N° d'ordre: 4616

▪ THÈSE ▪

présentée pour obtenir le grade de
Docteur de l'Université Louis Pasteur - Strasbourg I

École doctorale	:	Sciences pour l'ingénieur
Discipline	:	Électronique, électrotechnique, automatique
Spécialité	:	Traitement d'images et vision par ordinateur

Modèles statistiques d'apparence non gaussiens. Application à la création d'un atlas probabiliste de perfusion cérébrale en imagerie médicale

English title: "Non-Gaussian Statistical Appearance Models.
Application to the Creation of a Probabilistic Atlas of Brain Perfusion in Medical Imaging."

Soutenue publiquement
le 21 septembre 2004
par

Torbjørn VIK

Membres du jury:

Mme.	Isabelle	BLOCH	Rapporteur externe
M.	Jack-Gérard	POSTAIRE	Rapporteur externe
M.	Ernest	HIRSCH	Rapporteur interne
M.	Philippe	RYVLIN	Examinateur
M.	Fabrice	HEITZ	Directeur de thèse
M.	Jean-Paul	ARMSPACH	Directeur de thèse

Thanks

Working and writing this thesis during four years has been an incredible experience to me. It has given me the opportunity to learn a great many things, as well on the professional level as on the personal level. It has further given me the opportunity to work with many competent and resourceful persons. As with most research I believe, the achievement of this thesis has followed a route of ups and downs on which I have not been alone. The interaction with and the help from my collegues, friends and family has been indispensable and I would hereby like to express my gratefulness to everybody.

First of all, I would like to thank Isabelle Bloch, Jack-Gérard Postaire and Ernest Hirsch for their careful evaluation and judicious comments concerning the manuscript of this work. I would also like to thank Ernest Hirsch for his guidance as a pedagogical mentor and for his engagement in organizing bilingual scientific seminars which has permitted me to enlarge my professional vision. I would further like to thank Philippe Ryvlin for the interest he has shown in my work, an extremely difficult task for a non-specialist.

I am very grateful for having had the opportunity to work under the guidance of Fabrice Heitz and Jean-Paul Armspach who, through engagement and dedication, have created a rich and dynamic working environment for me and others. Furthermore, I would like to thank both for their support and the many discussions we have had during these years.

I thank Daniel Grucker and Jean-François Dufourd for having received me at their respective laboratories. A special thank to the personnel and co-workers at both laboratories.

I would also like to thank the persons I have met during my teaching duties, Sophie Kohler, Yoshi Takakura, Jean Martz, Laurent Thoraval and others.

I would like to thank my fellow doctoral students, Marcel Bosc, Sylvain Faisan, Vincent Noblet, Rozenn Dahyot, Farid Flitti, Aicha Far and others. In particular, I would like to thank Marcel Bosc, who taught me linux/unix and who took me to a new level of C++ programming. His concern and engagement for out-of-work issues has further been an immense personal enrichement to me. I would also like to thank Sylvain Faison who I could easily win for scientific discussions (mostly futile) through which I again caught pleasure in my work after a rude first year. I hope we shall have time again for our scientific and non-scientific bike-rides when stuff has calmed down after the thesis. During these years, I have also enjoyed the companionship of a long list of internships and temporary workers, most lately Thomas Berst, Nicolas Wiest-Daesslé and Samuel Sinapin.

A warm thanks goes to my wife Aude and to my son Emil. You make me a very happy person. I would also like to thank our families and friends for their support and for many pleasant occasions during these years and in the years to come.

Abstract

Single Photon Emission Computer Tomography (SPECT) is a 3D functional imaging technique that yields information about the blood flow in the brain (also called brain perfusion). This imaging technique has found application in the diagnostics of head trauma, dementia, epilepsy and other brain pathologies. To this end, SPECT images are analyzed in order to find abnormal blood flow patterns. For localized abnormalities such as stroke, this characterization remains an accessible task, whereas for diffuse and variable abnormalities such as beginning dementia, near-drowning episodes and toxic substance exposure, characterization is difficult. It is therefore necessary to develop quantitative methods in which computer-aided statistical analysis can take advantage of information present in a database of normal subjects.

This work deals with the construction and evaluation of a probabilistic atlas of brain perfusion in normal subjects as observed in SPECT images. The goals of such an atlas are twofold: (1) to describe perfusion patterns of the population represented by the atlas in a compact manner, and (2) to identify statistically significant differences between an individual brain perfusion pattern and the probabilistic atlas. The successful creation of a computerized, probabilistic atlas may have far-reaching impact on clinical applications where qualitative (visual) analysis of SPECT images is current practice.

Three issues have been central in this work: the statistical models that actually describe brain perfusion, the image processing tools used to make brains "comparable" and the experimental evaluation of the atlas. For the first issue, we have explored so-called appearance-based approaches. These have been developed in computer vision where they have also been widely adopted. Recent developments have given these models a proper statistical basis. In this work, we have introduced an original non-linear model based on principal component analysis (PCA) and Bayesian estimation theory.

The second issue is related to the spatial normalization of images (i.e. image registration) and intensity normalization. In order to compare brain images coming from different subjects, their relative positions must be found. This is done by calculating a non-linear mapping between corresponding anatomical regions. A registration scheme specifically adapted to SPECT images had to be developed for this task. Furthermore, since the gray values in SPECT images only represent relative measures of blood flow, the observed values must be normalized to allow for comparison between images. For this we devised an efficient, joint distribution-based intensity normalization scheme.

Finally, because of the lack of absolute knowledge about the brain perfusion in a normal population, an elaborate evaluation scheme had to be developed. The scheme is based on the detection of simulated abnormalities combined with a leave-one-out strategy. This scheme was used to evaluate and compare the different models and normalization schemes considered in this work. For evaluation on a clinical application, the atlas was also applied to characterize seizure foci in patients with epilepsy.

Résumé

Cette thèse a été rédigée en anglais. Voici un résumé détaillé de la thèse en français.

La tomoscintigraphie par émission mono-photonique (TEMP) est une méthode d'imagerie fonctionnelle 3D qui apporte des informations sur le débit sanguin cérébral (également appelé perfusion cérébrale). Cette méthode d'imagerie, par la détection visuelle d'anomalies de perfusion caractérisées par des zones hypo- ou hyper-intenses, est utilisée pour le diagnostic chez des patients atteints d'accidents vasculaires cérébraux, de démence, d'épilepsie ou d'autres pathologies cérébrales. La détection d'anomalies focalisées observées chez les patients ayant une attaque cérébrale est relativement aisée, alors que les anomalies diffuses, observées en début de démence, lors d'un accident entraînant une oxygénation insuffisante du cerveau ou suite à une exposition à une substance toxique, sont plus difficilement observables. Dans ces cas, une analyse quantitative des images, utilisant un atlas et des outils statistiques s'appuyant sur une base d'images de cas normaux, peut apporter une aide précieuse au diagnostic.

Le travail présenté dans cette thèse est centré sur la problématique de la construction et de l'évaluation d'un atlas probabiliste de perfusion cérébrale à partir des images TEMP de sujets dits normaux. Les objectifs d'un tel atlas sont doubles : (1) création d'une cartographie statistique de la perfusion cérébrale d'une population normale, décrite de manière compacte, et (2) identification des différences de perfusion cérébrale qui sont statistiquement significatives entre une image TEMP d'un individu et l'atlas probabiliste. L'utilisation d'un atlas devrait avoir un impact important sur les applications cliniques où l'analyse qualitative d'images TEMP est pratique courante.

Afin d'atteindre ces objectifs, trois points ont été abordés : le développement de modèles statistiques qui décrivent de façon fidèle la perfusion cérébrale, les outils de traitement d'images utilisés pour rendre les cerveaux « comparables », et enfin, l'évaluation expérimentale de l'atlas.

Pour le premier point, nous avons exploré les approches dites « par modèles d'apparence ». Ceux-ci ont été développés dans le domaine de la vision par ordinateur où ils ont été largement appliqués. Des développements récents ont redéfini ces modèles dans un cadre statistique. Dans ce travail, nous avons introduit un modèle original non linéaire et non gaussien, basé sur l'analyse en composantes principales (ACP) et la théorie de l'estimation bayesienne.

Le second point est lié à la fois à la normalisation spatiale de l'image (ou recalage d'images) et à la normalisation d'intensité des images. La création d'un atlas impose de mettre en correspondance les différentes structures anatomiques (qui doivent occuper le même emplacement dans l'espace). Ceci est réalisé à l'aide d'un recalage non rigide (dit « déformable », c.à.d. une transformation spatiale non linéaire. Une méthode de recalage spécifiquement adaptée aux images TEMP a dû être développée à cet effet. De plus, puisque les niveaux de gris dans les images TEMP représentent des mesures relatives à la perfusion, les valeurs observées doivent être normalisées afin de permettre une comparaison entre images. Pour cela, nous avons developpé une technique de normalisation d'intensité basée sur l'histogramme conjoint des images 3D.

Le dernier point concerne l'évaluation de l'ensemble de la chaîne de traitement. L'absence d'une vérité terrain relative à la perfusion cérébrale d'une population ou d'un individu, nous a amené à développer une procédure évoluée d'évaluation. Cette procédure est basée sur la détection d'anomalies simulées, combinée avec une stratégie de validation croisée. La procédure a été utilisée pour évaluer et comparer les différents modèles et techniques de normalisation

développés dans ce travail. Des résultats cliniques préliminaires ont été obtenus en utilisant l'atlas pour la caractérisation des foyers épileptogènes chez des patients épileptiques.

Contenu du mémoire

Le mémoire de thèse est divisé en trois parties. La première partie décrit le contexte et l'objectif principal des travaux réalisés. La seconde partie est consacrée aux développements théoriques de ce travail et la troisième partie concerne la construction et l'évaluation d'un atlas de perfusion cérébrale en utilisant le cadre de modélisation statistique décrit dans la seconde partie.

Première partie

Dans le **premier chapitre**, le cadre de travail est présenté ainsi qu'une introduction générale du problème abordé dans cet thèse, les principales difficultés rencontrées pour résoudre ce problème et les approches choisies. Le travail a été effectué en collaboration entre deux laboratoires de recherche, d'un côté le laboratoire des sciences de l'image, de l'informatique et de la télédétection (LSIIT – UMR 7005 CNRS), et de l'autre côté l'institut de physique biologique (IPB – UMR 7004 CNRS). Ceci explique aussi en partie la dualité de ce travail qui cherche d'une part à dévélopper des modèles mathématiques généraux et d'autre part à attaquer une application spécifique avec les problèmes pratiques qui lui sont liés. Le but de ce travail, la création d'un atlas probabiliste de perfusion cérébrale, nécessite des solutions adaptées aux problèmes suivants :

1. Le problème de rendre les images de cervaux provenant de différents sujets, comparables. Il est nécessaire d'utiliser des outils de traitements d'images tel que le recalage d'images.
2. Le problème de la définition d'un modèle statistique qui décrit la perfusion cérébrale d'une population et qui permet la comparaison d'une image avec cette population. Comme nous allons l'expliquer dans le chapitre 2, les images TEMP sont difficiles à interpréter, même pour un expert. Ceci nous a poussé à utiliser des approches par apprentissage statistique, largement employées en vision par ordinateur.
3. Le problème de l'évaluation de la qualité de l'atlas après sa création. Comme la perfusion cérébrale d'une population normale est inconnue, ce problème est particulièrement difficile.
4. Finalement, le problème de gestion des bases de données, les outils informatiques nécessaires ainsi que le développement des outils de traitement d'images.

Ce chapitre contient une liste des contributions originales de l'auteur et le chapitre est conclu par une description de l'organisation du manuscrit.

Le **chapitre 2** est une introduction à l'imagerie médicale cérébrale et en particulier à l'imagerie TEMP (single photon emission computer tomography – SPECT en anglais). Après quelques précisions sur la distinction entre l'imagerie dite anatomique (ou morphologique) et fonctionnelle, un rappel des méthodes utilisées pour l'imagerie fonctionnelle du cerveau est donné. Ensuite, une description plus détaillée sur les procédures d'acquisition des images TEMP est fournie. Nous décrivons d'abord la procédure telle qu'elle est perçue par le patient. Ensuite, on décrit le traceur radiopharmaceutique utilisé, ces propriétés physiques et biocinétiques. On décrit aussi le système d'acquisition et de reconstruction de l'image à partir des projections et on rapelle les différentes sources d'erreurs qui influencent la qualité de

l'image. Finalement, nous décrivons la façon dont les images sont actuellement interprétées par le médecin, en pratique.

Nous continuons en décrivant la charge de radiation que le patient subit par la technique de TEMP. Celle-ci est presque équivalente à la radiation qu'une personne reçoit normalement en deux ans par le rayonnement naturel. Ensuite, nous décrivons les applications cliniques de l'imagerie TEMP ainsi que les questions scientifiques qui sont abordées avec ce type d'imagerie. Cette introduction permet de situer les difficultés rencontrées, en particulier concernant le recalage et la normalisation d'intensité, détaillées dans les chapitres 6 et 7.

Deuxième partie

La seconde partie est consacrée aux développements théoriques de ce travail, basés sur les modèles d'apparence probabilistes. Cette partie n'est pas spécifique à l'application médicale, mais traite de la modélisation des images d'une façon générale. Le modèle original proposé est donc aussi bien utilisable pour des applications en vision par ordinateur classique (détection, reconnaissance, suivi) que pour la création d'un atlas probabiliste de perfusion cérébrale qui, quant à lui, est décrit dans la troisième partie.

Dans le **chapitre 3**, nous présentons un état de l'art des modèles d'apparence, leurs extensions et leurs variantes. Un rappel des techniques d'estimation statistique robuste (en particulier la théorie semi-quadratique) et des méthodes d'estimation de densité non paramétrique (en particulier le « mean shift ») est aussi présenté. Le but de ce chapitre est de donner au lecteur les éléments nécessaires pour comprendre et apprécier la contribution originale de ce travail, présentée dans le chapitre suivant.

Dans l'introduction de ce chapitre, on établit le lien entre la prédiction, la reconnaissance des formes et la necessité de chercher des structures dans les données (corrélations, regroupements, etc.). Une façon de chercher ce type de structures est d'imposer un modèle paramétrique et d'estimer les paramètres de ce modèle à partir des échantillons (apprentissage). Dans les approches ou les modèles dits *d'apparence*, nous prenons l'image brute comme donnée. Nous discutons les avantages et inconvénients d'une modélisation probabiliste d'une façon générale et finalement nous montrons qu'il est nécessaire et possible de modéliser des données multivariées de très grande dimension en utilisant des techniques de réduction de dimension.

Dans cet état de l'art on prête beaucoup d'attention aux modèles d'apparence qui sont à la base du nouvau modèle proposé plus loin. Ces modèles ont été développés pour des applications en vision par ordinateur, en particulier pour la reconnaissance des visages. Une revue chronologique et incrémentale de ces modèles est proposée. D'abord on présente le modèle de base de Sorovitch et Kirby [195]. Ce modèle est nommé « eigenfaces » (visages propres) car il repose sur l'analyse des valeurs et vecteurs propres de la matrice de covariance des images observées (base d'apprentissage). En rejetant les vecteurs propres correspondant aux valeurs propres les moins fortes, on obtient une représentation parsimonieuse de la base d'apprentissage où les variations les plus fortes de la base d'apprentissage sont retenues. Cette technique est connue sous le nom *d'analyse en composantes principales* (ACP) et elle est aussi utilisée pour la compression des données. Le modèle résultant est un modèle linéaire qui introduit un *sous-espace visuel*, aussi appelé *espace propre*. L'idée est d'introduire une variable cachée (non observable, latente) qui vit dans l'espace propre et qui gouverne l'apparence d'un objet (ou des visages). L'observation dépend donc d'une façon linéaire de la variable latente et les vecteurs propres. La tâche de reconnaissance (détection, suivi) se fait en se basant sur cette variable latente. Elle permet par exemple de différentier (reconnaître) les visages des différentes personnes. Pour cela, Sirovitch et Kirby modélisent sa distribution comme une

simple gaussienne pour chaque classe (visage).

Ensuite, nous présentons l'extension proposée par Murase et Nayar [163]. Ils montrent dans leur article que la distribution de la variable latente de certains objets est loin d'être gaussienne. Ceci mène les auteurs à proposer une modélisation non linéaire et non gaussienne de la variable latente. Leur méthode permet d'améliorer le taux de reconnaissance d'une façon considérable. Finalement, on présente les extensions probabilistes du modèle de base. Ces modèles sont à nouveau limités à considérer des distributions gaussiennes de la variable latente mais ils permettent une modélisation probabiliste de l'observation (l'ACP n'est pas en soi probabiliste). Le premier modèle est celui de Moghaddam et Pentland [162]. La superiorité de ce modèle par rapport aux modèles non probabilistes a pu être montrée dans un concours de reconnaissance de visages (FERET [186]). Dans ce modèle il n'y a pas que la variable latente qui est modélisée, mais l'observation (l'image) elle même est modélisée avec une distribution gaussienne (ou bien avec des extensions vers des mélanges de gaussiennes).

Le deuxième modèle probabiliste est l'ACP probabiliste (ACPP) de Tipping et Bishop [212, 211]. Ce modèle, qui est basé sur le modèle d'analyse factorielle (AF), introduit une séparation de la modélisation probabiliste entre la variable latente et le bruit d'observation. Ceci est intéressant car cela permet de considerer d'autres types de distributions que des gaussiennes (notamment des bruits d'observation non gaussiens. Pour compléter cet état de l'art sur les modèles d'apparence on propose une revue des méthodes générales de réduction de dimension. Dans cette revue on fait une distinction entre les méthodes linéaires et non linéaires. Les méthodes linéaires décrites sont la poursuite de projection, l'analyse discriminante de Fischer, l'analyse factorielle, et l'analyse en composantes indépendantes. Les méthodes non linéaires décrites sont les mélanges des modèles linéaires, les courbes et les surface principales, l'ACP de noyaux (ces méthodes sont toutes des extensions non linéaires de l'ACP), le carte d'auto-organisation de Kohonen et les résaux de densité de probabilité de MacKay.

Deux autres sujets sont ensuite traités dans ce chapitre : l'estimation robuste et la technique « mean shift ». Un rappel de l'estimation robuste en utilisant la théorie semi-quadratique [77, 78, 34, 36] est fait. Dans des travaux précedents dans notre équipe, Dayhot *et al.* [47, 46, 48] ont proposé une extension du modèle probabiliste de Tipping et Bishop qui a permis d'améliorer la performance de reconnaissance dans des situations où des occlusions ou d'autres artefacts dans l'image rendaient l'hypothèse d'un bruit gaussien non valable. Un modèle similaire a été développé par Black et Jepson [13]. La technique de Dahyot *et al.* est décrite plus en détail dans le chapitre suivant. Finalement, l'estimation non paramétrique en utilisant la technique « mean shift » [40] est décrite. Cette technique permet d'une façon efficace de considérer des distributions quelconques. Une technique originale d'apprentissage et d'inférence statistique basée sur le « mean shift », la théorie semi-quadratique et les modèles d'apparence probabilistes est la contribution théorique importante de cette thèse.

Dans le **chapitre 4**, nous présentons ce nouveau modèle. Nous avons décrit trois extensions des modèle d'apparence de base (Sorovitch et Kirby [195]) qui chacune apporte une amélioration des performances : l'extension non linéaire de Murase et Nayar [163], l'extension probabiliste de Moghaddam et Pentland [162] ainsi que celle de Tipping et Bishop (l'ACPP) [212, 211], et finalement l'extension aux bruits non gaussiens de Dahyot *et al.* [47, 46, 48]. Nous proposons un modèle qui combine ces trois extensions dans un cadre unifié et mathématiquement rigoureux. Pour pouvoir appliquer un tel modèle (dans un but de reconnaissance), il est nécessaire de résoudre deux problèmes : (1) le problème de reconstruction et (2) le problème d'apprentissage.

Le problème de reconstruction consiste à trouver la variable latente (non observable) à par-

tir d'une observation à condition de connaître les autres paramètres du modèle. Il est nécessaire de résoudre ce problème avant de passer au deuxième problème qui consiste à déterminer tous les paramètres du modèle à partir d'un ensemble d'observations (base d'apprentissage). Ces deux problèmes ne sont pas analytiquement solubles. Néanmois, nous avons pu développer une méthode efficace qui permet de résoudre le premier problème. Cette solution qui est basée sur un développement original du « mean shift » permet d'améliorer les performances de reconnaissance comme nous le montrerons dans le chapitre suivant. Ceci est possible même en utilisant une solution pragmatique et approximative du problème d'apprentissage.

Nous procédons d'une façon incrémentale dans la présentation de notre modèle et dans le développement de la solution au problème de reconstruction. Tout d'abord, nous reprenons le modèle de base de l'ACPP [212, 211] avec un résumé de ces propriétés. Une solution analytique du problème de reconstruction existe dans ce cas. Ensuite, nous détaillons les extensions faites au sein du laboratoire avec la thèse de Dahyot qui porte sur l'extension vers des bruits non gaussiens. Notre extension originale du « mean shift » permet de considérer des modèles avec une distribution non gaussienne dans l'espace propre et un bruit gaussien ou bien non gaussien. Le modèle final est alors non gaussien, non linéaire. Nous présentons d'abord l'extension avec un bruit gaussien (le « mean shift modifié »). L'extension finale vers un bruit non gaussien se fait avec la théorie semi-quadratique.

Avant de conclure ce chapitre, nous présentons un résumé sous forme de tableau de tous les modèles considérés lors de la présentation de notre modèle. Ce tableau permet d'avoir une vue d'ensemble sur les hypothèses de plus en plus générales ainsi que sur les solutions au problème de reconstruction pour chacun de ces modèles. Dans la conclusion, plusieurs voies d'investigations enviseagables sur notre modèle sont décrites. Notamment, des pistes pour attaquer le problème d'apprentissage, sont présentées.

Nous montrons ensuite l'avantage de notre modèle en menant des expériences de reconnaissance de formes sur une base standard d'images connue dans le domaine de vision par ordinateur. Il s'agit de la base COIL (Columbia Object Image Library) qui contient 1440 images de 20 objets différents. Pour chaque objet, des images selon 72 angles de vue différents ont été acquises. Chaque point de vue diffère de 5 degrés, ce qui fait qu'au total une rotation complète de l'objet est observée. Notre expérience, qui est repétée pour chaque objet, consiste à reconnaître l'angle de vue d'un objet dans une image. Pour évaluer la performance du système de reconnaissance, on introduit des dégradations contrôlées dans l'image qui lui est présentée.

Ces dégradataions sont des « occlusions » de tailles différentes, situées dans l'image. Avec des occlusions de plus en plus grandes, le système est donc confronté à des situations de plus en plus difficiles pour accomplir son objectif. Pour évaluer l'influence de ces occlusions sur les performances nous considèrons à la fois le taux de bonne reconnaissance (classification par le plus proche voisin dans l'espace propre) et la distance euclidéenne entre une image reconstruite (projection dans l'espace propre) et la vérité-terrain. Trois méthodes sont comparées : (1) le méthode classique de Sirovitch et Kirby [195] (ML), (2) l'extension robuste de Dahyot [47, 46, 48] (RML), et (3) notre nouvelle méthode (RMMS) présentée dans le chapitre précédent. Les résultats de ces expériences sont décrits avant que nous les discutions.

Les résultats obtenus sont résumés sous forme de graphiques « box-whisker ». Sur les 20 objets considérés, les résultats sont très hétérogènes. Néanmoins, il y a une claire tendance en faveur de notre méthode par rapport à la méthode robuste classique. Avec la méthode non robuste nous obtenons des résultats nettement moins bons. Nous présentons en particulier une analyse fine des résultats obtenus avec les occlusions les plus importantes (une couverture de

40% de l'image). La différence n'étant pas très marquée pour 13 objets, elle est importante (> 7 bonnes reconnaissances) pour les 7 objets restants quant au nombre de bonnes classifications. En analysant ces résultats plus finement, on constate que les objets qui sont le mieux décrits par le modèle (en termes de variance absolue non tronquée par l'ACP) sont ceux qui gagnent le plus en utilisant notre nouveau modèle. Nous remarquons aussi que l'amélioration des résultats avec notre modèle est plus nette quand on analyse les distances dans l'espace propre des images reconstruites. Ceci est important pour notre application de création d'un atlas (car nous nous servirons du résidu entre l'image reconstruite et l'observation, comme le nous décrivons plus bas).

Troisième partie

La troisième partie de la thèse concerne la construction et l'évaluation d'un atlas de perfusion cérébrale en utilisant le cadre de modélisation statistique décrit dans la deuxième partie. Elle concerne aussi les différents traitements qui sont spécifiques à cette application, en particulier le recalage d'images (normalisation spatiale) et la normalisation d'intensité. Dans le **chapitre 6**, nous présentons un état de l'art et une vue d'ensemble des méthodes statistiques de construction d'atlas, une revue des techniques de normalisation d'intensité ainsi qu'une description des techniques de recalage auxquelles nous faisons appel dans notre travail.

Il existe beaucoup de travaux sur la modélisation des images fonctionnelles du cerveau (TEP, TEMP, IRMf et d'autres). Pour mieux cerner la problématique que nous avons souhaité traiter, nous distinguons la comparaison d'un individu à un atlas (une image avec J images) de la comparaison entre deux groupes d'images (J avec J' images), et nous distinguons entre les études d'activation (détection) et les études dites paramétrique (régression, détermination d'une relation entre une cause et l'observation). Chacune de ces problématiques nécessite des approches adaptées. La création d'un atlas se situe comme un problème de comparaison J à 1 pour obtenir une détection.

Une revue, structurée selon la nature de l'analyse statistique (univariée ou multivariée) et sous-divisée selon les propriétés qui sont modélisées (voxel, région, autre) est ensuite présentée. La revue s'appuie sur des tableaux qui résument les caractéristiques principales. Ceci permet d'acquérir plus facilement une vue d'ensemble de ces méthodes qui sont très hétérogènes quant aux différentes techniques utilisées et l'application étudiée. Par exemple, dans une étude l'âge peut être considéré comme une variable de confusion, une variable d'intérêt, ou bien pas modélisée du tout.

Parmi les méthodes décrites, nous nous intéressons en particulier à la méthode de Houston *et al.* [104, 105] qui est en essence une méthode identique à un modèle d'apparence. Elle correspond à la méthode ML (non robuste), décrite dans les chapitres 3 et 4. Les auteurs introduisent une technique puissante de comparaison statistique d'une image avec l'atlas. Cette technique permet de caractériser le résultat de cette comparaison en termes de zones locales avec un z-score associé. La limitation des méthodes multivariées qui ne peuvent que caractériser une observation d'une façon globale est ainsi résolue avec cette méthode originale. Nous la reprenons dans nos travaux.

