Figure 5.25: Deviations from the optimal results for both the translation parameters (in mm) and the rotation parameters (in radians).

registration (different step sizes l_i) is usually around 2 to 3 minutes, with extremes up to 10 minutes. It is exactly the valley-shaped metric surfaces as in the lower picture of Fig. 5.24 that lead to long convergence times. Once the metric is stuck in the valley, it advances very slowly and as shown, to incorrect optima. The manual registration in the I-SPACE is fully interactive and takes about 10 minutes per registration.

5.6.1 Discussion

Our goal was to present a registration algorithm to align periventricular brain regions in the preterm US brain with their corresponding MRI regions. Due to the difference in both image modality and dimensionality this registration problem is one of the most difficult ones to handle. As such, we had to incorporate both extensive preprocessing and interaction through initialization.

The MRI volumes are interpolated to compensate for the low resolution in the Z-direction. US images are filtered using the adapted GenLik filter. As long as the medical protocol remains unchanged, the MRI interpolation can not be avoided. Moreover, it is not simple to increase the MRI scanning times to get a better Z-resolution since preterms can not be immobilized nor sedated so

movement artifacts are more likely to appear during longer scans.

The initialization is the interactive part of the algorithm. Although this usually takes no more than a minute, it is vital for the registration to perform well. Initializing well, i.e., at least in the neighborhood of the expected solution, guarantees good registration results. If we would start the registration procedure at any random place in the volume, the algorithm is bound to converge to sub-optimal solutions. It is difficult to quantify how close this initialization exactly has to be to the optimal solution as the brain structure tends to differ significantly from patient to patient.

The scan angle was not set to 45 degrees as suggested by the acquisition protocol, since through visual image inspection we found 35 degrees more likely to be the right scan angle. Although this correction resulted in good registration results, only probe tracking information could provide us the exact scan angle. In that case however, when calibrated correctly, the image registration algorithm is also no longer needed since we could switch to a non-image based registration. Note that for the images used in this registration study, probe tracking information was unfortunately not available.

Regarding the actual registration scheme, an intensity-based registration using the Mutual Information metric was found to perform well. In the master thesis of [Vandemeulebroucke et al., 2005] a normalized Mutual Information and Mutual Information proposed by [Viola and Wells, 1997], were compared to the Mutual Information we used yet no better results were obtained.

Other registration approaches, such as using the brain ventricle segmentation algorithm developed in the previous Chapter, can be an alternative to intensity-based methods, yet both practical and structural problems arise. Practically, we possess no 3D US information for this set of 28 patients so ventricle reconstruction is impossible here. Another problem is that currently no preterm 3D MRI ventricle segmentation algorithm was developed yet. Structurally, the choroid plexus, located inside of the ventricle, has a different appearance in the US images and MRI volume. This is shown in almost all pictures of Fig. 5.17 and Fig. 5.18. It will have to be investigated whether this disturbs the quality of the 3D ventricle segmentation and reconstruction.

Concerning the optimizer, we choose a simple, deterministic regular step gradient descent algorithm. In [Vandemeulebroucke et al., 2005] this optimizer was compared to a normal gradient descent optimizer and a genetic optimization algorithm and no better results were found.
As both the image quality and content differ from image to image, so does the optimal initialization step. Therefore, we applied the regular step descent for 10 different step sizes which increases the computation time. However, in most of our registrations we found the optimal solutions corresponded to a subset of 5 step lengths. If desired we could optimize the technique, in terms of computation time, by reducing the number of initializations.

Through qualitative inspection, physicians scored 79% of our registrations as correct. A quantitative evaluation of the same registration results, i.e., by

comparing them to the manual expert registrations, showed that the optimal results on average do not differ more than 1.5 mm in translation and 2.5 degrees in rotation. An experiment with two different users showed that the technique is reproducible and stable w.r.t. initialization (in the vicinity of the solution expected).

In terms of execution time, the interactive part of our algorithm only takes about 1 minute whereas the alternative, i.e., the I-SPACE, requires around 10 minutes per registration. On the other hand our registration algorithm is not optimized in terms of execution time and can take as long as 20 to 30 minutes to complete all 10 runs (AMD 1.8 GHz). Since off-line cross-validation is for now the main application of this registration method, it is mainly the interactivity that counts. This means that if a physicians wants to register 10 different image sets, it suffices to initialize all of them in a span of 10 minutes and let the algorithm compute the solutions while in the mean time he can do other tasks. The procedure for registering the same image sets in the cave takes up to 100 minutes of total interaction.

Note however that we do have some ideas on how to speed up our registration procedure. The problem now is that for small step lengths and small gradients pointing in the same direction, convergence is usually slow. A possible modification is to enlarge the step length again once the optimizer is found to move with small steps in the same direction for too long, which is related to Newton optimization algorithms.

5.7 Flaring segmentation comparison

As we mentioned in the beginning of this Chapter, the actual goal of our registration is to be able to cross-validate our flaring segmentation algorithm. Moreover, we want to investigate if flaring patterns manifest themselves at the same positions and to the same extent in both US and MRI. Now that we are able to both segment flaring and register US images and MRI volumes, cross-validation can begin.

At the moment, we are only able to compare our segmentation results qualitatively. This is because there exists no quantitative flaring segmentation algorithm on preterm MRI yet. Consequently, we can only compare to manual expert delineations on the MR images. Note however that, contrary to US delineations, manual MRI delineations are considered as a more reliable golden standard than manual US delineations.

Fig. 5.26 shows some remarkable results of the comparison of flaring for the images of Fig. 5.18. In each of the four cases, the US flaring segmentation based on our algorithm is shown on the left and the manual expert MRI flaring delineation on the right[3]. We notice that in the upper two images, flaring is

[3]The physician could not see the US images while manually segmenting the MR image.

picked up in US where it is not visible in MRI. In the lower two images, flaring seems to be spread out further in the US images than in the MR images.

Given that we only used optimal registration results, in terms of correspondence to the manual expert registration, for this segmentation comparison we can be confident about the correctness of our alignments. As such, the fact that flaring is not visible in the upper two MR images contrary to in the US images, can be considered as the first published evidence for a conjecture some neonatologist have been claiming for a while. Namely that in VLBW preterm infants PVL is not always picked up to its full extent in MRI. Hence, although MRI is irrefutably the golden standard at term and on later ages, US imaging can indeed be an asset in the very early diagnosis of the pathology.

Where flaring is picked up in MRI, we notice that the US flaring areas are consistently larger. This could indicate that some structural differences in white matter might not be picked up by the MRI scan of VLBW preterms.

Note however that, besides the T2 MRI information, physicians believe the Anisotropic Diffusion Coefficient (ADC) maps, acquired simultaneously with the MRI scans, contain a lot of valuable PVL information. Consequently, a continuation of this research consists in comparing PVL regions in US to the corresponding ADC maps. However, this demands for the ADC to MRI volume registration which is not performed yet.

5.8 Conclusions and hints for future work

We presented a first interactive 2D US to 3D MRI brain image registration scheme as a cross-validation for the characterization of affected periventricular white brain matter. With the necessary interaction, we succeed in registering about 79% of our test images correctly, the main cause of failing registrations being a degraded image quality. The quality of our automatic registrations is shown to be comparable to the manual expert registrations. Besides that, our method demands for only 10% of the user-interaction needed for a manual registration.

We are aware that the method has been designed in a very application-specific way, yet, as mentioned in the introduction, registering US to any other modality is very difficult. Consequently, specific choices were made in view of all information at hand. The ability of using this method in other US/MRI problems is therefore yet to be investigated.

By comparing our segmentation results to the MR images, we concluded that flaring is always picked up in US images whereas in some cases, it is not in the corresponding MR image. This proves that while MRI is indisputably the golden standard on PVL at term and on later ages, US imaging is an important diagnostic tool in the very early detection of PVL. Besides the problem of prevalence, we also noticed that flaring in the US images seems to be spread

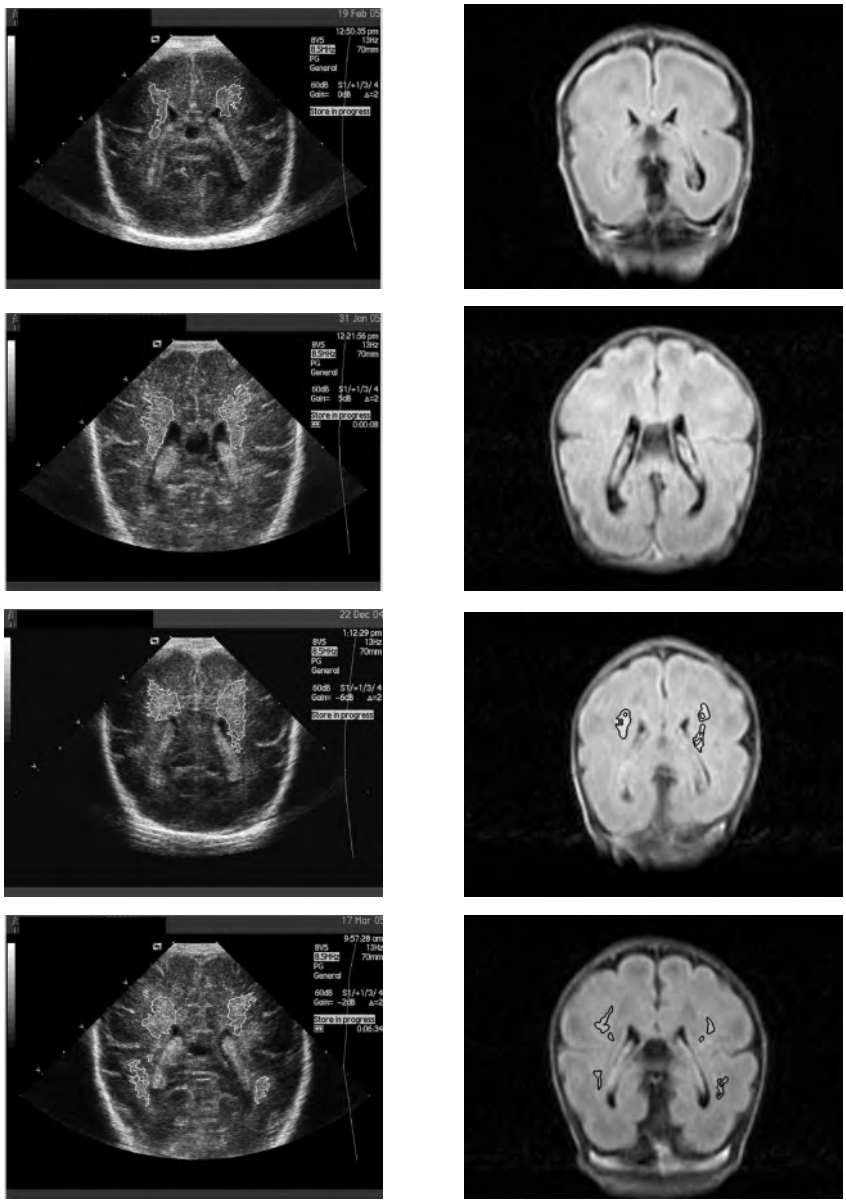


Figure 5.26: Segmentation results for the registrations of the T2 MRI volumes of Fig. 5.18. Left: the segmentations based on our algorithm. Right: manual expert delineations in the registered MRI images.

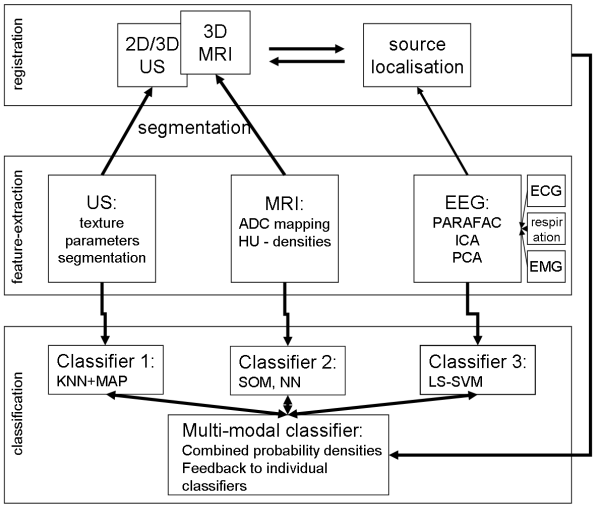


Figure 5.27: Flowchart of a future project where multimodal preterm brain data fusion is taken to the next level, incorporating US, MRI, EEG, ECG and EMG data in a classification, segmentation and registration framework.

out further than in the MR images. This could mean that structural tissue differences related to PVL manifest themselves better in US images.

Future work in the first place comprises the evaluation of the imaging protocols used for registration. Higher-resolution MRI volumes would no longer require B-spline interpolation and probably result in even more accurate results yet these have the drawback of increasing the acquisition time. 3D US would provide us with volumetrical ventricle information for a possible 3D US to 3D MRI registration. Besides altering the image acquisition, we believe hybrid-based registration methods could further improve the accuracy of the results. The combined use of the volumetric ventricle information and the Mutual Information criterion in a multivariate metric is one possible hybrid approach.

Another point for future investigation is the quantitative description of the flaring areas in preterm MRI. Given that MRI is of a higher image quality, e.g., contains clearer edge-information, this should be easier than in the case of US and would ultimately allow for the complete quantitative comparison of the flaring areas.

Finally, plans are already made to take the whole multimodal way of preterm brain analysis to the next level. In a forthcoming project we will be fusing both the MRI and US information, as presented so far in this thesis, to other modalities such as the electro encephalogram (EEG), the electromyogram (EMG), the electrocardiogram (ECG) and respiration data. The final goal there is to create a complete network of classification, registration and segmentation that

fuses the most complementary data, see Fig. 5.27.

Chapter 6

Psychophysical experiments on image quality

This Chapter differs from the former ones in that it is not a direct continuation of the PVL story. Here, we present a psychophysical approach to image quality. Through two psycho-visual experiments we assess the overall (image) quality of state-of-the-art noise-reduction filters as well as the relevance of speckle-reduction in qualitative medical US diagnosis.

6.1 Introduction

As image processing evolves, more and more competitive algorithms with similar functionality arise. To the end-user, the choice in favor of one algorithm is based mostly on its perceived quality. In this Chapter, we concentrate on how *subjective* quality impressions can be *measured* in a psycho-visual experiment. More precisely, we investigate the quality of noise-reduction filters in greyscale images and the effect of speckle-reduction on the image quality of medical US images. In that aspect, we follow the line of qualitative versus quantitative image analysis held throughout the thesis.

Measuring here implies that considered image sensations can be quantified in a consistent, meaningful and reproducible way. This means that the differences (in quality) we want to quantify are noticeable and that the expected outcome of the measurements does not change significantly in a repeated experiment.

The term subjective here refers to the fact that human subjects play an essential role in generating the responses. As such, in each experiment we normally have to justify not only why subjects are sensitive to what we want to measure

but also determine how subject responses relate to the sensations we want to measure.

Since arguing why subjects are able to come to (or agree on) a rating of certain sensations is very difficult, commonly the existence of such a relevant mechanism is *postulated*. Consequently, we will adopt the following mechanism used in former similar applications [Martens, 2003].

Although people act based on their own perception their opinions often agree remarkably well when being compared explicitly. For example, people agree to a great degree on typical matters as how colorful or bright an image is. Conversely, this perceptual agreement does not prohibit people in differing on more cognitively related aspects as overall image quality, e.g., image content might be a differing decisive factor to determine whether or not edges and details are preserved in an image.

In that aspect, we use *multidimensional scaling* (MDS) models that reflect both the common aspects of and the differences between individual judgments. The rationale underlying MDS is twofold:

- First of all, the concept of *homogeneity of perception* should hold, meaning that a *single* (multidimensional) geometrical configuration of the presented stimuli (being pictures, objects, etc..) underlies the judgments of *all* subjects.

- Secondly, the concept of overall image quality is rarely one-dimensional, meaning that multiple aspects or *attributes* such as blur or artifacts all influence the perceived image quality.
  Differences in attribute judgment of subjects are accommodated in MDS by the fact that mappings from the joint multidimensional stimulus configuration to the one-dimensional attribute judgments may vary per attribute and per subject.

As we show later on, the construction of a multidimensional geometrical stimulus configuration is also the basis for better *instrumental* quality measures. An instrumental quality measure is a mathematical formulation of the inter-stimuli similarity. The Root Mean Squared Error (RMSE) or Peak Signal-to-Noise Ratio (PSNR) are typical examples of such instrumental measures used in noise-reduction to express how well a filtered image resembles the noise-free image.

The advantage of these instrumental measures is that they are usually easy to compute (as compared to a psycho-visual experiment) and are considered benchmarks in many applications. The drawback of these measures is that they usually incorporate little to no information on how the overall quality is perceived. Consequently, one of the ultimate goals of quality assessment experiments is to define improved instrumental measures, i.e., measures that reveal the same amount of information as the psycho-visual experiment.

The structure of this Chapter is as follows: first, the fundamentals of the MDS framework are discussed in Section 6.2. Then, we show how to use this framework to its full extent in a first experiment where we compare 7 state-of-the-art noise reduction filters in Section 6.3. Subsequently, in Section 6.4, we perform a second experiment where we use MDS to assess the diagnostic value of a modified GenLik speckle-reduction filter in US images. As usual, we end with the conclusions on both experiments and hints for future work in Section 6.5.

6.2 Multidimensional scaling

As could already be deduced from our introduction, we use the following terminology in our psycho-visual experiments. The people involved in the experiments are called the *subjects*. The images they actually score are called the *stimuli*. The way these stimuli are scored and processed is called the *methodology*. Here, we restrict ourselves mainly to the methodology, as this is the common part of our experiments conducted. Subjects and stimuli differ from experiment to experiment and will be described in the following sections.

In both our experiments, subjects score image sensations or *attributes* (such as overall quality, noisiness, blurriness,...) on discrete scales. This discrete *scaling* can either be *nominal* or *ordinal*. Nominal scaling means that responses are assigned to categories that have a natural order, e.g., animals can be scored as mammals, birds, fish... Although nominal responses are often used in social sciences, they are rarely used in the psycho-visual experiments on image quality. Ordinal attributes have, as their name suggests, ordered labels. For example, the use of qualifications bad, poor, fair, good and excellent to designate levels of quality is an ordinal scoring. Usually these ordinal categories can be mapped onto integers, e.g., bad (1), poor (2), fair (3), good (4) and excellent (5). Ordinal integer scaling is also what we apply in our experiments.

Discrete scaling can be subdivided further into *metric* and *non-metric* scaling. Metric scaling means that equal distances on the discrete scale correspond to equal differences in attribute sensations. In non-metric scoring this is not the case. In general, it is very hard to prove why a certain scoring grade can be considered metric. Therefore, we usually postulate there is an unknown non-linear response function, assumed to be monotonically increasing or decreasing, that relates the internal sensations to the (metric) scaling.