Le deuxième sujet traité dans ce chapitre est le recalage (normalisation spatiale) de deux images. Le but du recalage est de ramener les structures anatomiques dans un même système de réference spatial. Nous constatons que la plupart des approches utilisent un recalage inter-individus affine (nous utiliserons un recalage déformable, décrit ultérieurement). Une revue complète de ce domaine vaste n'est toutefois pas fournie. Nous décrivons uniquement quelques techniques importantes, ainsi que les techniques auxquelles nous avons recouru dans notre

travail. Ces techniques n'ont pas été développées par l'auteur, mais par d'autres membres du groupe de recherche. Concernant le recalage d'images médicales, nous distinguons le recalage inter- et intra-individu, ainsi qu'inter- et intra-modal. Dans notre travail, nous nous sommes servi du recalage intra-sujet, inter-modal (TEMP sur IRM) ainsi que du recalage inter-sujets, intra-modal (IRM sur IRM). Chaque type de recalage nécessite une approche adaptée et est donc décrit séparément.

En résumé, on peut dire que le recalage consiste (1) à définir un modèle géométrique et paramétrique qui décrit la transformation entre deux images (rigide, affine, « déformable »), (2) à définir une fonction de coût qui mesure la similarité entre deux images, et (3) à développer un schéma d'optimisation qui permet l'estimation des paramètres de la transformation en minimisant la fonction de coût. Quand on recale deux images d'un même patient venant de deux modalités différentes (par exemple TEMP et IRM), on considère typiquement une transformation rigide (rotation et translation). Cela signifie qu'on considère le cerveau comme un objet rigide et que le système d'acquisition n'introduit pas de distortions géométriques. La fonction de coût qui s'est avérée la mieux adaptée au recalage multi-modal est l'information mutuelle. Avec un schéma d'optimisation basé sur l'algorithme de *simplexe* avec des initialisations multiples, on obtient des résultats acceptables pour toutes nos images.

Le recalage inter-individus, IRM-IRM est un sujet complexe qui fait encore objet de recherches intensives. Il n'existe pas une solution parfaite car personne ne sait mettre en correspondance les structures de deux cerveaux différents. La très grande variabilité anatomique, inter-individus n'est pas encore suffisamment bien comprise par les chercheurs et les neuroanatomistes. Néanmoins, nous utilisons une technique développée au sein du laboratoire qui permet de décrire une large variété de déformations. Cet algorithme est basé sur une représentation multi-échelle (B-splines d'ordre 1) du champ de transformation. Cette décomposition impose une certaine régularisation du champ et elle permet de choisir le niveau de précision (ou détails) de la transformation. La similarité entre deux images est mesurée avec une fonction de coût robuste.

La dernière partie de ce chapitre est une revue des techniques de normalisation d'intensité des images TEMP. C'est surtout l'intensité qui est porteuse d'informations dans ces images. Une normalisation est nécessaire pour pouvoir comparer la perfusion d'un sujet à un autre car l'intensité dans l'image dépend de la quantité de traceur injectée et d'autres facteurs liés à l'acquisition. Pourtant, ce sujet est controversé et un consensus sur une méthode n'existe pas dans la littérature. Les différentes approches se distinguent alors selon le type de fonction de transfert qui est utilisée (toujours linéaire, mais parfois additive, multiplicative ou bien les deux ensemble), et selon la façon dont cette fonction est estimée. On remarque que certains auteurs préfèrent utiliser l'histogramme conjoint qui, quant à lui, code la relation entre les paires de niveaux de gris de tous les voxels de deux images.

Tout au long ce chapitre, nous avons consideré le recalage et la normalisation d'intensité comme des étapes de traitement d'images préliminaires à la modélisation statistique proprement dite. Une autre approche altérnative consiste à intégrer ces deux étapes dans le modèle statistique. L'intégration du champ de transformation dans le modèle permettrait d'ajouter des informations morphologiques à l'analyse des images. L'intégration de la normalisation d'intensité permettrait d'adapter celle-ci à l'image analysée et pourrait se faire avec un modèle de type ANCOVA. Ces deux approches sont brièvement décrites. Notons finalement, que le chapitre 6 est un chapitre qui décrit les techniques existantes et nous ne décrivons pas notre approche et nos contributions originales. Ce chapitre permet d'obtenir une vue d'ensemble sur le processus de la création d'un atlas et les techniques impliquées. Ces techniques sont suffi-

samment décrites pour que le lecteur puisse se rendre compte des difficultés et des limitations liées à la création de l'atlas.

Dans **chapitre 7** nous procédons à une description détaillée de notre approche. Celle-ci est basée en partie sur les méthodes de recalage décrites dans le chapitre 6 et sur les modèles statistiques décrits dans le chapitre 4. Il s'agit d'une description des méthodes choisies et des résultats de la création de l'atlas. L'évaluation de l'atlas et son application dans un cadre clinique sont décrites dans les chapitres suivants. Après avoir présenté la base d'images que nous possédons à l'IPB, nous continuons par une description détaillée du schéma de recalage qui commence par une discussion sur le choix du référentiel. Nous avons choisi d'utiliser le standard appelé ICBM (International Consortium of Brain Mapping) qui est une image moyennée de 452 cerveaux (après recalage non rigide). Avant de recaler les images de notre base d'images sur cette référence, nous utilisons une référence intermédiaire qui est une image IRM issue du même appareil que les images de la base. L'utilisation d'une telle image améliore la pertinence du recalage déformable (inter-sujets). La référence intermédiaire est ensuite recalée sur l'image d'ICBM avec une transformation rigide.

Le recalage proprement dit est mené en trois étapes. D'abord, l'image TEMP est recalée sur l'image IRM du même sujet en utilisant la technique du recalage rigide décrit dans le chapitre 6. Ensuite, le recalage inter-individus est mené sur les images IRM : d'abord un recalage affine, suivi d'un recalage déformable (multi-résolutions, spline). Finalement, les champs issus de chaque étape de recalage sont combinés avant d'être appliqués à l'image TEMP d'origine de sorte qu'une seule interpolation est nécessaire. Une originalité dans ce schéma est l'introduction d'un filtrage du champ issu du recalage déformable (IRM-IRM). Ce filtrage a été nécessaire car l'application directe du champ s'est avérée infructueuse car elle introduisait des artefacts dans l'image TEMP. Pour évaluer la pertinence de ce filtrage, on a mené des expériences avec différentes approches de recalages :

1. uniquement un recalage affine ;
2. un recalage déformable à basse résolution ;
3. un recalage déformable à haute résolution mais sans filtrage ;
4. et un recalage déformable à haute résolution avec filtrage.

Ces quatre approches ont été comparées visuellement (en simulant des images TEMP) et dans l'étude de comparaison quantitative décrite dans le chapitre 8.

Après avoir ramené toutes les images dans un même référentiel spatial, il est nécessaire de segmenter le cerveau du fond (bruit) des images TEMP. Ceci est nécessaire à la fois parce que nous ne modélisons que la perfusion dans le cerveau et également parce que la normalisation d'intensité se base sur la perfusion du cerveau. Une simple seuillage est suffisent à ce but, aisément obtenu avec des techniques automatiques. La normalisation d'intensité est faite à partir de l'histogramme conjoint en calculant une fonction de transfert avec deux paramètres (multiplicatif et additif). Contrairement aux méthodes usuelles, nous proposons d'estimer ces paramètres avec une estimation de type « total least squares » (TLS). Cette estimation permet de prendre en compte des erreurs dans l'image de référence ainsi que dans l'image à normaliser.

La dernière partie de ce chapitre rappelle les modèles statistiques que nous avons évalués (chapitre 8) et nous présentons les résultats de l'apprentissage de ces modèles. Nous avons comparé trois modèles :

1. Un modèle gaussien « local ».
2. Le modèle « global » de Houston *et al.* [104, 105], qui est un modèle d'apparence non robuste, « ML » dans le chapitre 4.

3. Notre nouveau modèle « global robust » avec *a priori* dans l'espace propre, « RMMS » dans le chapitre 4.

L'apprentissage des deux dernières modèles se fait de la même manière, ce qui signifie que nous faisons une distinction entre les deux modèles uniquement lors de la phase de comparaison d'une image avec l'atlas. Lors de l'estimation du modèle on considère que la base d'apprentissage est distribuée selon une loi gaussienne. Par contre, lors de la phase de comparaison nous considérons un bruit non gaussien car l'image à comparer à l'atlas (analysée) peut contenir des lésions. La comparaison d'une image avec l'atlas est réalisée avec la méthode de Houston qui est rappelée encore une fois ici. Celle-ci consiste à analyser le résidu d'une image après l'avoir reconstruite avec le modèle statistique (soit ML soit RMMS).

Finalement, nous présentons l'image moyenne, l'image d'écart-type, les premiers vecteurs propres, les valeurs propres et la projection de la base d'apprentissage dans l'espace propre défini par les cinq premiers vecteurs propres. Ceux-ci sont difficiles à interpréter. Néanmoins, on voit que :

- l'écart-type du modèle local est moins homogène que l'écart-type issu de l'apprentissage des modèles globaux ;
- le premier vecteur propre semble refléter une variance liée à un recalage imparfait au niveau du cortex ;
- il est difficile de voir une distribution particulière (gaussienne) de la base d'apprentissage dans l'espace propre ;
- il est difficile de voir s'il y a des données (images) aberrantes dans l'espace propre puisque la base n'est pas suffisament grande pour dégager des tendances claires.

Le **chapitre 8** concerne l'évaluation de l'atlas. Nous comparons les différentes techniques de recalage et de normalisation d'intensité, ainsi que les différents modèles statistiques qui ont été décrits dans le chapitre 7. Avant de commencer la description de notre méthode d'évaluation, nous menons d'abord quelques réflexions sur la nécessité d'évaluer un tel atlas et les difficultées qui y sont liées. Ces dernières sont dues à la fois au fait que nous possédons un nombre limité d'échantillons (ce qui rend l'application des tests d'adéquation des données classiques impossible) et au manque de connaissances sur la véritable distribution (théorique) de ces données. On considère donc un schéma d'évaluation basé sur des simulations.

Les études dans la littérature qui utilisent des simulations sont très limitées dans le sens où elles ne prennent pas en compte toute la variabilité réelle qui existe dans une population. Nous utilisons une technique de validation croisée en combinaison avec des lésions synthétiques pour prendre en compte cette variabilité. Avec ce schéma d'évaluation on obtient une mesure de performance qui dépend de (1) l'adéquation des données au modèle et (2) de la sensibilité du modèle à détecter des lésions. Comme cette mesure de qualité dépend du recalage et de la normalisation d'intensité, nous pouvons l'utiliser pour comparer et évaluer ces prétraitements. La mesure est basée sur des courbes caractéristiques opérationnelles de récepteur (courbes COR) sur lesquelles nous effectuons des tests de significativité. Nous étudions aussi l'effet de lésions de différentes tailles et d'intensité, situées à différentes positions dans le cerveau.

Nous sommes capables de mener une large gamme d'analyses sur les résultats en fonction des différents facteurs qui ont une influence sur la performance de l'atlas. Tout d'abord, nous nous attachons à comparer les différents modèles statistiques. Pour des lésions de petites tailles nous montrons (pour la première fois) que le modèle global est significativement meilleur que le modèle local. Ensuite, nous montrons que pour des lésions de grandes tailles, un modèle robuste global est meilleur que les autre modèles. Par contre, pour les petites lésions, les deux modèles globaux sont équivalents. Nous cherchons aussi le nombre optimal de vecteurs propres

à utiliser dans les modèles globaux. Nous trouvons un nombre assez faible de 3 ou 4 vecteurs propres. Ceci peut être lié à la représentativité de la base d'aprentissage et est susceptible d'augmenter avec une base plus importante.

Ensuite, la supériorité du recalage déformable avec filtrage ainsi que la forte influence des méthodes de normalisation sur les résultats sont mises en évidence. Nous vérifions que les différentes étapes d'amélioration s'accumulent. Finalement, nous avons obtenu un résultat un peu surprenant et inconnu dans la littérature en montrant une forte dépendance de la performance sur la position de la lésion dans le cerveau. En particulier, la partie frontale droite montre des performances moins bonnes que les autres régions étudiées. Ceci est dû à la fois à une plus forte variabilité dans certaines régions que d'autres et à la limitation des modèles à décrire cette variabilité. Cette partie du travail a donné lieu à une publication dans *NeuroImage* [221].

Le **chapitre 9** est un chapitre expérimental. Il s'agit des résultats préliminaires en épilepsie. Nous avons choisi de traiter des cas pathologiques en épilepsie car une technique de référence existe, qui permet une comparaison avec la technique de l'atlas. Notons néanmoins que l'intérêt majeur de la technique de l'atlas reste surtout dans son application à des maladies où une technique de référence n'existe pas (la maladie d'Alzheimer, de Parkinson etc.).

Nous commençons avec une brève description de l'épilepsie. Un patient épileptique est une personne qui subit des crises épileptiques. Ces crises sont souvent dues à un fonctionnement anormal d'un ou plusieurs « foyers épileptogènes » qui déclenchent une crise en envoyant une avalanche de signaux d'une façon non controlée. Lors du déclenchement de la crise, ces foyers sont hyperperfusés et il est possible d'acquérir des images TEMP en injectant le produit radioactif au bon moment. L'acquisition de l'image se fait dans le 30 minutes qui suivent, et on obtient alors ce qu'on appele une image « ictale ». Une deuxième image, dite « interictale », est obtenue en dehors des crises et sert comme image de référence, à laquelle l'image ictale est comparée (simple soustraction). Pour comparer ces deux images, un recalage entre elles suivi par un recalage avec l'image d'IRM du patient sont nécessaires. Cette technique, connue sous l'acronyme « SISCOM », est la technique de référence à laquelle nous avons comparé notre technique d'atlas.

Nous avons fait évaluer six patients par un expert médecin. Celui-ci a constaté que les deux techniques donnent des résultats globalement similaires. Pourtant, quelques différences sont observables. Nous analysons ces différences et nous examinons les difficultés et limites liées à la technique de SISCOM. Ces problèmes sont liés à la difficulté d'obtenir des images qui sont véritablement ictales et interictales. Si l'injection du produit se fait trop tard par rapport au déclenchement de la crise, le foyer épilétogène risque d'être épuisé et, en conséquence, hypoperfusé à la place d'être hyperperfusé. Cela peut fausser l'analyse SISCOM. Souvent il est nécessaire d'acquérir plusieurs images ictales avant de pouvoir bien déterminer le ou les foyers. Au contraire, avec l'atlas il est possible de savoir si l'image ictale contient des zones hypoperfusées ou si l'image interictale contient des zones hyperperfusées. Nous montrons des exemples de ce type de résultats, ainsi qu'un résultat où l'analyse avec l'atlas permet de mettre en évidence une zone qui n'est pas détectée avec la méthode SISCOM.

À la fin de ce chapitre nous comparons aussi les résultats obtenus avec les différentes techniques de recalage et les différents modèles statistiques. Les différences entre les modèles sont moins importantes, mais il y a une différence marquante entre le modèle local et les modèles globaux. Il est difficile de constater quel est le meilleur, mais nous savons après nos études d'évaluation que les modèles globaux ont une meilleure précision. Concernant le recalage affine et déformable, nous constatons que le résultat obtenu avec le SISCOM et le

résultat obtenu avec l'atlas sont plus proches avec le recalage déformable qu'avec le recalage affine.

Dans le **chapitre 10** nous résumons la thèse. Quelques élements plus critiques sur le travail sont présentés et nous discutons les perspectives de notre travail. Celles-ci concernent d'un côté le modèle statistique (initialisation multiple et utilisation pour la détection en plus de la reconnaissance), et de l'autre côté les extensions pour prendre en compte la normalisation d'intensité et l'âge dans le modèle statistique ainsi que l'application de l'atlas à d'autre études cliniques.

En dehors des développements théoriques sur les modèles statistiques, une partie importante de ce travail a concerné le développement logiciel. L'auteur a contribué de manière significative au développement d'un logiciel libre de traitement d'images en C++, ImLib3D, qui est disponible gratuitement pour les autres chercheurs dans ce domaine [19]. Une autre librairie, gslwrap, qui facilite les calculs d'algèbre linéaire en C++ a également été développée dans un cadre collaboratif de logiciel libre[1]. L'atlas probabiliste a été implémenté en tant que module dans la plate-forme logicielle Medimax de l'IPB (Institut de Physique Biologique) et du LSIIT (Laboratoire des Sciences de l'Image de l'Informatique et de la Télédétection). Cette plate-forme est accessible aux médecins et chercheurs de ces 2 équipes de recherche. Une description de ces développements n'est pas incluse dans la thèse.

[1] `http://gslwrap.sourceforge.net`

Contents

Notation

We keep a coherent notation throughout this thesis. Mathematical symbols that occur recurrently are tabulated below to facilitate reading. As a general rule, matrices and vectors are typeset in bold, whereas scalars have a normal typeface. We do not make a notational distinction between random variables and their realizations apart from the case of noise/residual (see below). Transposes are denoted by the superscript T.

Vectors

Symbol	Signification	First reference
$\boldsymbol{y} = (y_1 \cdots y_D)^T$	images, observations in observation space	Sec. 3.3.1, p. 34
$\boldsymbol{x} = (x_1 \cdots x_Q)^T$	subspace variable, transformed variable	
$\boldsymbol{w}_q$	eigenvector or subspace vector	
$\boldsymbol{\epsilon}$	random noise variable	
$\boldsymbol{e} = (e_1 \dots e_J)^T$	residual (i.e. realization of $\boldsymbol{\epsilon}$)	
$\boldsymbol{t}$	z-score image	Sec. 6.3.3, p. 100
$\boldsymbol{b}$	robust weights in half-quadratic optimization	Sec. 3.5.3, p. 53
$\boldsymbol{\Theta}$	general parameter vector	
γ and β	parameters for linear regression and intensity normalization	Sec. 3.5.1, p. 51
c_j	constants in modified mean shift	Sec. 4.5.2, p. 68

Matrices

Symbol	Signification	First reference
$\boldsymbol{\Sigma}$	covariance matrix	Sec. 3.3.1, p. 34
$\boldsymbol{W} = [\boldsymbol{w}_1 \dots \boldsymbol{w}_Q]$	matrix of orthogonal vectors (e.g. eigenvectors or scaled eigenvectors)	
$\boldsymbol{Y}$	mean-free sample matrix	
$\boldsymbol{A}$	general purpose matrix	
$\boldsymbol{R}$	rotation matrix	Sec. 3.3.5, p. 39
$\boldsymbol{I}_Q$	$Q \times Q$ identity matrix	
$\boldsymbol{B} = diag(\boldsymbol{b})$	robust weights in half-quadratic optimization	Sec. 3.5.3, p. 53

Functions

Symbol	Signification	First reference
$\psi(\cdot)$	half-quadric expansion 1-dimensional	Sec. 3.5.3, p. 53
$\Psi(\cdot)$	half-quadric expansion D-dimensional	
$Q(\cdot)$	quadratic expansion in half-quadratic theory	

Indexes

Symbol	Signification	First reference
i	general purpose	
j=1...J	index of subjects, samples or images	Sec. 3.3.1, p. 34
d=1...D	observation space index (voxel index)	
q=1...Q	subspace index	
p=1...P	patches of local linear models	Sec. 3.4.5, p. 48
k=1...K	mixture components in mixture model, activation in activation studies	Sec. 3.4.5, p. 48

Abbreviations

Commonly used abbreviations used in thesis are tabulated below.

Abbreviation	Signification
IPB	Institut Physique Biologique
SPECT	Single Photon Emission Computer Tomogropaphy
MRI or MR	Magnetic Resonance Imaging
ECD	Ethyl Cysteinate Dimer
HMPAO	Hexamethyl Propylamine Oxime
FWHM	Full Width at Half Maximum
pdf	probability density function
i.i.d.	identically and independently distributed
LS	Least Squares
ML	Maximum Likelihood
MAP	Maximum A Posteriori
PCA	Principal Component analysis
PPCA	Probabilistic Principal Component analysis
EM-algorithm	Expectation-Maximization algorithm
ANOVA	Analysis of variance
ANCOVA	Analysis of covariance
MANCOVA	Multivariate analysis of covariance
ROC	Receiver Operating Characteristics
i.e.	that is
e.g.	for example

Part I

Introduction

Chapter 1

Introduction and overview

Single Photon Emission Computer Tomography (SPECT) is a 3D functional imaging technique that yields information about the blood flow in the brain (also called brain perfusion). This imaging technique has found application in the diagnostics of head trauma, dementia, epilepsy and other brain pathologies. To this end, SPECT images are analyzed in order to find abnormal blood flow patterns. For localized abnormalities such as stroke, this characterization remains an accessible task, whereas for diffuse and variable abnormalities such as beginning dementia, near-drowning episodes and toxic substance exposure, characterization is difficult. It is therefore necessary to develop quantitative methods in which computer-aided statistical analysis can take advantage of information present in a database of normal subjects.

The goal of this work has been the construction and evaluation of a probabilistic atlas of brain perfusion in normal subjects as observed in SPECT images. The purposes of such an atlas are twofold: (1) to describe perfusion patterns of the population represented by the atlas in a compact manner, and (2) to identify statistically significant differences between an individual brain perfusion pattern and the probabilistic atlas. The successful creation of a computerized, probabilistic atlas may have far-reaching impact on clinical applications where qualitative (visual) analysis of SPECT images is current practice.

1.1 Research environment

This thesis has been pursued in a cooperative setting between the two research groups *Models, Images and Vision* (MIV) at the *Laboratoire des Sciences de l'Image, de l'Informatique et de la Télétection* (LSIIT – UMR 7005 CNRS), and the image processing group at *Institut de Physique Biologique* (IPB – UMR 7004 CNRS). Both LSIIT and IPB are joint laboratories of the University of Strasbourg and the french national scientific research organization *Centre National de la Recherche Scientifique* (CNRS).

MIV is a group that is specialized in image processing and interpretation in general. It pursues fundamental research in computer vision and cooperates with other laboratories such as IPB or *Laboratoire Régional des Ponts et Chaussées* (i.e. regional roads and bridges laboratory – LRPC). The image processing group at IPB pursues image processing research applied to medical imaging. It bridges the gap between fundamental vision research and its application in medical research and clinical use. IPB is an interdisciplinary institute where not only physicians and biologists work together, but also psychologists, neurologists and other medical experts cooperate. In particular, there is a nuclear medicine facility affiliated with the institute that provides the nuclear imaging services at the university hospital of Strasbourg.

The SPECT images that have been studied in this thesis have been acquired at this nuclear medicine service.

Other works have been achieved in this setting. In particular, we would like to mention those of C. Nikou [171], O. Musse [165], and M. Bosc [17], as well as the ongoing works of V. Noblet (thesis) and S. Sinapin (technology transfer project, Plamaivic). Together with the works of Hamdan [87] (MIV) and R. Dahyot [46] (cooperation MIV – LRPC), these works form the precursors and the building elements on which the developments in this thesis are based.

1.2 Approach and general overview

Fig. 1.1 shows a problem-oriented view that can be used to describe the global approach taken in this thesis. As we shall see in the next chapter, SPECT images are difficult to interpret, even for a trained person. This is because these images are diffuse in appearance and because the anatomical variation between subjects is large, whereas the variation in perfusion (image intensities) at a given region of the brain may be subtle and difficult to describe and quantify. This is the reason why we have attacked the statistical modeling problem with methods for unsupervised learning. The main focus of this work has thus been the development of such methods, which are not limited to our particular problem, but also find use in general pattern recognition and image analysis applications.

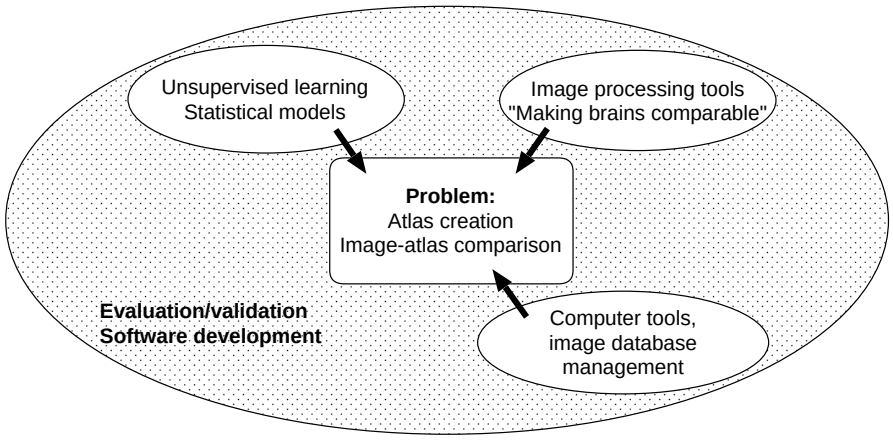


Figure 1.1: A problem oriented view of our approach.

Based on earlier work in our team (Dahyot [46], Hamdan [87]), we have been interested in a particular class of global linear models called *appearance-based* models. However, to prepare the images for statistical modeling, it is necessary to "make brains comparable", that is, to perform spatial registration and intensity normalization on brain images of different subjects. Image processing tools which makes this possible have been developed in earlier works (Nikou [171], Musse [165]) and ongoing works (Noblet, Sinapin). However, as we experienced, special adaptations had to be made.

Because of the large number of images and intermediate images resulting from different image processing steps, specific tools for managing these had to be developed. Finally, the last axis that has been central in this work is the systematic evaluation and comparison of the

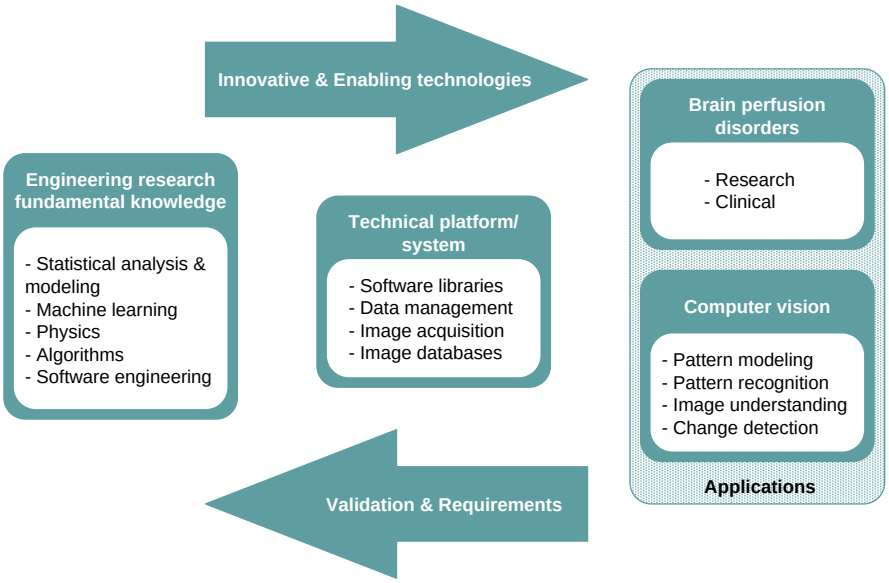


Figure 1.2: A system view of research and application, seen from the point of view of a computer scientist.

developed methods. Since no absolute knowledge about normal brain perfusion exists, this issue has been particularly difficult.

The importance of validation and evaluation can also be seen in Fig. 1.2. This figure shows an overview that relates fundamental research, technical platform and applications. This provides an alternative view for situating the analyses, developments and contributions that have been achieved during this thesis. In this overview, we see that an application, such as the creation of a brain perfusion atlas, is enabled through theoretical and fundamental methods and is realized by means of a technical platform/system. This relationship is the same for the application of technologies and knowledge to problems in computer vision. The arrows show the relationships between groups. For example, the clinical application of a method necessitates its validation. The validation again imposes requirements to the method. The method in turn requires a validation of the underlying algorithm (e.g. proof of convergence). Imposing such requirements can be very helpful in the development of new algorithms and methods. It may even give rise to new theoretical ideas. In the other direction we have that theoretical developments can enable new applications, and the circle is closed.

The basic idea in this work has been to use and eventually extend techniques that have been successfully applied in computer vision (appearance-based models) as well as algorithms, which existed as software libraries (registration), to address our application in brain perfusion disorders.

1.3 List of contributions

The original contributions of this work are:

- Bibliographical/theoretical:
  - An in-depth review of appearance-based models for image modeling. We develop the exact relationship between two popular and similar probabilistic models.
  - A novel, non-Gaussian appearance-based model and the associated algorithms. This model can be used for pattern recognition and image modeling in general.
  - A review of statistical models used for SPECT/PET brain atlases.
- Methodological:
  - A sophisticated registration scheme for inter-subject, SPECT-MRI matching, based on existing algorithms.
  - An intensity normalization technique for SPECT images.
  - A comprehensive evaluation study for comparing atlases and image processing techniques.
- Applicative:
  - The application of a non-Gaussian statistical model to the creation of a brain perfusion atlas.
  - Contributions to an open-source platform for medical image processing developed using modern software engineering techniques.

A more detailed description of these contributions is given in the conclusion of this manuscript (Ch. 10).

1.4 Organization of this document

The manuscript is divided into three separate parts. In Ch. 2 we present an introduction to SPECT imaging, the ECD radiotracer and their clinical application. This introduction familiarizes the reader with this imaging modality, some of its possibilities and limitations. This familiarity helps him to understand some of the later discussions on the difficulties encountered, particularly the registration and intensity normalization issues.

The **second part** is devoted to the theoretical developments of this work. These are based on appearance-based models. In Ch. 3, we present a state of the art of such approaches, their extensions and variations. We also review techniques for robust estimation (i.e. half-quadratic theory) and non-parametric density estimation/mode detection (i.e. Mean Shift). These elements compose the foundations of our original non-Gaussian model, which is presented in Ch. 4. Finding the exact maximum likelihood estimates of the model parameters of this model is difficult. Furthermore, a prerequisite to this phase of *learning* is what we call the *reconstruction problem*. This problem is defined and then solved for a series of increasingly more general models. This solution opens up the perspective of solving the learning problem to

which currently an approximative solution is used. To assess the model, a comprehensive classification experiment is performed on a standard computer vision database. This experiment is described in Ch. 5.

In **part three**, we turn our attention to the construction and evaluation of a probabilistic brain perfusion atlas using the above described statistical modeling framework. In Ch. 6, we first present an overview of the main processing steps: statistical modeling, image registration and intensity normalization. There are only few reports of the use of general-purpose, probabilistic atlases in the nuclear medicine literature. We propose a review of these together with some related methods that share concerns that are similar to ours.

In Ch. 7, we then proceed to a detailed description of our approach. An overview of the registration scheme and possible variations is presented. We detail the non-linear, inter-subject registration of SPECT images by co-registering each subject with its associated MR image and we describe the total least squares method used for intensity normalization. Finally, the different statistical models on which we have performed our evaluation studies are summarized at the end of chapter 7. They correspond to the appearance-based models presented in Ch. 4, part two.

Ch. 8 describes the atlas evaluation. We first discuss some general aspects of validation in medical image processing and review other work in quantitative SPECT image validation. We then detail our proposed evaluation scheme: leave-one-out, synthesized images and evaluation criterion (receiver operating characteristics - ROC analysis). The evaluation study was used to compare different models, registration schemes, intensity normalization and the dependency of the atlas performance to brain location. We present and discuss the results of these comparisons at the end of Ch. 8. Part three is concluded by Ch. 9, where we present preliminary results obtained by comparing images of patients with epilepsy to the atlas.

The thesis is summarized and concluded in Ch. 10 where we also discuss some possible paths for future research.

Chapter 2

Introduction to medical imaging and single photon emission computer tomography imaging

The purpose of this chapter is to acquaint the reader with the characteristic properties of SPECT images. We begin with a brief introduction to medical imaging and, more specifically, techniques for imaging the brain. This allows us to situate SPECT imaging and better understand its particular position among the many techniques that exist. We then give an introductory description of the procedure of acquiring SPECT images, how the radiotracer is distributed in the body and to the brain, how the emitted gamma-rays are transformed to voltage pulses, and how the transversal slices are reconstructed from projections. We then touch upon other issues like image interpretation, clinical applications of SPECT imaging with examples, as well as the important issue of radiation burden.

2.1 Medical imaging and imaging the brain

Medical imaging is an extraordinary example of multidisciplinary research. It originated with radiation physics and with the discovery of X-rays in 1895 by Roentgen. The application for medical purposes was immediate. Today, additional knowledge from medicine (anatomy, histology, physiology, pathology), biology and cytology (tissue, cells and their interaction), chemistry (radiopharmaceuticals) mathematics, signal processing and computer science, as well as high precision mechanics (rotating cameras) and electronics (computers, superconductors) have led to the development of a multitude of different techniques that are used on a routine basis. Today medical imaging represents a series of ubiquitous tools for diagnosis, treatment and medical research.

The multitude of techniques does not however mean that there is no need to continue research on medical imaging. Many questions still remain open concerning the way the body functions and how pathologies develop. In particular, the last 15 years have shown many advances in brain research and the associated development of brain imaging techniques. Such development aims at creating new imaging techniques that make it possible to "see" new or alternative aspects of the brain, improving the resolution of existing techniques, reducing costs of imaging, or reducing patient and personnel inconvenience (time of scan, radiation burden, invasiveness etc.). Furthermore, they bring new insights in physiology, biology and pathology.

In this presentation we do not intend to give a review of all imaging techniques applied in

Figure 2.1: Progress in medical imaging. On the left: Mrs. Roentgen's left hand (1895). On the right: modern X-rays.

medicine. For this, we refer the reader to [37] for a more technical understanding of the physical principles of MRI, SPECT, PET and ultrasound as well as the mathematical foundations for tomographic reconstruction of images from projections. For an introduction to the application and interpretation of such images, course material from radiology is adequate [98].

Imaging the brain helps us to improve our understanding of how the brain works and how pathologies develop. Research in this domain has led to more accurate diagnosis and better treatments of brain pathologies. The understanding of how the brain works has fascinated man for a long time, but it also poses philosophical and ethical questions. Consider for example the combination of marketing and brain research, "Neuroeconomy"[1], where a better understanding of how the brain works will be used to influence the habits of consommation.

2.2 Volumetric brain imaging

Modern medical imaging began with the development of computer tomography in 1972 by G. N. Hounsfield [103]. The reconstruction of slices and volumes from projections obtained from different angles around the body makes it possible to "see" inside the body. We can observe structures, physiological parameters and their relative positions in space. All images of the brain in this work are three-dimensional, either from stacking transaxial slices (SPECT) or from true reconstructed images in three-dimensions (MRI). See Fig. 2.2 for an explanation of *multiplanar* visualization and notation.

2.3 Anatomical and functional brain imaging

We can distinguish between anatomical and functional techniques for brain imaging. Anatomical imaging techniques make images of the brain tissue which mainly yield information as to the relative position of brain structures and organs. Functional techniques make images of some physiological process, and thus yield additional information on how the brain works. We shall describe functional techniques shortly in Sec. 2.5. Anatomical imaging comprises X-ray tomography and MR (magnetic resonance) imaging. By intraveneously injecting a contrast

[1]Frankfurter Allgemeine Zeitung, 05.11.2003, Nr. 257. See also: http://www.neuroeconomy.org

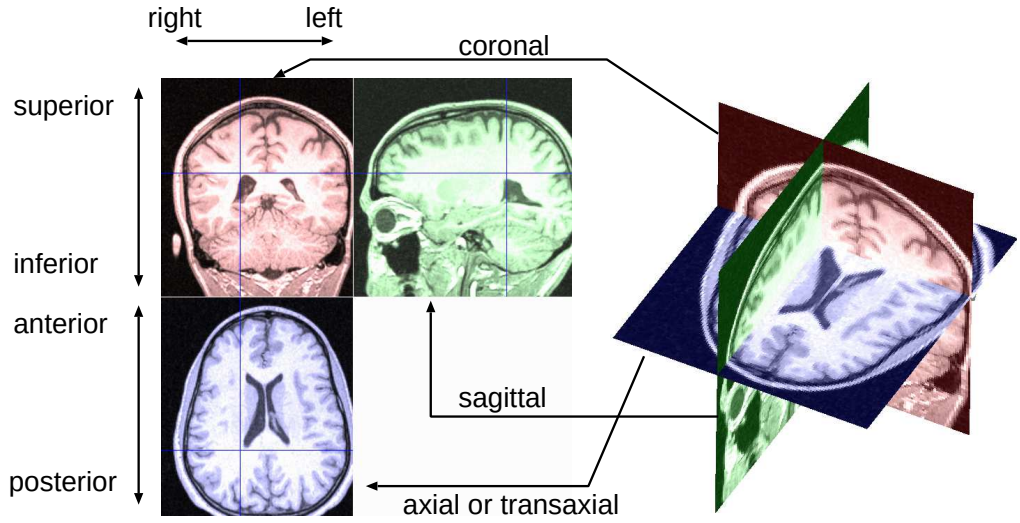


Figure 2.2: Multiplanar visualization of three-dimensional brain images and some basic notation for orientation. The three slices or planes are denoted coronal, sagittal and axial (or sometimes transaxial) slices. A cursor marks the intersection between the three planes.

agent, X-ray images can be obtained with enhanced contrast of certain features, for example, arteries. In MR imaging one can adjust a large number of parameters (impulse sequence) to obtain different contrasts between different molecules (water, lipides etc.). Fig. 2.3 shows an example T2-weighted[2] MR image, which are the type of MR images used in this work. The resolution of these images is $1\,\mathrm{mm}^3$. This is far from capable of imaging cells and neurons (~ 0.01 mm), but we can distinguish gray and white matter as well as the different macrostructures of the brain.

Anatomical images such as MRI provide important information in many pathologies such as multiple sclerosis, cerebrovascular diseases, dementia and cancer. This is particularly true when the disease leads to tissue changes (atrophies). However, some pathologies are not associated with morphological changes, or only at a late stage of the disease. Instead changes in physiological parameters such as regional cerebral blood flow (rCBF) can be symptomatic of a disease. In such cases functional images provide important insight. For example, in order to decide on the proper medication for patients with early signs of dementia, it is important to distinguish whether they have Alzheimer's disease or not. This can be done with SPECT (single photon emission tomography) or PET (positron emission tomography) imaging since characteristic patterns of rCBF in Alzheimer's disease distinguish the disease from other forms of dementia.

Functional images do contain some morphological or structural information about the brain, but typically at a much lower resolution than MR images and X-ray (CT) images. Thus, in order to precisely locate abnormal functional lesions, the functional and anatomical images are often fused together (superimposition), see Fig. 2.4. This is possible due to modern computer algorithms that find the relative locations of corresponding brain structures in two images obtained by different imaging techniques.

[2]T2 signifies a particular parameter setting for acquiring MR images.

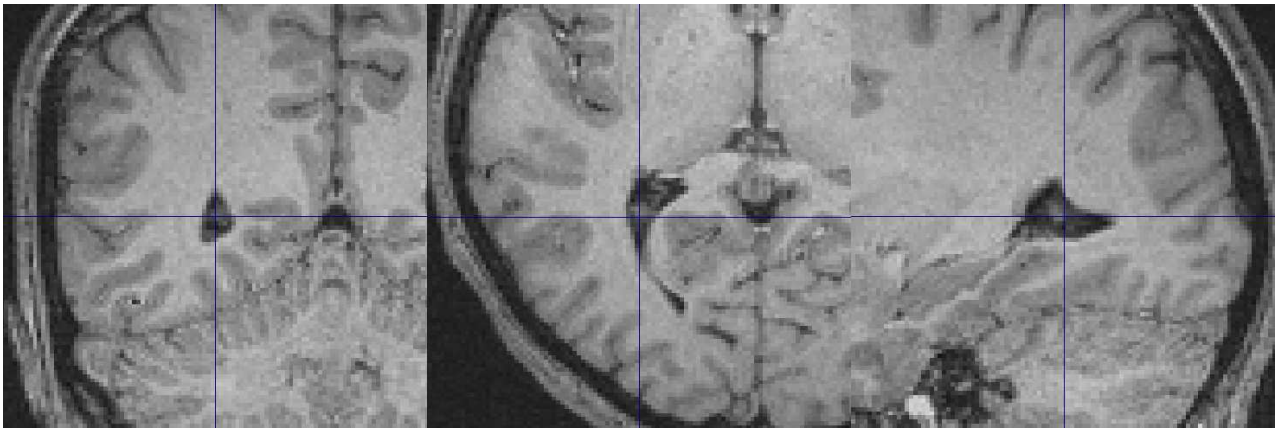


Figure 2.3: Example of a T2-weighted MR image of $1\,\mathrm{mm}^3$ resolution. Only a magnified part of the image is shown. This is what we call an anatomical image, gray and white matter as well as different macrostructures of the brain can be identified.

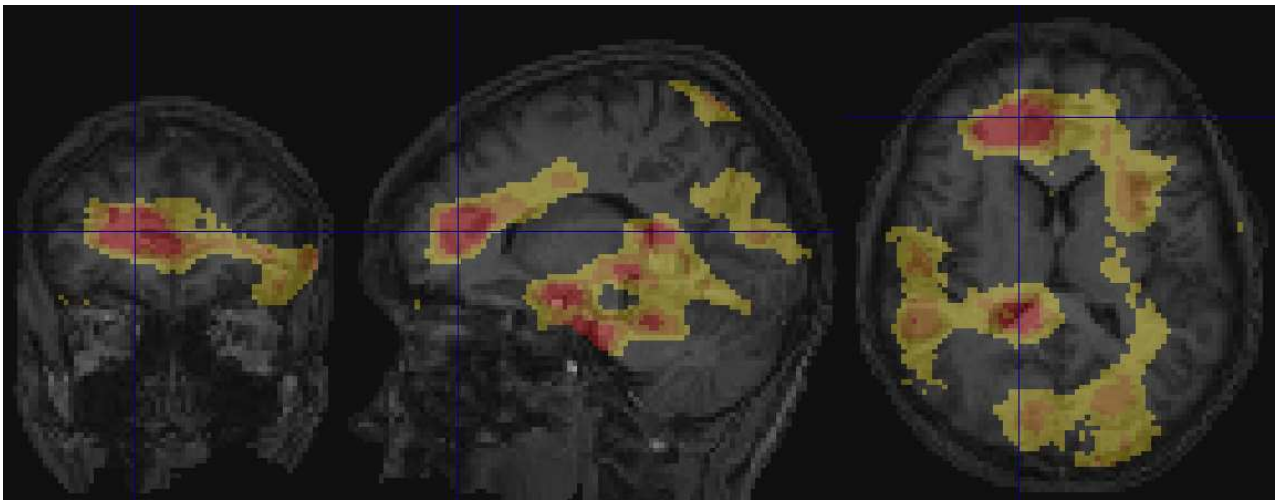


Figure 2.4: Example of an abnormal perfusion patterns (hot color coded) detected using SPECT. The pattern is superimposed on the T2-weighted MR image of the same patient for precise localization. A patient with epilepsy.

2.4 Brain function

In the literature, the term "brain function" is somewhat loosely applied to mean one or more of several things: (1) cognitive brain function (memory, planning, etc.) or sensomotorical tasks (audiovisual tasks, walking, etc.), (2) neurons and neuronal networks (topological organization, signaling etc.) or (3) biochemical or electrical activity (metabolism, glucose consumption, blood oxygenation level, etc.). When reading brain imaging literature, it is useful to be aware of these different levels of brain functions.

In this work we shall consider a functional image to be a macroscopic (with respect to neurons), spatiotemporal measure of one of the activities associated with the third group. These measures are related to cognitive or sensomotorical functions in different ways. For example the regional cerebral brain flow (rCBF) is believed to be linearly related to neuronal activity around a level of resting state [180, p.11]. The neuronal activity will depend on the way a cognitive or a sensomotorical task is organized. Here, two models exist: that of functional segregation (specialization) and that of functional integration [67]. Simplifying, one could say that the former considers that the execution of a specific task (for example finger tapping) is a result of neuronal activity in a cluster of neurons. The latter considers that a specific task results from the *mediation* of remotely located neurons. Which model is actually "correct", probably depends on the task at hand (e.g. most sensomotorical tasks are of the former type). There could also be a combination of both. The difference has consequences for the statistical analysis in so-called activation studies and will be further discussed in Ch. 6.

Higher cognitive tasks are modeled at a higher level of interconnected "areas" (language area, audio area, etc.) and their cooperation. The connections between image analysis, statistics and neurology leads to the different usages of the expression "brain function". However, from a medical viewpoint, what is important is that different pathologies can manifest themselves as abnormal functioning of one or more of these types. Furthermore, such abnormalities can either be seen in the images of brain function, in neuropsychological tests of brain function, or (mostly) in a combination of both.

2.5 Overview of functional brain imaging techniques

To situate brain SPECT imaging, we briefly describe the different *non-invasive* imaging techniques used for functional brain imaging. Many techniques are complementary in their clinical accessibility and the biochemical characteristics they image. Furthermore, there is a tendency to trade-off between spatial and temporal resolution with the different techniques. A comparison of different functional imaging techniques and their temporal and spatial resolving power is shown in Fig. 2.5. We can distinguish between ionizing techniques (PET, SPECT), functional magnetic resonance imaging (fMRI) and techniques based on measuring electromagnetic potentials (MEG/EEG).

Positron emission tomography (PET) and single photon emission computer tomography (SPECT) are ionizing techniques. Here, the image is created by detecting nuclear radiation which is emitted from the brain after injecting a radioactive pharmaceutical. The pharmaceutical is a molecule that imitates a substance which is implicated in a specific biochemical process, for example glucose or oxygen consummation. The pharmaceutical contains a radionuclide that emits either positrons (PET) or gamma rays (SPECT). These are then detected with a gamma camera. Examples of radiopharmaceuticals used in brain SPECT is HMPAO (Hexamethyl Propylamine Oxime) and ECD (Ethyl Cysteinate Dimer), the latter will be described

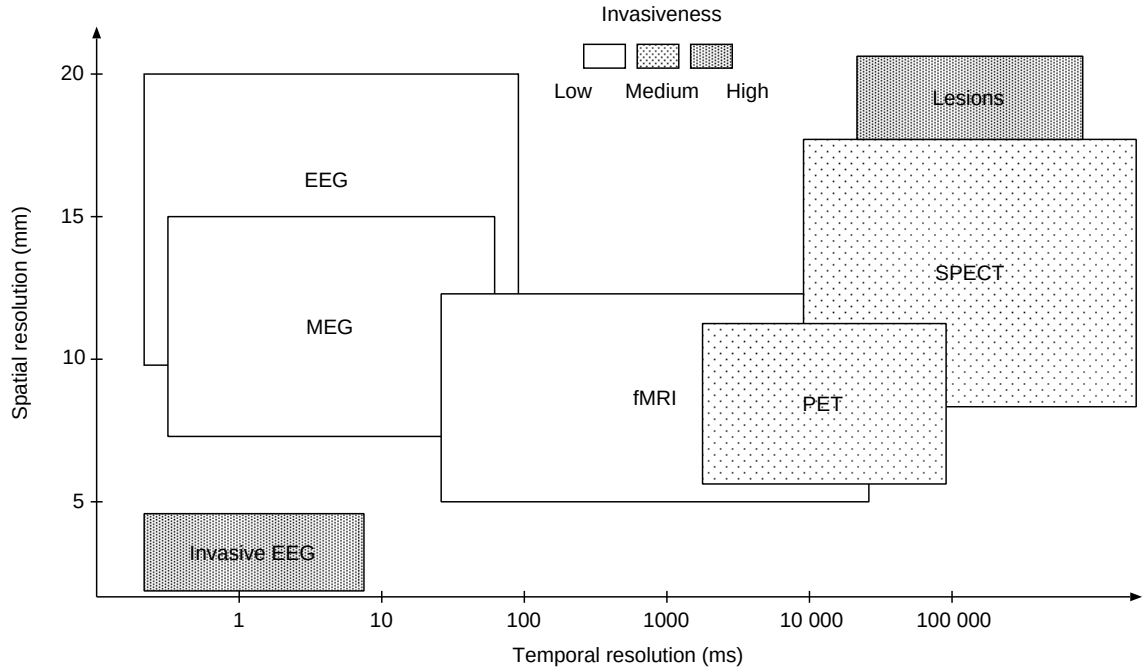


Figure 2.5: Overview of functional imaging techniques of the brain and their spatial and temporal resolving power (from [93]). The term *functional* can be interpreted at different levels. It can mean a biochemical functioning (often in resting state), or it can mean a cognitive function, such as recognizing words or images. Typical lesions shown for reference.

in more detail in 2.6.2. Because of better resolution, sensitivity and homogeneity, PET imaging is considered to be superior to SPECT imaging. There are however distinct applications where SPECT imaging is more appropriate than PET. In addition, a small cyclotron is necessary to create PET images because the radionuclides used in PET are short-lived The instrumentation for PET is thus substantially more complex and more costly and as such limited to a few imaging centers.

Functional MRI (fMRI) is based on the measurement of the oxygenation level in the blood supplying the neurons. Activity in a group of neuronal cells augments the consumption of oxygen, which leads to changes in the concentration of oxygen carrying haemoglobin in the blood. Since the magnetic properties of the haemoglobin with and without oxygen are different, these changes can be measured using magnetic resonance imaging. Functional MRI is most widely used in research in order to map cognitive brain function in the normal and pathological brain. Because of the low signal-to-noise ratio, such studies are done by acquiring a series of images where the subject repeats a cognitive task. This time series is then analyzed with statistical tools. This is a so-called activation study, to which we shall have more to say in Sec. 6.2.3.

Magnetoencephalography (MEG) and electroencephalography (EEG) can measure magnetic induction outside the scalp or electric potentials on the scalp produced by electrical activity in groups of neural cells. This activity is again a result of the functional level in the brain. MEG and EEG have very high temporal resolution, but their spatial resolution is less than or equal to that for SPECT.

As a new emerging technique, diffuse optical tomography (DOT) [15], might also find application in measuring brain blood volume and fast changes therein, however it has lower

Figure 2.6: Photo of the gamera camera (left) used at our institute and the control room (right). The camera has two heads that are controled by a motor (behind). The patient is placed on the bed with an additional support for his head. The camera heads are then positioned as close as possible to the patient's head. During acquisition, the camera heads rotate around the axis of the patient and the patient is asked to remain still. The acquisition procedure can be programmed and for brain SPECT, two 180 degrees rotations of the cameras are made with acquisitions (projections) at every 4 degrees for 10 seconds. The total acquisition time is about 20 minutes.

spatial resolution than fMRI. For this, DOT has still to be finalized for clinical application.

2.6 Description of the SPECT imaging procedure

2.6.1 Overview, the procedure

In order to better understand what the SPECT images represent, let us describe in some detail how they are obtained and the physics behind their creation. The brain SPECT procedure follows a standard protocol (similar to the guideline proposed by the Society of Nuclear Medecine [177]) to assure that the image acquisition conditions rest as stable as possible between scans.

A picture of the gamma camera that is installed at the service of nuclear medicine at the institute of physics and biology (IPB) is shown in Fig. 2.6, and an overview of the different time-steps of the imaging procedure is shown in Fig. 2.7. The injection itself is administered when the subject is seated in a comfortable position in a quiet room, keeping his eyes open and relaxing. After injection, the subject remains seated for five minutes. Image acquisition only starts after about 20-30 minutes. The subject lays down on a table with a support for the head. The double-headed camera is positioned so that the rotating center allows the cameras to be as close as possible to each other without touching the patient. During the acquisition itself, which takes another 20-30 minutes, the subject is asked to move as little as possible. The whole procedure takes about an hour for the patient. After the projections have been acquired, the image is reconstructed by the computer before it can be transfered to the

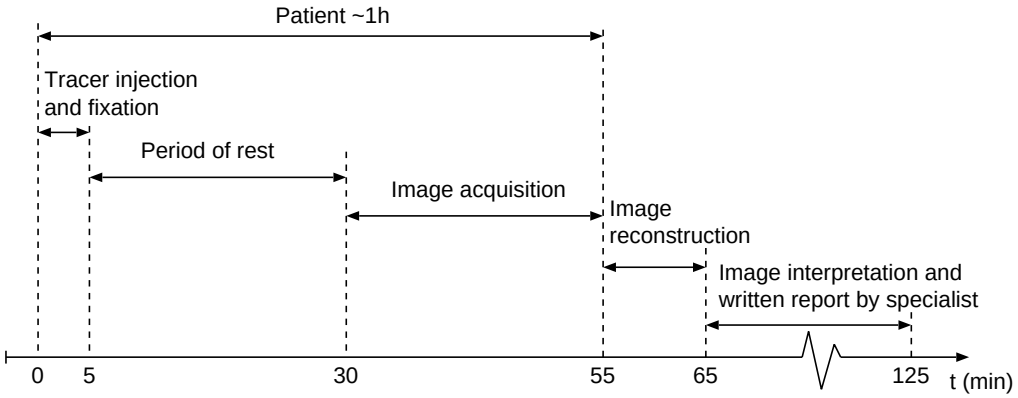


Figure 2.7: Timetable (approximate) and overview explaining the SPECT imaging procedure. For the patient, the procedure takes about an hour after the tracer has been injected. It is a prerequisite that the patient is capable of remaining immobile during the time of image acquisition (mostly 20 minutes in clinical practice, sometimes 30 minutes for research protocols).

physicians workstation for interpretation.