Furthermore, in our experiments images are presented on a computer display and that in two different ways. Images are either shown one at the time, in which case we speak of a *single-stimulus* experiment, or in pairs, in which case we speak of a *double-stimulus* experiment. Double-stimuli scoring is usually easier than single-stimulus scoring since subjects are better at assigning the relative difference of an attribute over an image pair, than the absolute attribute score when stimuli are shown one after the other.

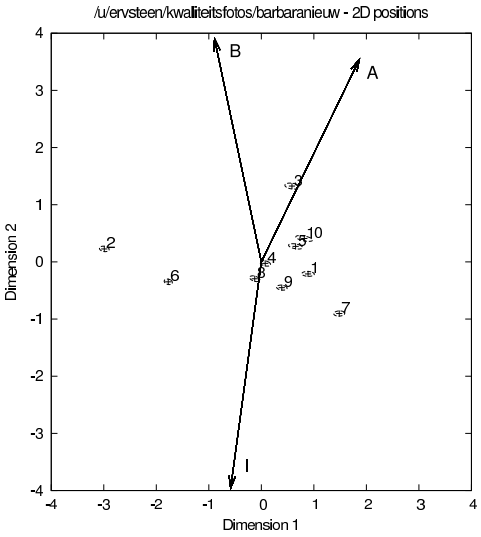


Figure 6.1: Example of a 2-dimensional perceptual space obtained with the MDS XGms software [Martens, 2003] for 8 denoised versions (1,4,5,6,7,8,9,10) of a noisy image (2), shown together with the original noise-free (3) image. A preference axis for image impairment (I), and attribute axes image blur (B) and image artifacts (A) are also shown.

In order to simultaneously model the scoring results from *multiple* stimuli and subjects concretely, *multidimensional* models are used. In such models, stimuli are represented by points in a multidimensional geometrical space and all observations are related to the distances between points and coordinates of point projections on specific axes.

Fig. 6.1 shows an example of a two-dimensional space (which we will also encounter later on). The 10 points in this space correspond to 8 denoised versions of one noisy image together with the noise-free and the noisy image. The points are placed so that the distances between them correspond to the perceived perceptual differences between the images as scored by 37 subjects. Consequently, the further two points are apart, the more they are perceived dissimilar.

The three axes in the diagram correspond to the perceived impairment (**I**), blur (**B**) and artifacts (**A**) as scored by the same subjects. The orthogonal projection of the points onto these axes allows us to rank the stimuli according to increasing impairment or decreasing amount of blur or artifacts.

In the rest of this section, we describe the mathematical model used to obtain these geometrical spaces. We start by shortly describing the 3 different kinds of responses that are gathered in our experiments: *dissimilarity* data, expressing how dissimilar two images in an image pair are, are discussed in Subsection

6.2.1. *Preference* data, expressing which image (of an image pair) is preferred over the other based on a certain attribute, such as blur or quality, are presented in Subsection 6.2.2. *Attribute* data, expressing the strength of a specific attribute, such as noise or blur, for each individual image, are discussed in Subsection 6.2.3. Finally, in Subsection 6.2.4 we explain the maximum-likelihood optimization criterion used to estimate the multidimensional model parameters from the data.

6.2.1 Dissimilarity data

Denote the subjects (people) involved in our experiment by $k = 1, \ldots, K$ and the different stimuli (images) by $i = 1, \ldots, N$. The first kind of data gathered expresses for each subject k the judged dissimilarity for all unique stimulus pairs $(i, j),\ i < j$. This results in an upper-triangular dissimilarity matrix D_k per subject k where the entry $D_{k,i,j}$ corresponds to how dissimilar i and j are judged by subject k.

The goal of MDS is to construct a stimulus configuration in an n-dimensional vector space (as shown in Fig. 6.1 for a two-dimensional example) based on these dissimilarity data. In that vector space, stimuli are represented by points $\mathbf{x}_i,\ i = 1, \ldots, N$ that must be organized in such a way that the dissimilarity scores $D_{k,i,j}$ are monotonically related to the inter-stimulus distances d_{ij} calculated as

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_l \triangleq \Big[\sum_{m=1}^{n} |x_{im} - x_{jm}|^l\Big]^{1/l}, \tag{6.1}$$

where the norm is the Minkowski norm with power l. The default value $l = 2$ corresponds to an Euclidean norm.

More precisely, a linear relationship is pursued between transformed dissimilarity scores $f_{dk}(D_{k,i,j})$ and the inter-stimuli distances d_{ij} since, as stated in the introduction, it is hard to prove that the perceived dissimilarities scores are metric. However, since we want to compare the dissimilarity scores to metric distance we have to assure they are metric.
So, we postulate there exists a monotonic transformation f_{dk} (different for each subject k)[1] that maps non-metric observed dissimilarities $D_{k,i,j}$ into metric dissimilarities $f_{dk}(D_{k,i,j})$. In our framework, both a generalized optimum power-like transformation, a generalized Kruskal transformation and an optimum spline transformation can be selected, as suggested by [Martens, 2003].

The pursued linear relationship between the (transformed) experimental data and the inter-stimulus distances can then be expressed as

$$f_{dk}(D_{k,i,j}) \approx\ r_k \cdot d_{ij}. \tag{6.2}$$

[1] The index d in f_{dk} denotes that the transformation is related to the dissimilarity scores.

where the $\mathbf{x}_i,\ i = 1, \ldots, N$ and resulting distances d_{ij}, the regression coefficient r_k as well as the transformation f_{dk} all are to be estimated.

More concretely, this linear relationship is calculated as follows. As said, the only information at hand are the $D_{k,i,j}$ gathered from the experiment. Both the (optimal) d_{ij}, r_k and f_{dk} are unknown and have to be determined by our MDS framework.
Suppose now that we consider as a starting assumption a particular geometrical configuration of the stimuli $\mathbf{x}_i,\ i = 1, \ldots, N$ and specific instances of r_k and f_{dk}. The value $\widehat{f_{dk}(D_{k,i,j})}$ defined as the estimate of $f_{dk}(D_{k,i,j})$ according to this particular configuration and specific r_k and f_{dk} can then be calculated as

$$\widehat{f_{dk}(D_{k,i,j})} \triangleq r_k \cdot \|\mathbf{x}_i - \mathbf{x}_j\|_l. \tag{6.3}$$

We compare the estimate $\widehat{f_{dk}(D_{k,i,j})}$ obtained for this configuration to the $f_{dk}(D_{k,i,j})$ as obtained from the experiment and use the estimation error as a measure for how close the configuration is to the "truly" optimal configuration.

In MDS approaches the correspondence is calculated through a *cost* function ϕ. This function links the scores of the experiment to the distances in a specific configuration by expressing the probability $p(f_{dk}(D_{k,i,j}))$ that the experimental score $f_{dk}(D_{k,i,j})$ results from the configuration and instances of r_k and f_{dk}

$$p(f_{dk}(D_{k,i,j})) = \phi[f_{dk}(D_{k,i,j}) - \widehat{f_{dk}(D_{k,i,j})}]. \tag{6.4}$$

Here, ϕ is chosen to be a zero-mean generalized Gaussian PDF computed as

$$\phi(x) = \frac{r}{2\rho\Gamma(1/r)} e^{-|\frac{x}{\rho}|^r} \tag{6.5}$$

with $\rho = \sigma\sqrt{\frac{\Gamma(1/r)}{\Gamma(3/r)}}$ and [2]

$$\Gamma(a) = \int_0^\infty z^{a-1} e^{-z} dz \tag{6.6}$$

the Gamma function.

However, we want the probability $p(D_{k,i,j})$ of obtaining the original dissimilarity $D_{k,i,j}$ and not the transformed dissimilarity $f_{dk}(D_{k,i,j})$. Since the transformation f_{dk} is assumed monotonic, this probability $p(D_{k,i,j})$ can be derived from the probability $p(f_{dk}(D_{k,i,j}))$ as

$$p(D_{k,i,j}) = p(f_{dk}(D_{k,i,j})) \cdot |f'_{dk}(D_{k,i,j})| \tag{6.7}$$

[2]Note that there is a σ in the definition of ρ which also differs from subject to subject, i.e., $\sigma(k)$, and is estimated from repeated $D_{k,i,j}$ scores on the same stimulus pair. To not overcomplicate our description we refer to [Martens, 2003] for more details.

where f'_{dk} denotes the derivative of f_{dk}.

Now, $p(D_{k,i,j})$ only expresses the probability for one stimulus pair (i, j) and one specific subject k. According to the principle of homogeneity of perception the *entire* stimulus configuration should be shared by *all* subjects. As such

$$P_d = \prod_{k=1}^{K} \prod_{(i,j)} p(D_{k,i,j}) \tag{6.8}$$

is the overall probability, according to that particular configuration and parameters r_k and f_{dk} (per subject), of finding the specific dissimilarity responses observed in the experiment.

Finding the model parameters, i.e., the positions of the specific configuration $\mathbf{x}_i,\ i = 1, \ldots, N$, the r_k and f_{dk} that maximize this probability P_d or, equivalently, that minimize the inverse of the log-likelihood function

$$L_d = -\log P_d = -\sum_{k=1}^{K} \sum_{(i,j)} \log p(D_{k,i,j}) \tag{6.9}$$

corresponds to adjusting the model such that the experimentally observed responses are the most likely. Note that it is easy to see this ML-criterion is invariant to linear translations or rotations of the stimulus coordinates $\mathbf{x}_i$ since those transformations preserve distances between points.

In practice, finding the optimal configuration is performed iteratively and each iteration consists of a 3-step process. In the first step, stimulus positions are optimized, assuming fixed values for the regression parameters and the monotonic transformations. In a second stage, the regression parameters are optimized for the fixed stimulus configuration and fixed monotonic transformations on the data. In a third stage the monotonic transformations are updated assuming a fixed configuration and regression parameters.

In our experiments this iterative optimization is incorporated completely in the XGms and XGobi software packages, developed and provided by Prof. Jean-Bernard Martens of the Technical University of Eindhoven, The Netherlands. Fig. 6.3 shows an example of the program's interface, where all parameters discussed in the previous sections can be tuned. For the details on how the initialization and optimization of the regression parameters and monotonic transformations is done exactly in this program we refer again to [Martens, 2003].

6.2.2 Preference data

A second kind of data are preference scores for each subject k on each couple (i, j). These scores again result in a preference matrix P_k per subject where the entry $P_{k,i,j}$ in the matrix corresponds to the preference score of subject

k on a stimulus pair (i, j). More precisely $P_{k,i,j}$ expresses whether subject k prefers stimulus i over j (or vice versa) according to a specific attribute, such as quality, blurriness or amount of artifacts. The ML criterion for the double-stimulus preference data can be derived in a more or less similar way as in the case of dissimilarity data, and is equal to

$$L_p = -\sum_{k=1}^{K} \sum_{(i,j)} \log p(P_{k,i,j}) \tag{6.10}$$

where $p(P_{k,i,j})$ is derived from $p(f_{pk}(P_{k,i,j}))$, and transformation functions f_{pk}[3] similar to the ones for the dissimilarity data (called f_{dk}) are used

$$p(P_{k,i,j}) = p(f_{pk}(P_{k,i,j})) \cdot |f_{pk}^{'}(P_{k,i,j})| \tag{6.11}$$

and

$$p(f_{pk}(P_{k,i,j})) = \phi[f_{pk}(P_{k,i,j}) - \widehat{f_{pk}(P}_{k,i,j})]. \tag{6.12}$$

We notice that the (transformed) experimental preference responses $f_{pk}(P_{k,i,j})$ are again compared to $\widehat{f_{pk}(P}_{k,i,j})$ obtained from a specific geometrical stimulus configuration using the same cost function ϕ as mentioned earlier. The estimates $\widehat{f_{pk}(P}_{k,i,j})$ are now defined as

$$\widehat{f_{pk}(P}_{k,i,j}) = m_k \cdot [< \mathbf{p}_k, \mathbf{x}_i > - < \mathbf{p}_k, \mathbf{x}_j >], \tag{6.13}$$

where m_k and $\mathbf{p}_k$ are the parameters associated to a specific geometrical configuration. Equation (6.13) actually expresses that the estimate $\widehat{f_{pk}(P}_{k,i,j})$ is derived from the stimulus positions $\mathbf{x}_i$, $\mathbf{x}_j$ and a preference vector $\mathbf{p}_k$ and is equal to the vector product between the difference $\mathbf{x}_i - \mathbf{x}_j$, and this preference vector $\mathbf{p}_k$

$$< \mathbf{p}_k, \mathbf{x}_i > - < \mathbf{p}_k, \mathbf{x}_j > = < \mathbf{p}_k, \mathbf{x}_i - \mathbf{x}_j > \tag{6.14}$$

$$= \sum_{m=1}^{n} p_{km}(x_{im} - x_{jm}). \tag{6.15}$$

The regression parameter m_k expresses the linear relationship between the transformed preference scores and the difference $\mathbf{x}_i - \mathbf{x}_j$. In other words, what we actually look for is a vector $\mathbf{p}_k$ so that the difference in (orthogonal) projections of $\mathbf{x}_i$ and $\mathbf{x}_j$ onto this vector relates directly to the preference of i and j according to the attribute scored, see Fig. 6.2. As an example, the $\mathbf{I}$ axis in Fig. 6.1 corresponds to a quality preference axis.

[3]The index p in f_{pk} stands for preference scores.

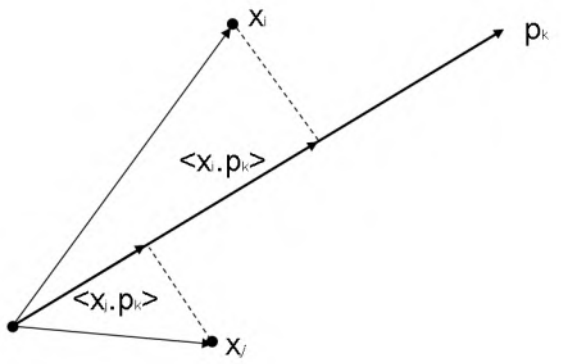


Figure 6.2: Visualization of what the difference in the left part of equation (6.14) represents in the geometrical configuration.

In our experiments, we consider all subjects as belonging to a homogeneous group, in which case a single vector $\mathbf{p}_k = \mathbf{p}$ is estimated for all subjects k in the group. In the actual XGms implementation preference data are added by also estimating m_k and $\mathbf{p}$ in the second step and f_{pk} in the third step of the optimization procedure as described earlier and the likelihood term of equation (6.10) is added to the likelihood term of equation (6.9) for optimization.

6.2.3 Attribute data

Finally, attribute data such as the degree of blur, amount of artifacts and image quality can be gathered for each subject and each individual stimulus. This results in an attribute array A_k where the entry $A_{k,i}$ corresponds to the attribute score of subject k on stimulus i. The ML criterion for the attribute data is

$$L_a = -\sum_{k=1}^{K} \sum_{i} \log p(A_{k,i}), \tag{6.16}$$

where again $p(A_{k,i})$ and $p(f_{ak}(A_{k,i}))$ are derived in a similar way as above.

$$p(A_{k,i}) = p(f_{ak}(A_{k,i})) \cdot |f'_{ak}(A_{k,i})| \tag{6.17}$$

and

$$p(f_{ak}(A_{k,i})) = \phi[f_{ak}(A_{k,i}) - \widehat{f_{ak}(A_{k,i})}]. \tag{6.18}$$

Now the transformed attribute scores $f_{ak}(A_{k,i})$ are compared against estimates $\widehat{f_{ak}(A_{k,i})}$ for a specific configuration, defined as

$$\widehat{f_{ak}(A_{k,i})} = c_k + n_k \cdot < \mathbf{a}_k, \mathbf{x}_i > . \tag{6.19}$$

These estimates for subject k are derived from stimulus positions $\mathbf{x}_i$ and the attribute vector $\mathbf{a}_k$. The regression parameters c_k and n_k determine the linear relationship between transformed attribute scores and predictions for subject k.

Here again the vector-product expresses the correlation between the transformed attribute scores $f_{ak}(A_{k,i})$ and the linear prediction

$$c_k + n_k \cdot < \mathbf{a}_k, \mathbf{x}_i >= c_k + n_k \sum_{m=1}^{n} a_{km} x_{im}. \tag{6.20}$$

Hence, the average strength of a scored attribute for stimulus i increases with the coordinate of the stimulus projection on a one-dimensional axis with scale factor n_k. As an example, the **A** and **B** axes in Fig. 6.1 correspond to the artifacts and blur attribute axes.

As in the case of the preference data, the subjects are considered to belong to one group, which again corresponds to estimating a single prediction vector $\mathbf{a}_k = \mathbf{a}$ for all subjects k in the group. In the actual XGms implementation preference data are added by also estimating c_k, n_k and $\mathbf{a}$ in the second step and f_{pk} in the third step of the optimization procedure as described earlier and the likelihood term of equation (6.16) is added to the likelihood terms of equations (6.9) and (6.10) for optimization.

6.2.4 Maximum-Likelihood estimation

Combining all the data from the former sections, our ML criterion

$$\begin{aligned} L &= L_d + L_p + L_a \\ &= \sum_{k=1}^{K} \sum_{i,j} \log p(D_{k,i,j}) + \sum_{k=1}^{K} \sum_{i,j} \log p(P_{k,i,j}) + \sum_{k=1}^{K} \sum_{i} \log p(A_{k,i}) \end{aligned} \tag{6.21}$$

is minimized as a function of the stimulus positions $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, the regression parameters r_k (for dissimilarity), m_k and $\mathbf{p}$ (for preference), c_k, n_k and $\mathbf{a}$ (for attribute scaling), and the monotonic transformations f_{dk}, f_{pk} and f_{ak} in the three step optimization as described earlier.

The following general result on ML estimation can be used to compare the goodness-of-fit of alternative configuration models at any stage. The estimation of the different parameters, such as r_k, m_k, c_k and n_k, leads to a number of

degrees of freedom (DOF) in the MDS.
Suppose now that L_1 is the optimized ML criterion value for a model with F_1 degrees of freedom, and L_2 ($> L_1$) is the optimized ML criterion value for a simpler model, i.e., with F_2 ($< F_1$) DOFs. It can then be shown that the statistic

$$G_{12}^2 = 2 \cdot (L_2 - L_1) \tag{6.22}$$

satisfies a χ^2-distribution (chi-squared) with $F_1 - F_2$ degrees of freedom, i.e., the probability that G_{12}^2 exceeds the value χ^2 is given by

$$P(G_{12}^2 > \chi^2; F_1 - F_2) = \frac{\Gamma(\frac{F_1-F_2}{2}, \frac{\chi^2}{2})}{\Gamma(\frac{F_1-F_2}{2})} \tag{6.23}$$

in case both models apply, with Γ the incomplete gamma function

$$\Gamma(a, x) = \int_x^\infty z^{a-1} e^{-z} dz. \tag{6.24}$$

Suppose that $\chi_\alpha^2(F_1-F_2)$ is the value for which $P(G_{12}^2 > \chi_\alpha^2(F_1-F_2); F_1-F_2) = \alpha$. The observed value $G_{12}^2 > \chi_\alpha^2(F_1-F_2)$ is then an indication that both models are not equivalent (and the simpler model does not hold). The probability that such a value occurs in the case both models are equivalent is less than α. In our application $\alpha = 0.05$ is chosen.