In the following, we will describe these different steps and their physical and physiological properties.

2.6.2 Injection: biodistribution and physical properties of the Tc-99m ECD radiotracer

A radiotracer used for medical imaging must have properties that fulfill several criteria. These can be separated into physical, biological and chemical. Physical properties concern aspects that touch upon technicalities of the imaging technique such as statistical properties of disintegration (half-life) and the wavelength of the emitted photons. These aspects are important for practical concerns since a long half-life makes it possible to prepare the radionuclide in dedicated centers, and photons are easier to detect at certain wavelengths than others. The biological properties concern the distribution of the product in the body and the target organ, how it is cleared out (and the associated burden of radiation for the patient), and finally how the uptake relates to the physiological function of the organ under study (e.g. whether there is a linear relationship or not). It is under these constraints that a chemically stable molecule must be found that can serve as a tracer.

The radiopharmaceutical used at IPB for brain SPECT imaging, Technetium-99m ethyl cysteinate dimer (ECD), also called bicisate, is one such molecule. The radionuclide Technetium-99m is the main tracer for clinical imaging, and it was among the first tracers to be clinically used. It has a half-life of 6.03h and emits photons of 140 keV that are easily detectable using a gamma camera [37]. Technetium-99m is obtained from desintegration of Molybden-99, which has a half-life of about 66h. Because of this relatively long half-life time, Molybden is delivered only once a week to the institute. Using another assembly kit (Neurolite® from DuPont Pharma), doses that are ready for injection are prepared on-site. Typically, two different quantities are used, one for adults and a smaller quantity for children.

The ECD radiotracer (Bicisate) is an indicator of cerebral blood flow. When injected, the tracer is rapidly distributed to the brain [216, 141]. This happens because the bicisate is trapped in the brain cell after metabolism, subsequent to its crossing of the blood brain

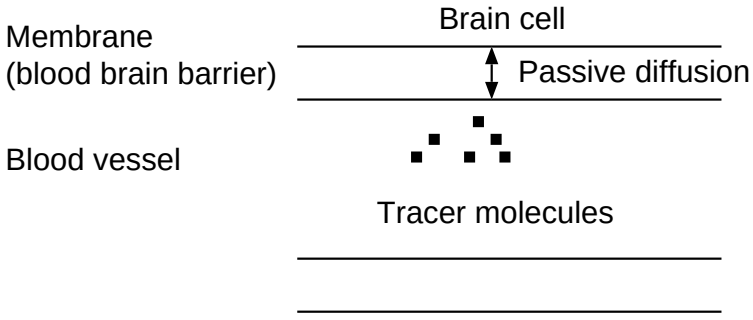


Figure 2.8: The brain vessels are equiped with a selective membrane (*filter* in technical terms) – called the blood brain barrier – through which only certain molecules can pass (glucose, certain proteins etc.). The molecules react chemically with a transporter molecule, water or fat, that passes through the barrier by passive diffusion. When the transporter molecule is a fat (water), the molecule that is transported is said to be a lipophilic (hydrophilic). The TC-99m ECD tracer is a lipophilic. Once the ECD tracer has entered the brain cell it undergoes a biochemical transformation (ester hydrolisis) and becomes a charged acid metabolite. This metabolite is unable to exit the brain – it becomes trapped.

barrier (by passive diffusion), see Fig. 2.8. The concentration of the tracer in the blood drops to less than 10% of the inital dose after only one minute.

Bicisate has the desirable property of washing out very slowly from the brain (biexponential with half-lives of 1.3 hr (40%) and 42.3 hr (60%)), contrary to other tissues, from which it is cleared out quickly. In particular, face tissue, neck and scalp is cleared rapidly so that the brain to crane signal is high. This is why image acquisition only starts about 20-30 minutes after injection. The product is finally cleared from the body by the renal and the hepatobiliary system.

Studies have shown that the uptake of bicisate is proportional to the regional cerebral blood flow (rCBF) [42]. The uptake is normally greater in the cortical grey matter where the blood flow is higher than in the white matter. The ratio of grey matter to white matter uptake is normally greater than 2:1.

Radioactive disintegration: Poisson noise

The radioactive disintegration is a random process and the detected number of gamma photons within a period of time follows a Poisson distribution [90]. The number of photons captured over a constant period of time follows the equation:

$$P(k) = \frac{\mu^k e^{-\mu}}{k!} \tag{2.1}$$

Here, $P(k)$ is the probability of emitting k photons and μ is the (usually unknown) mean. The variance of the Poisson distribution, σ^2, is equal to the mean μ and implies that the signal to noise ratio (SNR) is equal to the square root of the mean:

$$SNR = \frac{\mu}{\sigma} = \frac{\mu}{\sqrt{\mu}} = \sqrt{\mu}$$

This means that the image SNR is higher for higher count levels (gray levels in the image) than for lower levels. It is therefore desirable to increase the photon counts as high as possible. This

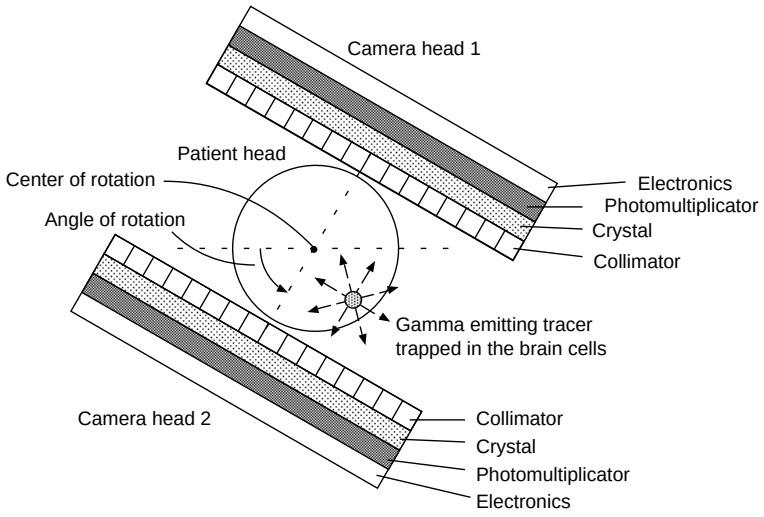


Figure 2.9: Block-diagram illustrating image acquisition. The two camera heads rotate around the subject's head in a step-wise fashion, collecting photons for a fixed amount of time at each angle. Since the tracer emits photons isotropically in all directions, a certain number of photons are not detected by the camera. Positioning the camera heads as close as possible to the patient's head, limits this loss of sensitivity.

is done by keeping the acquisition time as long as possible – at the risk of patient motion – and by keeping the amount of the administered dose as high as possible, without dangering the patient's and personels' health.

Other tracers

Many other tracers also exist for imaging the brain. Every tracer has its advantages and disadvantages. For an overview see for example [99]. Besides the ECD tracer the most common tracer in routine clinical use for brain SPECT imaging is the Tc-99m hexamethyl propylamine oxime (HMPAO) tracer. It has quite similar properties to the ECD tracer. The main advantage of the ECD tracer over the HMPAO tracer is that it is stable for a much longer time which facilitates on-site preparation.

2.6.3 Image acquisition

The image is constructed from a series of *projections*. A projection is obtained by capturing gamma-rays for a fixed amount of time using a scintillation camera that is positioned with a specific angle of view with respect to the subject, see Fig. 2.9. Projections at successive angles are obtained so that all 360 degrees are covered. For routine SPECT examinations, two projections (by rotating the cameras twice around the patient) are acquired at every 4 degrees of view for 10 seconds. Compared to an fMRI acquisition, a SPECT image acquisition is more quiet and more comfortable for the patient.

Scintillation camera

The main constituents of the camera are shown in Fig. 2.9. In the camera, the energy of the gamma ray is transformed into a voltage pulse, which in turn is measured by an analog

electronic circuit. The camera consists of a collimator, a scintillating phosphor (crystal) and photomultiplicating tubes that are connected to the electronic circuits. The purpose and functioning of the camera can briefly be described as follows. For more in-depth material on the physical properties of the scintillating camera and other types of gamma-cameras, the reader is referred to [37].

1. The photons that are emitted isotropically from within the subject are mechanically collimated. The collimator is usually a plate made of lead that absorbs photons that are not aligned with the holes drilled in it. Collimation is necessary in order to know the direction from which the photon was emitted.

2. The high-energy (gamma) photons that pass the collimator and enter the scintillating phosphor, lose some of their energy in the collision and excitation of molecules in the crystal. These in turn, emit optical photons (visible light, scintillation) when they return to the ground state. The intensity of the scintillation is proportional to the energy lost by the gamma ray in the crystal.

3. The scintillation light is then guided toward the cathodes of the photomultiplier tubes where they are converted to electrons by means of the photoelectric effect. The electrons are multiplied in their flight toward the anode where they give rise to a voltage pulse.

4. The analog and digital electronic circuits measure the output voltage pulses from the photomultiplier anodes and estimate the position of the incoming gamma-ray.

Image quality

There are many sources that influence the quality of a SPECT image. These have been tabulated in Tab. 2.1. For more detailed description of these factors, we again refer to [37], or alternatively [124]. Let us just briefly describe two. First, attenuation, which is caused by the absorption of photons in the head of the subject, depends on the distance of the cameras from the source of radiation. This attenuation is therefore to some degree compensated by the fact that two cameras are used. Further compensation can be made by using specialized reconstruction filters or reconstruction algorithms.

Second, Compton scattering, where a gamma ray interacts with a free electron in the brain and changes direction (after loosing some of its energy), leads to lack of sharpness in the images. To reduce the effect of Compton scattering it is necessary to have a camera with a high energy resolution: photons with less than 140 keV (such those issued from a scattering event) can then be filtered out. It is also possible to model the Compton scattering during image reconstruction when using iterative image reconstruction schemes (see below).

Modeling the image acquisition process

In a first approximation and for convenience, the image acquisition filter is often modeled as a Gaussian curve distribution. The images acquired at IPB have a full width at half maximum (FWHM) of about 8mm. However, a certain number of simulators have been developed that offer more accurate modeling of the image acquisition process. Efforts to standardize and compare these can be found in [26].

Principal source	**Category**	**Factor or effect**
Patient	Anatomy	Body Size
		Anatomical structures
	Time dependency	Tracer distribution
		Movements
Physical phenomenon		Attenuation (absorption)
		Compton scattering
Technical	Instrumentation (Camera)	Response
		Efficiency
		Time resolution
		Energy resolution
		Spatial resolution
		Uniformity, linearity
	Acquisition	Number of projections
		Time of acquisition
		Mechanical precision
		Distance from object
	Reconstruction	Algorithm
		Error compensation
		Image processing

Table 2.1: An overview of factors influencing the quality of SPECT images (from [124]).

2.6.4 Image reconstruction

After the acquisition of projections, it is necessary to reconstruct the transversal image slices. The most widely used technique in clinical practice is the classical filtered backprojection algorithm [37]. It has the advantage of being fast and necessitates little user interaction. The operator only defines the orientation of the slices that are to be reconstructed as well as the low-pass filter that is used in connection with the reconstruction filter (ramp or derivative filter). Other methods also exist. For example so-called algebrical or discrete methods that lead to iterative reconstruction algorithms, [124, 25]. The advantage of these algorithms is that some of the error sources (such as attenuation, Compton scatter and camera response) can be explicitly modeled and taken into account during reconstruction. All the SPECT images that we consider in this thesis have been reconstructed using the filtered backprojection algorithm.

2.6.5 Image interpretation

Once the images are reconstructed, they are analyzed by a physician. For this, all 32 slices of 64×64 image size are displayed on a computer screen simultaneously using a color palette for coding the image intensities (see Fig. 2.10). This color scale can be regulated interactively, thus presenting an image intensity normalization. When interpreting the image, the physician is guided by his experience, knowledge about anatomy and the pathology under suspicion, as well as the patient record. He will sometimes judge that certain lesions of high or low values in the image are significant, whereas others are not.

Whereas gross anomalies in the images are readily detected, subtle deviations from normal intensity (and blood flow) values are more difficult to detect. The physician therefore often looks for asymmetries in the right and left hemispheres. Using the counterpart of a structure in the opposed hemisphere as a reference, he can often find abnormal blood flow levels. However, it may be difficult to know whether asymmetrical values are caused by a deficit in one side or an augmentation in the other. When abnormalities are present on both sides (bilaterally), the abnormality may go undetected. In these cases, it is important to have absolute measures of the blood flow in order to compare to normal values. The absolute gray value of a voxel value can have an equivalent physiological interpretation such as regional cerebral brain flow as measured in ml/min/g (blood per time per tissue). This is known as *quantitative imaging*. Several measures are aimed at making the image intensity values as comparable as possible: (1) keeping imaging parameters as constant as possible from scan to scan (acquisition time, time between injection and acquisition etc.), (2) reducing the influence of acquisition errors during or after image reconstruction (see above), and (3) calibrating the images by using an intensity normalization technique. The important issue of intensity normalization shall be discussed in detail in Secs. 6.6, 7.5 and 10.2.2 of this thesis.

2.7 Radiation burden

Radioprotection is an important issue in SPECT imaging. The imaging facilities are regularly inspected and controled by the authorities in order to monitor the radiation burden to which the personel and patients are exposed. SPECT imaging is however a safe procedure. The injected dose for brain SPECT is 925 Bq (Becquerel, disintegrations per second). The effective dose (the energy (J/kg) of the radiation corrected for the biological effect of this radiation) that is received by a patient that weights 70 kg is for this injection 5.4 mSi (Sievert). This is less than twice the annual dose received from the natural background radiation, which is

usually (depending on the geographic location) between 2-3 mSi. This is also about the same dose that a patient receives during a X-ray lung scan.

2.8 SPECT atlases for educational purposes

In order to classify a SPECT image of brain perfusion as normal or abnormal, one has to know what normal brain perfusion is. Vice versa, it is difficult to say what normal is without relative comparison to what is not normal. For the education of nuclear medicine specialists, SPECT brain atlases with examples of normal SPECT images and SPECT images in different pathologies have been created (Fig. 2.10 is taken from one such atlas). Some of these atlases are accessible on the Internet[3] [99]. The description of these images and their variability is however only qualitative. We shall come back to statistical models and atlases used for functional brain imaging in Ch. 6.

2.9 Clinical applications

There are many clinical applications where brain SPECT imaging is used. These include depression, lyme disease, chronic fatigue syndrom, Alzheimer and other dementia, epilepsy, stroke, drug and alcohol abuse [99, 3, 2]. Two examples of typical findings in epilepsy and depression are shown in Fig. 2.10.

2.10 SPECT studies for brain research

SPECT imaging is also used in brain perfusion research. Here, the situation is often difficult because there is less a priori knowledge about the function or pathology that is being studied, and of the particular, subtle changes that may be involved (large changes are of course easily found). The need for statistical pooling is therefore even more important here. Fig. 2.11 shows a diagram taken from the introduction to the international symposium on quantification of brain function using PET, Oxford, England, 1995 [116]. The diagram recapitulates the many factors that interrelate in the image formation process aimed at answering research and diagnostic questions.

Though designed for PET images, the issues arising in quantitative SPECT studies are the same (with the addition of "Collimator" to "Collection of Scan Data", and changing "Kinetic studies" to "Multi-subject studies"). It is clear that all these issues influence the image quality and may show mutual dependencies. For example, the choice of tracer or reconstruction algorithm has an influence on the statistical properties of the images and thus the statistical evaluation. Given the high number of influencing factors, it is also clear that any statistical model can only be approximative. Ideally, such models must take into consideration as much as possible the different physical and physiological factors affecting the images and the study.

[3] `http://www.brainplace.com/bp/atlas/default.asp`, `http://brighamrad.harvard.edu/education/online/BrainSPECT/BrSPECT.html` and `http://www.unice.fr/html/ATLAS/`

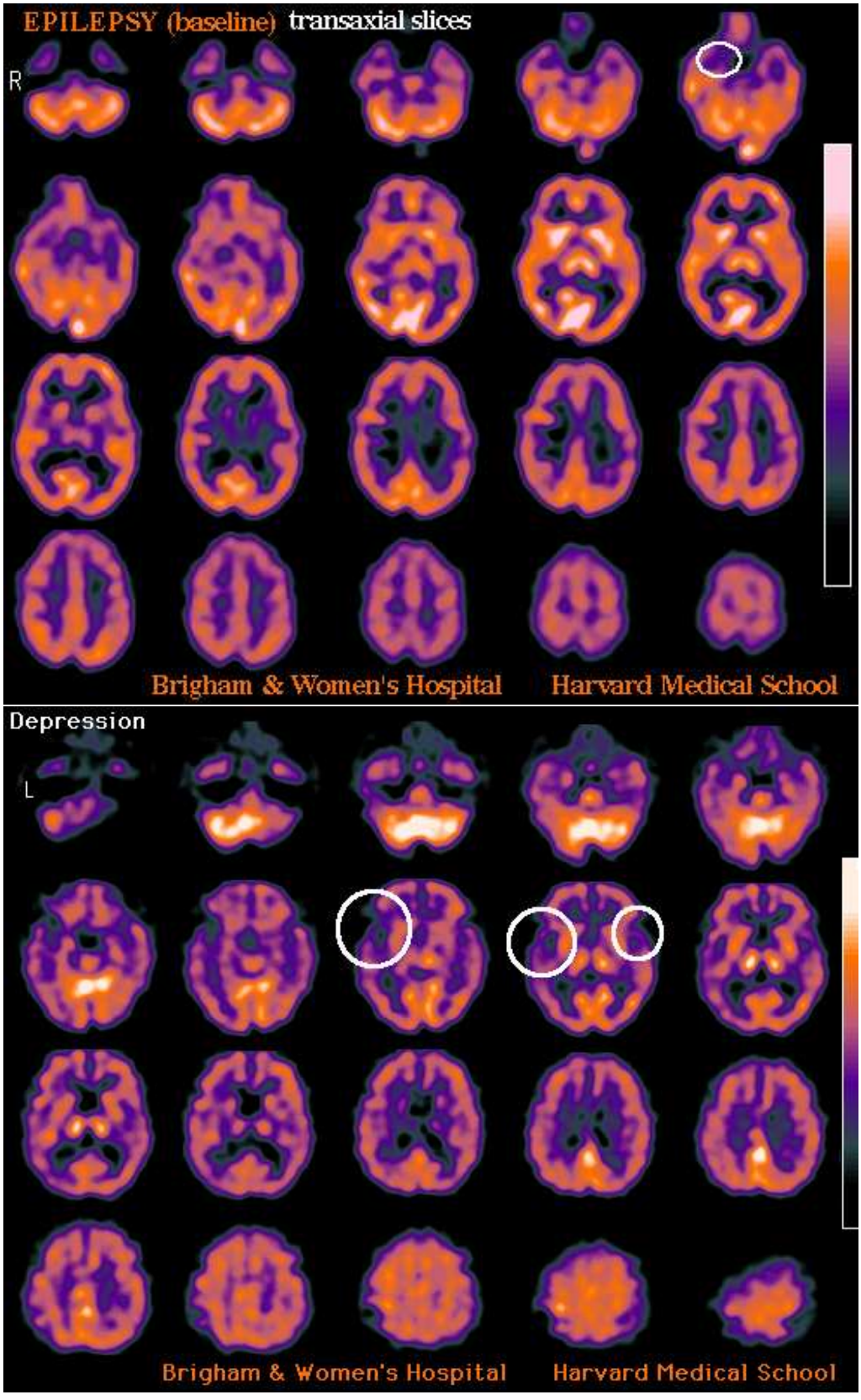


Figure 2.10: Examples of blood flow deficit as seen in HMPAO SPECT images of a patient with epilepsy (top) and a patient with depression (bottom) (taken from [99]). The images are qualitatively equivalent to ECD SPECT images.

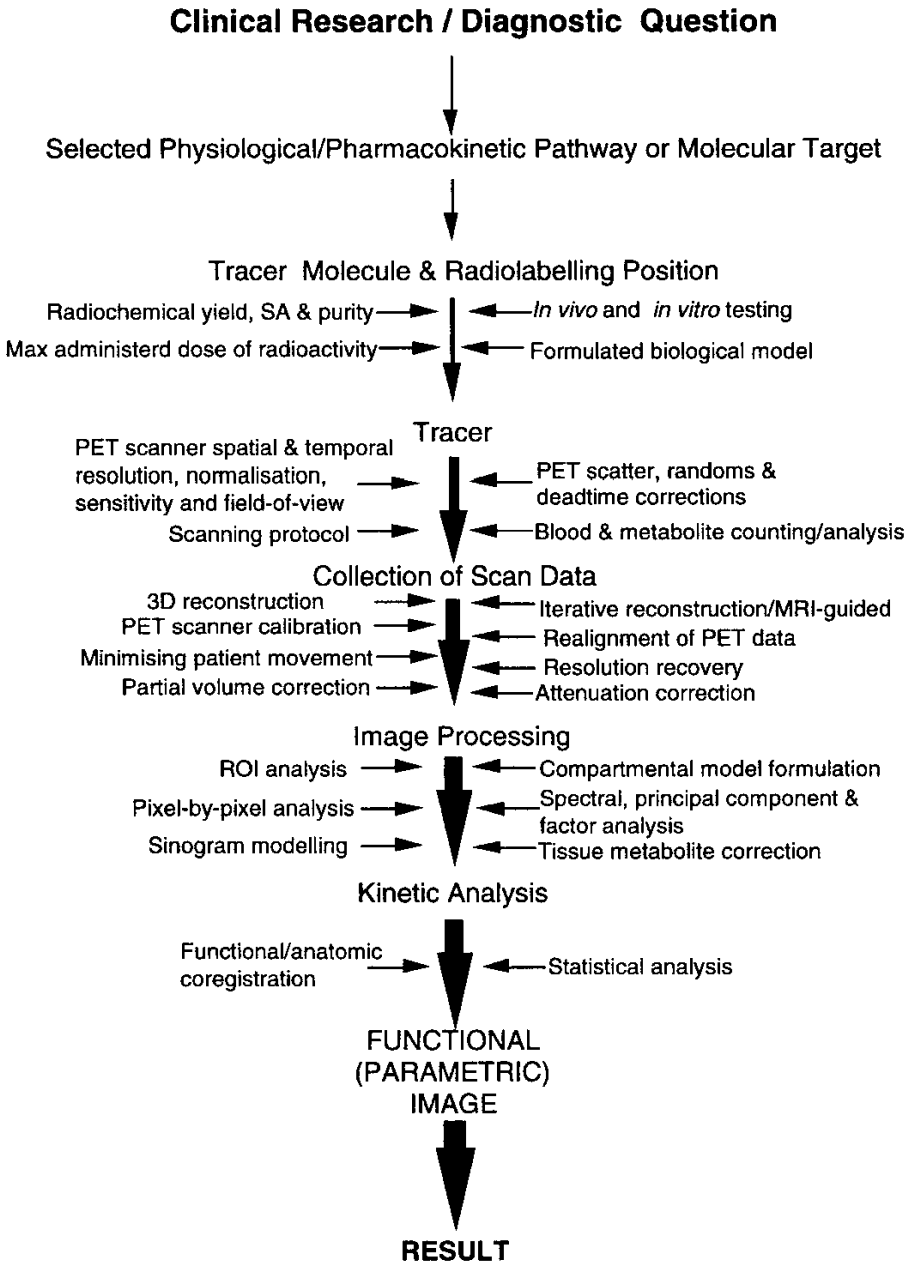


Figure 2.11: Overview of methodological flow, starting with the need to answer clinical research and diagnostic questions, ending with the formation of functional images (from [116]). The figure was devised for PET imaging, but is equally valid for SPECT imaging.

2.11 Conclusion

In this introductory chapter, we have tried to give an overview of the SPECT image formation process. The purpose of this chapter is to give a reader that is unfamiliar with this image modality some basic notions about these images and what they represent. An understanding of the image formation process is necessary in order to understand what we actually see in SPECT images. Contrary to usual image processing problems, these images are difficult to interpret. This is because they have a low spatial resolution and because their interpretation necessitates additional expert knowledge. These difficulties complicate SPECT image processing, but also make the application of exploratory data analysis interesting. There is information to discover in SPECT images that cannot be easily found by a human.

Because of the low resolution of the SPECT images, we shall in a later chapter use the high-resolution MR images in order to register images from different subjects (spatial normalization). How this is done is explained in Chs. 6 and 7. The fact that the ratio of the image intensity of grey to white matter in SPECT images normally is greater than 2:1 makes it possible to perform a simplified simulation of SPECT images from MR images. First, the MR image is segmented into white and grey matter, then intensity values of for example 100:50 is attributed. Finally a low pass filter can simulate the image acquisition process. This kind of simulation is used in Ch. 7 to evaluate the influence of difference registration schemes and to better understand the difference of anatomical and functional variation.

Furthermore, because image intensities observed in a SPECT image are not quantitative, the issue of image intensity normalization is particularly delicate. The information contained in the superimposition of two images (spatial normalization) can be used with advantage to improve such normalization as explained in Chs. 6 and 7 where this issue is also further discussed.

Finally, it would be desirable in future work to sources of image errors into account into the model (Poisson noise, scatter and attenuation errors, etc.). This has not been done in this work. Typically, these effects are modeled during tomographic reconstruction. We have only treated the reconstructed images as is, using learning-based approaches to capture observed variation as described in the part II of this thesis.

In this part II, we shall turn our attention to statistical models for image modeling and pattern recognition. We develop a novel statistical model that we first assess on a standard computer vision database. In Part 3, we then come back to the processing and modeling of SPECT images for the purpose of creating an atlas of brain perfusion.

Part II

Appearance-Based, Probabilistic Image Modeling

Chapter 3

State of the art

3.1 Introduction, overview

We have been interested in using appearance-based models for modeling images. The methods for such modeling have been applied with much success since the nineties, beginning with the works of Sirovich and Kirby [195] for compression of face images and Turk and Pentland [213] for recognizing faces in images. These methods are based on a Karhunen-Loeve transformation of the observed images. This transformation is found by solving the eigenproblem for the covariance matrix of representative images, which yields the name of the method: *eigenfaces*. Since the eigenvalues to this eigenproblem can be interpreted as the variances along the corresponding eigenvector, one retains the principal energy of the representative images by keeping only information along the eigenvectors of the highest eigenvalues. This is called principal component analysis and is effectively a dimension reduction technique since the retained eigenvectors describe a subspace of the original observation space.