6.3 Comparing state-of-the-art noise-reduction filters

To demonstrate the power of this multidimensional scaling framework we conducted a first (non-medical) experiment on noise-reduction filters. Image denoising, or the removal of artificial and/or system related image degradations, has been a hot topic for many years in image analysis tasks ranging from image restoration over segmentation to broadcasting and video-applications [Zlokolica, 2006].

Numerous denoising filters have been presented in recent literature based on advanced techniques such as locally adaptive filters in a multi-resolution representation [Pizurica et al., 2003, Portilla et al., 2003, Sendur and Selesnick, 2002], shape-adaptive transforms [Foi et al., 2006], block-matching with 3D transforms [Dabov et al., 2006], steerable filter pyramids [Guerrero-Colon and Portilla, 2005, Rooms, 2005] and fuzzy logic [Van De Ville et al., 2003]. All of these filters try to suppress the noise present while preserving as much image content, structures and detail information, as possible.

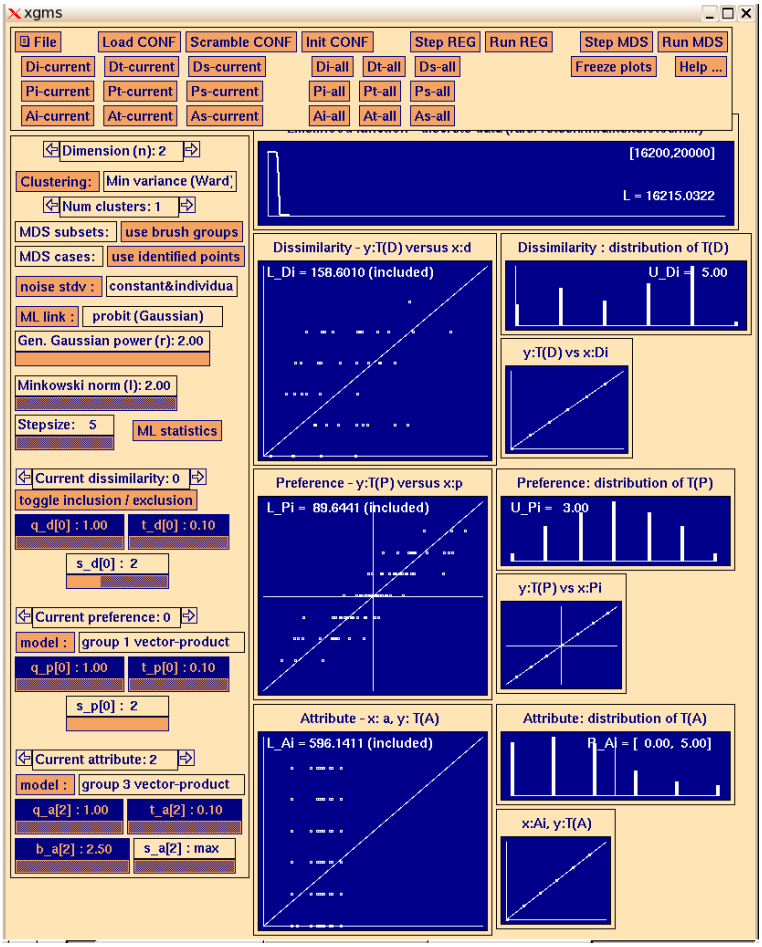


Figure 6.3: Example of the XGobi software interface that allows us to tune the parameters of the underlying XGms software package.

At our department in particular, we have a long tradition in wavelet-based denoising schemes.

As mentioned in the introduction, different well-known instrumental measures such as the RMSE or PSNR are used to compare filter performance. Although these measures are well-suited to determine a *distance* from the filtered image to the original noise-free image, and accordingly to rank the filters, these distances do not necessarily correspond to the overall image quality. Also, as filters become more and more specialized and performant they converge more and more in terms of classical PSNR and RMSE. As such, we can question the significance of slight improvements in PSNR and RMSE in terms of overall image quality.

In this psycho-visual experiment, 7 state-of-the-art denoising schemes were compared and ranked according to their perceived overall image quality, with as main goals to determine why subjects prefer one filter over the other and to investigate fuzzy logic alternatives to PSNR.

In the following Subsection, 6.3.1, we start by reviewing the filters used. The set-up of the experiment is discussed in Subsection 6.3.2 and the results are presented in Subsection 6.3.3. Subsequently, we show how fuzzy similarity measures correspond to the experimental outcome in Subsection 6.3.4. We end with a small discussion in Subsection 6.3.5.

6.3.1 State-of-the-art filters

From literature, the following filters were selected based on the fact that they are either considered as benchmarks or because of high PSNR-values when compared to their peers. Note that this list is not exhaustive since in view of the length of the experiment we had to restrict the number of filters. We also only present the general idea behind the filter frameworks. For technical details on the exact implementations of the filters we refer to the papers in the references:

- **The GOA filter** [Van De Ville et al., 2003]: this is a two-step filter where first a fuzzy image derivative for eight different directions is computed, which is then used to perform a fuzzy smoothing by weighting the contributions of neighboring pixel values. Both stages are based on advanced fuzzy rules and membership functions.

- **The SA-DCT filter** [Foi et al., 2006]: the Shape-Adaptive Discrete Cosine Transform scheme uses an over-complete transform-domain filter in conjunction with anisotropic Local Polynomial Approximation (LPA) and local Intersection of Confidence Intervals (ICI) techniques, which - for every location in the image - adaptively define an appropriate shape for the transform's support.

- **The 3D-DFT filter** [Dabov et al., 2006]: the block-matching and 3D filtering approach exploits the possible correlation among similar blocks within an image by filtering in the 3D-transform domain. The third dimension corresponds to stacking together the blocks which are matched as similar.

- **The ProbShrink filter** [Pizurica and Philips, 2006]: this adaptive spatial filter shrinks wavelet coefficients, in a multi-resolution representation, according to the probability of the presence of a signal of interest conditioned on a local spatial activity indicator.

- **The BLS-GSM filter** [Guerrero-Colon and Portilla, 2005]: the Bayesian Least Squares - Gaussian Scale Mixtures method extends filtering in the steerable pyramid domain based on Gaussian scale mixtures [Portilla et al., 2003] by employing a two-level (coarse-to-fine) local adaptation to spatial image features.

- **The Bishrink1, Bishrink2 filters** [Sendur and Selesnick, 2002]: this method applies a bivariate shrinkage of wavelet coefficients using the interscale dependencies and the local spatial variance estimation. Two variants were provided corresponding to different noise estimation techniques.

- **The SPERRIL filter** [Rooms, 2005]: this Steerable Pyramid-based Estimation and Regularization of Richardson-Lucy filter is essentially an image restoration method, where the regularization (denoising) part is done in the steerable pyramid domain employing the inter-scale (parent-child) relationships between the coefficients.

Fig. 6.4 shows the results obtained from applying these filters to a greyscale image with artificial zero-mean white noise of $\sigma = 15$ added.

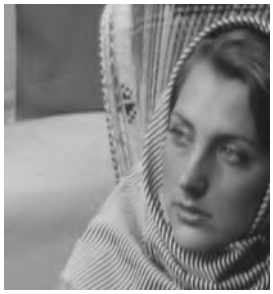
6.3.2 Psycho-visual experiment

A psycho-visual experiment for the assessment of perceived image quality has been described in detail in [Kayagaddem and Martens, 1996] for images artificially degraded by noise and blur. We constructed our own experiment that focused on more subtle image distortions as compared to the existing experiment.

Stimuli. Three 512×512 pixels 8-bit scenes (Barbara, Face and Hill) containing different content information ranging from texture over fine details to uniform backgrounds were used in the experiment, see Fig. 6.5. These images were degraded by additive zero-mean white Gaussian noise with a standard deviation of $\sigma = 15$ and sent to the authors of the filters who were asked to denoise them blindly, i.e., without any information on the noise level.


(1) BLS-GSM filter


(2) noisy image


(3) original image


(4) ProbShrink filter


(5) SA-DCT filter


(6) GOA filter


(7) SPERRIL filter


(8) Bishrink1 filter


(9) Bishrink2 filter


(10) 3D-DFT filter

Figure 6.4: The test images for Barbara as presented in our psycho-visual experiment.

(a) Barbara (b) Face (c) Hill

Figure 6.5: The test scenes used in our psycho-visual experiment.

The original image, the noisy one and the 8 denoised images were then presented on a 19 inch BELINEA high-resolution LCD display, under fixed lighting conditions. The viewing distance was 80 cm, which is about three times the height of the monitor. All images were displayed at their full resolution on a white background and by means of the keyboard the subjects themselves could decide when to switch to the next image(s). Fig. 6.6, 6.7 and 6.8 show the denoised versions of the three scenes used in the experiment.

Methodology. By inspecting the denoised images qualitatively we notice almost all filters succeed in reducing the image noise completely. This is not surprising since most of the filters are considered state-of-the-art benchmarks. However, all noise-reduction filters have in common that with the reduction of noise both high-frequent information is lost and artifacts appear. Consequently, we asked our subjects to score the images for the following three attributes: blur, artifacts[4], and overall quality. Besides these attribute scores, we also asked for overall dissimilarity scores and preference scores in quality.

In practice, the data were collected in 3 sessions. In a first session, dissimilarity scores and quality preference scores for all pairs of stimuli were collected in a double-stimulus set-up. All unique couples of different stimuli were presented, one on the left and one on the right of the display. The subjects were instructed to rate the dissimilarity between two images using an integer score between 0 and 5. A score of 5 indicated the greatest dissimilarity and a score of 0 implied no perceived difference.
Next, the subjects were asked to assign preference scores for image quality on the same couples, this on an integer scale ranging from -3 to +3. Here -3 corresponded to the greatest preference for the left hand image, +3 to the greatest preference for the right hand image, 0 to no preference in perceived quality.

In a second session perceived blur, artifacts and quality were judged using a nu-

[4]Our definition of an artifact is every artificial *structure* in the image that disturbs the image quality, apart from blurring that is. As such, the noise in the degraded non-filtered image is also considered as an artifact.

(1) BLS-GSM filter

(2) noisy image

(3) original image

(4) ProbShrink filter

(5) SA-DCT filter

(6) GOA filter

(7) SPERRIL filter

(8) Bishrink1 filter

(9) Bishrink2 filter

(10) 3D-DFT filter

Figure 6.6: The test images for Barbara as presented in our psycho-visual experiment.

(1) BLS-GSM filter

(2) noisy image

(3) original image

(4) ProbShrink filter

(5) SA-DCT filter

(6) GOA filter

(7) SPERRIL filter

(8) Bishrink1 filter

(9) Bishrink2 filter

(10) 3D-DFT filter

Figure 6.7: The test images for Face as presented in our psycho-visual experiment.

(1) BLS-GSM filter

(2) noisy image

(3) original image

(4) ProbShrink filter

(5) SA-DCT filter

(6) GOA filter

(7) SPERRIL filter

(8) Bishrink1 filter

(9) Bishrink2 filter

(10) 3D-DFT filter

Figure 6.8: The test images for Hill as presented in our psycho-visual experiment.

merical category single-stimulus scaling procedure. Each of the subjects scored the attributes mentioned for all scenes presented separately and the numerical category scale ranged from 0 to 5. The stronger the perceived attribute, the higher the score.

In a third session a follow-up experiment took place where, based on the MDS outcome, 5 well-chosen triples of images were shown to the subjects who were asked to retain the 2 best images, in terms in overall quality, and describe *in words* why they had retained them.

All subjects took part in a training session involving 10 stimuli, from a fourth scene, covering the entire range of distortions to adjust the sensitivity of their scale. The combined three sessions of the experiment took about 75 minutes and short pauzing was allowed.

Subjects. 37 subjects, ranging between 22 and 45 years of age, took part in the first two sessions of the experiment. Since only grey scale images were used, no screening for color-blindness was conducted. 10 subjects from the same group took part in the third session of the experiment. All subjects were familiar with numerical category scaling and the concepts of image quality, blur and artifacts.

6.3.3 Results

Once the data are collected, the XGms package is used to process them. Fig. 6.9, 6.10 and 6.11 show the geometrical spaces as optimized by the MDS framework under the assumption the 37 subjects constitute a homogeneous group (attribute and preference vectors are shared by all). The stimulus configurations were computed for spline interpolation with two kernel knots as monotonic transformations for the dissimilarity, preference and attribute data.

Each point in these figures corresponds to one of the filters shown in Fig. 6.6, 6.7 and 6.8. The 95% confidence intervals on the positions are also plotted as the little ellipses. We notice that overall similar configurations were obtained for all three scenes, apart from a rotational variance inherent to the MDS framework.

The arrows in the figures point out the directions along which the different scored attributes should be measured. The **I**-axis stands for the impairment preference vector and is the opposite direction of perceived overall image quality: the further along the axis the more quality degrades. The **B**-axis stands for the perceived blur attribute axis: the further along the axis the less blur is perceived. The **A**-axis stands for artifacts attribute axis, the further along the axis the less artifacts perceived.

The orthogonal projection of the points onto these axes gives us a relative ranking in attributes. For example, Fig. 6.12, 6.13 and 6.14 represent the projection of the perceived quality for all three scenes. On the X-axis the

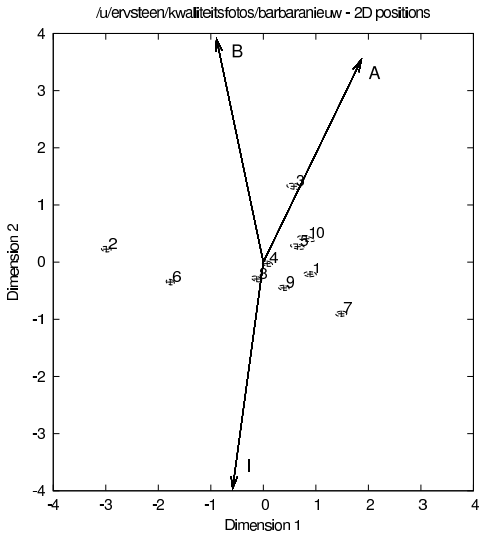


Figure 6.9: The multidimensional scaling output of the experiment for the Barbara scene. The numbers of the filters correspond to the numbers in Fig. 6.6.

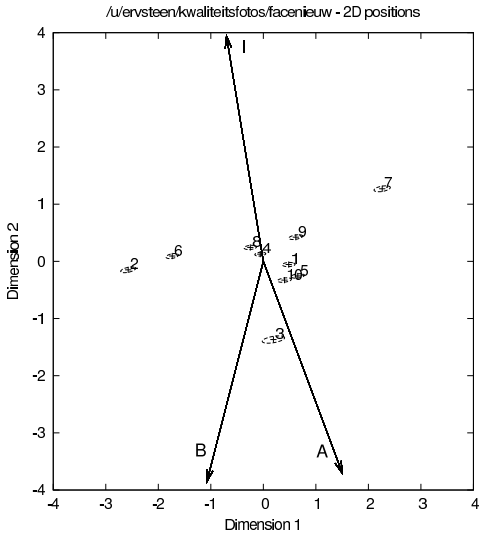


Figure 6.10: The multidimensional scaling output of the experiment for the Face scene. The numbers of the filters correspond to the numbers in Fig. 6.7.

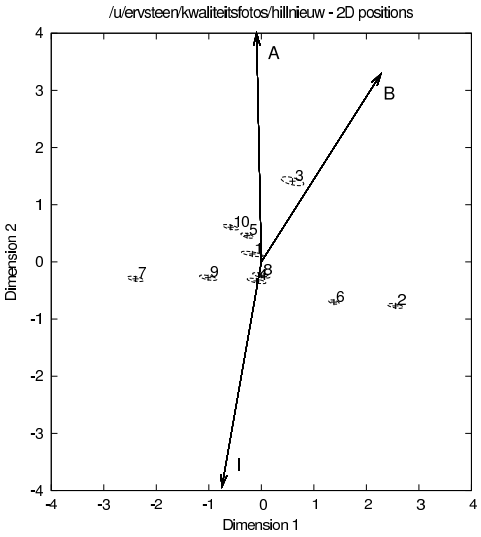


Figure 6.11: The multidimensional scaling output of the experiment for the Hill scene. The numbers of the filters correspond to the numbers in Fig. 6.8.

numbers of the filters are presented, the perceived overall quality score on the Y-axis.

The outcome of the different plots in Fig. 6.12, 6.13 and 6.14 is also comprised in Table 6.1. In the MDS columns of this table, we see that for all three scenes the original image (3) always comes out best, consistently followed by the 3D-DFT (10) and SA-DCT (5) filters. The BLS-GSM (1), ProbShrink (4) filters and Bishrink1 (8) follow in that order except for Barbara, where BLS-GSM and ProbShrink switch place. The GOA filter (6) and SPERRIL filter (7) and Bishrink2 (9) filters are consistently ranked as worst, even below the noisy image (2).

The power of MDS is that a connotation to the distances along the quality axes in Fig. 6.12, 6.13 and 6.14 can be given through an interpretation of the attribute axes in Fig. 6.9, 6.10 and 6.11. For example, to investigate the cause of the perceived difference between the original image (3) and the best-two performing filter 3D-DFT (10) and SA-DCT (5), we can inspect the perceptual geometry.

In Fig. 6.15, on the left the projections on the impairment-axis are shown. In the middle, the projections on the artifacts-axis are shown and on the right the projections on the blur-axis. From these figures it is clear that although the SA-DCT (5) filter results in a slightly less blurred image, it is particularly the absence of artifacts that adds up to the overall image quality.

If we compare the BLS-GSM (1) and Probshrink (4) filter, see Fig. 6.16, we again notice that it is mostly the presence or absence of artifacts that influences

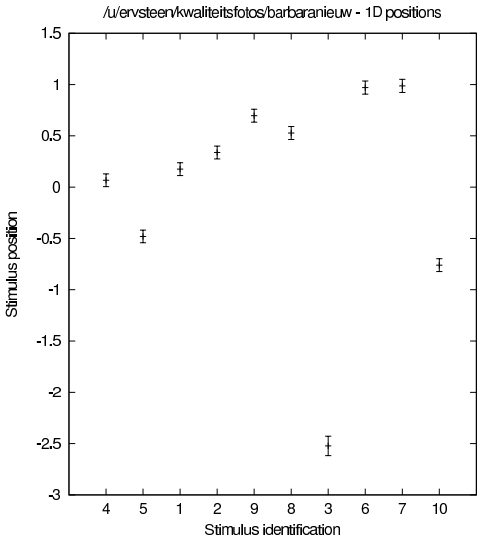


Figure 6.12: The 1D-geometrical output of the MDS framework for the Barbara scene. On the X-axis the different filters as numbered in Fig. 6.6. On the Y-axis the perceived quality.