We thus have an alternative to the classic approach of modeling objects as seen in images. Instead of extracting features and modeling these, one instead tries to capture interesting structure in the data using a dimension reduction technique directly on the entire image. These methods are therefore called global methods. The work of Turk and Pentland inspired many new ideas in this domain: other applications (object detection, object tracking in image sequences), other techniques for dimension reduction (linear and non-linear), behavior of object appearance (object rotation, illumination changes), additional modeling (robust, subspace modeling) as well as probabilistic bases for such models.

Two leaps forward in this domain are particularly relevant to our work. The first is the work of Murase and Nayar [163], where the authors introduced a non-linear modeling of the images in the subspace spanned by the eigenvectors. This modeling permitted the authors to boost the performance of recognizing objects (and object pose). The second is the work of Moghaddam and Pentland [162], where a complete probabilistic model based on principal component analysis is developed. This complete model permitted the authors to boost, among others, face detection performance. Another major contribution came from Black and Jepson [13]. They introduced robust estimation techniques that made the models robust toward occlusions and cluttered backgrounds, which is often encountered in natural (outdoor) scenes.

Because of the central position of these methods in our work, we devote a section to the *eigenfaces era* where we review the chronological developments associated with these models. We then proceed by presenting an overview of dimension reduction techniques in general, before other important developments relative to our work - robust noise modeling and non-

parametric density estimation - are discussed at the end of this chapter. Global appearance modeling, principal component analysis, non-parametric density modeling and robust noise modeling are the building blocks of the model presented in the next chapter.

Note that in this part, we are not concerned with the creation and application of a brain perfusion atlas. We shall see in part III how appearance-based models can be used for this purpose. The model we develop in Ch. 4 is quite general in scope. In order to illustrate the advantages and the versatility of this model, we perform experiments on a standard computer vision database in Ch. 5.

3.2 General background

Before describing the eigenfaces methods in detail, we will touch on some of the ideas that lie behind the approaches we have pursued in our work. These ideas are of a more general nature. Concrete and specific cases follow in the next section. For the sake of clarifying the vocabulary, we begin with a definition of the word *appearance*.

3.2.1 Appearance

What is the appearance of an object? Cambridge advanced learner's dictionary provides the following definition: *the way a person or thing looks to other people*. Appearance is the visible aspect of an object or a person. This aspect depends on the object itself and how it reflects light. The appearance of an object, as we capture it with a camera, is therefore a combined effect of object shape, reflectance properties, pose, as well as illumination conditions. The appearance is a notion that is related to an object in its totality, e.g. the appearance of a face, a hand, a car etc. We therefore consider an appearance to be a global property of the object. However, we often distinguish between local and global object appearance in images. By *global appearance* we understand the appearance of the object in its entirety, by *local appearance* we understand the appearance of a part of the object.

3.2.2 Representing appearance

In computer vision, we try to teach computers how to interpret scenes and recognize objects and people from their appearances[1]. How can we represent, model and interpret appearance? When we point a camera at an object, we obtain a matrix of *pixels*, each with a particular *gray value* (or even three values for a color camera). We can think of this two-dimensional image as a surface in a three dimensional space, where the two first axes are the pixel coordinates and the third axis is the gray value. To interpret the image, we need to analyze this surface. Such interpretation often relies on recognizing an object in the image, i.e. pattern recognition. To do this, we need to find a way of representing images mathematically. We can distinguish "classical" methods and appearance-based methods. The classical approaches are based on geometrical object models and the analysis of local image characteristics such as edges and corners (i.e. local appearances, features). The image is represented by a set of features. These approaches are well suited for the analysis of artificial (human made) environments. Another class of more recent approaches are the so-called *appearance-based* approaches and are based

[1]Many people consider that one should *not judge* people by their appearance. As long as computers remain impersonal, this is of no big concern to computer scientists.

on techniques for *learning* the appearance of an object from examples. These approaches proceed in two steps:

- The image is coded into a vector with the same number of elements as there are pixels in the image, say D. The image is therefore coded as a point in the vector space $\mathbb{R}^D$.
- A set of example images (learning images) form together a cloud of points in this space. This cloud is described in mathematical terms, typically using a model.

A simple model could, for example be the representation an object by its average appearance. A simple form for interpretation could then be to compare a new image to this average image by calculating the euclidean distance between the two points in $\mathbb{R}^D$ - this is the well known *template matching* technique. More sophisticated techniques use dimension reduction techniques and statistical modeling, which are the methods that we have chosen in this thesis. The term appearance-based might seem a bit unclear since any computer vision system is, in some sense, appearance-based. The term however, refers more to the ability of these methods to *learn models* from appearances than the fact that the signal we are processing is an appearance.

In Part III of this thesis we use appearance-based models to model three-dimensional SPECT images. The "scene" or "object" we are looking at is the distribution of radiomarkers in the brain cells. The appearance of this scene is a hyper-surface in a 3+1 dimensional space. This is clearly a much wider interpretation of the term appearance than the definition given above, but both classical approaches and appearance-based methods can also be used to interpret three-dimensional images. An explicit geometrical model (whether it exists or not) is not known to us and we would like to employ a method that can learn the model from examples. This motivates the use of appearance-based modeling.

3.2.3 Structure in data

The ability to recognize complex patterns embedded in large quantities of information (such as images), is a remarkable faculty of animals in general and the human in particular. The renowned biologist Konrad Lorenz described how in one week he became capable of distinguishing between the plentiful families of colorful fishes that he encountered when scubadiving between coral reefs in Florida [142]. The ability to recognize nutritious fruits, recognize pray and predators and their patterns of movement, etc. is probably a very important element in increasing the chances of a species to survive in the game of nature where mutation and selection play a decisive role. The important thing about recognizing patterns is that it improves the ability to *predict* the behaviour of, say, predators. Now, why would one like to teach such capabilities to computers? The first answer is curiosity. Humans are interested in recursively understanding how they themselves recognize patterns - recognizing how a human recognizes a pattern. Reproducing pattern recognition under controlled conditions on computers is a powerful way for doing this. The answer is, one could say, quite biologically inspired: humans cannot help themselves for searching new patterns in data since they have always done so. The second answer is related to the first, we would like to use the calculating power of a computer to recognize patterns in quantities of data that surpasses the capabilities of the human brain. Even though humans are excellent pattern recognizers, the limits are soon encountered when trying to understand and predict complex systems: sea and air dynamics, human influence on the nature, social behaviour, how the brain functions etc. If we find principled ways of learning patterns, we bear hope that this can be used to better understand such complex systems.

A third answer to the above question is more pragmatic, but probably the one that makes pattern recognition research thrive the most: the goal of "making robots and computers hear and see". This has indeed many applications such as the replacement of tedious or dangerous processes which are done, or cannot be done, by man today (spam filtering, maintenance tasks in dangerous environments, etc.).

The problem of recognizing patterns is one of finding structure in data. Without structure, *we cannot predict anything*. Structure in the data can be [30]:

- Linear: correlations between variables
- Non-linear: clustering or multimodality, skewness or kurtosis (non-Gaussianity), discontinuities and, in general, concentration along nonlinear manifolds.

If we reason in terms of points in high dimensional spaces, we can find such structures automatically, using methods that are generally denoted *dimension reduction techniques*. For the modeling of high dimensional data, such techniques are, on one hand *necessary*, and on the other hand *possible*. First, dimension reduction is necessary because of a phenomenon called the *curse of dimensionality* (also *empty space phenomenon*). This refers to the fact that, in the absence of simplifying assumptions, the sample size needed to estimate a function of several variables to a given degree of accuracy grows exponentially with the number of variables. This is because the hyper-volume of such spaces are indeed vast (see for example [58, 72, 110] or [30]). Second, dimension reduction is possible because many physical phenomena are indeed governed by a few variables. Our measurement systems (e.g. microphone, CCD or gamma camera) do not measure these directly but instead measure a set of redundant variables. This redundancy may be caused by:

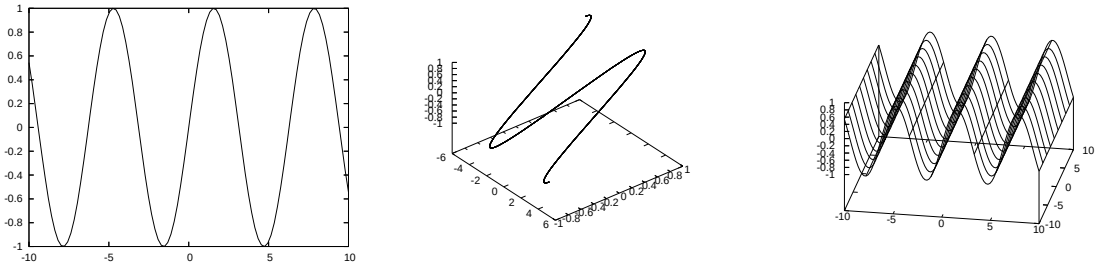
- Variation in the variables that is smaller than the measurement noise. These are therefore irrelevant.
- Variables that are correlated (either through linear combinations or other functional dependencies).

As an example, consider a set of appearances that are representative of a particular object. These form a cloud of points in the *image space*, $\mathbb{R}^D$ (observation space). This image space spans all possible gray value combinations of pixels. It is therefore clear that the cloud of appearances of the object will only occupy a lower dimensional subspace (manifold) in the observation space. This is the subspace that we try to determine using dimension reduction techniques. Some illustrative examples of low-dimensional manifolds embedded in spaces of higher dimensions are shown in Fig. 3.1.

In conclusion, dimension reduction techniques provide a way of automatically learning models from examples that can be used for recognition and prediction. These techniques find structure in data. In practice, principal component analysis (PCA) is the single most important such technique [30]. We shall discuss PCA in Sec. 3.3 and we review other dimension reduction techniques used in visual modeling in Sec. 3.4. Dimension reduction is necessary to allow for probabilistic modeling because of the curse of dimensionality.

3.2.4 Modeling for classification and detection

All problems of pattern recognition can in general be reduced to a problem of *classification*: assigning a semantic class label to a pattern. An example of such classification is the task of



(a) (b) (c)

Figure 3.1: Examples of low-dimensional manifolds embedded in higher dimensional spaces: (a) one-parametric manifold in 2-D (dimensions), (b) one-parametric manifold in 3-D, and (c) two-parametric manifold in 3-D.

identifying a person from a passphoto (also called *recognition*). However, for the particular problem of deciding whether a person is present in an image or not, the term *detection* is generally used. The detection problem is often more difficult to solve than the recognition problem. This is because we do not in general have a specific model for the non-person class. Methods that learn from examples do not deal easily with this problem because it is difficult to obtain a complete sample of all non-person appearances (though such methods exist [80]). One possible way to deal with this problem is to build a complete, realistic and *generic* model of the object class and then base the detection on thresholding some similarity measure between the model and the observation. The disadvantage of such an approach over discriminating approaches is that we have to model many aspects that may be irrelevant for detection or classification. An advantage of generic modeling is however that training is reduced to a set of images of the object class. If the modeling is correct, there is hereby no loss in the discriminative performance.

In Part II of this thesis, where we try to model normal brain perfusion, a generic model is preferable because of the difficulty to define what is abnormal brain perfusion. An exact, realistic model is furthermore necessary in order to characterize abnormal perfusion patterns. A famous citation, often mentioned when discussing class and non-class distributions, is the opening sentence of Leo Tolstoy's Anna Karenina: "All happy families are alike; each unhappy family is unhappy in its own way."

3.2.5 Probabilistic modeling

As in many physical problems, exact mathematical modeling of object appearance in real scenes is not feasible, however stochastic modeling has been applied with much success in many domains. The key to a statistical approach is to have probability models that capture the essential variability and yet are tractable. In a statistical approach, we can model uncertain (i.e. *random*) effects. Other potential advantages of using probabilistic models are:

- They provide principled methods for model estimation and comparison, as well as classification of new observations (e.g. Maximum Likelihood principle and, in particular, Bayesian inference).

- They provide a natural way of combining different measurements (information fusion).

- There are principled ways for dealing with missing data in multivariate models (e.g. by marginalizing over the missing observations).
- Statistical decision theory can be employed (i.e. hypothesis testing).
- We can generate new samples from the model.

A word of moderation is required however, since statistical assumptions are often approximative. This means that many of the above mentioned advantages of statistical modeling are no longer valid, in which case we return to "random thresholding". However, even in cases where model assumptions are clearly wrong, statistical approaches are often attractive because (1) they provide a framework for modeling and interpretation, and (2) statisticians and pattern recognitioners "feel" better when their method has at least *some relation* to well established mathematical principles. This is indeed a difficult issue!

3.2.6 Probability density estimation

In pattern recognition we are interested in developing methods that learn a model (e.g. of object appearance) from observations. For a probabilistic model this is the same as estimating the probability density function. The estimation can be done based on different principles such as least-squares minimization, maximum likelihood principle, maximum-a-posteriori principle, Bayesian inference, minimum description length principle or structural risk minimization [58, 12, 72, 217]. High accuracy of this estimation is indeed important to obtain good recognition performance. For the high dimensional data that we are considering in this thesis, density estimation is not possible without using dimension reduction techniques (because of the empty space phenomenon).

The model that we have chosen will be presented in the next chapter. On the theoretical level, this thesis has been less concerned with the estimation of the density function of this model than with making extensions to the model when considering that the model is approximately justified (sufficiently well estimated). For this we needed to develop a new algorithm (presented in Ch. 4). The importance of this algorithm is that it makes it possible to make inferences under the extended model. For model estimation we have used principal component analysis (PCA - Sec. 3.3.1). This choice was motivated by the development of probabilistic PCA (PPCA - Sec. 3.3.5), which provides the link between a complete multivariate probabilistic model, dimension reduction (by PCA) and maximum likelihood density estimation.

3.2.7 Partial conclusion

After this section on general aspects of appearance, pattern recognition, probabilities and high-dimensional data, we shall now become more concrete and describe the eigenface methods in detail.

3.3 The Era of eigenfaces

3.3.1 Principal component analysis (PCA)

In most appearance-based models, principal component analysis (PCA) is still the basic workhorse. In fact, appearance-based techniques have probably gained a lot of popularity

because of their conceptual simplicity combined with readily available algorithms of low computational complexity for performing PCA. PCA in its original form is *not* based on any model and is therefore a method for *exploratory* data analysis. We can summarize PCA as follows. Let $\{\boldsymbol{y}\}_{j=1}^{J}$ be J samples from a stochastically distributed variable of D components. If we denote the mean-free sample matrix by $\boldsymbol{Y} = [\boldsymbol{y}_1 - \boldsymbol{\mu} \dots \boldsymbol{y}_J - \boldsymbol{\mu}]$, we can estimate the covariance of the data as:

$$\hat{\boldsymbol{\Sigma}}_y = \frac{1}{J}\boldsymbol{Y}\boldsymbol{Y}^T.$$

As with any symmetric real matrix, we can diagonalize the covariance matrix (or the estimate thereof) by solving the eigenproblem and we obtain:

$$\boldsymbol{\Sigma}_y = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^T = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_D] \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_D \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1^T \\ \vdots \\ \boldsymbol{w}_D^T \end{bmatrix}, \tag{3.1}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues and $\boldsymbol{W}$ is a $D \times D$ rotation matrix of eigenvectors ($\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$). This decomposition has several interesting properties (see for example [58, 72] or [12]):

- The eigenvectors define a new coordinate basis into which the original samples can be transformed as (Karhunen-Loeve transformation): $\boldsymbol{x} = \boldsymbol{W}^T(\boldsymbol{y} - \boldsymbol{\mu})$. The covariance of the transformed variable is $\boldsymbol{\Sigma}_x = \boldsymbol{W}^T\boldsymbol{\Sigma}_y\boldsymbol{W} = \boldsymbol{\Lambda}$ since the eigenvectors are orthonormal. This means that the components of the transformed variable are decorrelated. Furthermore, if the distribution of $\boldsymbol{y} \sim p(\boldsymbol{y})$ is Gaussian, then the components of $\boldsymbol{x}$ are statistically independent. We can also think of this transformation as a projection operator which projects the observation vector into the transformed space.

- The eigenvalues are the variances (signal energies) of the observation variable along the corresponding eigenvectors. The eigenvectors that belong to the highest eigenvalues are the *principal components* of the data.

- Principal component analysis is a signal compression (or dimension reduction) scheme where we define a truncated version of the Karhunen-Loeve transformation defined above: If we reorder the eigenvectors and eigenvalues so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$ we obtain an optimal *signal compression* scheme in the least squares sense by defining the transformation $\boldsymbol{W}_Q = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_Q]$, where $Q < D$. Compression is done by *projecting* the original signal into the *eigenspace* or *subspace* as

  $$\boldsymbol{x}_Q = \boldsymbol{W}_Q^T(\boldsymbol{y} - \boldsymbol{\mu}). \tag{3.2}$$

  The *reconstruction* of the signal is then defined as:

  $$\hat{\boldsymbol{y}}_Q = \boldsymbol{W}_Q\boldsymbol{x}_Q + \boldsymbol{\mu}. \tag{3.3}$$

  This compression is lossy and the *residual* of the reconstructed signal is defined as:

  $$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}_Q. \tag{3.4}$$

  The compression has the property that it minimizes the square sum reconstruction error $\boldsymbol{e}^T\boldsymbol{e}$, and retains a percentage of the original signal energy as:

  $$\frac{\sum_{q=0}^{Q}\lambda_q}{\sum_{q=0}^{D}\lambda_q}. \tag{3.5}$$

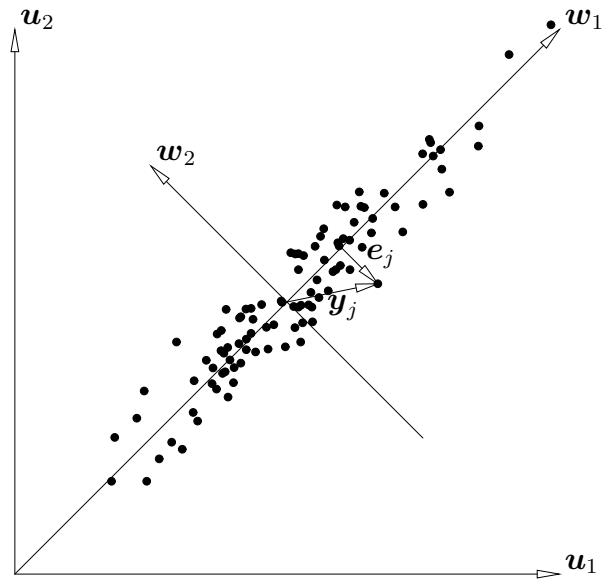


Figure 3.2: A cloud of points in a two dimensional space spanned by the canonical basis vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. The components y_1 and y_2 of the random variable $\boldsymbol{y}$ are strongly correlated. PCA yields a transformation into new uncorrelated variables expressed in the base of $\boldsymbol{w}_q$, $q = 1, 2$. By ordering the eigenvalues in decreasing order, the first principal component, $\boldsymbol{w}_1$, corresponds to the direction in which the variance is the largest. If we approximate the original (observed) data point $\boldsymbol{y}_j$ by its reconstruction (Eq. 3.3) using $\boldsymbol{w}_1$, we introduce an approximation error $\boldsymbol{e}_j$.

In Fig. 3.2 an illustration of PCA in two dimensions is shown. Note that when the number of samples is smaller than the number of dimensions, which is often the case for images, we can invert the roles of dimensions and samples to find the principal components of the data. We explain how this is done in App. B. In the rest of this thesis we shall consider the matrix $\boldsymbol{W}$ to be $D \times Q$, and we only include the Q subscript when this is necessary for clarity.

3.3.2 Face recognition with PCA

In the last section we have seen the mathematical properties of PCA. In [213], Turk and Pentland used PCA to make a system that learned in an unsupervised manner to recognize faces. The goal of the recognition task is to identify a person from an image of the person. This system is however equally well suited to recognize objects in images. A database of people, each with several representative face images is considered. First, each image is associated with a vector $\boldsymbol{y}$ by lexicographically ordering the pixels into the vector components. These vectors thus form the sample matrix $\boldsymbol{Y}$, which is analyzed using PCA. Turk and Pentland were motivated by the work of Sirovich and Kirby [195] who used PCA to identify such a subspace for compression purposes. Turk and Pentland interpreted this subspace as a feature space that could be used for recognition and classification purposes. This feature space is itself thus *automatically learned* and does not in general correspond to isolated features such as ears, eyes, and noses.

The method for face recognition is illustrated in Fig. 3.3 for a simple case of a two-dimensional subspace and three individuals[2]. The authors denote the principal eigenvectors

[2]This is a common trick among mathematicians, they think in $\mathbb{R}^3$ and write out the results in $\mathbb{R}^D$, where

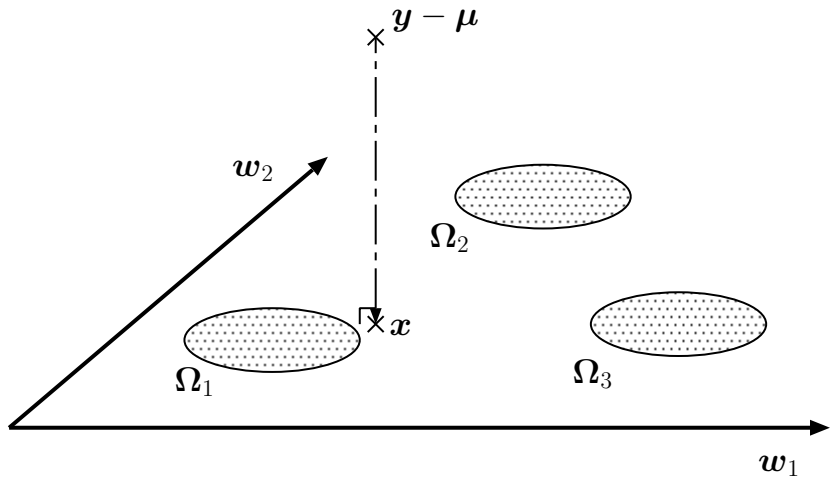


Figure 3.3: A geometric illustration of the subspace ("face space") for two eigenvectors, $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ ("eigenfaces"), and three known classes (individuals). New images are projected into the subspace and associated with the nearest class in this subspace.

"eigenfaces", which span a subspace they denote "face space". This face space is then used for recognition. A new image is transformed (projected) into the face space and classified to the class that has the closest projections. Several cases are distinguished: by looking at the residual - or the reconstruction error - they first decide whether the image is a face or not (detection). A large residual means the distance from face space is far. This is the distance that Moghaddam later denotes *distance from feature space* (DFFS). Second, if an image is far from any class, but close to the face space, it is considered to be an unknown face which can eventually be added to the database. Third, if an image is close to face space and close to a face class, the observation is associated with that class.

3.3.3 Non-linear subspace modeling

Whereas Turk and Pentland modeled every class (person) in the subspace with a Gaussian distribution, Murase and Nayar [163] introduced a more complex parametric subspace model for object recognition. In an experiment, they acquired images of the same object under different views (poses): they placed an object on a turntable and obtained an image at every 5 degrees. This was repeated for several objects. Two types of eigenspaces were built: one for object identification, where all the images were included in the learning set, and another subspace for each object intended for pose classification. They noticed that the rotated objects formed particular spiral-like patterns in the subspace and modeled these using cubic spline interpolation, see Fig. 3.4. The authors denoted this model *parametric eigenspace representation*. Pose classification consisted of projecting the observation into the subspace and finding the closest points on the manifold described by the parametric function.

The method is non-linear, but we can think of the model as an initial linear dimension reduction followed by non-linear modeling. This is done in many other dimension reduction techniques as well, where an initial PCA reduction is followed by some other methods (described in Sec. 3.4). We note that in [31], Chalmond and Girard develop a method for

the size of D is not even mentioned.

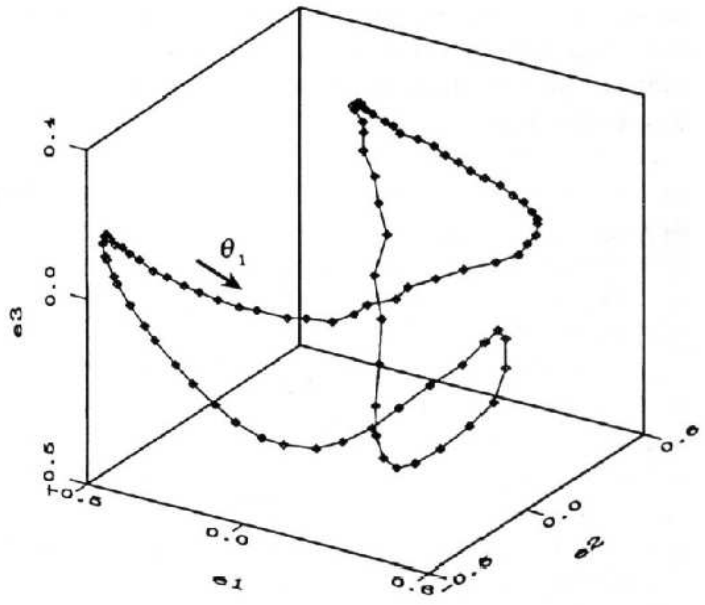


Figure 3.4: Non-linear, parametric subspace modeling using cubic spline curve interpolation of the projected subspace images. "e1", "e2" and "e3" are the three first eigenvectors and θ_1 is the parameter of the spline curve (from [163].

automatically building a similar model, but this time for unordered data points - a much more difficult problem (see also Sec. 3.4.4).

3.3.4 Probabilistic modeling with PCA

The main drawback of the eigenspace methods that have been presented until now, is that they do not possess a probabilistic model. A complete probabilistic model has several advantages as pointed out in Sec. 3.2.5. In particular it implies that we can calculate the likelihood of observing a particular image of an object class: $p(\boldsymbol{y}|\Omega)$, where $\boldsymbol{y}$ is the observed image and Ω the object class. Even though the links between PCA and the factor analysis model (which is a probabilistic model) have been known in the statistics literature for some time (see for example [4, p.567]), Moghaddam and Pentland [162] were the first to take advantage of a probabilistic interpretation of PCA for visual object modeling. They found a clever way of writing out a complete Gaussian in the observation space when only a marginal part of this Gaussian can be estimated. In a multi-center large-scale evaluation study of face recognition algorithms (the FERET evaluation study [186]), they showed the superiority of their model over competing non-probabilistic methods.

The data is simply considered to be distributed as a global Gaussian[3]:

$$p(\boldsymbol{y}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_y|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right). \tag{3.6}$$

This Gaussian is separated into two marginal distributions, one in-eigenspace distribution, $p_{\mathcal{F}}(\boldsymbol{y})$ - defined by PCA - and one out-of-eigenspace distribution, $p_{\bar{\mathcal{F}}}(\boldsymbol{y})$: the orthogonal,

[3]We drop the conditional in $p(\boldsymbol{y}|\Omega)$ to alleviate notation.

isotropic noise distribution. Since the latter distribution is not known, it is estimated as $\hat{p}_{\bar{\mathcal{F}}}(\boldsymbol{y})$ by considering the noise orthogonal to the subspace to be isotropic. The separation into $p_{\mathcal{F}}$ and $\hat{p}_{\bar{\mathcal{F}}}$ is equivalent to a partitioning of the observation space. The following estimation of the complete Karhunen-Loeve transformation is then obtained:

$$p(\boldsymbol{y}) = \left[\frac{\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Lambda}_Q^{-1}\boldsymbol{x}\right)}{(2\pi)^{Q/2}|\boldsymbol{\Lambda}_Q|^{1/2}}\right] \cdot \left[\frac{\exp\left(-\frac{\boldsymbol{e}^T\boldsymbol{e}}{2\sigma^2}\right)}{(2\pi\sigma^2)^{(D-Q)/2}}\right] = p_{\mathcal{F}}(\boldsymbol{y})\hat{p}_{\bar{\mathcal{F}}}(\boldsymbol{y}), \tag{3.7}$$

where $\boldsymbol{\Lambda}_Q$ is the truncated diagonal matrix of eigenvalues, $\boldsymbol{x}$ is the projection of the image into the subspace (Eq. 3.2) defined by the eigenvectors $\boldsymbol{W}_Q$, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{W}_Q\boldsymbol{x}$ is the residual vector from Eq. 3.4, and σ is the estimated, isotropic noise variance in the orthogonal directions of the eigenspace. The first term is associated with the Mahalonobis distance $\boldsymbol{x}^T\boldsymbol{\Lambda}_Q^{-1}\boldsymbol{x}$ which the authors term *Distance in Feature Space* (DIFS). The second term is computable because the noise is considered to be isotropic in the orthogonal space: We can ignore the exact direction of the residual, all that matter is that it lies in this orthogonal space. To see this we write the distribution in Eq. 3.7 as:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_y), \quad \text{where} \quad \boldsymbol{\Sigma}_y = [\boldsymbol{W}_Q\boldsymbol{W}_{D-Q}] \begin{bmatrix} \Lambda_Q & & & 0 \\ & \sigma^2 & & \\ & & \ddots & \\ 0 & & & \sigma^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_Q^T \\ \boldsymbol{W}_{D-Q}^T \end{bmatrix}. \tag{3.8}$$

Here, the eigenvectors of the orthogonal space, $\boldsymbol{W}_{D-Q}$, do not need to be computed[4]: the Mahalonobis distance in the complementary space can be calculated directly from the reconstruction errors (residuals) as $\frac{\boldsymbol{e}^T\boldsymbol{e}}{\sigma^2}$ (Eq. 3.7). The authors call this distance the *Distance from Feature Space* (DFFS). The DFFS is particularly important for detection, which necessitates the calculation of the likelihood of an observation. However, it can be ignored for recognition [161] (verified in personal communication). Fig. 3.5 illustrates the distances DIFS and DFFS.

The noise variance of the "noise-space" is estimated by minimizing the Kullback-Leibler divergence between the true and estimated densities, which yields:

$$\hat{\sigma}^2 = \frac{1}{D-Q}\sum_{q=Q+1}^{D}\lambda_q. \tag{3.9}$$

In practice, many of these eigenvalues may not be available. One possibility, as proposed by Moghaddam and Pentland, is to estimate these by fitting a $1/f$ function to the spectrum of computed eigenvalues (which are declining).

Note that the separation of the Gaussian into the product of two marginal distributions is different from a model where the systematic part and the noise have been separated, as is the case in "true" probabilistic PCA and in factor analysis. In this kind of probabilistic PCA the noise and the systematic part are treated equally. We shall come back to this point in the next section.

3.3.5 Probabilistic Principal Component Analysis (PPCA)

The so-called probabilistic principal component analysis (PPCA) model that is based on the factor analysis model was developed by Tipping and Bishop [212, 211] and independently by

[4]Often these eigenvectors cannot be calculated because the sample size is too small.

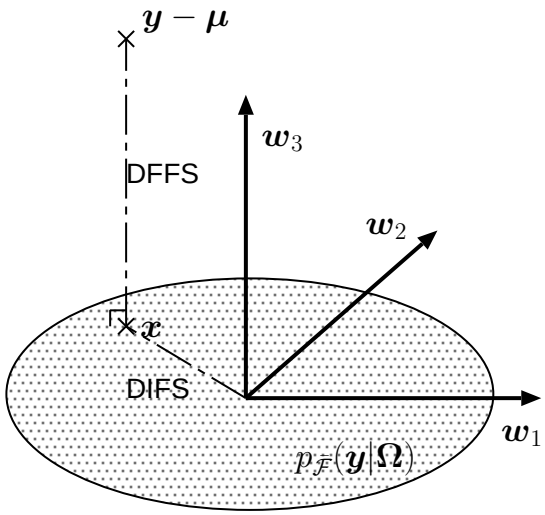


Figure 3.5: In illustrative example of the two distances DFFS and DIFS (see the text). The subspace is spanned by two principal components $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, the complementary space is spanned by $\boldsymbol{w}_3$. The inspace distribution of the object is drawn as a dotted ellipse. For the observed vector $\boldsymbol{y}$ both DFFS and DIFS are evaluated to determine whether an object is close to its class distribution or not.

Roweis [188, 189] (under the name of sensible PCA). This model has not to our knowledge been used for face modeling, but has been applied to build class hierarchies of objects [57] as well as for analysis and visualization of multispectral data [11]. The PPCA model has the advantage that it leaves more room for interpretation and extensions than the model proposed by Moghaddam and Pentland. We now describe the PPCA model and will later describe its relation to the model of Moghaddam and Pentland.

The model is a linear latent variable model of the form

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{3.10}$$

where a new *random variable* has been introduced: the Q dimensional latent variable (subspace variable), $\boldsymbol{x}$. The components of this variable are considered to be *identically and independently distributed* (i.i.d) as a Gaussian with unit variance, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_Q)$. The noise distribution is also Gaussian and isotropic, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_D)$. Here, the $D \times Q$ dimensional *factor loadings* or generation matrix, $\boldsymbol{W}$ is in general *not* the eigenvectors of the covariance matrix, but it is orthogonal. Under the model assumptions we can calculate the following distributions ([212] and our calculations):

- The distribution of the observation is Gaussian, $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_y)$, with the variance (see for example [4, p.553]):

$$\boldsymbol{\Sigma}_y = \sigma^2 \boldsymbol{I}_D + \boldsymbol{W}\boldsymbol{W}^T. \tag{3.11}$$

  If we let $\boldsymbol{W}_Q = \boldsymbol{W}$ and denote by $\boldsymbol{W}_{D-Q}$ a matrix of vectors that spans the space that is orthogonal to the space defined by $\boldsymbol{W}$, we can rewrite the variance as (analog to

Eq. 3.8):

$$\boldsymbol{\Sigma}_y = [\boldsymbol{W}_Q \boldsymbol{W}_{D-Q}] \begin{bmatrix} \sigma_1^2 + \sigma^2 & & & & & 0 \\ & \ddots & & & & \\ & & \sigma_Q^2 + \sigma^2 & & & \\ & & & \sigma^2 & & \\ & & & & \ddots & \\ 0 & & & & & \sigma^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_Q^T \\ \boldsymbol{W}_{D-Q}^T \end{bmatrix}, \tag{3.12}$$

where σ_q, $q = 1 \dots Q$ are the variances induced by the systematic part of the model, $\boldsymbol{W}\boldsymbol{x}$.

- The conditional distribution of the observation given the latent variable is $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_D)$. This is also the key motivation for the model: the components of the observed image $\boldsymbol{y}$ are conditionally independent given the latent subspace variable $\boldsymbol{x}$.

- The posterior distribution of the latent variable given the observation is Gaussian with $p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{x|y})$, where

$$\boldsymbol{\mu}_x = (\sigma^2 \boldsymbol{I}_Q + \boldsymbol{W}^T \boldsymbol{W})^{-1} \boldsymbol{W}^T (\boldsymbol{y} - \boldsymbol{\mu}), \tag{3.13}$$

  and

$$\boldsymbol{\Sigma}_{x|y} = \sigma^2 (\sigma^2 \boldsymbol{I}_Q + \boldsymbol{W}^T \boldsymbol{W})^{-1}.$$

  Note that the last term in Eq. 3.13 is the same as the PCA projection, Eq. 3.2. In PPCA the expectation of the posterior, $\boldsymbol{\mu}_x$, is indeed the most appropriate choice for projecting an observation into the subspace. We shall discuss in depth this kind of posterior distribution under more general model assumptions in the next chapter.

In [212], Tipping and Bishop develop the exact maximum-likelihood estimate of this model, when the covariance $\boldsymbol{\Sigma}_y$ is estimated by the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_y$. The resulting estimators are:

- The mean vector

$$\hat{\boldsymbol{\mu}} = \frac{1}{J} \sum_{j=1}^{J} \boldsymbol{y}_j. \tag{3.14}$$

- The generation matrix

$$\hat{\boldsymbol{W}} = \boldsymbol{U}_Q (\boldsymbol{\Lambda}_Q - \sigma^2 \boldsymbol{I}_Q)^{1/2} \boldsymbol{R}, \tag{3.15}$$

  where the Q column vectors in the $D \times Q$ matrix $\boldsymbol{U}_Q$ are the eigenvectors of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_y$, with corresponding eigenvalues in the $Q \times Q$ diagonal matrix $\boldsymbol{\Lambda}_Q$. The matrix $\boldsymbol{R}$ is an arbitrary $Q \times Q$ orthogonal rotation matrix. This arbitrariness is explained by an ambiguity in the model Eq. 3.10: the distribution $p(\boldsymbol{y})$ is invariant to rotation of the generation matrix, i.e. the model variance does not change (Eq. 3.11).

- The noise variance

$$\hat{\sigma}^2 = \frac{1}{D - Q} \sum_{q=Q+1}^{D} \lambda_q, \tag{3.16}$$

  where $\lambda_{q+1}, \dots, \lambda_d$ are the smallest eigenvalues of $\hat{\boldsymbol{\Sigma}}_y$. This maximum likelihood estimate is exactly the same as the estimate found by Moghaddam and Pentland by minimizing the Kullback-Leibler divergence, Eq. 3.9.

With these estimates, the distribution of observations, $p(\boldsymbol{y})$, becomes exactly the same as the distribution found by Moghaddam and Pentland (i.e. with the estimated generation matrix Eq. 3.15, the global covariance matrix Eq. 3.12 becomes equal to the global covariance matrix in Eq. 3.8). We can thus explain the relation between the two models: since Moghaddam and Pentland do not model the subspace variable as an independent random variable, we obtain their model from the PPCA model as the marginal distribution at the maximum likelihood estimate $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{W}}, \hat{\boldsymbol{\mu}}, \hat{\sigma})$ of the joint distribution $p(\boldsymbol{y}, \boldsymbol{x})$ as:

$$p(\boldsymbol{y})|_{\hat{\boldsymbol{\Theta}}} = \int p(\boldsymbol{y}, \boldsymbol{x})|_{\hat{\boldsymbol{\Theta}}} \mathrm{d}\boldsymbol{x}.$$

As an illustrative example, consider an object whose appearance only changes in the lower region of the image. Furthermore, consider that both models could identify this region as the principal subspace with, say, the same number of principal axes as there are pixels in this region. Now, whereas the model of Moghaddam and Pentland only considers the noise to be present in the *upper region* of the image (the orthogonal, complementary space), the PPCA model considers an additive noise *everywhere* in the image. This is also seen in the Eqs. 3.12 and 3.8. The relation between these two models has to our knowledge not been clearly stated in the literature, however, Moghaddam points out in [161] that their model is a special case of the PPCA model.

3.3.6 Analytical PCA

The experiments of Murase and Nayar [163], described in Sec. 3.3.3, showed that the distribution of rotated objects formed characteristic clouds in the subspace. Several authors have therefore investigated the exact relationship between the subspace distribution (as found by PCA) of appearances issued from controlled changes such as illumination changes [9, 187], rotation of images [215] and panoramic images [113]. These studies undermine the idea that the images of objects have intrinsic dimensions that are largely inferior to the dimension of the observation space $\mathbb{R}^D$. In particular, Uenohara and Kanade [215] showed that the distribution in subspace of rotated images of an object will lie on a one-dimensional circular-like curve, i.e. the projections of the learning images into a two-dimensional subspace spanned by two successive eigenvectors form circles. We have used this knowledge to design experiments with learning set data that form non-linear, non-Gaussian subspace distributions (Ch. 5).

3.3.7 Partial conclusion

We have presented a detailed and chronological tour of PCA-based methods that have been used for modeling objects and faces in the last decade. The tour was rounded off with the probabilistic PCA model, which establishes a probabilistic framework for appearance-based models. These particular methods were presented because they form the natural ancestry of the model developed in this work. We now review other methods that have been, or can be, used for appearance modeling. We then come back to some recent developments concerning robust noise modeling in the PPCA framework.

Before continuing, let us just mention a little curiosity. Whereas many researchers try to find compact manifolds that capture facial variations, a complementary approach for *visualizing* high-dimensional variables (for exploratory purposes) was developed by Chernoff[5].

[5]See for example `http://people.cs.uchicago.edu/~wiseman/chernoff/`

The method is based on generating drawings of faces. Since humans are highly specialized in recognizing faces and facial expressions, Chernoff made a mapping between a ten-dimensional subspace onto different facial features such as size of the nose, eccentricity of the head, distance between eyes etc. The idea is that when mapping the (high-dimensional) samples onto these faces, a human could be capable of recognizing structure (clusters) in the data. We do not know if someone has tried (or succeeded) in creating photorealistic Chernoff-faces by learning a subspace from training samples. This could certainly be envisaged by means of appearance-based models!

3.4 Other dimension reduction techniques

We have explained in Sec. 3.2.3 why dimension reduction is necessary and why it is (often) possible. In Sec. 3.3 we saw our method of choice, PCA, in action. In this section we provide a brief outline of other dimension reduction techniques in image modeling. These can be divided into linear and non-linear methods. Let us begin by providing a formal definition of the problem and what the *intrinsic dimension* of a phenomenon is.

3.4.1 Problem statement

The problem of dimension reduction can be stated as follows; We have a given sample of observations, say $\{\boldsymbol{y}_j\}_{j=1}^{J}$, of D-dimensional real vectors drawn from an unknown probability distribution. The fundamental assumption that justifies the dimension reduction is that the sample actually lies, at least approximately, on a manifold of smaller dimension than the data space. The goal of dimension reduction is to find a representation of that manifold (a coordinate system) that will allow us to project the data vectors on it and obtain a low-dimensional, compact representation of the observed data.

3.4.2 The intrinsic dimension of a sample

We have already seen examples of low dimensional manifolds embedded in higher dimensional spaces in Fig. 3.1. The figure in the middle could for example describe the path of a butterfly through the room, which is governed by one single independent variable. In practice, many phenomenon that are governed by a few independent variables will *appear* as having many more degrees of freedom due to the influence of a variety of factors: noise, imperfection in the measurement system, addition of irrelevant variables, etc. The *intrinsic dimension* of a phenomenon can be defined as the number of independent variables that explain satisfactorily that phenomenon, e.g. a single independent variable would suffice to describe the butterfly's flight.

The determination of the intrinsic dimension is central to the problem of dimension reduction, knowing it eliminates the risk of over- or underfitting. Furthermore, all dimension reduction techniques take the intrinsic dimension as a parameter. In practice, either a trial-and-error process is necessary, or a regularization term which contains a priori information on the amount of smoothing to be done must be introduced to determine the intrinsic dimension. The problem is ill-posed because we can fit any data sample given enough free parameters[6]. In Ch. 8 we determine this dimension by evaluating the generalization error of a detector of

[6]The problem is closely related to the problem of determining the number of mixtures in density mixture models.

abnormal perfusion patterns. In principal component analysis one often defines a percentage of total variance to be explained by the reduced space, see Eq. 3.5. Minka has proposed an automatic method to determine the number of principal components in PPCA using Bayesian inference [156]. Other methods have been proposed in [91] and [137].

3.4.3 Classification of techniques

An obvious way of classifying dimension reduction techniques is to distinguish linear and non-linear methods. Another useful view is proposed by Carreira-Perpiñán in [30]:

Hard dimension reduction problems, in which the data have dimensions ranging from hundreds to perhaps hundreds of thousands of components (e.g. image modeling), and usually a drastic reduction is sought. PCA is one of the most widespread techniques in most practical cases.

Soft dimension reduction problems, in which the data is not too high-dimensional (a few tens of components), and the reduction is not very drastic. Typically, the components are observed or measured values of different variables which have a straightforward interpretation. In this class, we find the usual multivariate methods: PCA, factor analysis, linear discriminant analysis, multidimensional scaling etc.

Visualization problems, in which the data doesn't normally have a very high dimension in absolute terms, but we need to reduce it to 2, 3 or 4 at most in order to plot it. In this class we find projection pursuit, PCA, multidimensional scaling, self-organizing maps and density networks, as well as interactive programs.

Svensén [201] makes a distinction between generative (probabilistic) and non-generative methods. Other possible distinctions are static/time-dependent, and discrete/continuous data. We only consider static, continuous methods, however both generative and non-generative.

3.4.4 Linear methods

Linear models are by far the most applied models in dimension reduction and for many other signal processing applications as well. Advantages of linear models over non-linear models are: mathematical tractability, computationally less expensive, conceptually simpler, and facilitated interpretation. We find again linear models in numerous domains. An interesting unifying view of linear models (PCA, factor analysis, ICA, Kalman filters, hidden Markov models,...) can be found in [189], and links between these as graphical models in [164].

Projection pursuit

Projection pursuit is not one single method for linear dimension reduction, but rather a general framework in which other methods can be seen as special cases. It was introduced as a method for visualization purposes and exploratory data analysis by Friedman and Tukey [63]. Nice introductions can be found in [30] and in [137]. Projection pursuit is an unsupervised technique that picks *interesting* low-dimensional linear orthogonal projections of a high-dimensional point cloud by maximizing a objective function called the *projection index*. This can be computationally tedious. The projection index is a real functional defined on the space of distributions in $\mathbb{R}^Q$ (Q is the subspace dimension). A projection is considered

as being interesting if it contains structure. Because of the following two results, the normal distribution is considered to be the least interesting (the least structured) density:

- For fixed variance, the normal distribution has the least information (it has the lowest negative entropy [58, p.631]) among all probability distributions.
- For most high-dimensional clouds, most low-dimensional projections are approximately normal [86].

Several projection indices are therefore based on measures of higher order statistics (cumulants), the Fisher information or negative entropy. These criteria coincide with the criteria used for estimating the independent components in independent component analysis (ICA), which we shall see shortly (Sec. 3.4.4). Taking variance as the projection index, the principal components of the data are found. This is why PCA and ICA can be seen as special cases of projection pursuit. Another method for dimension reduction that can be formulated as a projection pursuit problem is multidimensional scaling [193, 31]. Here, projections are sought, for which the euclidean distances between the projected data remains the same as for the data in the original space. Closely related, Chalmond and Girard [31] uses a an index that conserves the neighborhood between observations in order to find a linear subspace that is well suited for spline interpolation of the projected data. Their method therefore provides an automatic way of obtaining a non-linear model similar to the one of Murase and Nayar [163] (Sec. 3.3.3) in the case of non-ordered observations.

Linear discriminant analysis (LDA)

Linear discriminant analysis (also called Fisher discriminant analysis) is a method for supervised classification. Here optimal projections of the high dimensional, pre-classified patterns are chosen so that the projected data is optimally separable. The criterion for optimal separability is a quotient that maximizes the inter-class variance and at the same time minimizes the intra-class variance. LDA is optimal as a method for dimension reduction when

- the data is linearly separable,
- pre-classified learning samples exist,
- and the samples of each class are normally distributed.

In these cases LDA may indeed be a much more powerful technique for discrimination high-dimensional data than PCA (see also the section on modeling for classification, Sec. 3.2.4). An illustrative example where this is the case is shown in Fig. 3.6. However, only a subspace of $C-1$ dimensions can be found, where C is the number of classes. In [8], Belhumeur *et al.* showed that *Fisherfaces* were superior to eigenfaces for recognition. However, this is not always true as later demonstrated by Martinez *et al.* in [149]. Furthermore, in practice one often applies LDA after an initial PCA. Another method for finding linear separation planes between high-dimensional data is the linear support vector machine (SVM) which optimizes a different objective function. SVM are more often used in connection with non-linear discrimination (kernel-methods) which we shall see in the next section.

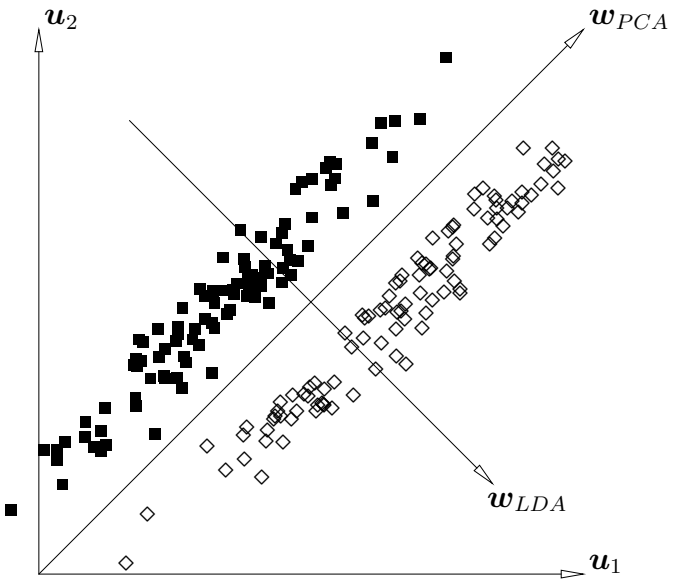


Figure 3.6: When the data is pre-classified like here (diamonds and filled boxes represent two distinct classes), LDA can find low-dimensional projections $\boldsymbol{w}_{LDA}$, that have higher discriminating power than the projections found by PCA, $\boldsymbol{w}_{PCA}$.

Factor analysis

Strictly speaking, factor analysis is not considered to be a dimension reduction technique, but it can be used for this purpose. Contrary to the other models in this section, it is a true probabilistic model, originally developed in psychology [4] and it is the basis for the PPCA model presented in Sec. 3.3.5. The idea in factor analysis is to partition the observation into an unobserved systematic part and an unobserved error part. This is different from PCA, where directions are sought that *explain* the observed variance. The model can be written as the PPCA model in Eq. 3.10:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{3.17}$$

however with a small but crucial difference: the noise follows a non-isotropic Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi}$ diagonal. The generation matrix $\boldsymbol{W}$ is in factor analysis called the *factor loadings*. The elements of $\boldsymbol{\Psi}$ are often called the *uniquenesses* of the model, which points to the unique, independent noise on each observation axis (pixel). The difference between PCA (and also PPCA) and factor analysis is given by the following invariances [211]:

- PCA is covariant to rotation of the observation axes, i.e. multiplication from the left with an orthogonal matrix $\boldsymbol{R}$ ($\boldsymbol{R}^T = \boldsymbol{R}^{-1}$) in Eq. 3.17 simply leads PCA to find the rotated eigenvectors $\boldsymbol{RW}$.
- Factor analysis is covariant to scaling of the axes, i.e. multiplication from the left with a diagonal matrix $\boldsymbol{A}$ in Eq. 3.17 leads to rescaling of the rows of $\boldsymbol{W}$ and rescaled uniquenesses.

This also means that factor analysis will find directions of correlation (as opposed to those of high variance) between the variables [97]. These directions are independent of the variance which is unique to each variable. This is a desirable property for measurements of variables that are not believed to have the same noise variance properties (e.g. different mental test scores), but is probably less appropriate for images where we assume the pixel noise variance

to be the same for every pixel. Furthermore, we would expect a rotation of our camera system to lead to a rotation in the projection axes, not a completely different model.

Another specificity to the factor analysis model is the interpretation of the factor loadings and the projected variables. Factor analysis was originally used to find latent variables with a meaning, e.g. "intelligence" (psychology), "social class" (sociology[7]), time curve of the left heart chamber in dynamic SPECT studies [10], and others. For this it is necessary to interpret the factor loadings. To facilitate this interpretation it is further (often) necessary to rotate the axes of $\boldsymbol{W}$ – ideally to obtain directions with a few high loadings. This rotation is called *oblique analysis* and much research is dedicated to the development of criteria for this (the *varimax* criterion being one of the best known [118]). Oblique analysis can be applied to any set of axes, also to those found by PCA. This was done for brain SPECT studies in [115], but has not been an issue in the computer vision community.

Factor analysis is typically used for observation variables of "reasonable" dimensions, say on the order of 10. However, there have been works on face and digit modeling using mixture of factor analyzers. Such mixtures are possible because the factor analysis model is a probabilistic model. We shall come back to this in the next section on non-linear models.

Independent component analysis (ICA)

Independent component analysis is another model for finding linear subspaces that are not necessarily orthogonal but where the projected variables are statistically independent [117, 41, 107, 108]. The motivation behind ICA is the same as for projection pursuit and was first developed in blind source separation. The idea is that (1) sources of signals that are of interest (objects in images, speakers at a cocktail party) emit stochastic signals that are non-Gaussian ("colored"), and (2) these sources are statistically independent. The observation model is stated as follows:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{y}$ is the observation, $\boldsymbol{A}$ is a mixing matrix that mixes the statistically independent components of $\boldsymbol{x}$ together. There also exist a "noisy ICA model"

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon},$$

with additive random noise (also called probabilistic ICA in [7]). However, most research has been done on the former due to the difficulty of estimating the latter. Estimating the demixing matrix $\boldsymbol{A}^{-1}$ can be quite difficult and several algorithms have been developed for this [117, 41, 107, 108]. The main properties of ICA are:

- The column vectors of the mixing matrix are (in general) not orthogonal.
- There is no natural ordering of the components (as in PCA).

ICA has been used for modeling local features [236] and for modeling face appearances [6]. Due to contradictory claims in the literature on the relative performance of PCA and ICA, Draper *et al.* [56] recently made a comparative study where they concluded that PCA performed well, but not as well as ICA. In brain activation studies – both fMRI [150] and PET [183] studies – ICA has been used as an exploratory alternative to model-based analysis in order to find regions with statistically independent temporal behavior.

[7]For example, Pierre Bourdieu used factor analysis in his famous book "Distinction" [20].