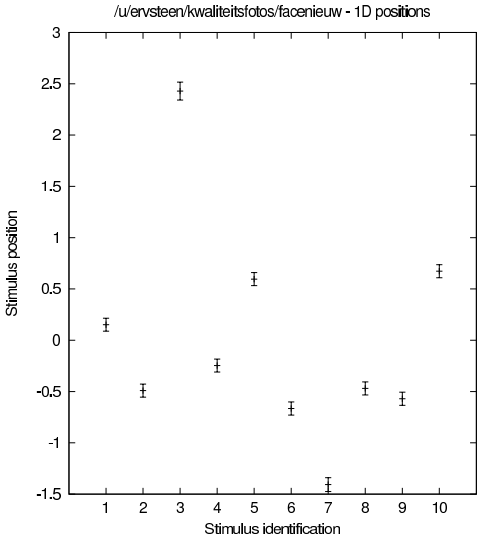


Figure 6.13: The 1D-geometrical output of the MDS framework for the Face scene. On the X-axis the different filters as numbered in Fig. 6.7. On the Y-axis the perceived quality.

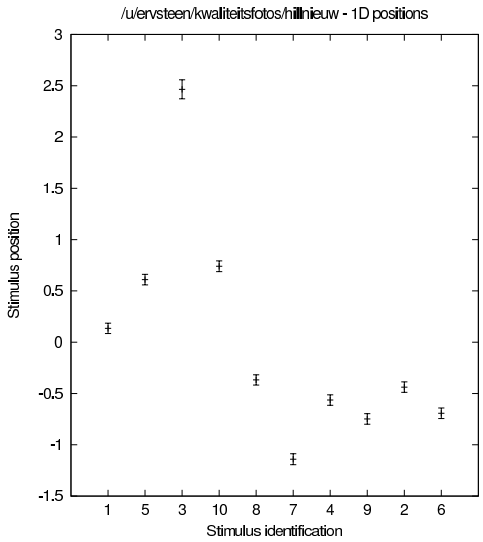


Figure 6.14: The 1D-geometrical output of the MDS framework for the Hill scene. On the X-axis the different filters as numbered in Fig. 6.8. On the Y-axis the perceived quality.

Table 6.1: Quality ranking of the filters for each of the 3 scenes based on the outcome of the MDS as well as PSNR. 1 corresponds to best, 10 corresponds to worst.

rank	**Face**		**Barbara**		**Hill**	
	MDS	*PSNR*	*MDS*	*PSNR*	*MDS*	*PSNR*
1	Original	Original	Original	Original	Original	Original
2	3D-DFT	3D-DFT	3D-DFT	3D-DFT	3D-DFT	3D-DFT
3	SA-DCT	SA-DCT	SA-DCT	BLS-GSM	SA-DCT	SA-DCT
4	BLS-GSM	BiShrink1	Probshrink	SA-DCT	BLS-GSM	BLS-GSM
5	ProbShrink	BLS-GSM	BLS-GSM	Bishrink1	Bishrink1	Bishrink1
6	Bishrink1	Bishrink2	Noisy	ProbShrink	Noisy	ProbShrink
7	Noisy	ProbShrink	Bishrink1	Bishrink2	ProbShrink	Bishrink2
8	Bishrink2	SPERRIL	BiShrink2	SPERRIL	Bishrink2	SRERRIL
9	GOA	Noisy	GOA	Noisy	GOA	Noisy
10	SPERRIL	GOA	SPERRIL	GOA	SPERRIL	GOA

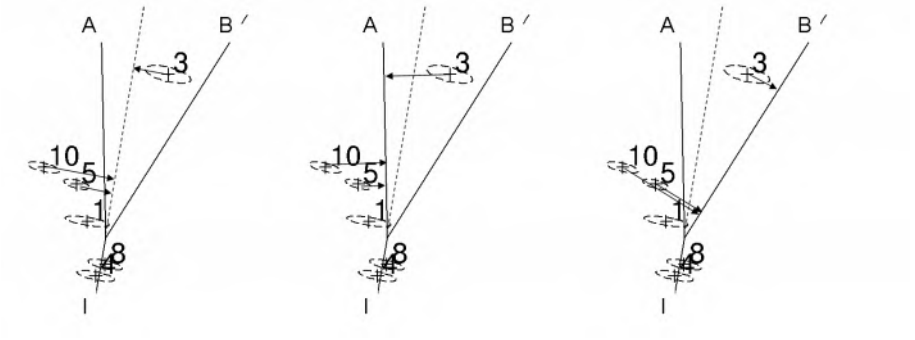


Figure 6.15: Comparison of the Original image (3) and the best performing filter, i.e.,, the 3D-DFT filter (9) based on the MDS geometry.

the overall image quality.

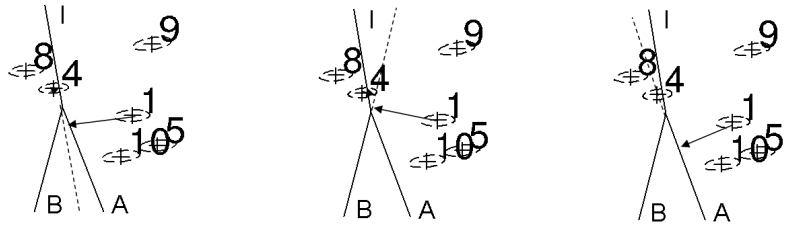


Figure 6.16: Comparison of the Probshrink (4) and the BLS-GSM filter (1) based on the MDS geometry.

However, for some filters, the projections on the quality axes are either too close to one another or equally far apart which makes it difficult to distinguish what exactly influences the quality most. Therefore in the third session of the experiment, we performed the follow-up experiment to validate the perceptual differences for well-chosen filter couples.

The results from this experiment for the Face image are presented in Table 6.2. This table shows the most decisive criteria considered by the different subjects in pointing out the actual difference between the images and should be interpreted as follows: the filter in row i outperforms the filter in column j, mainly because of table entry (i, j).
If we now again look at the 2 best-performing filters, 3D-DFT (10) and SA-DCT-filter (5), we see that although these are very close to one another in the perceptual space, their main difference lies in the amount of artifacts and details. This confirms the result we got from interpreting the projection in the perceptual space.

Another result from the follow-up experiment is that the noisy image is pre-

Table 6.2: This table shows the main attributes by which the filter in row i is chosen over the filter in column j, based on the follow-up experiment. B = blur, A = artifacts, D = detail information.

	ori.	3D-dft	SA-dct	bls-gsm	ProbShr.	Noisy	Bishr.2	GOA
ori.	/	A	A	B	/	N	/	$N+B$
3D-DFT	/	/	$A+D$	$B+A$	/	/	/	/
SA-DCT	/	/	/	$A+D$	$B+A$	/	/	/
BLS-GSM	/	/	/	/	$B+D$	/	/	/
ProbShr.	/	/	/	/	/	B	B	/
Noisy	/	/	/	/	/	/	B	B
Bishr.2	/	/	/	/	/	/	/	N
GOA	/	/	/	/	/	/	/	/

ferred above the Bishrink1 (8) and GOA (6) filter mainly because of the lesser blur in the images. This means that although there is a lot of noise present, subjects tend to prefer the preservation of high-frequency information and sharpness of edges in the images. This is in line with earlier psycho-visual findings on blur and noise [Martens, 2003].

Finally, from Fig. 6.9, 6.10 and 6.11 we consistently see that perceived overall quality axis is a (linear) combination of the blur and artifacts axes, meaning that as expected they are not uncorrelated. Also, artifacts seem to contribute slightly more to image quality since the angle of the artifacts-axis to the impairment-axis is smaller than that of the blur-axis. Furthermore, since the MDS likelihood criterion is optimized for a 2D geometrical configuration, no other attributes (than blur and artifacts) seem to amount to the overall quality in the case of these filters.

6.3.4 Similarity measures

Performing and evaluating a psycho-visual experiment is very time-consuming. Also, once an experiment is finished, it is not straightforward to fit in a new filter since we would have to do the entire experiment over including the filter. Therefore, as mentioned earlier, instrumental similarity measures were developed that quantify how (dis)similar two images are in terms of image quality. The most common similarity measure in image denoising is the Peak Signal-to-Noise Ratio (PSNR) defined as follows: suppose I is an original greyscale image and I' an either noisy or filtered version, then the PSNR is computed as

$$PSNR(I, I') = 20 \log_{10} \frac{255}{\sqrt{\sum_x \sum_y (f_{I'}(x,y) - f_I(x,y))^2}} \tag{6.25}$$

where $f_I(x, y)$ corresponds to the grey value of pixel in I at position (x, y). PSNR is expressed in dB and, as can easily be seen from the formula, the higher the PSNR the higher the pixel-to-pixel correspondence between two images.

Table 6.3: PSNR values (dB) for all three scenes and all 7 filters used in the first experiment.

PSNR	**Face**	**Barbara**	**Hill**
3D-DFT	36.99	33.30	31.69
SA-DCT	36.82	31.38	31.54
BLS-GSM	36.43	32.04	31.42
Bishrink1	36.47	31.20	31.01
Bishrink2	36.27	29.76	29.95
ProbShrink	35.53	31.19	30.87
SPERRIL	31.09	28.99	28.50
Noisy	24.65	24.62	24.65
GOA	21.25	23.45	22.26

In almost all state-of-the-art papers, PSNR is the benchmark for the quality performance of a filter. When images differ significantly, i.e., show well-pronounced artifacts or noise-levels, PSNR is shown to be a valid measure for image quality. However, as state-of-the-art filter performance improves, PSNR differences amongst the filters usually become very small, see Table 6.3. Consequently, we can ask ourselves how small improvements in PSNR actually relate to image quality.

Therefore, we also plotted the PSNR-ranking in Table 6.1 as a comparison to the MDS ranking. We see that the 3D-DFT filter still performs best in terms of PSNR but now the BLS-GSM filter comes in second in case of Barbara, followed by the SA-DCT filter. If we look at the bottom of the table we also see some changes. Besides that, we notice a bigger shift in PSNR-ranking through the scenes than in our psycho-visual experiment where the top and bottom 3 images were always ranked consistently. Finally, if we compare the SA-DCT (5) and BLS-GSM (1) filter visually in the case of Barbara, see cutouts (5) and (1) in Fig. 6.4, it is clear the SA-DCT filter has less quality-disturbing artifacts, e.g., around the nose, although their PSNR-ranking suggests the opposite. As such, our experiment shows that when it comes to small image distortions, PSNR is not the most reliable overall measure.

Consequently, different instrumental measures have been presented, with the aim of relating better to psycho-visual rankings. We compared the PSNR to 7 fuzzy similarity measures proposed by the Fuzziness and Uncertainty Modeling Research Unit of our university [Van der Weken et al., 2001]. To understand these measures we first need to introduce the concept of a *fuzzy set*.

Fuzzy sets are an extension of classical set theory. In classical set theory the membership of elements in relation to a set is assessed in binary terms according to a crisp condition, i.e., an element either belongs or does not belong to the set. Fuzzy set theory, on the contrary, permits the gradual assessment of the

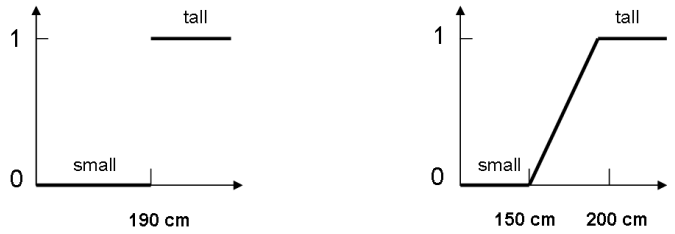


Figure 6.17: Example of the human length as a fuzzy set. Left: a sharp threshold for small and tall is put as 190 cm. Right: a more gradual transition from small to tall is proposed.

membership of elements in relation to a set. This is usually described with the aid of a *membership function*.
Suppose x is an element of a universe X. A fuzzy set A over the universe X is then characterized by a membership function γ_A, which takes on a value in the interval [0,1]:

$$A: \ X \to [0,1], x \longmapsto \gamma_A(x). \tag{6.26}$$

The value $\gamma_A(x)$ is called the degree of membership of x in the fuzzy set A.

A simple intuitive example of a fuzzy set is "human length". It is difficult to put an exact threshold on when someone is tall or small. To show this, suppose you would consider 190 cm sharp as the threshold for being tall, see Fig. 6.17 (left). Someone measuring 189 cm would then be considered small, whereas someone measuring 191 cm would be tall although in fact there is only a 2 cm difference between the two. This can be solved by allowing a more gradual change in length. Consider 150 cm as being really small and 200 cm as really tall and everything in between as increasing in length, see Fig. 6.17 (right). The gradual assessment in membership of this fuzzy set then relates to the concept of human length in a more natural way.

Consider now a digital image A containing greyscale values in the [0,255] interval, 0 being corresponding to white, 255 to black. By normalizing this image we obtain the same greyscale image, yet now in the interval [0,1], where 0 is still black, 1 is white, and all the values in between are considered different greyscales. Given the definitions mentioned before, we have now transformed the digital image A into a fuzzy set over the universe of pixels.

Next, we define the *intersection* and *union* of two fuzzy sets. The intersection $A \cap B$ and union $A \cup B$ of fuzzy sets A and B are computed as

$$A \cap B : X \to [0,1], x \longmapsto \min(\gamma_A(x), \gamma_B(x)) \tag{6.27}$$

and

$$A \cup B : X \to [0,1], x \longmapsto \max(\gamma_A(x), \gamma_B(x)) \tag{6.28}$$

respectively.

Now that these basic fuzzy concepts are clear, we can define similarity measures between two images based on similarity measures between two fuzzy sets. A well-known example of a fuzzy set similarity measure, also applicable to images, was presented by [Chen et al., 1995]:

$$M_6(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.29}$$

$$= \frac{\sum\limits_{(i,j)\in X} \min(\gamma_A(i,j), \gamma_B(i,j))}{\sum\limits_{(i,j)\in X} \max(\gamma_A(i,j), \gamma_B(i,j))}. \tag{6.30}$$

with A and B two digital images over the universe X of pixels (i,j) and the intersection and union defined as above.

Similarity measures which are applied directly to the normalized digital images on a pixel to pixel basis, just as the PSNR, do not perform very well in the case of small image distortions. Therefore, in [Van der Weken et al., 2001], more advanced similarity measures were proposed. First of all, pixel-based similarity measures were applied to image regions in order to construct neighborhood-based similarity measures. Suppose the image is divided into n image regions of size 8×8 pixels and the similarity between the image region A_i of image A and the image region B_i of image B is denoted by $M(A_i, B_i)$, then the similarity between the two images A and B is given by the average of the similarities in the corresponding disjoint image regions. So, we have that

$$M^p(A,B) = \frac{1}{n} \sum_{i=1}^{n} M(A_i, B_i), \tag{6.31}$$

Next, weighted neighborhood-based similarity measures were constructed. Suppose again the image is divided into n image parts of size 8×8 pixels, and the similarity between the image part A_i of image A and the image part B_i of image B is denoted by $M(A_i, B_i)$, then the similarity between the two images A and B is given by the weighted average of the similarities in the corresponding disjoint image parts. So, we have that

$$M^h(A,B) = \frac{1}{n} \sum_{i=1}^{n} w_i \cdot M(A_i, B_i), \tag{6.32}$$

where the similarity $M(A_i, B_i)$ is calculated using the pixel-based similarity

measures restricted to the image parts A_i and B_i and the weight w_i is defined as the similarity between the homogeneity h_{A_i} of image part A_i and the homogeneity h_{B_i} of image part B_i.
The homogeneity h_{A_i} of image part A_i is computed as the similarity between the pixel in the image part with maximum intensity and the pixel in the image part with minimum intensity, using the similarity function s defined as

$$s(x,y) = \begin{cases} 1 - \frac{|x-y|}{a}, \ if \ |x-y| < a \\ 0, \ if \ |x-y| \geq a \end{cases} \tag{6.33}$$

where x and y correspond to pixel grey values and where 0.5 is a typical value for a. As such, we define

$$h_{A_i} = s(\max_{A_i} x, \min_{A_i} y) \tag{6.34}$$

and the weight w_i is calculated as:

$$w_i = s(h_{A_i}, h_{B_i}). \tag{6.35}$$

Finally, besides applying the similarity measures directly to the pixels of the considered images or neighborhoods of pixels, similarity measures were applied to the image histograms. An image histogram can also be considered a fuzzy set when it is again normalized. Also, before normalizing values of a histogram can be ordered in such a way the least occurring grey value is placed in the first position of the histogram and the remaining frequencies are ordered in increasing order. We can then apply the different similarity measures to these ordered and normalized histograms.

In this thesis, we selected 7 particular measures. We considered the similarity measure M_6 of equation (6.29) applied in different ways to a pair of images:

- applied neighborhood-based (M_6^p);
- applied neighborhood-based and incorporating homogeneity (M_6^h);
- applied to normalized histograms (H_6);
- applied to ordered normalized histograms (OH_6);

Furthermore, we considered the neighborhood-based similarity measures M_1^h, M_{18c}^h and M_{i3}^h, that also incorporate homogeneity. Again we refer to [Van der Weken et al., 2001, Van der Weken et al., 2002, Van der Weken et al., 2003, Van der Weken et al., 2004, Van der Weken, 2004] for more technical details.

What we are interested in, is to investigate which of these fuzzy similarity measures corresponds best to the obtained psycho-visual configuration. From

the geometrical multidimensional space, the Euclidean distance matrix for the stimuli can be computed. Using the similarity measures, we can also compute the different distance matrices directly from the input images. By comparing these matrices to the experimental Euclidean distance matrix we obtain the similarity measure that fits the experiments best, or in other words that is closest related to human visual quality assessment.

Let D_p present the inter-picture distance matrix from the 2D geometrical MDS configuration and suppose D_s the distance calculated through the fuzzy similarity measures. If a similarity measure predicts the geometrical configuration well, their respective distance matrices should be equivalent.

This equivalence between two matrices can be computed by the *Spearman Rank Order Correlation* (SROC) coefficient. Let D_p be the inter-picture distance matrix and D_s be a $N \times N$ similarity measure matrix, N being the number of input images used. Let

$$R_{ps} = 1 - 6\frac{\sum_{i=2}^{N}\sum_{j=1}^{i-1}(Rank[D_p(i,j)] - Rank[D_s(i,j)])^2}{N_D(N_D^2 - 1)}$$

where $Rank[D(i,j)]$ stands for the rank of matrix entry $D(i,j)$ which is a number between 1 and $N_D = N(N-1)/2$ when we order all matrix elements ascendingly. A derivative of the Spearman Rank Order Correlation coefficient D_{ps} is then given by

$$D_{ps} = \sqrt{1 - R_{ps}^2}.$$

We can easily see that this value ranges in the interval $[0, 1]$ and that the smaller the value, the bigger the equivalence is between the matrices. Table 6.4 shows the Spearman coefficients for our psycho-visual experiment.

Table 6.4: Derived Spearman Rank Order Correlation coefficients for our psycho-visual experiment.

Spearman's Rank Order coefficient D_{ps}	**Barbara**	**Face**	**Hill**
M_1^h	0.622	0.509	0.622
M_{18c}^h	0.627	0.505	0.538
M_6^p	0.607	0.506	0.514
M_{i3}^h	0.589	**0.497**	0.516
H_6	0.738	0.584	0.68
M_6^h	**0.478**	**0.531**	**0.516**
OH_6	0.588	0.59	0.69
PSNR	0.586	0.642	0.634

From the results in Table 6.4 we conclude that for the Barbara image the measure M_6^h performs best, that the measures M_{18c}^h, M_6^h and M_{i3}^h perform best for the Face image and the measures M_6^p, M_6^h and M_{i3}^h perform best for the Hill image. In all of the scenes the fuzzy measures outperform the PSNR.

6.3.5 Discussion

Using the MDS framework we were able to rank 7 state-of-the-art noise-reduction filters on perceptual image quality, based on a psycho-visual experiment on three different scenes, including 37 subjects. Although one can claim three scenes is too little to achieve acceptable results, for time constraints we restricted ourselves to scenes where all important types of (greyscale) content information was visible, i.e., texture, edges, highly detailed as well as uniform regions.

From the MDS geometrical models we see that the stimulus configurations are similar for all three scenes, which allows us to conclude that subjects are able to assess the overall filter quality independent of the scene content and thus that adding more scenes will probably not result in more or different information.

The number of subjects used in this test is relatively high, in comparison to other similar tests varying usually between 10 and 20 subjects. If we look at the resulting confidence intervals of the stimuli (small ellipses) we consider our population big enough to achieve significant results. Note however that we assumed all subjects to belong to the same population, i.e., sharing the same attribute and preference vectors. As such, we did not comment on the inter-subject differences here.

Concerning the quality ranking, it is clear the 3D-DFT and SA-DCT filters outperform their peers and can as such be considered the filters of highest image quality. The MDS ranking also showed that filtering in some cases degrades the overall quality as compared to the noisy image. The third session of our experiment learned us that this is because artifacts are considered more decisive for image quality when images are only slightly blurred. On the contrary, in case of a high degree of artifacts, blur becomes the decisive criterion for image quality.

Compared to PSNR, our MDS ranking is more robust and true to image quality. Therefore, we conclude that PSNR is not the most suited criterion for image quality comparison when small image distortions are involved. As such, we investigated fuzzy similarity measures as an alternative to PSNR and found at least one measure, M_6^h, outperforming the PSNR over all scenes. Note however that fuzzy similarity measures are not the *only* alternatives to PSNR. Other similarity measures can be considered/constructed.

We believe that similarity measures that account for multiple attributes, e.g., that measure the amount of blur and artifacts separately and then either combine these numbers into one number or use these as a 2D similarity measure, will

outperform the fuzzy similarity measures tested. Sticking to fuzzy measures, we believe that replacing the homogeneity criterion in the weighted neighborhood measures by a criterion that measures the amount of blur or artifacts should also result in significant improvements. The plan is to investigate such measures in future.

Finally, we note that we only performed the experiment for one noise-level. The level was chosen so that the differences between the images were visible yet not too obvious. The authors of the filter papers were also asked to denoise the images corrupted with noise of $\sigma = 35$. However this resulted in far too many artifacts and severely degraded images as none of the filters succeeded in denoising well. Our own test on noise of $\sigma = 5$ resulted in images that were too alike to really tell the visual difference. As such, our choice of $\sigma = 15$ appeared the most realistic one to compare the filter performance.

6.4 Diagnostic value of speckle-reduction

In Chapter 2 we showed that speckle is the building block of all US images. We also mentioned there are different kinds of speckle, according to the scatterers they result from. The most relevant speckle for our research was speckle resulting from low-density structural scatterers since this allowed us to characterize (pathological) white matter tissue properties. Although in our tissue characterization we did not filter out possible redundant or irrelevant speckle there are applications where speckle-reduction is useful.

We used speckle-reduction as a *preprocessing* to image segmentation and registration. In Chapter 4, Section 4.4, we used a speckle-reduction filter called the modified GenLik filter, to suppress speckle around the ventricle borders without blurring the borders itself. In Chapter 5, Section 5.3, we reduced speckle in the US images, using the same modified GenLik filter, as a preprocessing step to our registration algorithm.

As such, the benefit of speckle-reduction in actual quantitative image processing has already been shown. However, enhancing image structures or features and suppressing irrelevant information should also alter the qualitative perception of the images. As such, we can ask ourselves if speckle-reduction would also be an asset to the visual image diagnosis. Are diagnostically relevant structures of certain pathologies indeed enhanced by speckle-reduction or not? Are physicians by training and expertise used to looking at the unpreprocessed speckle images and do they accordingly conceive speckle-reduction as artificial? Does the quality of a filter depend on the image content or how exactly a pathology is pronounced in the US images?

These are all questions that can not be answered by again just computing standard instrumental measures. Therefore, the main goal of this section is to conduct a small psycho-visual experiment involving multiple physicians, images

of multiple pathologies and a modification of the GenLik filter, to investigate if speckle-reduction improves diagnostic image quality.

The structure of this section is as follows: we start by explaining the fundamentals of the modified GenLik filter in Subsection 6.4.1. In Subsection 6.4.2, we present the set-up of our psycho-visual experiment. The results based a one-dimensional scaling framework are presented in Subsection 6.4.3. We conclude with a discussion in Subsection 6.4.4.

6.4.1 The modified GenLik filter

The original GenLik filter [Pizurica et al., 2003] was developed at our department and is a multi-resolution denoising method that guarantees a good preservation of clinically interesting features due to both an efficient spatial adaptivity and its adaptivity to a given preference.

The method uses the non-decimated wavelet transform of [Mallat, 1998] to create a multi-resolution image representation and subsequently shrinks each wavelet coefficient according to the probability that it presents a signal (or a feature) of interest, given the observed coefficient value and a Local Spatial Activity Indicator (LSAI) calculated from the surrounding wavelet coefficients.

Let y_k and w_k respectively denote the noise-free and the observed wavelet coefficient (of a certain scale) at (pixel) position (x, y) in the image and let z_k denote the LSAI at the same position. The LSAI is the locally averaged magnitude of the observed wavelet coefficients, defined as

$$z_k = \frac{1}{n} \sum_{i \in N_k} |w_i|, \tag{6.36}$$

where N_k is the square 5×5 window centered at position k and $n = 25$. Further on, let X_k denote a binary random variable being a *significance label* for w_k. The event $X_k = 1$ reads: w_k represents a signal of interest (hypothesis H_1) and the event $X_k = 0$ denotes the opposite hypothesis H_0. The GenLik denoiser is then defined as

$$\hat{y}_k = P(X_k = 1 | w_k, z_k) = \frac{r \xi_k \eta_k}{1 + r \xi_k \eta_k} w_k, \tag{6.37}$$

where $r = P(X_k = 1)/P(X_k = 0)$ is the prior ratio and ξ_k, η_k the likelihood ratios, $\xi_k = p_{W_k|X_k}(w|1)/p_{W_k|X_k}(w|0)$ and $\eta_k = p_{Z_k|X_k}(z|1)/p_{Z_k|X_k}(z|0)$, which are estimated empirically from the input image.

The characteristic parts of this method are shown in Fig. 6.18: in a first stage, inter-scale products are compared against a threshold T in order to locate the significant coefficients, i.e., edge coefficients in our case. This threshold can be tuned to define the notion of a significant feature. This preliminary classification yields a *mask* $\hat{\mathbf{x}}$, which is then used for empirical estimation of

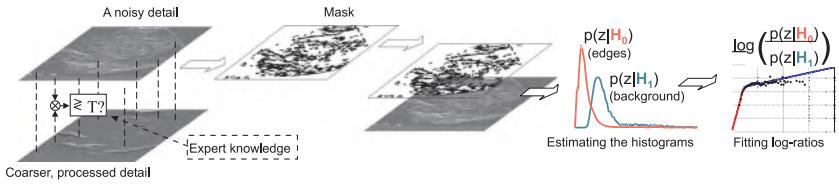


Figure 6.18: Characteristic parts of the GenLik denoising scheme that uses a generalized likelihood ratio in the wavelet domain (source: [Pizurica, 2002]).

the conditional probability density functions. As Fig. 6.18 pictorially shows, the likelihood ratios ξ_k, η_k are finally subjected to a piece-wise linear fitting in a logarithmic representation. The prior ratio is estimated as $\hat{r} = \sum_{k=1}^{N} \hat{x}_k / (N - \sum_{k=1}^{N} \hat{x}_k)$, where N is the number of the coefficients in a given sub band.

We made the following adjustments to adapt this method to US images. A first adaptation is trivial but crucial for practice: medical US devices generate images that contain some text on an ideally flat background. We need to take care that the parameter estimation procedure from Fig. 6.18 is applied to the actual image part of the display only and not to the text.
Secondly, we identify and resolve the limiting cases which do not permit a reliable estimation of the involved likelihood ratios:

- *Case 1*: the estimated mask from Fig. 6.18 contains too few *edge* labels. This may happen when the noise level is extremely high. In this case there are not enough data points to fit the log-likelihood ratios. We zero all the coefficients in the corresponding sub band.

- *Case 2*: the empirically estimated density $p_{W_k|X_k}(w|0)$ (and $p_{Z_k|X_k}(z|0)$) shows a delta-pulse at zero. This may happen when the noise level is extremely low and the image background relatively flat. The log-likelihood ratios cannot be estimated (due to divisions by zero). We leave all the coefficients in the corresponding subband intact.

With the above adjustments the resulting denoising method yields a speckle-reduced image which not only preserves edges but often enhances features. Fig. 6.19 shows the result of our adapted GenLik filtering for threshold values $T = 1, \ldots, 9$.

6.4.2 Psycho-visual experiment

Stimuli. US images of four different preterm brain pathologies were included in the experiment: two images contained well pronounced pathologies or structures (a subcortical infarct in a festering meningitis and a cerebral hemisphere of a preterm) and two images contained less pronounced PVL flaring. The

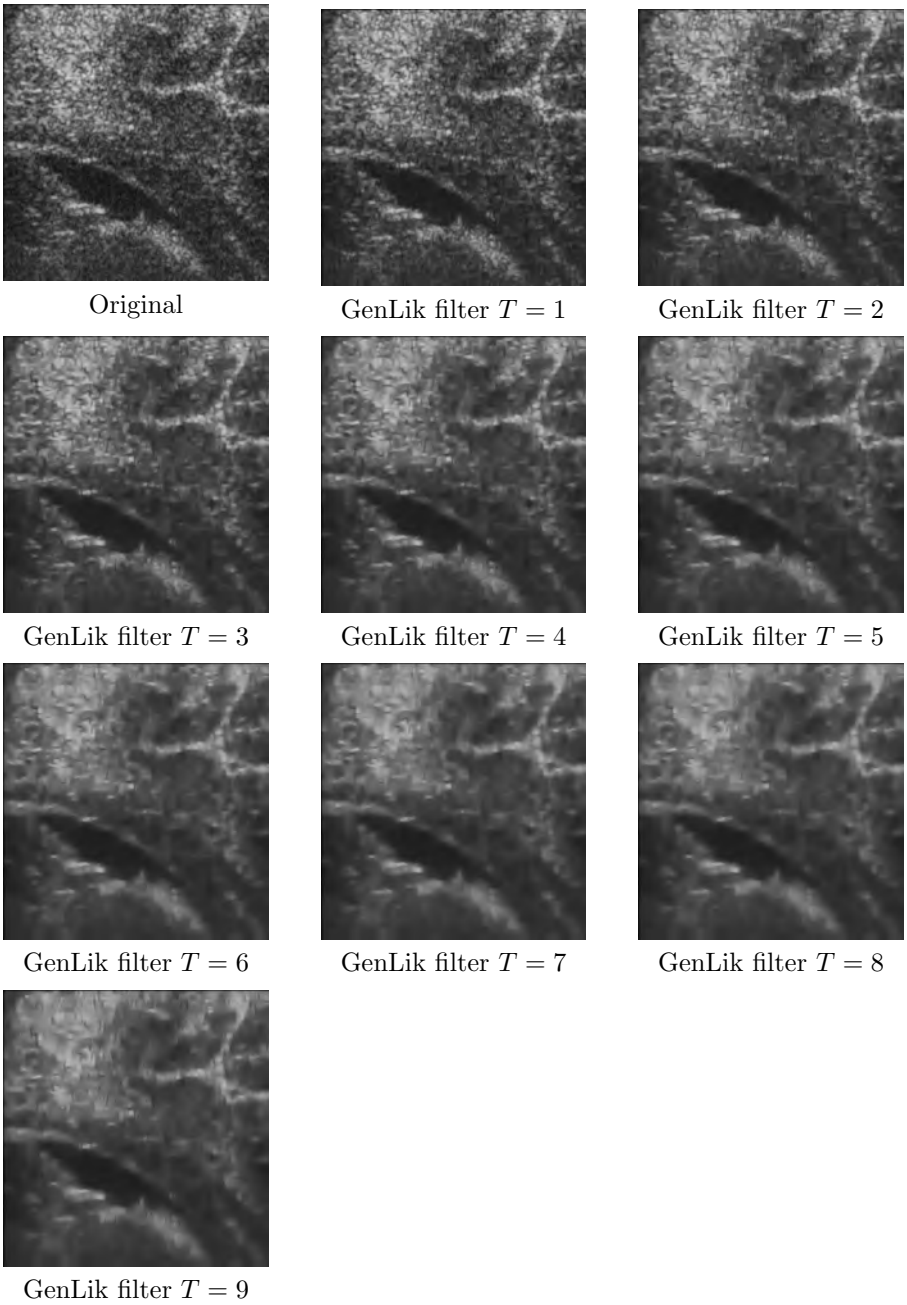


Figure 6.19: An original US image is shown (upper left) together with 9 speckle-reduced versions for varying settings of T.

choice for both well pronounced and less pronounced images was to see if this would result in a different scoring behavior.

All four scenes were filtered using different T-values of the modified GenLik filter. Initially, images were filtered up to $T = 9$, as shown in Fig. 6.19. However, evaluating all combinations of filtered images for all four US images would make the experiment too time-consuming. As we wanted as many expert physicians as possible to take part in the experiment, we reduced the number of images in the following way.

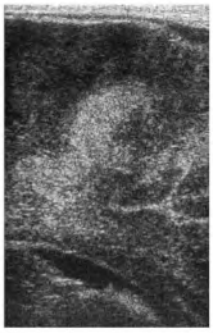
By examining the images in Fig. 6.19 from close by, we notice that for high values of T, the images are filtered strongly and reconstruction artifacts appear. Consequently, it is assumed that this kind of heavy filtering does not really enhance the image quality anymore. Therefore, we restricted ourselves to filtering $T \leq 4$. Fig. 6.20, 6.21, 6.22 and 6.23 show both the original US images as their filtered versions. In that way, in total 20 different images were included in this experiment.

Methodology. Since US images are of a poor visual quality, it is more difficult to define specific individual attributes to score than in the first experiment. Just to name one problem, how do we define what is (visual) noise or artifact when in fact all information in the US images is shown in the form of speckle. As such, instead of scoring attributes, we restricted ourselves to only measuring the preference in overall diagnostic quality. Although this approach deprives us from correlating diagnostic quality to specific attributes, a follow-up or feedback experiment similar to the third session in the previous experiment should provide us with that kind of information.

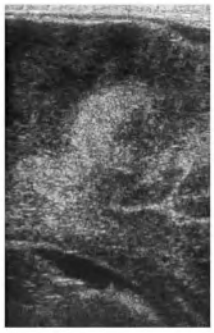
In addition to the uncertainty of individual attributes actually producing valuable results, another important consideration is the duration of the experiment. By only scoring for preferences, the experiment could be done in 10 minutes per subject. Adding dissimilarity and attribute scores would easily lead to an experiment that would have taken more than 30 minutes up to one hour. Our physicians are unlikely to be willing to spend that time. Therefore, we restricted us to information strictly necessary to draw significant conclusions.

As such, in practice all unique combinations of image couples were presented to the physicians through a web-interface and preference scores were asked for diagnostic quality as an integer on a scale ranging from -3 to +3. As in the first experiment, -3 corresponds to the strongest preference for the left hand image, +3 to the greatest preference for the right hand image, 0 to no preference. We opted for a web-interface instead of a completely calibrated display because this allowed the experts to perform the experiment remotely.

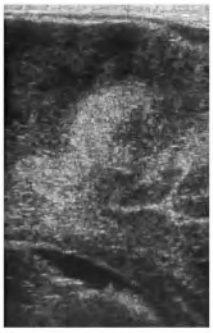
Subjects. Seven expert physicians took part in the experiment, based in two different medical institutions. Six experts are working at the neonatology department of the Erasmus Medical Center, Rotterdam, one expert is working at the neonatology department of the University Hospital, Ghent. All physicians were familiar with the images and pathologies shown.



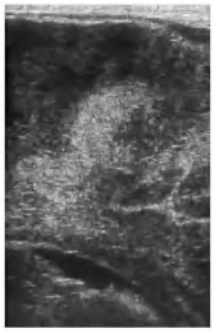
(1) original image



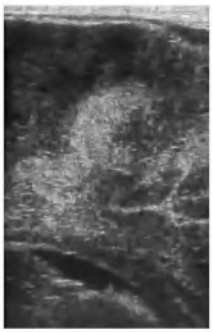
(2) GenLik filter $T = 1$



(3) GenLik filter $T = 2$

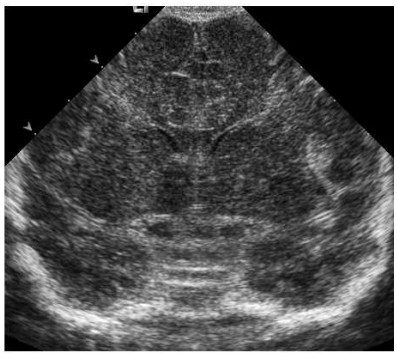


(4) GenLik filter $T = 3$

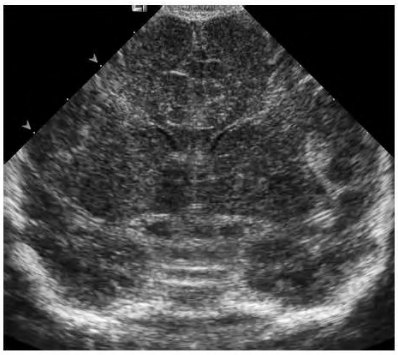


(5) GenLik filter $T = 4$

Figure 6.20: Original US image showing a subcortical infarct (white brush) in a festering meningitis together with 4 filtered versions.