3.4.5 Non-linear methods

In this section we consider briefly non-linear models for dimension reduction. These have the potential to capture structure of complex form, but can be more sensitive to overfitting and outliers than linear models. This is because they normally fit many more parameters than linear models, which also makes them more suitable when there is ample data. Non-linear dimension reduction models can all be fitted using neural networks - so called autoencoder (or -associative) networks[8] - for which there exists a multitude of learning (estimation) paradigms. A successful paradigm has been the support vector machine concept [217], which also perform well for building classifiers even from few data samples. We consider mixtures of linear models as being non-linear. This is because they are (almost always) globally non-linear even though locally linear. Many of the non-linear models can be seen as generalizations of either PCA or factor analysis.

Local linear models

A special form of a mixture of linear models was proposed in [235], where the authors built one PCA model for each partition of the image. The image is simply divided into P patches, each with D_p pixels and Q_p subspace variables so that the resulting model can be written as[9]:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_P \end{bmatrix} = \begin{bmatrix} \boldsymbol{W}_1 & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{W}_P \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_P \end{bmatrix} + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

with $\boldsymbol{y}_p \in \mathbb{R}^{D_p}$, $\boldsymbol{x}_p \in \mathbb{R}^{Q_p}$, and the $D_p \times Q_p$ dimensional matrices $\boldsymbol{W}_p$, and with the constraints

$$D = \sum_p^P D_p \quad \text{and} \quad Q = \sum_p^P Q_p.$$

Contrary to the global PCA model, one can choose the number of degrees of freedom of the model to be much larger than the number of images in the learning sample (however with the restriction that the number of latent variables for each patch is less than the number of images). The reconstructed images can therefore have a lower mean square error and be more visually pleasing than for global PCA methods, but the partitioning typically remains visible (see the blocking artifacts in Fig. 3.7 b and c). All the linear models presented in the next chapter can of course be combined in this manner to obtain globally non-linear models.

Mixture of linear models

More general than local linear models are mixtures of linear models, or short *mixture models*. There exist probabilistic mixture models (mixtures of factor analysis or PPCA models) [97, 62, 211, 233] and non-probabilistic mixture models (PCA eigenspaces combined in some manner) [119, 120, 97, 28, 139] (only references for models that have been used for image modeling). The former are particularly interesting because they can be estimated using the EM-algorithm [55, 151]. Probabilistic mixture models are models that estimate the probability density function of the observed variable and are in the one-dimensional case often considered

[8] These can of course also fit linear models.

[9] Reformulated from the authors' original formulation

(a) (b) (c)

Figure 3.7: Examples of a reconstructed image from the COIL database (see Ch. 5). In (a) the image has been reconstructed with 10 eigenvectors, in (b) and (c), the image has been reconstructed with 9 and 25 patches respectively, each patch with 10 eigenvectors.

as semi-parametric density estimators. Like kernel density estimation methods (see also Sec. 3.6), these are therefore potentially capable of modeling (and reducing) any linear and non-linear manifolds. Their estimation remains difficult however since the problem of determining the intrinsic dimension is more complex and the EM-algorithm can only find local minima. Mixture distributions are modeled as a sum of linear distributions (typically Gaussian). For example a mixture of factor analyzers is given by:

$$p(\boldsymbol{y}) = \sum_{k}^{K} \pi_k p_k(\boldsymbol{y}), \quad \text{with} \quad p_k(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k + \boldsymbol{W}_k \boldsymbol{W}_k^T),$$

where each factor is a linear model of the form

$$\boldsymbol{y} = \boldsymbol{W}_k \boldsymbol{x} + \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_k.$$

Different variants can then be formulated by fixing the same noise variance across components or assuming the noise variance of each component to be isotropic, which yields the mixture PPCA model [211].

Principal curves and principal surfaces

Principal curves and surfaces have been suggested as non-linear generalizations to PCA and were proposed by Hastie and Stuetzle [92]. Intuitively, a principal curve is a smooth, one-dimensional curve that passes through the "middle" of a cloud of data points in the embedding observation space. For dimension reduction the points in observation space are projected onto this curve. The estimation of this curve is based on a heuristic algorithm (for which no convergence proof exists) and the model is not generative. Later Tibshirani [210] proposed a probabilistic extension to principal curves that can be trained by the EM algorithm. Principal surfaces are extensions to principal curves, a first definition proposed in the original paper of Hastie and Stuetzle [92] and later as an alternative definition by LeBlanc and Tibshirani in [133]. Probabilistic (generative) extensions to principal surfaces has been proposed by Chang and Ghosh [33][10].

[10]This model again is related to the generative topographic model (GTM) shortly to be mentioned.

Kernel PCA

Like support vector machines, kernel PCA [192] makes use of the "kernel trick" to model non-linear data. Kernel based methods map an observation vector to a high (maybe even infinite) dimensional feature space where inner products (dot products) between vectors are calculated. The mapping need not be linear, but it is assumed that the mapped data is linear in the feature space. The trick consists in that the mapped vectors are never explicitly computed, only their inner product is calculated using the so-called kernel function. Since the covariance matrix in feature space (high-dimensional space) is calculated from dot products, one can perform standard linear PCA in this space. It might seem strange to reduce dimensions by first passing to a higher dimensional space, but this has indeed been done for face modeling [232]. A problem with kernel PCA is that we cannot get a hold of the actual principal components. A projection onto the component can be calculated, but not the reconstruction of this projection. This makes it difficult to study the variance along the principal components in the observation space.

Self-organizing maps, density networks, and generative topographic mapping

Finally, let us briefly mention "topological" methods for dimension reduction (mostly applied for visualization purposes), where nearby points in the observation space are mapped to nearby points in the subspace (much like multidimensional scaling). To be topological, the mapping is continuous. The best known method for this is the self-organizing map (SOM) of Kohonen [123], which, albeit its successful application, still lacks a sound theoretical foundation (e.g. no proofs of convergence exists). In curvilinear component analysis, Demartines [54] combines a SOM input network with a subsequent output network in order to obtain "double" non-linear mapping (with dimension reduction) from input space to output space.

MacKay [144] has developed a very general framework denoted *density networks* where nonlinear feed-forward neural networks and latent variable models are merged together, forming a complete probabilistic model. A particular density network is the generative topographic mapping (GTM) method put forward by Svensén and Bishop [201] as a principled view of the self-organizing maps of Kohonen. We do not know of any image modeling applications where density networks or GTMs have been applied, but we mention them here for completeness since they offer many possibilities for interpretation and have many relations to almost all the other techniques we have mentioned in this section, see in particular [201, 30] and [12].

3.4.6 Partial conclusion

Dimension reduction techniques are central in many domains such as statistics, information theory and pattern recognition. Linear models are simple and easy to interpret, but may not be realistic for real data. Non-linear models can potentially fit any manifold, but are more difficult to estimate and are prone to overfitting. The major problem with all dimension reduction techniques is the verification of the methods. The properties of high-dimensional spaces become quite counter-intuitive and it may be difficult to verify that the dimension reduction actually captures the structure of interest.

3.5 Robust estimation

The most widespread paradigm for parameter estimation in regression problems is the maximum likelihood (ML) paradigm combined with an assumption of additive Gaussian noise. In this setting, the estimation is equivalent to least-squares (LS) estimation. The estimate can be calculated analytically, which is convenient and fast. However, it has been shown that the estimate is sensitive to *outliers* in the data and can become arbitrarily wrong [152]. We can have outliers in the data because of several factors: the assumption of Gaussian noise is approximative and not really justified, or the data has for some reason been corrupted by noise that was not considered in the model. In visual scene analysis, corrupted data is often encountered either as occlusions or cluttered background. Since such noise is indeed difficult to model accurately, another approach is often taken that is based on making the LS estimate robust to outliers. The goal in robust estimation is to obtain an estimate that has a high *breakdown point*, which is the smallest amount of outlier contamination that may force the value of the estimate outside an arbitrary range. For example, the (asymptotic) breakdown point of the mean is 0 since a single large outlier can corrupt the result. The median on the other hand remains reliable as long as less than half of the data are contaminated, yielding a (asymptotically) maximum breakdown point of 0.5.

We will not provide a complete review of robust estimation techniques here and point the reader to [152, 106, 14] or [234]. However, we will briefly introduce the family of M-estimators and their optimization based on half-quadratic theory. For this we consider a simple linear regression problem. We then review robust appearance-based models. There are two distinct cases of robust methods in this context, one for learning (also called "robust PCA"), and one for projecting the observed image into the eigenspace ("robust reconstruction").

3.5.1 Least-squares regression

Consider the linear regression problem:

$$y_j = \gamma x_j + \beta, \quad j = 1, \ldots, J \tag{3.18}$$

of J data pairs $\{(y_j, x_j)\}_{j=1}^{J}$, where $\boldsymbol{x}$ is the *predictor* variable and $\boldsymbol{y}$ the *dependent* variable. The least-squares (LS) estimate of the parameters, $(\hat{\gamma}, \hat{\beta})$ minimizes the square sum of residual errors:

$$(\hat{\gamma}, \hat{\beta}) = \arg\min_{(\gamma,\beta)} \sum_{j}^{J} e_j^2 = \sum_{j}^{J} (y_j - \gamma x_j - \beta)^2,$$

where $e_j = y_j - \gamma x_j - \beta$ are the residuals. This estimate is optimal (lower Cramer-Rao bound) for Gaussian noise in the measurements $\{y_j\}_{j=1}^{J}$. However a corrupted data pair $(y_j,\, x_j)$ may force this estimate outside an arbitrary range [152]. The goal of robust estimators is to yield an estimate that is not fooled by contaminated measurements. The estimate should furthermore be close to the LS (optimal) estimate when the noise is actually Gaussian (known as *relative efficiency*).

3.5.2 M-estimators

M-estimators are a family of robust estimators that are known to have good relative efficiency and that yield a breakdown point close to $\frac{1}{1+p}$, where p is the number of parameters in the

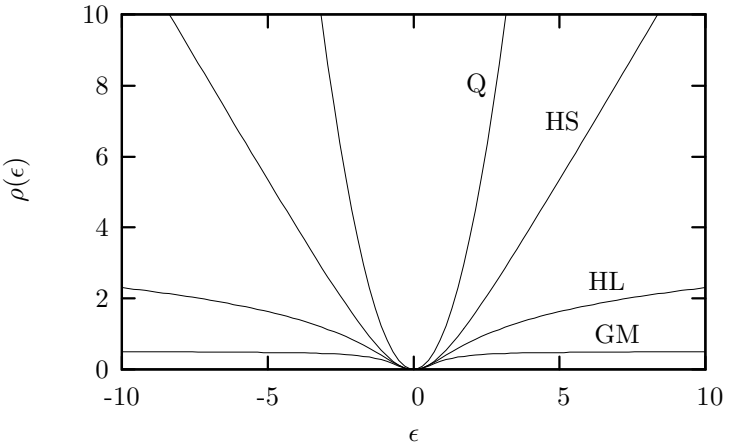
regression (e.g. $p = 2$ in Eq. 3.18). M-estimators minimize an energy function, $J : \mathbb{R}^p \to \mathbb{R}$ defined as the sum of a symmetric, positive-definite function $\rho(e_j)$ of the residuals e_j as:

$$(\hat{\gamma}, \hat{\beta}) = \arg\min_{(\gamma,\beta)} J(\gamma, \beta) = \arg\min_{(\gamma,\beta)} \sum_{j}^{J} \rho(e_j) = \sum_{j}^{J} \rho(y_j - \gamma x_j - \beta). \tag{3.19}$$

Choosing the quadratic function, i.e. for $\rho(e) = e^2$, this becomes the LS estimate. As the function $\rho(\cdot)$ is a function that penalizes the residuals, one can think of the residuals as exercising influence on the total energy that is to be minimized. In order to reduce the influence of outliers on the estimate, one therefore chooses a penalty function $\rho(\cdot)$ that takes on lower values than the quadratic function for large residuals.

Cost functions

Some possible cost functions are shown in Fig. 3.8 (see [234] for others). These are quadradic for small residuals and become linear for large residuals. A hyperparameter, σ_ρ, that acts as a scale or control parameter is introduced into the cost function, $\rho(e_j/\sigma_\rho)$. This scale parameter defines the point of transition between the quadratid and linear parts of the cost function.



name	$\rho(\epsilon)$
Q	ϵ^2
HS	$2\sqrt{1+(\epsilon^2)} - 2$
HL	$\log(1+\epsilon^2)$
GM	$\frac{\epsilon^2}{1+\epsilon^2}$

Figure 3.8: Examples of penalty (weight) functions $\rho(\cdot)$: *Q*, *HS*, *HL* and *GM* (from [35, 34]). As the residual error grows, the influence on the total cost function in Eq. 3.19 diminishes compared to the quadratic *Q* function. The resulting estimator becomes less sensitive to gross errors. The point of transition between the quadratic and linear part is regulated by a scale parameter σ_ρ.

In order to find the optimum of Eq. 3.19, an iterated reweighted least-squares (IRLS) algorithm is most often employed [234]. The resulting estimate depends on the choice of the penalty function as well as the initialization. For convex cost functions, a global optimum exist, but this is not the case for non-convex cost functions (such as the *HL* or *GM* functions in Fig. 3.8). However, these reject outliers much more efficiently than convex cost functions do. This is why an approach of using cost functions sequentially (in continuation) is often used. This is for example favored in [47, 46] where the *HS*, *HL* and *GM* cost functions are applied in succession. Each minimization is in this case initialized with the result of the preceding optimum.

3.5.3 Optimization with half-quadratic theory

Whereas optimization of M-estimators is possible with the IRLS algorithm, half-quadratic (HQ) theory offers more room for interpretation and extensions [77, 78, 34, 36]. The theory provides an elegant way of linearizing the energy function J (Eq. 3.19) for optimization . The basic idea is to rewrite the original energy function into an augmented cost function that involves an auxiliary variable. This function is introduced in such a way that:

- The optimization of the parameters of interest is quadratic when the auxiliary variable is fixed.
- The optimization of the auxiliary variable is explicit when the parameter of interest is fixed.
- An alternate optimization of these two converges towards
  - a global minimum for convex cost-functions $\rho(\cdot)$,
  - a local minimum for non-convex cost-functions.

Note the similarity to the Expectation-Maximization (EM) framework for maximum likelihood estimation. Here an augmented (hidden or missing) variable is introduced to simplify the maximization of the likelihood. The crucial difference is that the augmented variable in this case is considered to be a *random* variable. Instead of alternate optimization in the parameters and the hidden variable, a (marginal) likelihood function is calculated over the hidden variable with fixed parameters (by taking the expectation of the complete-data log-likelihood with respect to the hidden variable) which in turn is maximized in the parameters.

Let $\boldsymbol{\Theta} = (\gamma, \beta)^T$ denote the parameter vector. We can then formulate the half-quadratic optimization as follows. First, the energy function to optimize, $J(\boldsymbol{\Theta})$ (Eq. 3.19), is augmented with the auxiliary variable $\boldsymbol{b} = (b_1 \dots b_J)^T$:

$$\mathcal{J}(\boldsymbol{\Theta}, \boldsymbol{b}) = \sum_j^J [Q(e_j, b_j) + \psi(b_j)]$$

where $Q(\cdot)$ is a quadratic function and $\psi(\cdot)$ is a *dual potential* function. The augmentation is done in such a manner that the original energy function is recovered as the minimum in the auxiliary variable:

$$J(\boldsymbol{\Theta}) = \min_{\boldsymbol{b}} \mathcal{J}(\boldsymbol{\Theta}, \boldsymbol{b}). \tag{3.20}$$

Two types of expansions exist, one multiplicative and one additive, respectively:

$$Q(e_j, b_j) = e_j^2 b_j, \quad j = 1, \dots, J$$

and

$$Q(e_j, b_j) = (b_j - e_j)^2, \quad j = 1, \dots, J$$

for which the minimum with respect to the parameters $\boldsymbol{\Theta}$ is easily obtained. Each of these expansions lead to a different alternating optimization algorithm, in [35, 34, 36] denoted ARTUR and LEGEND. Furthermore, the *dual potential* function $\psi(\cdot)$ (which is different for the two types of expansion) need not be explicitly known, but must only fulfill certain conditions in order to yield the following minima with respect to $\boldsymbol{b}$ (which yields the relation Eq. 3.20):

$$b_j = \frac{\rho'(e_j)}{2e_j}, \quad j = 1, \dots, J$$

for the first expansion, and

$$b_j = e_j \left(1 - \frac{\rho'(e_j)}{2e_j}\right) \quad j = 1, \ldots, J$$

for the second expansion. The conditions imposed on the dual potential functions were originally defined by Geman *et al.* [77, 78], but were later extended by Charbonnier *et al.* [34, 36]. The interested reader is referred to these references for more details. We also note that Huber came to the same results [106]. The two algorithms ARTUR and LEGEND for reconstruction (projection) under PCA and under the PPCA model will be detailed in the next chapter.

3.5.4 Robust methods with PCA

In combination with PCA there are two distinct cases of robustness, one for learning (outliers in the learning data), and one for recognition. We treat the latter first, which is further separated into robust reconstruction with standard PCA and robust reconstruction under a probabilistic PPCA model.

Robust reconstruction by minimizing the reconstruction error in PCA

In PCA (Sec. 3.3.1), we have seen that the projection $\hat{\boldsymbol{x}} = \boldsymbol{W}^T(\boldsymbol{y} - \boldsymbol{\mu})$ (Eq. 3.2) onto the principal components is chosen so as to minimize the square residual of the reconstructed signal $\boldsymbol{e}^T\boldsymbol{e}$, where $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{W}\hat{\boldsymbol{x}}$, (Eq. 3.4). This projection is therefore a least-squares estimate and, hence, sensitive to outliers (occlusions) in the observation $\boldsymbol{y}$. Robustification using M-estimators is straightforward:

$$\hat{\boldsymbol{x}} = \min_{\boldsymbol{x}} \sum_{d=1}^{D} \rho(e_d) = \min_{\boldsymbol{x}} \sum_{d=1}^{D} \rho((\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x})_d).$$

That is, the residual at each pixel is weighted with the robust cost function instead of the quadratic cost function. Black and Jepson [13] explored this for reconstruction and tracking. An alternative to M-estimation was presented by Leonardis and Bischof [138]. They used a robust approach similar to the RANSAC method ([152]), combined with a complexity measure based on the minimum description length principle to compare hypotheses.

Robust reconstruction under a linear model

In [47], Dahyot *et al.*, used half-quadratic theory for robust reconstruction of color images under a PCA-based model. The authors later made a reformulation of the reconstruction problem for the PPCA model [46, 48]. We shall detail this formulation in the next chapter together with our original contribution. This reformulation has wide-reaching impact because it allows the introduction of a prior distribution on the subspace variable. With the prior, the reconstruction problem can be solved using a maximum a posteriori paradigm. Dahyot *et al.* considered a uniform distribution (for the sake of completeness and to show the connection to the ML estimate), a Gaussian distribution, and a non-parametric distribution as priors. The non-parametric distribution was motivated by the findings of Murase and Nayar [163] (see Sec. 3.3.3), but the reconstruction could only be solved approximatively. We present a solution to this latter reconstruction problem in the next chapter together with the MAP estimates for Gaussian and uniform prior subspace distributions. This solution could be found by combining half-quadratic theory with an extension of the mean shift procedure (presented shortly).

Robust PCA

One can distinguish two types of outliers in the learning images used for the calculation of principal components:

- A complete image has been sneaked into the learning set, for example the image of a duck among the images of faces.
- In some of the images there are a limited number of corrupted pixels.

The first case can be tackled by using M-estimators to estimate the covariance matrix, which is then diagonalized. A method that effectively does this without actually calculating the covariance matrix has been proposed by Xu and Yuille [230]. Kamiya and Eguchi [121] extends this idea to a class of methods. The second case has been addressed in [53, 52] and in [196]. De la Torre and Black [53, 52] minimizes the robust reconstruction error based on M-estimators over all the learning images:

$$\min_{\boldsymbol{W},\boldsymbol{\mu},\boldsymbol{x},\boldsymbol{\sigma}} \sum_{j=1}^{J}\sum_{d=1}^{D} \rho\left(\frac{(\boldsymbol{y}-\boldsymbol{\mu}-\boldsymbol{W}\boldsymbol{x})_d}{\sigma_d}\right), \quad \text{where} \quad \boldsymbol{\sigma} = (\sigma_1 \dots \sigma_D)^T.$$

This energy is minimized using a gradient descent scheme. In another approach, Skočaj *et al.* [196] employ an EM-algorithm for calculating the principal components in which outliers can be treated as missing data. These are determined by outlier rejection and can hence be marginalized out of the calculation. Outlier rejection (equivalent to "regression diagnostics" in [234]) consists in first fitting a component using all data and then consider as outliers, data that lie far from the component. The (theoretical) breakdown point remains 0 however [234].

3.5.5 Evaluation of robust techniques

Let us finish this section with a remark on the evaluation of robust techniques. Basically, we design by robust estimation techniques methods that are insensitive to outliers in the data, i.e. robust estimation deals with *unforeseen* working conditions. In order to experimentally evaluate the benefit of robust techniques, it is however necessary to introduce *controlled unforeseen* data, which is paradoxal. Since the unforeseen events have not been explicitly modeled, it is difficult to produce "realistic" (i.e. depending on the application) average results of robust estimation techniques on databases. The theoretical breakdown point does not necessarily provide any help since these are *worst case* performances. In practice we often observe that a method is more robust to "realistic" outliers than would suggest the theoretical breakdown point of the method.

3.6 Non-parametric density estimation and the Mean Shift

A general class of widely used methods for non-parametric density estimation, feature space clustering, kernel-regression and general machine learning are so-called kernel density estimation methods (sometimes also called Parzen windowing). An efficient method for gradient ascent-based optimization of this class of estimates is provided by an old pattern recognition procedure called the *mean shift* [73]. However, this procedure was not widely applied until

recently, and in [40], Comaniciu and Meer presented the mean shift in a general context along with several applications such as discontinuity preserving filtering and image segmentation.

Let us resume briefly the method for Gaussian kernels. For a more general description, the reader is referred to [40]. We consider the Q-dimensional (random) feature variable $\boldsymbol{x}$ with the distribution $p(\boldsymbol{x})$ along with a sample of this variable $\{\boldsymbol{x}_j\}_{j=1}^{J}$. The kernel estimate of the distribution $p(\boldsymbol{x})$ is given by

$$\hat{p}(\boldsymbol{x}) = \sum_{j=1}^{J} \Gamma_j(\boldsymbol{x}), \tag{3.21}$$

with the Gaussian kernels $\Gamma_j(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}_j, \boldsymbol{\Sigma}_x)$, $j = 1, \ldots, J$. Identical, radial symmetric kernels are chosen, which means $\boldsymbol{\Sigma}_x = h^2 \boldsymbol{I}_Q$, where h denotes the bandwidth. The choice of bandwidth is an important issue: the larger the bandwith, the larger the smoothing performed on the data. If it is chosen too large, modes in the distribution will be lost. If it is chosen too small, the estimated density will be "noisy" (irregular, see also [58, p.169-170]).

The gradient of $p(\boldsymbol{x})$ can be estimated by taking the gradient of the estimate $\hat{p}(\boldsymbol{x})$. Since the gradient of a Gaussian is Gaussian it is fairly easy to find the following relation:

$$\nabla \hat{p}(\boldsymbol{x}) = \hat{p}(\boldsymbol{x}) \boldsymbol{\Sigma}_x^{-1} ms(\boldsymbol{x}), \tag{3.22}$$

with the mean shift term:

$$ms(\boldsymbol{x}) = \left[\frac{\sum_{i=1}^{N} \Gamma_i(\boldsymbol{x}) \boldsymbol{x}_i}{\sum_{i=1}^{N} \Gamma_i(\boldsymbol{x})} - \boldsymbol{x} \right]. \tag{3.23}$$

That is, the mean shift term is proportional to the *normalized* (estimated) density gradient (Eq. 3.22). Gradient ascent optimization consists of iterating the mean shift of Eq. 3.23 until convergence. Since the gradient is normalized by the density of the distribution, $\hat{p}(\boldsymbol{x})$, the procedure is an adaptive gradient ascent method where large steps are taken for low probability density (i.e. far from a maximum of the density) and small steps when the density is high (i.e. close to a maximum). Note also that the density estimate itself $\hat{p}(\boldsymbol{x})$ need not be explicitly calculated. We finally note that several extensions have been proposed to the mean shift, such as a quasi-Newton mean shift [231] and variable bandwidth mean shift (i.e. $\boldsymbol{\Sigma}_{x_i}$ instead of $\boldsymbol{\Sigma}_x$) [39].

3.7 Conclusion

In this chapter we have tried to give a comprehensive overview of appearance-based models, dimension reduction techniques, robust estimation and the mean shift procedure. We have seen that global image modeling techniques have been applied with much success in the last decade. Here, PCA is the most widely used method for automatically learning the model parameters. More realistic modeling was proposed by Murase and Nayar [163], who introduced non-linear subspace modeling combined with a linear mapping *from* the subspace *to* the observation space.

In our atlas application we have chosen an appearance-based approach because:

1. We wanted to use a method that can learn from samples. This is to alleviate the problem of modeling explicitly brain geometry, its variance as well as (the unknown) normal brain perfusion.

2. These methods have been used with much success.

Hereby, we have chosen to use a linear model as the basis model since we only have a few samples at disposition. Using non-linear models could be interesting in a future work when more samples are available and when these can be validated based on more knowledge about normal brain perfusion.

In an effort to improve atlas performance, we have refined the linear basis model with a non-parametric, non-linear distribution in the subspace. This method is similar to the method of Murase and Nayar, but is fully non-supervised. The method was made possible by an original extension of the mean shift procedure and presents a natural extension of earlier work in our group.

Finally, we have considered the problem of comparing non-normal images (images with lesions) to the atlas. Since these might be very non-Gaussian, we have modeled these as outliers using robust techniques (half-quadratic theory). Robust modeling has also been a subject of research in our group. All these modifications of the basis linear model have led to the development of our original model which will be presented in the next chapter.

Chapter 4

An original non-Gaussian probabilistic appearance model

In the last chapter we have seen that there exist several variants of global appearance models (PCA-based and others) that have been applied with much success to problems of recognition, detection and tracking in computer vision. Beside their conceptual elegance, they simplify object modeling because they have the capacity to learn from examples. We have seen that important variants of the basic PCA model include versions with non-Gaussian subspace modeling and versions with robust noise modeling. In this chapter, we propose a global, probabilistic model that combines both within a unified mathematical framework.