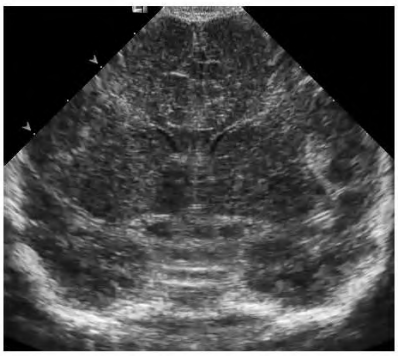


(1) original image

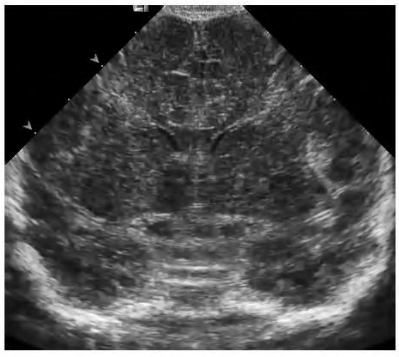


(2) GenLik filter $T = 1$

(3) GenLik filter $T = 2$



(4) GenLik filter $T = 3$



(5) GenLik filter $T = 4$

Figure 6.21: Original US images showing mild gPVL flaring and 4 filtered versions.

(1) original image

(2) GenLik filter $T = 1$

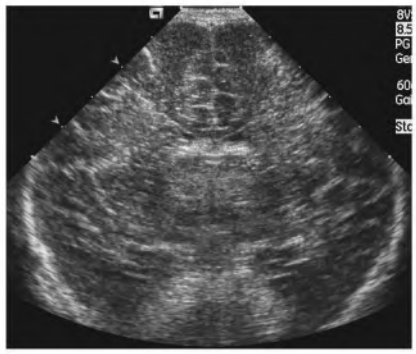
(3) GenLik filter $T = 2$

(4) GenLik filter $T = 3$

(5) GenLik filter $T = 4$

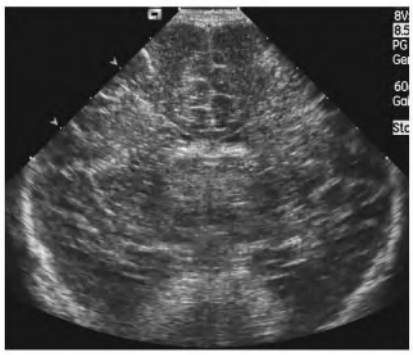
Figure 6.22: Original US image showing the preterm cerebral hemisphere and 4 filtered versions.

(1) original image

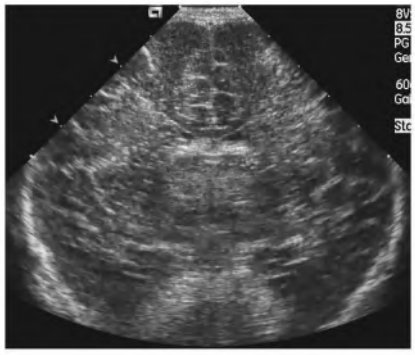


(2) GenLik filter $T = 1$

(3) GenLik filter $T = 2$



(4) GenLik filter $T = 3$



(5) GenLik filter $T = 4$

Figure 6.23: Original US images showing mild gPVL flaring and 4 filtered versions.

6.4.3 Results

Fig. 6.24 shows the ranking of the filtered images along the axis of impairment (in diagnostic quality). The plots should be read as follows: the higher the value on the Y-axis, the lesser the images are perceived suitable for visual diagnosis. We again notice a similar tendency along all 4 scenes. The original images are judged as being most suited and the perceived quality seems to drop as T increases.

From the follow-up experiment, where all physicians described in words why they preferred the non-filtered images, we could conclude there are two major reasons, shared by all physicians, for the poor diagnostic value of the filtered images. First of all, the filtered images are found to contain less feature structure and detail information. Secondly, the filtered versions of the images are perceived as "unnatural". Unnatural mainly being too smooth or containing small artifacts.

6.4.4 Discussion

Although speckle-reduction is shown to be very useful within the image processing framework, the results of our experiment point in the direction of it not being suited to improve the diagnostic value of the US images. Although our modified GenLik algorithm tends to preserve and enhance edges while smoothing homogeneous regions, physicians perceive this as a loss of both detail and structural information.

Note that although this is a disappointing result from the image enhancement point of view, it supports our approach in Chapters 1 and 2 where we did not denoise any of our US images for texture characterization since we would risk the loss of relevant structural information. Others have also followed that line, yet upto now this was not proven quantitatively.

Note that one of the reasons physicians gave for the degraded quality concerned denoised images to appear as unnatural. To us, this pinpoints the core-problem of US image denoising. By training and expertise, physicians are accustomed to looking at the granular speckle images. Smoothing the entire image, reducing speckle and enhance edge-contrast, is not part of their daily diagnostic routine. Although in the end all physicians agree and ultimately reject any form of speckle-reduction in their diagnosis, 5 out of the 7 physicians declare that in most cases the main reason for preferring the original image over the other was because they were simply used to diagnose the non-filtered ones. Consequently, we can ask ourselves if speckle-reduction is a good idea at all in qualitative US diagnosis.

We believe the answer to that question is still yes. US denoising should not be useless all together and we see the following possible application. Instead of filtering the entire US image, at the risk of loosing too much structural information, we could allow physicians to select certain regions of interest in

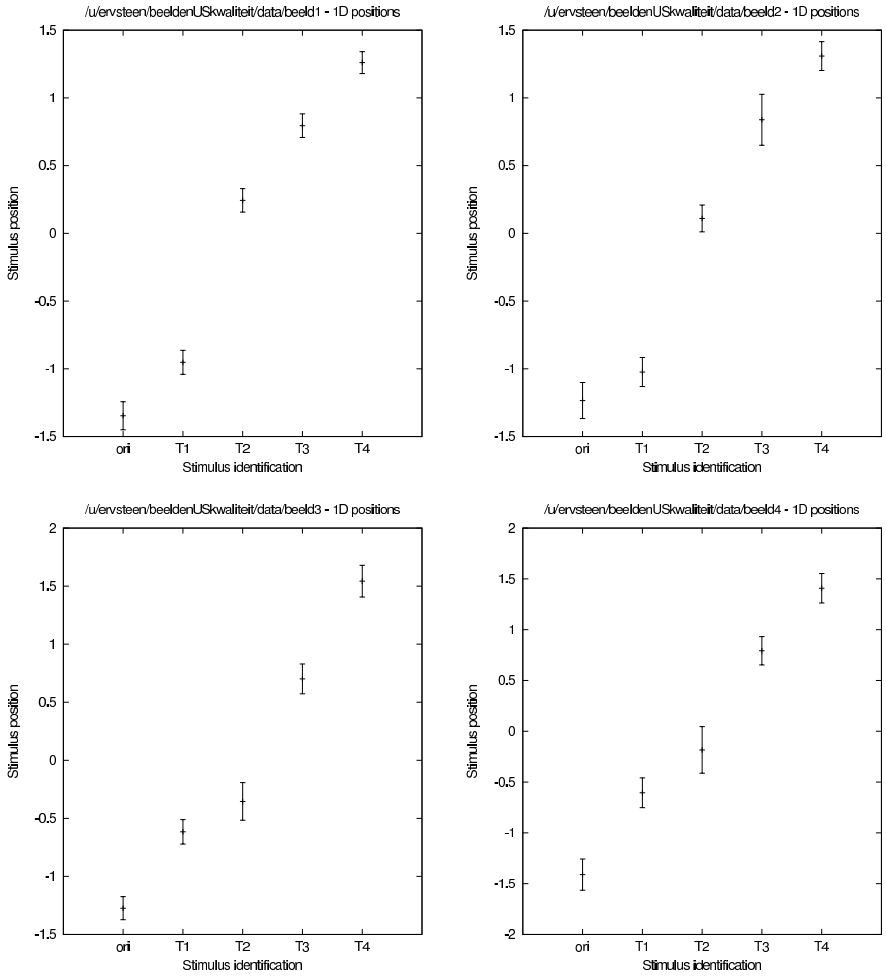


Figure 6.24: This figure shows the 1D-geometrical output of the MDS framework for the for images. On the X-axis the original and the 4 different filters, on the Y-axis the image impairment in terms of diagnostic quality. Upper left: festering meningitis. Upper right: gPVL flaring. Lower left: preterm hemisphere. Lower right: gPVL flaring.

the images, where the clinically relevant information is less pronounced. In those regions, we might then filter to the degree the physicians alter themselves in order to see if more or less (structural) information actually appears or not.

6.5 Conclusions and hints for future work

The aim of this Chapter was to present a psychophysical approach to image quality. We set up psycho-visual experiments to assess the overall image quality of state-of-the-art noise-reduction filters as well as the relevance of speckle-reduction in qualitative medical US diagnosis.

In the first experiment, we showed that using the MDS framework we achieved a filter ranking that is more reliable than the PSNR-ranking which is nowadays considered as the benchmark. Therefore, we conclude that when image distortions become rather small, PSNR should not be the (only) decisive criterion for image quality.

We integrated fuzzy similarity measures as an alternative to PSNR, yet a broad class of similarity measures is still left unstudied. This is what we consider the main point of interest for any future research. In addition, we could also investigate how our results in the MDS framework corresponds to other widely used and accepted psycho-visual models such as the National Telecommunication and Information Administration Video Quality Metric (NTIA VQM) or the Video Quality Expert Group (VQEG).

In a second experiment we showed that, contrary to preprocessing for quantitative image analysis, our modified GenLik speckle-reduction is unsuited for diagnostic purposes. Although our filter enhances edges and reduces speckle in homogeneous regions, it degrades the structural quality of the images and leads to unnatural images. Although the perception of loss of structural information is an agreement with the hypothesis of valuable information being present in the entire speckle pattern, we should also note that part of this qualitative judgment is inspired by the habit of physicians looking at the granular images.

Of course our GenLik filter is not necessarily the golden standard nor the only speckle-reduction filter available and consequently others could be tested too. However, given the reasons for the rejection of our filter we do not believe different results are to be expected for other filters.

Finally, although this US result might appear predominantly negative, we believe it teaches us two lessons. First of all, if the power of speckle-reduction filters is to be measured this should merely be done through how they influence, e.g., segmentation or registration results rather than on how they improve the overall image quality. Secondly, that there is no such thing as noise in the qualitative inspection of US images. Physicians are accustomed to look at *all* information present in the image, even if they don't necessarily use it for their diagnosis.

Chapter 7

Conclusions

To determine the degree of brain injury depends on many aspects of biomedicine. Clinical information surfaces in an erratic manner and data must be clustered in a quantitative way before meaningful diagnostic probabilities can emerge and treatment can be targeted to them. This thesis dealt with the currently most widely trusted tool of measuring preterm brain injury, i.e., US imaging.

With the study of Periventricular Leukomalacia as the main theme throughout this work, we developed several quantitative tools to assist physicians in a computer-aided diagnosis. The demand for these tools sprung from both the difficulty and subjectivity of a qualitative diagnosis in the case of less pronounced pathologies.

After presenting the basic physics of US imaging in Chapter 2, in Chapter 3 we presented a quantitative description of pathological white brain matter in VLBW infants. A comparative study of 7 different texture feature sets and 3 classifiers resulted in a tissue pattern recognition algorithm that quantifies the texture differences between pathological and non-pathological tissue. We showed how to construct an algorithm that classifies periventricular tissues with an accuracy above 90% and a sensitivity of 88%, outperforming all current qualitative descriptions. Besides that, the classifier has a relatively low-complexity and shows good generalization properties.

In Chapter 4 we continued this quantitative tissue characterization and showed how mathematical morphology combined with texture information and image thresholding can be applied to US segmentation. Primarily, a segmentation algorithm was developed to determine to which extent pathological white matter tissue spreads out. This algorithm started from the texture characterization of pathological tissue and incorporated quantified information on the echolucency of the choroid plexus. Combinations of morphological operations resulted in both an area estimate and a contour delineation of the pathological flaring regions.

The main result of Chapter 4 was that by this method we are now able to describe pathological flaring to its full extent. In a blind experiment, we showed that by measuring the areas of pathological white matter regions, we succeeded in improving the sensitivity for the recognition of pathological flaring. Besides that, we compared our method to both (constructed) ground truth information, obtained by averaging manual expert delineations, and a state-of-the-art active contour technique.

Although our method corresponds better to both ground truth information and the state-of-the-art flaring segmentation technique, we mentioned that we had to be cautious in interpreting these results. We namely also showed that our delineation corresponds better to the group of individual expert delineations than the expert delineations within the group correspond to themselves. This indicates that there is no clear agreement on ground truth information. Although we didn't really need this (ground truth) information to prove the power of our method, it was worthwhile to pursue this problem from another point of view, namely through multimodal image analysis.

This brings us to the main result of Chapter 5 where we presented a first interactive 2D US to 3D MRI registration scheme as a cross-validation for the characterization of affected periventricular white brain matter. With a limited user-interaction we succeeded in registering 79% of our test images correctly, the main cause of failing registrations being a sub-optimal image acquisition. The quality of our registrations was comparable to the manual expert registration, yet requires much less interaction. We are aware that the method has been designed in a very application-specific way, yet registering US to any other modality is very difficult. Consequently, specific choices were made in view of all information at hand and generalizations of this method to other US/MRI problem are to be investigated.

By comparing our segmentation results to the MR images, we concluded the following. In some cases, flaring seems to be visible on US where it is not on MRI. This is the first published evidence for a conjecture made by some neonatalogists, namely that although MRI is indisputably the golden standard on PVL at term and on later ages, US imaging is an important diagnostic tool in the early (first days of life) detection of PVL. Apart from the problem of prevalence, we also noticed that flaring in the US images seems to be spread out further than in the MR images. This suggests that structural tissue differences related to PVL are better manifested in the US images than in the MR images.

Finally, in Chapter 6 we presented two experiments on perceptual image quality. Although one of those was not directly related to US, both experiments followed the line of qualitative versus quantitative image processing. We presented a psychophysical approach to image quality by setting up experiments to assess the overall image quality of state-of-the-art noise-reduction filters as well as to investigate the relevance of speckle-reduction in qualitative medical US diagnosis. We showed that using the MDS framework we achieved a filter ranking that is more reliable than the PSNR-ranking which is nowadays

considered as the benchmark. Therefore, we concluded that when image distortions become small, PSNR should not be the (only) decisive criterion for image quality.

In a second experiment we were the first to prove that, contrary to preprocessing for quantitative image analysis, speckle-reduction is unsuited for diagnostic purposes. Although our modified GenLik filter enhances edges and reduces speckle in homogeneous regions, which is very suited for segmentation and registration purposes, it degrades the perceived structural quality and results in images judged as unnatural. Although the perception of loss of structural information is an agreement with the hypothesis of valuable information being present in the entire speckle pattern, we should also note that this qualitative judgment is inspired by the habit of physicians looking at the granular images.

Although this last result might appear predominantly negative, we concluded two things from it. First of all, if the power of speckle-reduction filters is to be measured it should be done through quantitative measurement, e.g., on segmentation or registration results, rather than on how they improve the overall image quality. Secondly, that there is no such thing as noise in the qualitative inspection of US images. Physicians are accustomed to look at *all* information present in the image, even if they do not necessarily use it in their diagnosis.

Overall, the power of this thesis lies in the application and clinical validation of quantitative image processing algorithms in the field of preterm US imaging. We have provided physicians with actual tools and took the classification, segmentation and registration algorithms to a level of validation that lies beyond what is common in this field. Our psycho-visual experiments also resulted in some critical notes on widespread concepts as PSNR, perceived image quality, and the effect of speckle-reduction on qualitative diagnosis.

The limitations of the algorithms proposed in this work are that all are developed for specific US machines. Also, given that this is one of the few quantitative studies on preterm US brain images, both cross-validation and the definition of ground truth information is not straightforward. Related to that, to fully evaluate the clinical significance of the algorithms, we need the follow-up of all patients over a period of at least 5 years as well as more accurate information on MRI data.

Consequently, to us the main research goal for the future is a continued multimodal investigation. It is unlikely that one imaging modality or one algorithm will suffice in the case of these difficult pathologies. The advent of more-performing imaging modalities results in a growing gamut of potentially complementary diagnostic data. In the case of US brain imaging the fusion of EEG, ECG, MRI, DTI and US data through segmentation, classification, fiber tracking and registration should result in an even higher-level quantitative description of the preterm brain.
To achieve this, in the near future the close collaboration between engineers and mathematicians on one side and physicians on the other, will become increas-

ingly important. Nowadays, in major institutions image processing groups are already embedded in the clinical environments, resulting in both quantitative (semi)-automated computer-aided diagnostic tools presented to the physicians in their natural environment and the necessary clinical feedback presented to the researchers in a fast and accurate way.

To summarize, this research has resulted in four main contributions: three CAD algorithms, all of which have been clinically validated, and one psycho-visual experiment to assist physicians in a more objective, quantitative analysis of preterm US brain images:

1. A classification algorithm to characterize pathological white brain matter based on texture features [Vansteenkiste et al., 2007a].
2. A flaring segmentation algorithm combining texture information and morphological operators to delineate flaring boundaries and estimate flaring areas [Vansteenkiste et al., 2007b].
3. A registration algorithm to align two-dimensional US images and three-dimensional MRI volumes [Vansteenkiste et al., 2006f].
4. A modification of an existing speckle-reduction filter and a psycho-visual experiment to assess the effect of speckle-reduction on the diagnostic ultrasound image quality [Pizurica et al., 2006].

Besides these, this research also lead to the following algorithms (related to the ones above):

1. An extension of the flaring segmentation algorithm to both a three-dimensional brain ventricle segmentation algorithm and a carotid artery segmentation algorithm.
2. An extension of the flaring segmentation algorithm for textiles that show a texture pattern similar to speckle [Morent et al., 2006].
3. A second psycho-visual experiment to investigate the quality of 7 state-of-the-art noise-reduction filters on non-medical images [Vansteenkiste et al., 2006c], [Vansteenkiste et al., 2006b].

This work resulted in 6 A1-publications [Vansteenkiste et al., 2006f], [Vansteenkiste et al., 2006c], [Pizurica et al., 2006], [Morent et al., 2006], [Vansteenkiste et al., 2006b], [Vansteenkiste et al., 2007a]. One A1-paper is currently still in review [Vansteenkiste et al., 2007b]. 14 papers were published in international conferences [Vansteenkiste et al., 2003a], [Huysmans et al., 2004a], [Zlokolica et al., 2006], [Vansteenkiste et al., 2005a], [Vansteenkiste et al., 2006e],[Huysmans et al., 2004b], [Conneman et al., 2005], [Vansteenkiste et al., 2005c],
[Vansteenkiste et al., 2006a], [Vansteenkiste et al., 2004b],

[Vansteenkiste et al., 2004a], [Vansteenkiste et al., 2005b], [Vandemeulebroucke et al., 2006] and 4 papers were presented in national conferences and symposia [Vansteenkiste et al., 2002], [Vansteenkiste et al., 2003b], [Vansteenkiste et al., 2006d], [Govaert et al., 2006].

Bibliography

[Abdel-Dayem et al., 2005] Abdel-Dayem, A., El-Sakka, M., and Fenster, A. (2005). Watershed segmentation for carotid artery ultrasound images. In *The 3rd ACS/IEEE International Conference on Computer Systems and Applications*, page 131, Cairo, Egypt.

[Abdel-Elmoniem et al., 2002] Abdel-Elmoniem, K., Youssef, A.-B. M., and Kadah, Y. (2002). Realtime speckle reduction and coherence enhancement in ultrasound imaging via nonlinear anisotropic diffusion. *IEEE Transactions on Biomedical Engineering*, 49(9):997–1014.

[Achim et al., 2001] Achim, A., Bezerianos, A., and Tsakalides, P. (2001). Novel bayesians multiscale method for speckle removal in medical ultrasound images. *Transactions on Medical Imaging*, 20(8):772–783.

[Amelung, 1995] Amelung, J. (1995). *Automatische Bildverarbeitung für die Qualitätssicherung*. PhD thesis, Technische Hochschule Darmstadt.

[Amin et al., 2003] Amin, D., Kanade, T., DiGioia, A., and Jaramaz, B. (2003). Ultrasound registration of the bone surface for surgical navigation. *Comput Aided Surgery*.

[Anuja et al., 2002] Anuja, N., Barry, D., Kuban, E., Tuzcu, M., Schoenhagen, P., Nissen, S., and Vince, D. (2002). Coronary plaque classification with intravascular ultrasound radiofrequency data analysis. *Circulation*.

[Avi-Itzhak et al., 1995] Avi-Itzhak, H., Diep, T., and Garland, H. (1995). High accuracy optical character recognition using neural networks with centroid dithering. *IEEE Transactions on Pattern Analysis and Machine Intellingence*, 17(2):218–224.

[Avianto and Ito, 2001] Avianto, E. and Ito, M. (2001). Speckle reduction for ultrasonic imaging using fuzzy morphology. In *IEICE Trans. Inf. Systems*, volume 84, pages 502–510.

[Babcock and Ball, 1983] Babcock, D. and Ball, W. (1983). Postasphyxial encephalopathy in full-term infants: Ultrasound diagnosis. *Pediatric Radiology*, 148(2):417–423.

[Basset et al., 1993] Basset, ., Sun, Z., Mestas, J., and Gimenez, G. (1993). Texture analysis of ultrasonic images of the prostate by means of co-occurrence matrices. *Journal of Ultrasonic Imaging*, 9:218–237.

[Bins, 2000] Bins, J. (2000). *Feature selection from huge feature sets in the context of computer vision*. PhD thesis, Colorado State University.

[Bland and Altman, 1986] Bland, J. and Altman, D. (1986). Statistical methods for assessing the agreement between two clinical measurements. *Lancet*, 1:307–310.

[Chalana and Ki, 1997] Chalana, V. and Ki, Y. (1997). A methodology for evaluation of boundary detection algorithms on medical images. *IEEE Transactions on Medical Imaging*, 16(5):642–652.

[Chen et al., 1995] Chen, S., Yeh, M., and Hsiao, P. (1995). A comparison of similarity measures of fuzzy values. *Fuzzy Sets and Systems*, 72(1):79–89.

[Chen and Abolmaesumi, 2005] Chen, T. and Abolmaesumi, P. (2005). A mutual information based registration algorithm for ultrasound-guided computer-assisted orthopaedic surgery. In *Proceedings of SPIE Medical Imaging*.

[Childs et al., 2001] Childs, A., Cornette, L., and Ramenghi, L. e. a. (2001). Magnetic resonance and cranial ultrasound characteristics of periventricular white matter abnormalities in newborn infants. *Clinical Radiology*, 56(8):647–655.

[Christodoulou et al., 2003] Christodoulou, C., Pattichis, C., Pantziaris, M., and Nicolaides, A. (2003). Texture-based classification of atherosclerotic carotid plaques. *IEEE Transactions on Medical Imaging*, 22(7):902–912.

[Coggins, 1982] Coggins, J. (1982). *A framework of texture analysis based on Spatial Filtering*. PhD thesis, Michigan State University, East Lansing, Michigan.

[Conneman et al., 2005] Conneman, N., Vansteenkiste, E., Lequin, M., and Govaert, P. (2005). Sonographic segmentation of preterm white matter damage. In *Proceedings of the Radiology Society of Nothern America Conference*, Chicago, IL, USA.

[Costello et al., 1988] Costello, A., Hamilton, P., Baudin, J., Townsend, J., Bradford, B., Stewart, A., and Reynolds, E. (1988). Prediction of neurodevelopmental impairment at four years from brain ultrasound appearance of very preterm infants. *Dev Med Child Neurology*, 30:711–722.

[Counsell et al., 2003] Counsell, S., Rutherford, M., Cowan, F., and Edwards, A. (2003). Magnetic resonance imaging of preterm brain injury. *Arch Dis Child Fetal Neonatal*, 88:269–274.

[Dabov et al., 2006] Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. (2006). Image denoising with block-matching and 3D filtering. In *Proc. SPIE Electronic Imaging: Algorithms and Systems V*, volume 6064, San Jose, CA, USA.

[Dammann and Leviton, 1997] Dammann, O. and Leviton, A. (1997). Duration of transient hyperechoic images of white matter in very-low-birthweight infants: a proposed classification. *Developmental Medicine and Child Neurology*, 39:2–5.

[Davis, 1997] Davis, D. (1997). Review of cerebral palsy. *Neonatal Network*, 16(3):7–12.

[De Backer, 2002] De Backer, S. (2002). *unsupervised pattern recognition: dimensionality reduction and classification*. PhD thesis, Visielab, University of Antwerp.

[De Bruin et al., 2002] De Bruin, P., Vos, F., Post, F., Vossepoel, A., and de Block, S. (2002). Interactive matching of ultrasound and mri for visualization during resection of myomata. In *Proceedings of SPIE Medical Imaging*, volume 4081.

[De Vries et al., 1992] De Vries, L., Eken, P., and Dubowitz, L. (1992). The spectrum of leukomalacia using cranial ultrasound. *Behavioural Brain Research*, 49:1–6.

[De Vries et al., 1993] De Vries, L., Eken, P., Groenedaal, F., van Haastert, I., and Meiners, L. (1993). Correlation between the degree of periventricular leukomalacia diagnosed using cranial ultrasound and mri later in infancy and children with cerebral palsy. *Neuropediatrics*, 24(5):263–268.

[De Vries et al., 1988] De Vries, L., Regev, R., Pennock, J., Wigglesworth, J., and Dubowitz, L. (1988). Ultrasound evolution and later outcome of infants with periventricular densities. *Early Human Development*, 16:225–233.

[DiPietro et al., 1986] DiPietro, M., Brody, B., and Teele, R. (1986). Peritrigonal echogenic "blush" on cranial sonography: pathologic correlates. *AJR Am J Roentgenology*, 146:1067–1072.

[Dussik, 1942] Dussik, K. (1942). On the possibility of using ultrasonic waves as a diagnostic aid. *Neurologic Psychiatr*, pages 153–168.

[Dutt, 1995] Dutt, V. (1995). *Statistical Analysis Of Ultrasound Echo Envelope*. PhD thesis, Mayo Graduate School, Rochester.

[Evans and Nixon, 1996] Evans, A. and Nixon, M. (1996). Biased motion-adaptive temporal filtering for speckle reduction in echocardiography. *IEEE Transaction on Biomedical Imaging*, 15(1):39–50.

[Fawer et al., 1985] Fawer, C.-L., Calame, A., Perentes, E., and Anderegg, A. (1985). Periventricular leukomalacia: a correlation study between real time ultrasound and autopsy findings. *Neuroradiology*, 27:292–230.

[Finette et al., 1983] Finette, S., Bleier, A., and Swindel, W. (1983). Breast tissue classification using diagnostic ultrasound and pattern recognition techniques: I. methods for pattern recognition. *Ultrasonic Imaging*, 5:55–70.

[Foi et al., 2006] Foi, A., Dabov, K., Katkovnik, V., and Egiazarian, K. (2006). Shape-adaptive dct for denoising and image reconstruction. In *Proceedings of the SPIE Electronic Imaging 2006, Image Processing: Algorithms and Systems V*, pages 203–214, San Diego, CA, USA.

[Frost et al., 1982] Frost, V., Stiles, J., Shanmugan, K., and Holtzman, J. (1982). A model for radar images and its applications to adaptive digital filtering and multiplicative noise. *IEEE Transactions on Pattern Analysis and Machine Intellingence*, 4(1):57–166.

[Glor, 2004] Glor, F. (2004). *Integrating Medical Imaging and computational fluid dynamics for measuring blood flow in carotid arteries.* PhD thesis, Ghent University.

[Govaert and De Vries, 1995] Govaert, P. and De Vries, L. (1995). *Geluiden uit de hersenen van de pasgeborene (in Dutch).* Saint-Luc, Nazareth, Belgium.

[Govaert et al., 2006] Govaert, P., Vansteenkiste, E., and Philips, W. (2006). Quantification of brain injury with in vivo imaging. In *IEEE/EMBS Benelux Symposium*, pages 16–19.

[Guerrero-Colon and Portilla, 2005] Guerrero-Colon, J. and Portilla, J. (2005). Two-level adaptive denoising using gaussian scale mixtures in overcomplete oriented pyramids. In *Proceedings of IEEE International Conference on Image Processing*, pages 105–108, Genova, Italy.

[Gupta et al., 2004] Gupta, S., Chauhan, R., and Sexana, S. (2004). Wavelet-based statistical approach for speckle reduction in medical ultrasound images. In *MBEC*, volume 42, pages 189–192.

[Haralick, 1979] Haralick, R. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 86:786–804.

[Haralick et al., 1976] Haralick, R., Shanmugam, K., and Dinstein, I. (1976). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621.

[Haris et al., 1998a] Haris, K., Efstratiadis, S., and Maglaveras, N. (1998a). Hybrid image segmentation using watersheds and fast region merging. *IEEE transactions on image processing*, 7(12):1684–1699.

[Haris et al., 1998b] Haris, K., Efstratiadis, S., and Maglaveras, N. (1998b). Watershed-based image segmentation with fast region mergin. In *IEEE International Conference on Image Processing*, volume 3, pages 338–342.

[Hawkins, 1969] Hawkins, J. (1969). *Picture Processing and Psychopictorics*. Academic Press, New York.

[Hernandez and Barner, 2005] Hernandez, S. and Barner, K.E. Yuan, Y. (2005). Region merging using homogeneity and edge integrity for watershed-based image segmentation. *Journal of Optical Engineering*, 44(1):14.

[Hope et al., 1988] Hope, P., Gould, S., Howard, S., Hamilton, P., Costello, A., and Reynolds, E. (1988). Precision of ultrasound diagnosis of pathologically verified lesions in the brains of very preterm infants. *Developmental Medicine and Child Neurology*, 30:457–471.

[Hope et al., 2004] Hope, T., Gregson, P., Linney, N., and Schmidt, M. (2004). Ultrasonic tissue characterization as a predictor of white matter damage: results of a preliminary study. In *IASTED International Conference on Biomedical Engineering*, volume 3, pages 2157–2160.

[Horsch et al., 2005] Horsch, S., Muentjes, C., Franz, A., and Roll, C. (2005). Ultrasound diagnosis of brain atrophy is related to neurodevelopmental outcome in preterm infants. *Acta Paediatr*, 94(12):1815–1821.

[Huang and Chen, 2004] Huang, Y. and Chen, D. (2004). Watershed segmentation for breast tumor in 2d-sonography. *Ultrasound in Medicine and Biology*, 30(5):625–632.

[Hughes, 2001] Hughes, S. (2001). Medical ultrasound imaging. *Physics Education*, pages 468–475.

[Huynen et al., 1994] Huynen, A., Giessen, R., de la Rosette, J., Aamink, R., Debruyne, F., and Wijkstra, H. (1994). Analysis of ultrasonic prostate images for the detection of prostatic carcinoma: the automated urologic diagnostic expert system. *Ultrasound in Medicine and Biology*, 20(1):1–10.

[Huysmans et al., 2004a] Huysmans, B., Vansteenkiste, E., Govaert, P., and Philips, W. (2004a). An evaluation of texture classifiers for the detection of periventricular leukomalacia. In *Proceedings of the IEE Medical Signal and Information Processing Conference*, pages 201–206, Sliema, Malta.

[Huysmans et al., 2004b] Huysmans, B., Vansteenkiste, E., and Philips, E. (2004b). A comparative study of white matter damage. In *Proceedings of the Signal Processing Symposium*, pages 61–64, Hilvarenbeek, The Netherlands.

[Ibanez et al., 2005] Ibanez, L., Schroeder, W., Ng, L., and Cates, J. (2005). The itk software guide, second edition. Technical report, Insight Software Consortium.

[Inder et al., 2003] Inder, T., Anderson, N., Spencer, C., Wells, S., and Volpe, J. (2003). White matter injury in the premature infant: A comparison between serial cranial sonographic and mr findings at term. *American Journal of Neuroradiology*, 24(5):805–809.

[Ishii et al., 2003] Ishii, H., Tsuruoka, S., Ibrahimy, M., Kimura, F., Wakabayashi, T., Ohyama, W., and Sekioka, K. (2003). Tissue characterization of local myocardium using phase frequency spectrum of ultrasonic rf-signal. In *IEEE EMBS Asian-Pacific Conference on Biomedical Engineering*, pages 142–143.

[Jain and Tuceryan, 1998] Jain, A. and Tuceryan, M. (1998). *Handbook of Pattern Recognition and Computer Vision*, chapter Texture Analysis. World Scientific Publishing Co.

[Jakeman and Pusey, 1976] Jakeman, E. and Pusey, P. (1976). A model for non-rayleigh sea echo. *IEEE Transactions on Antennas and Propagation*, 24:806–814.

[Jendoubi et al., 2004] Jendoubi, A., Zeng, J., and Chouikha, M. (2004). Segmentation of prostate ultrasound images using an improved snakes model. In *7th International Conference on Signal Processing*, volume 3, pages 2568–2571.

[Jeremias et al., 1999] Jeremias, A., Kolz, M., Ikonen, T., Gummert, J., Oshima, A., Hayase, M., Honda, M., Komiyama, D., Berry, G., Morris, R., Yock, P., and Fitzgerald, P. (1999). Feasibility of in vivo intravascular ultrasound tissue characterization in the detection of early vascular transplant rejection. *Circulation*.

[Kadah et al., 1996] Kadah, Y., Fara, A., Zurada, J., Badawi, A., and Youssef, A. (1996). Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images. *IEEE Transactions on Medical Imaging*, 15(4):466–478.

[Karaman et al., 1995] Karaman, M., Kutay, M., and Bozdagi, G. (1995). Adaptive speckle suppression filter for medical ultrasonic imaging. *IEEE Transactions on Medical Imaging*, 14(2):283–292.

[Kayagaddem and Martens, 1996] Kayagaddem, V. and Martens, J. (1996). Perceptual characterization of images degraded by blur and noise: experiments. *Journal of the Optical Sociecty of America*, 13(6):1178–1188.

[Keeney et al., 1991] Keeney, S., Adcock, E., and Mc Ardle, C. (1991). Prospective observations of 100 high-risk neonates by high-field (1.5 tesla) magnetic resonance imaging of the central nervous system: Ii. lesions associated with hypoxic-ischemic encephalopathy. *Pediatrics*, 87(4):431–438.

[Kuban, 2001] Kuban, K. e. a. (2001). Topography of cerebral white-matter disease of prematurity studied prospectively in 1607 very-low-birhtweight infants. *Journal of Child Neurology*, 16:401–408.

[Laub and Ingrisch, 1986] Laub, M. and Ingrisch, H. (1986). Increased periventricular echogenicity (periventricular halos) in neonatal brain: a sonographic study. *Neuropediatrics*, 17:39–43.

[Laws, 1980] Laws, K. (1980). Rapid texture identification. In *Proceedings of SPIE Image Processing for Missile Guidance*, volume 238, pages 376–380, San Diego, CA, USA.

[Ledda, 2006] Ledda, A. (2006). *Mathematical Morphology in Image Processing*. PhD thesis, Ghent University.

[Lee, 1980] Lee, J. (1980). Digital image enhancement and noise filtering by use of local statistics. *Pattern Analysis and Machine Intelligence*, 2(2):165–168.

[Levene et al., 1983] Levene, M., Wigglesworth, J., and Dubowitz, V. (1983). Haemorrhagic periventricular leukomalacia in the neonate: a real time ultrasound study. *Pediatrics*, 71:794–797.

[Maalouf et al., 2001] Maalouf, E., Duggan, P., and Counsell, S. e. a. (2001). Comparison of findings on cranial ultrasound and magnetic resonance imaging in preterm infants. *Pediatrics*, 107(4):719–727.

[Maintz and Viergever, 1996] Maintz, J. and Viergever, M. (1996). An overview of medical image registration methods.

[Mallat, 1998] Mallat, S. (1998). *A wavelet tour of signal processing*. Academic Press.

[Mann and Whitney, 1947] Mann, H. and Whitney, D. (1947). On a test of whether one of 2 random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

[Martens, 2003] Martens, J.-B. (2003). *Image Technology Design*, chapter Psychophysical measurement and modelling of image quality. Springer.

[Martin and Spinks, 2001] Martin, K. and Spinks, D. (2001). Measurement of the speed of sound in ethanol/water mixtures. *Ultrasound Med Biol*, 27(2):289–291.

[Matheron, 1975] Matheron, G. (1975). *Random sets and Integral Geometry*. John Wiley and Sons, Inc., New York.

[Mattes et al., 2001] Mattes, D., Haynor, D., Vesselle, H. Lewellen, T., and Eubank, W. (2001). Non-rigid multimodality image registration. In *Proceedings of SPIE Medical Imaging: Image Processing*, pages 1609–1620, San Diego, CA, USA.

[Mercier et al., 2005] Mercier, L., Lango, T., Lindseth, F., and Collins, D. (2005). A review of calibration techniques for freehand 3-D ultrasound systems. *Ultrasound Med Biol*, 31(4):449–471.

[Miller et al., 2003] Miller, S., Cozzio, C., Goldstein, R., Ferriero, D., and Partridge, J. e. a. (2003). Comparing the diagnosis of white matter injury in premature newborns with serial mr imaging and transfontanel ultrasonography findings. *American Journal of Neuroradiology*, 24:1661–1669.

[Mischi, 2004] Mischi, M. (2004). *Contrast Echocardiography for Cardiac Quantifications*. PhD thesis, Technical University Eindhoven.

[Mohamed et al., 2003] Mohamed, M., Abdel-galil, T., salama, M., El-saadany, E., Kamel, M., Fenster, A., Downey, D., and Rizkalla, K. (2003). Prostate cancer diagnosis based on gabor filter texture segmentation of ultrasound image. In *Canadian Conference on Electrical and Computer Engineering, IEEE CCECE*, volume 3, pages 1485–1488, Saskatchewan, Canada.

[Morent et al., 2006] Morent, R., De Geyter, N., Leys, C., Vansteenkiste, E., De Bock, J., and Philips, W. (2006). Measuring the wicking behavior of textiles by the combination of a horizontal wicking experiment and image processing. *Review of Scientific Instruments*, 77:093502–1/6.

[Mullaart et al., 1999] Mullaart, R., Thijssen, J., and Rotteveel, J. (1999). Quantitative ultrasonography of the periventricular white and grey matter of the developing brain. *Ultrasound in Medicine and Biology*, 25(4):527–530.

[Noble and Boukerroui, 2006] Noble, A. and Boukerroui (2006). Ultrasound image segmentation: A survey. *IEEE Transactions on Medical Imaging*, 25(8):987–1010.

[Paneth et al., 1994] Paneth, N., Nudelli, R., Kazam, E., and Monte, W. (1994). *Brain Damage in the preterm infant*. Mac Keith Press, London.

[Peng et al., 2003] Peng, H., Long, F., and Chi, Z. (2003). Document image recognition based on template matching of component block projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1188–1192.

[Pierrat et al., 2000] Pierrat, V., Duquennoy, C., van Haastert, I., Ernst, M., Guilley, N., and de Vries, L. (2000). Ultrasound diagnosis and neurodevelopmental outcome of localised and extensive cystic periventricular leukomalacia. *Arch Dis Child Fetal Neonatal*, 84:151–156.

[Pizurica, 2002] Pizurica, A. (2002). *Image Denoising Using Wavelets and Spatial Context Modeling*. PhD thesis, Ghent University.

[Pizurica and Philips, 2006] Pizurica, A. and Philips, W. (2006). Estimating probability of presence of a signal of interest in multiresolution single- and multiband image denoising. *IEEE Transactions on Image Processing*, 15(3):654–665.

[Pizurica et al., 2003] Pizurica, A., Philips, W., Lemahieu, I., and Acheroy, M. (2003). A versatile wavelet domain noise filtration technique for medical imaging. *IEEE Transactions on Medical Imaging*, 22(3):323–331.

[Pizurica et al., 2006] Pizurica, A., Wink, A., Vansteenkiste, E., Philips, W., and Roerdink, J. (2006). A review of wavelet denoising in mri and ultrasound brain imaging. *Current Medical Image Reviews*, 2(2):247–260.

[Portilla et al., 2003] Portilla, J., Strela, V., Wainwright, M., and Simoncelli, E. (2003). Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12(11):1338–1351.

[Prager et al., 1998a] Prager, R., Rohling, R., Gee, A., and Berman, L. (1998a). Automatic calibration for 3-d free-hand ultrasound. Technical report, Cambridge University Engineering Department.

[Prager et al., 1998b] Prager, R., Rohling, R., Gee, A., and Berman, L. (1998b). Rapid calibration for 3-d freehand ultrasound. *Ultrasound Med Biol*, 24(6):855–869.

[Richard and Keen, 1996] Richard, W. and Keen, G. (1996). Automated texture-based segmentation of ultrasound images of the prostate. *Comput. Med. Imaging. Graph.*, 20(3):131–140.

[Rooms, 2005] Rooms, F. (2005). *Nonlinear Methods in Image Restoration Applied to Confocal Microscopy*. PhD thesis, Ghent University.

[Sattar et al., 1997] Sattar, F., Floreby, L., Salomonnson, G., and Lövström, B. (1997). Image enhancement based on nonlinear multi-scale method. *IEEE Transactions on Image Processing*, 6:888–895.

[Schmitz et al., 1994] Schmitz, G., Ermert, H., and Senge, T. (1994). Tissue characterization of the prostate using kohonen maps. In *Proceedings of Ultrasonics Symposium*, pages 1487–1490.

[Schouman-Claeys et al., 1993] Schouman-Claeys, E., Henry-Feugeas, M., and Roset, F. e. a. (1993). Periventricular leukomalacia: correlation between mr imaging and autopsy findings during first 2 months of life. *Radiology*, 189:59–64.

[Sendur and Selesnick, 2002] Sendur, L. and Selesnick, I. (2002). Bivariate shrinkage with local variance estimation. *IEEE Signal Processing Letters*, 9(12):438–441.

[Seongjai and L., 2005] Seongjai, K. and L., H. (2005). A hybrid level set segmentation for medical imagery. In *IEEE Nuclear Science Symposium Conference Record*, volume 3.

[Serra, 1982] Serra, J. (1982). *Image Analysis and Mathematical Morphology*, volume 1. Academic Press, New York.

[Serra, 1988] Serra, J. (1988). *Image Analysis and Mathematical Morphology: Theoretical Advances*, volume 2. Acedemic Press, New York.

[Shekhar and Zagrodsky, 2002] Shekhar, R. and Zagrodsky, V. (2002). Mutual information-based rigid and nonrigid registration of ultrasound volumes. *IEEE Transactions on medical imaging*, 21(1):9–22.

[Sie et al., 2000] Sie, L., van der Knapp, M., van Wezel-Meijler, G., Taets van Amerongen, A., Lafeber, H., and Valk, J. (2000). Early mr features of hypoxic-ischemic brain injury in neonates with periventricular densities on sonograms. *American Journal of Neuroradiology*, 21(5):852–86.

[Simaeys et al., 2000] Simaeys, B., Philips, W., Lemahieu, I., and Govaert, P. (2000). Quantitative analysis of the neonatal brain by ultrasound. *Computerized Medical Imaging and Graphics*, 24(1):11–18.

[Sklansky, 1978] Sklansky, J. (1978). Image segmentation and feature extraction. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 237–247.

[Skranes et al., 1998] Skranes, J., Nilsen, G., Smevik, O., Vik, T., and Brubakk, A. (1998). Cerebral mri of very low birth weight children at 6 years of age compared with the findings at 1 year. *Pediatr. Radiol.*, 28(6):471–475.

[Sonotech, 2003] Sonotech (2003). The technology of ultrasound scanning gels. Technical report, Sonotech.

[Stippel, 2004] Stippel, G. (2004). *Speckle Suppression, Segmentation and Registration of Medical Ultrasound Images*. PhD thesis, Ghent University, Belgium.

[Stippel et al., 2001] Stippel, G., Duskunovic, I., Philips, W., Lemahieu, I., Zecic, A., and Govaert, P. (2001). Segmenting flares in ultrasound images using prior statistics. *Image Processing and Communications*, 7(1-2):41–54.

[Sun et al., 1996] Sun, Y.-N., Hong, H.-M., Lin, X.-Z., and Wang, J.-Y. (1996). Ultrasonic image analysis for liver diagnosis. *IEEE Transactions in Medicine and Biology*, 5(4):619–634.

[Tamisari et al., 1986] Tamisari, L., Vigi, V., Fortini, C., and Scarpa, P. (1986). Neonatal periventricular leukomalacia: diagnosis and evolution evaluated by real-time ultrasound. *Helv Paediatr Acta*, 41:399–407.

[Tamura et al., 1978] Tamura, H., Mori, S., and Yamawaki, Y. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 460–473.

[Tauber et al., 2004] Tauber, C., Batatia, H., Morin, G., and Ayache, A. (2004). Robust b-spline snakes for ultrasound image segmentation. In *Computers in Cardiology*, pages 325–328, Chicago, IL, USA.

[Theodoridis and Koutroumbas, 1999] Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition.* Academic Press.

[Thijssen and Oosterveld, 1990] Thijssen, J. and Oosterveld, B. (1990). Texture in tissue echograms: Speckle or information. *Journal of Ultrasound Med*, 9:215–229.

[Townsend et al., 1999] Townsend, S., Rumack, C., Thilo, E., Merenstein, G., and Rosenberg, A. (1999). Late neurosonographic screening is important to the diagnosis of periventricular leukomalacia and ventricular enlargement in preterm infants. *Pediatr Radiol*, 29:347–352.

[Unser, 1986] Unser, M. (1986). Sum and difference histograms for texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1):118–125.

[Unser, 1999] Unser, M. (1999). Splines: A perfect fit for signal and image processing. *IEEE Signal processing magazine*, pages 22–38.

[Unser et al., 1993] Unser, M., Aldroubi, A., and Murray, E. (1993). B-spline signal processing: Part ii - efficient design and applications. *IEEE Transactions on Signal Processing*, 41(2).

[Valckx and Thijssen, 1997] Valckx, F. and Thijssen, J. (1997). Characterization of echographic image texture by cooccurrence matrix parameters. *Ultrasound in Medicine and Biology*, 23(4):559–671.

[Van De Ville et al., 2003] Van De Ville, D., Nachtegael, M., Van der Weken, D., Kerre, E., Philips, W., and Lemahieu, I. (2003). Noise reduction by fuzzy image filtering. *IEEE Transactions on Fuzzy Systems*, 11(4):429–436.

[Van der Weken, 2004] Van der Weken, D. (2004). *The use and the construction of similarity measures in image processing.* PhD thesis, Ghent University.

[Van der Weken et al., 2001] Van der Weken, D., Nachtegael, M., and Kerre, E. (2001). The applicability of similarity measures in image processing. *Intellectual Systems*, 6(1-4):231–248.

[Van der Weken et al., 2002] Van der Weken, D., Nachtegael, M., and Kerre, E. (2002). An overview of similarity measures for images. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3317–3320, Orlando, FL, USA.

[Van der Weken et al., 2003] Van der Weken, D., Nachtegael, M., and Kerre, E. (2003). Using similarity measures for histogram comparison. *Lecture Notes in Artificial Intelligence*, 2715:396–403.

[Van der Weken et al., 2004] Van der Weken, D., Nachtegael, M., and Kerre, E. (2004). Using similarity measures and homogeneity for the comparison of images. *Image and Vision Computing*, 22(9):695–702.

[Vandemeulebroucke et al., 2005] Vandemeulebroucke, J., Vansteenkiste, E., and Philips, W. (2005). Registratie van echo en mr hersenbeelden van prematuren. Master's thesis, Ghent University.

[Vandemeulebroucke et al., 2006] Vandemeulebroucke, J., Vansteenkiste, E., and Philips, W. (2006). A multi-modal 2d/3d registration scheme for preterm brain images. In *Proceedings of the Engineering in Medicine and Biology Conference*, New York, USA.

[Vansteenkiste et al., 2007a] Vansteenkiste, E., Huysmans, B., Govaert, P., and Philips, W. (2007a). Texture-based classification of preterm periventricular leukomalacia. *accepted for Current Medical Imaging Reviews*.

[Vansteenkiste et al., 2005a] Vansteenkiste, E., Huysmans, B., and Philips, W. (2005a). Classifying affected brain tissue in uncompensated ultrasound images of neonates. In *Proceedings of EUSIPCO*, Antalya, Turkey.

[Vansteenkiste et al., 2005b] Vansteenkiste, E., Huysmans, B., and Philips, W. (2005b). Classifying affected brain tissue in uncompensated ultrasound images of neonates. In *Proceedings of the 3rd European Medical and Biological Engineering Conference*, volume 11, Prague, Czech Republic.

[Vansteenkiste et al., 2007b] Vansteenkiste, E., Philips, W., Conneman, N., Lequin, M., and Govaert, P. (2007b). Segmenting periventricular leukomalacia in preterm ultrasound images. *In review: Ultrasound in Medicine and Biology*.

[Vansteenkiste et al., 2005c] Vansteenkiste, E., Pizurica, A., and Philips, W. (2005c). Improved segmentation of ultrasound brain tissue incorporating expert evaluation. In *Proceedings of the Engineering in Medicine and Biology Conference*, page 524, Shanghai, China.

[Vansteenkiste et al., 2004a] Vansteenkiste, E., Schoutteet, A., Gautama, S., and Philips, W. (2004a). Analysing multispectral textures in very high resolution satellite images. In *Proceedings of IGARSS*, Anchorage, Alaska.

[Vansteenkiste et al., 2004b] Vansteenkiste, E., Schoutteet, A., Gautama, S., and Philips, W. (2004b). Comparing colour and textural information in high resolution ikonos image classification. In *Proceedings of the International Conference on Image Processing*, Singapore.

[Vansteenkiste et al., 2003a] Vansteenkiste, E., Stippel, G., Govaert, P., Ledda, A., and Philips, W. (2003a). Segmenting leukomalacia using textural information and mathematical morphology. In *Proceedings of STW/Prorisc*, pages 441–446, Veldhoven, The Netherlands.

[Vansteenkiste et al., 2002] Vansteenkiste, E., Stippel, G., and Philips, W. (2002). Classification and segmentation in different fields based on textural information. In *Proceedings of the second PhD symposium*, Ghent, Belgium.

[Vansteenkiste et al., 2003b] Vansteenkiste, E., Stippel, G., and Philips, W. (2003b). Tissue based segmentation and classification of medical us investigation of brain and liver tissue. In *Proceedings of VENEB, BIOMED flanders*, page 27, Aalst, Belgium.

[Vansteenkiste et al., 2006a] Vansteenkiste, E., Van der Weken, D., Philips, W., and Kerre, E. (2006a). Evaluation of fuzzy image quality measures using a multidimensional scaling framework. In *Proceedings of the Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, page on CD, Arizona, USA.

[Vansteenkiste et al., 2006b] Vansteenkiste, E., Van der Weken, D., Philips, W., and Kerre, E. (2006b). Evaluation of the perceptual performance of fuzzy image quality measures. In *LNCS: Knowledge-Based and Intelligent Information and Engineering Systems, KES*, volume 4251, pages 623–630, Bournemouth, England. Springer-Verlag.

[Vansteenkiste et al., 2006c] Vansteenkiste, E., Van der Weken, D., Philips, W., and Kerre, E. (2006c). Perceived image quality measurement of state-of-the-art noise reduction schemes. In *Lecture Notes in Computer Science ACIVS*, volume 4179, pages 114–124, Antwerp, Belgium.

[Vansteenkiste et al., 2006d] Vansteenkiste, E., Van der Weken, D., Philips, W., and Kerre, E. (2006d). Psycho-visual evaluation of fuzzy similarity measures. In *Proceedings of the Signal Processing Symposium (SPS-DARTS)*, pages 127–130, Antwerp, Belgium.

[Vansteenkiste et al., 2006e] Vansteenkiste, E., Van der Weken, D., Philips, W., and Kerre, E. (2006e). Psycho-visual quality assessment of state-of-the-art denoising schemes. In *Proceedings of EUSIPCO*, Florence, Italy.

[Vansteenkiste et al., 2006f] Vansteenkiste, E., Vandemeulebroucke, J., and Philips, W. (2006f). 2D/3D registration of neonatal brain images. In *Lecture Notes in Computer Science*, volume 4057, pages 272–279.

[Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley and Sons.

[Viola and Wells, 1997] Viola, P. and Wells, W. (1997). Alignment by maximization of mutual information. *International Journal on Computer Vision*, 24(2).

[Wagner et al., 1983] Wagner, R., Smith, S., Sandrik, J., and Lopez, H. (1983). Statistics of speckle in ultrasound b-scans. *IEEE Transactions of Speckle and ultrasonics*, 30(3):156–163.

[Wang and Shung, 1997] Wang, S.-H. and Shung, K. (1997). An approach for measuring ultrasonic backscattering from biologicaltissues with focused transducers. *IEEE Transactions on Biomedical Engineering*, 44(7):549–554.

[Wu et al., 1989] Wu, J., Liao, M., and Wang, S. (1989). Texture segmentation of ultrasound b-scan image by sum and difference histograms. In *Proceedings of the Annual International Conference of the IEEE Engineering in Engineering in Medicine and Biology Society*, pages 417–418.

[Wydooghe et al., 2004] Wydooghe, D., Vansteenkiste, E., Stippel, G., and Philips, W. (2004). Ontwerp van een tool voor het vergelijken van echografiebeelden. Master's thesis, Ghent University.

[Xiao et al., 2002] Xiao, G., Brady, M., and Noble, A. (2002). Segmentation of ultrasound b-mode images with intensity inhomogeneity correction. *IEEE Transactions of Image Processing*, 21(1).

[Yiqiang and Dinggang, 2003] Yiqiang, Z. and Dinggang, S. (2003). Automated segmentation of 3d us prostate images using statistical texture-based matching method. In *Lecture Notes in Computer Science*.

[Yu, 2002] Yu, Y. (2002). Speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 11(11):1260–1270.

[Zhu and Tian, 2003] Zhu, F. and Tian, J. (2003). Medical image segmentation using level set and watershed transform. In *Advanced Biomedical and Clinical Diagnostic Systems, Proceedings of the SPIE*, volume 4958, pages 294–302.

[Zijdenbos et al., 1994] Zijdenbos, A., Dawant, B., Margolin, R., and Palmer, A. (1994). Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE Transactions on medical imaging*, 13:716–724.

[Zlokolica, 2006] Zlokolica, V. (2006). *Advanced Non-linear Methods for Video Denoising*. PhD thesis, Ghent University.

[Zlokolica et al., 2006] Zlokolica, V., Pizurica, A., Vansteenkiste, E., and Philips, W. (2006). Spatio-temporal approach for noise estimation. In *Proceedings of The international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 146–148, Toulouse, France.

[Zu and Prince, 1997] Zu, C. and Prince, J. (1997). Gradient vector flow: a new external force for snakes. In *Proceedings of the Conference on computer Vision and Pattern Recognition*, pages 66–77, San Juan, Puerto Rico.

[Zucker and Kant, 1981] Zucker, S. and Kant, K. (1981). Multiple level representations for texture discrimination. *Proceedings of the IEEE Pattern Recognition and Image Processing Conference*, pages 609–614.