To apply such a model, two problems need to be solved: (1) the reconstruction problem when the model is known (making inferences under the model), and (2) model parameter estimation (model identification). The first problem is a prerequisite to the second. We take a pragmatic approach to the second problem and use an approximate solution. For the first problem, however, we develop a new algorithm that solves the reconstruction problem using the MAP paradigm. This algorithm has been developed by deriving a procedure for gradient ascent optimization based on the mean shift [40] and combining this procedure with half-quadratic theory [77, 36]. The algorithm makes MAP image reconstruction feasible for high-dimensional images as is demonstrated in later chapters.

This chapter is organized as follows. First, we recall the factor analysis model and the linear PPCA model (probabilistic principal component analysis model) from Ch. 3. These form the basis for our developments. We then pose the image reconstruction problem and recall the solutions to this problem under the basis models. The solutions to the reconstruction problem are then presented for increasingly more general model hypotheses, the final assumptions being a model with non-parametric subspace distribution and non-Gaussian noise. The different models and algorithms that are progressively developed are summarized at the end of the chapter before we discuss some possible paths for future research. Experiments with these models are presented in the next chapter and lengthy calculations are left to the appendix.

4.1 The basis: global linear model with additive noise

Our model is based on the linear image generative model (factor analysis or PPCA model) [4, 212, 211]

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{4.1}$$

which describes the relationship between a Q-dimensional subspace variable $\boldsymbol{x}$ and the D-dimensional observed image $\boldsymbol{y}$ with $Q < D$. The variables of the model are:

- $\boldsymbol{y}$: The $D \times 1$ dimensional, randomly distributed observation (images).
- $\boldsymbol{x}$: The $Q \times 1$ dimensional, randomly distributed subspace variable (latent or hidden variable). We have that $Q < D$ (generally $Q \ll D$).
- $\boldsymbol{W}$: The $D \times Q$ dimensional generation matrix of orthogonal column vectors that define the subspace (feature space).
- $\boldsymbol{\mu}$: The $D \times 1$ dimensional mean.
- $\boldsymbol{\epsilon}$: The $D \times 1$ dimensional, randomly i.i.d. observation-/pixel-noise.

In this model we have two independent random variables: the subspace variable $\boldsymbol{x}$ and the observation noise $\boldsymbol{\epsilon}$. The primary model parameters are given by $\boldsymbol{W}$ and $\boldsymbol{\mu}$. The properties of the model depend on the distributions assumed for the independent random variables, $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$. We shall refer to the factor analysis assumptions (Sec. 3.4.4, [212, 211]) where $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\epsilon$ are diagonal covariance matrices as the *basic assumptions*:

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_x), \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon), \tag{4.2}$$

which becomes the PPCA model [212, 211] for

$$\boldsymbol{\Sigma}_x = \boldsymbol{I}_Q, \quad \text{and} \quad \boldsymbol{\Sigma}_\epsilon = \sigma^2 \boldsymbol{I}_D. \tag{4.3}$$

Let us recall from Sec. 3.3.5 the different probability densities under the *basic assumptions* (slightly reformulated for diagonal covariance matrices, $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_\epsilon$). These densitites are all Gaussian:

- The distribution of the observation is given by:

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_y), \quad \text{with} \quad \boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_\epsilon + \boldsymbol{W}\boldsymbol{W}^T.$$

- The conditional distribution of the observation $\boldsymbol{y}$ given the subspace variable $\boldsymbol{x}$ is given by:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu}, \boldsymbol{\Sigma}_\epsilon). \tag{4.4}$$

- The posterior distribution of the latent variable given the observation is

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{x|y}), \tag{4.5}$$

  where

$$\boldsymbol{\mu}_x = \boldsymbol{\Sigma}_{x|y}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{y} - \boldsymbol{\mu}), \tag{4.6}$$

  and

$$\boldsymbol{\Sigma}_{x|y} = (\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W})^{-1}.$$

There is an intuitive way to think about this model (from [189]). First, white noise is used to generate a spherical ball of density in the Q-dimensional subspace. This ball is then stretched and rotated into the D-dimensional observation space by the matrix $\boldsymbol{W}$ where it looks like a Q-dimensional pancake. The pancake is then convolved with the covariance density of $\boldsymbol{\epsilon}$ (described by $\boldsymbol{\Sigma}_\epsilon$) to get the final distribution of $\boldsymbol{y}$.

Let us summarize the characteristics of this model which forms a basis for our model:

- The model is linear
- It is a factor analysis model (Sec. 3.4.4, [212, 211])
- Correlations between the observation variables y_d are "explained" by the term $\boldsymbol{W}\boldsymbol{x}$. We can think of the matrix $\boldsymbol{W}\boldsymbol{W}^T$ as a correlation matrix.
- Equivalently, the components y_d of the observation are statistically independent given the subspace variable $\boldsymbol{x}$.
- The model can serve as a compact representation of high dimensional images ($Q \ll D$).
- The model is generative, i.e. we can generate new images from the same distribution.

For convenience, we shall in the following denote the mean-free observation by: $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\mu}$.

4.1.1 Image reconstruction under the model[1]

We have seen in Ch. 3 that the subspace spanned by $\boldsymbol{W}$ can serve as a feature space for, among others, classification (recognition). For this, learning images and images to classify must be projected into this subspace. Under the probabilistic model, this is a problem of *inference*: given fixed model parameters, $\{\boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\Sigma}_x\}$, we want to estimate the hidden variable (subspace variable) $\boldsymbol{x}$ that generated the observation. This is also known as the reconstruction problem and can be formulated as either a maximum likelihood (ML) or a maximum a posteriori (MAP) estimation problem, i.e.

$$\hat{\boldsymbol{x}}_{ML} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x}), \quad \text{or} \quad \hat{\boldsymbol{x}}_{MAP} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}),$$

respectively, where $\boldsymbol{y}$ is fixed. These estimates can be solved analytically under the *basic assumptions* (Eq. 4.2). For the ML estimate it suffices to minimize the negative logarithm of Eq. 4.4

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}}(-\log p(\boldsymbol{y}|\boldsymbol{x})) = \arg\min_{\boldsymbol{x}} \left((\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x})^T \boldsymbol{\Sigma}_\epsilon^{-1} (\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x})\right),$$

which is solved as

$$\hat{\boldsymbol{x}}_{WML} = (\boldsymbol{W}^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^T \boldsymbol{\Sigma}_\epsilon^{-1} \tilde{\boldsymbol{y}}, \tag{4.7}$$

by setting the derivation to zero. This estimate is also known as the weighted least squares estimate (WLS) (we have chosen the subscript WML). For isotropic noise $\boldsymbol{\Sigma}_\epsilon = \sigma^2 \boldsymbol{I}_D$, the WML estimate becomes the (unweighted) least squares (LS) estimate (subscript ML)[2]:

$$\hat{\boldsymbol{x}}_{ML} = (\boldsymbol{W}^T \boldsymbol{W})^{-1} \boldsymbol{W}^T \tilde{\boldsymbol{y}}, \tag{4.8}$$

[1] Note, that we use of the word *reconstruction* in a different manner than in tomographic imaging where it denotes the process of calculating a volumetric image from two-dimensional projections.

[2] The term $(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$ is also known as the *pseudoinverse* of the non-square matrix $\boldsymbol{W}$, i.e. $(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_Q$.

which is also the solution in the PPCA case (Eq. 4.3).

The MAP estimate is simply given by the mean of the posterior distribution in Eq. 4.5:

$$\hat{\boldsymbol{x}}_{GWMAP} = \boldsymbol{\mu}_x = (\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}. \tag{4.9}$$

For isotropic noise (PPCA), this estimate becomes:

$$\hat{\boldsymbol{x}}_{GMAP} = (\sigma^2\boldsymbol{I}_Q + \boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\tilde{\boldsymbol{y}}. \tag{4.10}$$

For a uniform prior distribution of the subspace variable, the MAP and ML estimates become equal, as is simply seen from:

$$\arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x}).$$

A uniform distribution is also called a *non-informative* prior [58], and is equivalent to assuming that the subspace variable is a deterministic parameter and not a random variable. This is the way the reconstruction problem is solved in standard PCA.

For orthonormal subspace vectors ($\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_Q$), there are intuitive geometric interpretations to these estimates. The ML estimate above is simply the orthogonal projection into the subspace. This is seen by taking the scalar product between the reconstructed observation and the residual, which yields zero. For the weighted ML estimate, the observation axes are scaled according to the noise variance matrix before orthogonal projection. The GMAP and GWMAP projections are skewed toward the origin (the prior distribution on $\boldsymbol{x}$ having zero mean) with respect to the ML and WML estimates respectively.

4.1.2 Model estimation

Whereas for reconstruction, or inference, the model parameters are assumed to be known, learning or *system identification* consists in estimating all model parameters. This is a more general inference problem where the goal is to estimate $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{\mu}, \Sigma_\epsilon\}$ from the observations $\{\boldsymbol{y}_j\}_{j=1}^J$. ML estimation is more common for this than MAP estimation. For the PPCA model the analytic ML estimate of the model parameters were derived by Tipping and Bishop [212, 211]. The results were presented in Sec. 3.3.5 p. 39. For the factor analysis model the estimate must be determined using an optimization algorithm, for example the EM-algorithm [189][3]. Note that it is necessary to solve the reconstruction (inference) problem before one can proceed to the model estimation problem.

We have chosen to estimate the image generating matrix, $\boldsymbol{W}$, by the whitened eigenvectors (as found by PCA) $\boldsymbol{W} = \boldsymbol{U}_Q\boldsymbol{\Lambda}_Q^{1/2}$ ($\boldsymbol{U}_Q$ is the matrix of the Q principal components of the sample covariance matrix, $\boldsymbol{\Lambda}_Q$ the Q first eigenvalues). This way we assure a compact representation. For the PPCA model, this estimate is close to the ML estimate as can be seen from (Eq. 3.15, p. 3.15):

$$\hat{\boldsymbol{W}} = \boldsymbol{U}_Q(\boldsymbol{\Lambda}_Q - \sigma^2\boldsymbol{I}_Q)^{1/2}\boldsymbol{R},$$

which asymptotically approaches the whitened eigenvectors for $\sigma \to 0$ (recall that $\boldsymbol{R}$ is an arbitrary rotation matrix that does not change the model). For the mean, $\boldsymbol{\mu}$, we have used the sample mean, and for the noise variance σ^2 we have used the variance of the pixel residuals of the learning images:

$$\hat{\sigma}^2 = \frac{1}{D-Q-1}\mathrm{tr}\left(\frac{1}{J}\sum_j^J \boldsymbol{e}_j\boldsymbol{e}_j^T\right), \quad \text{or equivalently,} \quad \hat{\sigma}^2 = \frac{\sum_{q=Q+1}^J \lambda_q}{D-Q-1}, \tag{4.11}$$

[3]For PCA and PPCA, the solution can also be found using an EM-algorithm, see [188] and [212].

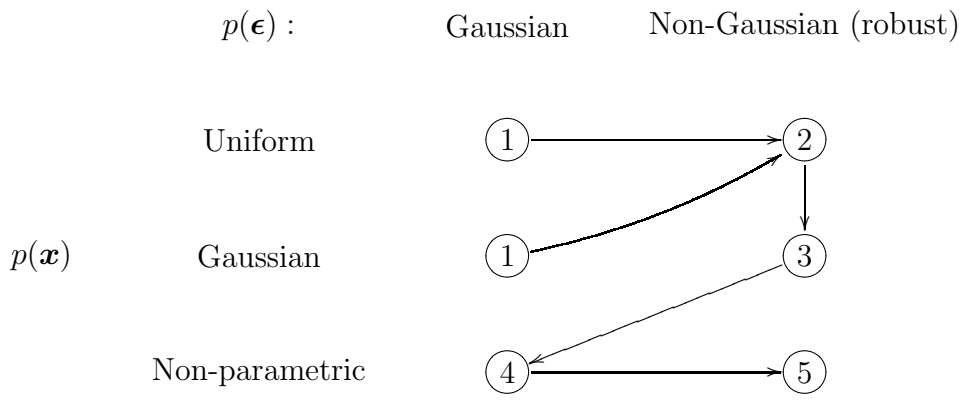


Figure 4.1: Overview of model assumptions and how we proceed in the derivation of the most general model.

with the sum of the truncated, non-zero eigenvalues found in the learning stage. This estimate depends on the number of eigenvectors used for reconstruction. Note that the estimate of $\boldsymbol{W}$ by the eigenvectors is *not* the ML estimate for the non-Gaussian model introduced in the following. This is no limitation, however, because the derived inference algorithms are valid for any full rank matrix $\boldsymbol{W}$. We shall come back to the issue of model parameter estimation when we discuss paths of future work at the end of this chapter.

4.2 Generalizing the hypotheses

Motivated by the good experimental results obtained by Dahyot *et al.* [46, 47, 48], we have sought to make inference under the model in the case of more general model assumptions than the *basic assumptions* Eq. 4.2. We consider generalizations to the distribution of the subspace variable $\boldsymbol{x}$ and the observation noise $\boldsymbol{\epsilon}$. In the final model we consider a non-parametric distribution of $\boldsymbol{x}$ and a non-Gaussian (robust) noise distribution of $\boldsymbol{\epsilon}$.

In order to derive the final model, we proceed in a step-wise fashion as described in the flow chart in Fig. 4.1. A uniform distribution of $\boldsymbol{x}$ is considered for completeness, recall that in this case the MAP estimate becomes equivalent to the ML estimate. In Fig. 4.1 the estimates for the *basic assumptions* ① are the standard ML and MAP estimates already described in Sec. 4.1.1, (Eqs. 4.7 - 4.10). We therefore proceed to a non-Gaussian (robust) noise distribution, case ② in the next section. This model has been considered by Black and Jepson [13] and an estimation scheme based on half-quadratic theory was proposed by Dahyot *et al.* [47, 46, 48] (see Sec. 3.5.4). Dahyot *et al.* also proposed a solution for case ③ (non-Gaussian noise, Gaussian subspace modeling), which is simply obtained by modifying one step of the robust ML estimation algorithm. We will then present the modified mean shift algorithm which provides an elegant solution to the reconstruction problem under the model hypothesis marked as ④ in the diagram (Gaussian noise, non-parametric subspace modeling). With this modified mean shift algorithm it is further straightforward to extend the algorithm to non-Gaussian noise using half-quadratic theory, case ⑤. By keeping the general notation of a diagonal noise covariance matrice, $\boldsymbol{\Sigma}_\epsilon$, we can at every step consider a non-isotropic or an isotropic noise distribution, the estimates in the former case will contain an extra subscript W - for "weighted". The complete calculations can be found in App. A.

4.3 Case 2: non-Gaussian noise, uniform subspace distribution

In order to render the estimates robust to non-Gaussian noise (outliers), we reformulate the noise distribution as

$$p(\boldsymbol{\epsilon}) \propto \exp(-\frac{1}{2}J(\tilde{\boldsymbol{\epsilon}})) = \exp(-\frac{1}{2}\sum_{d=1}^{D}\rho(\tilde{\epsilon}_d)), \quad \tilde{\boldsymbol{\epsilon}} = \frac{1}{\sigma_\rho}\boldsymbol{\Sigma}_\epsilon^{-1/2}\boldsymbol{\epsilon}, \tag{4.12}$$

where the explicit proportional factor will not be needed. The function $\rho(\cdot)$ is one of the robust cost functions described in Sec. 3.5.2, p. 52, the noise is scaled by the scale factor σ_ρ and the standard deviations $\boldsymbol{\Sigma}_\epsilon^{-1/2}$. To solve the reconstruction problem using the ML paradigm (which is equivalent to a uniform prior on $\boldsymbol{x}$), we need to solve

$$\arg\min_{\boldsymbol{x}}(-\log p(\boldsymbol{y}|\boldsymbol{x})) = \arg\min_{\boldsymbol{x}} J\left(\frac{1}{\sigma_\rho}\boldsymbol{\Sigma}_\epsilon^{-1/2}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{x})\right) = \arg\min_{\boldsymbol{x}}\sum_{d=1}^{D}\rho(\tilde{e}_d),$$

which is exactly the M-estimator as described in Sec. 3.5.2. Furthermore, as described in Sec. 3.5.3, half-quadratic theory provides two different algorithms for optimizing this target function. This is done by introducing an auxiliary variable, $\boldsymbol{b}$, and then performing an alternating coordinate descent in the auxiliary variable and the subspace variable. The auxiliary variables are also called weights. We shall detail the two algorithms in the following two subsections. Note that there is no longer any simple geometric interpretation for the robust ML estimates - the projection is non-orthogonal.

4.3.1 ARTUR (multiplicative expansion)

With the notation, $\boldsymbol{b} = (b_1 \dots b_D)^T$ and $\boldsymbol{B} = diag(\boldsymbol{b})$, the multiplicative expansion [77, 36] of the energy function $J(\boldsymbol{\epsilon})$ is given by the expression

$$\mathcal{J}(\boldsymbol{\epsilon}, \boldsymbol{b}) = \tilde{\boldsymbol{\epsilon}}^T\boldsymbol{B}\tilde{\boldsymbol{\epsilon}} + \Psi(\boldsymbol{b}), \tag{4.13}$$

with the residual

$$\tilde{\boldsymbol{\epsilon}} = \frac{1}{\sigma_\rho}\boldsymbol{\Sigma}_\epsilon^{-1/2}(\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x}). \tag{4.14}$$

We have from half-quadratic theory that the minimum in the weights is given by (with the realization $\tilde{\boldsymbol{e}}$ of $\tilde{\boldsymbol{\epsilon}}$):

$$b_d = \frac{\rho'(\tilde{e}_d)}{2\tilde{e}_d}, \quad d = 1, \dots, D. \tag{4.15}$$

With the weights $\boldsymbol{b}$ fixed, the minimum in $\boldsymbol{x}$ is simply calculated by minimizing the first term in Eq. 4.13, which becomes the WML-estimate (Eq. 4.7) with modified noise variance:

$$\boldsymbol{x} = (\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{B}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{B}\tilde{\boldsymbol{y}}. \tag{4.16}$$

Robust weighted ML (non-isotropic noise) reconstruction using ARTUR is thus done by iterating Eqs. 4.14–4.16. For robust ML (isotropic noise) reconstruction the last step is simplified to:

$$\boldsymbol{x} = (\boldsymbol{W}^T\boldsymbol{B}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{B}\tilde{\boldsymbol{y}}.$$

4.3.2 LEGEND (additive expansion)

The additive expansion [78, 35] of the energy function $J(\boldsymbol{\epsilon})$ is given by the expression

$$\mathcal{J}(\boldsymbol{\epsilon}, \boldsymbol{b}) = (\tilde{\boldsymbol{\epsilon}} - \boldsymbol{b})^T(\tilde{\boldsymbol{\epsilon}} - \boldsymbol{b}) + \Psi(\boldsymbol{b}), \tag{4.17}$$

again with the residual

$$\tilde{\boldsymbol{\epsilon}} = \frac{1}{\sigma_\rho}\boldsymbol{\Sigma}_\epsilon^{-1/2}(\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x}). \tag{4.18}$$

We have from half-quadratic theory that the minimum in the weights is given by:

$$b_d = \tilde{e}_d\left(1 - \frac{\rho'(\tilde{e}_d)}{2\tilde{e}_d}\right), \quad d = 1, \ldots, D. \tag{4.19}$$

With the weights $\boldsymbol{b}$ fixed, the minimum in $\boldsymbol{x}$ is simply calculated by minimizing the first term in Eq. 4.17, which after derivation and equating to zero becomes:

$$\boldsymbol{x} = (\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}(\tilde{\boldsymbol{y}} - \sigma_\rho\boldsymbol{\Sigma}_\epsilon^{1/2}\boldsymbol{b}). \tag{4.20}$$

Robust weighted ML (non-isotropic noise) reconstruction using LEGEND is thus done by iterating Eqs. 4.18–4.20. For robust ML (isotropic noise) reconstruction the last step is replaced by:

$$\boldsymbol{x} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T(\tilde{\boldsymbol{y}} - \sigma_\rho\sigma\boldsymbol{b}),$$

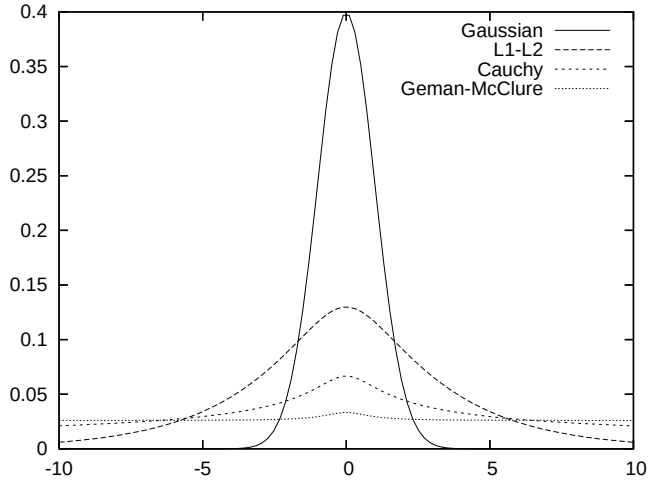
with the square root of the isotropic noise variance σ^2.

4.3.3 Probabilistic interpretation

We claimed that the function in Eq. 4.12 is a probability density function. This is only justified when the function is integrable. The cost functions we have used (Fig. 3.8, p. 52), were chosen because they efficiently reject outliers. However, only the first function (i.e. *HS*) is integrable. Thus, rigorously speaking, we only have a true probabilistic model for this cost function, but not for the two others. The distributions are depicted in Fig. 4.2 (the *HL* and *GM* functions are represented with an arbitrary scaling). Note also that the robust ML solutions are exactly the same as for standard robust PCA reconstruction (Sec. 3.5.4, p. 54). However, there is a subtle difference: since there is no noise model in standard PCA, the scaling of the residual, Eqs. 4.14 and 4.18, is regulated by only one parameter that fuses together σ_ρ and the model noise. This makes the choice of this parameter somewhat more arbitrary.

4.3.4 Interpretation of weights

The auxiliary variable introduced in the half-quadratic expansion has an interpretation. Every element of $\boldsymbol{b}$ can only take on values between 0 and 1. As can be seen in Eq. 4.16, for $b_d = 0$, the corresponding y_d is set to zero and thus has no influence on the estimate. The pixel d is therefore an outlier. On the other hand, when all $b_d = 1, d = 1, \ldots, D$, we obtain the usual WML estimate: all the corresponding $y_d, d = 1, \ldots, D$ flow into the estimate.



name	$\rho(x)$
Q	x^2
HS	$2(\sqrt{1+(x^2/2)}-1)$
HL	$\frac{c^2}{2}\log(1+(x/c)^2)$
GM	$\frac{x^2/2}{1+x^2}$

Figure 4.2: The resulting probability density functions and quasi-probability density functions resulting from equation Eq. 4.12.

4.3.5 Computational issues

Note that for the ARTUR algorithm, we have to perform a matrix inversion at each iteration (Eq. 4.16). For the LEGEND algorithm, it is only necessary to perform the matrix inversion once as seen in Eq. 4.20. This makes each iteration of ARTUR more costly than for LEGEND when Q becomes larger. However, investigations on the convergence rate of these two algorithms have shown that the ARTUR algorithm converges in fewer iterations than the LEGEND algorithm, [106, 46] and more recently [170]. A good compromise, proposed by Dahyot [46] and that we practice, is therefore to iterate a few times with the ARTUR algorithm, and then switch to the LEGEND algorithm.

4.4 Case 3: non-Gaussian noise, Gaussian subspace distribution

Robust MAP reconstruction with a Gaussian prior is a straightforward extension that was already derived by Dahyot in [46]. The function to optimize is the posterior probability $p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{p(\boldsymbol{y})}p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$, from which $p(\boldsymbol{y})$ can be left out. Furthermore, it suffices to perform the same half-quadratic expansion on $p(\boldsymbol{y}|\boldsymbol{x})$ as in the last section. The minimum of the expanded function $\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{b})p(\boldsymbol{x})$ in $\boldsymbol{b}$ with $\boldsymbol{x}$ fixed is given by the same formulas as above (Eqs. 4.15 and 4.19). For optimization in $\boldsymbol{x}$ with $\boldsymbol{b}$ fixed, we only need to minimize $-\log\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{b})|_{\boldsymbol{b}=\hat{\boldsymbol{b}}}\ p(\boldsymbol{x})$, which is

$$-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x} - \frac{1}{2}\tilde{\boldsymbol{e}}^T\boldsymbol{B}\tilde{\boldsymbol{e}} + cst,$$

for the multiplicative expansion (ARTUR). Optimization is done by replacing Eq. 4.16 with

$$\boldsymbol{x} = (\sigma_\rho^2\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{B}\tilde{\boldsymbol{y}},$$

and iterating until convergence, which yields $\hat{\boldsymbol{x}}_{\mathrm{RGWMAP}}$. For the additive expansion (LEGEND), we obtain

$$\boldsymbol{x} = (\sigma_\rho^2\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}(\tilde{\boldsymbol{y}} - \sigma_\rho\boldsymbol{\Sigma}_\epsilon^{1/2}\boldsymbol{b}),$$

which is substituted for Eq. 4.20 to do robust MAP reconstruction with a Gaussian prior.

4.5 Case 4: Gaussian noise, non-Gaussian subspace distribution

We now proceed to present our own contribution to the subspace-based image estimation problem. An estimation scheme, relying on a modified version of the mean shift algorithm [40] combined with half-quadratic theory [77, 36] is derived in the general case of non-Gaussian noise and non-Gaussian subspace models. As the transition from Gaussian noise to non-Gaussian noise can be solved by half-quadratic theory, it is necessary to solve the estimation problem for Gaussian noise first, case ④ in Fig. 4.1.

4.5.1 Non-Gaussian subspace distribution

In order to model any non-Gaussian distribution in the subspace, we make use of kernel density estimation with Gaussian kernels[4]:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{J}\sum_{j=1}^{J} p_j(\boldsymbol{x}) = \frac{1}{J}\frac{1}{(\sqrt{2\pi})^Q|\boldsymbol{\Sigma}_x|^{1/2}}\sum_{j=1}^{J}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_j)^T\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{x}-\boldsymbol{x}_j)), \tag{4.21}$$

where the (symmetric) covariance matrix Σ_x is fixed across all samples. The samples of $\boldsymbol{x}$ are taken as the ML estimates (orthogonal projections) of the learning samples (images) into the subspace. We now consider Gaussian noise which leads to the "Modified Mean Shift" operator for image reconstruction.

4.5.2 Densities under Gaussian noise

Density of the observation

The probability density of the observation is of no immediate interest for deriving the modified mean shift, but it is derived for completeness and interpretation. With Eqs. 4.4 and 4.21, we obtain:

$$\begin{aligned} p(\boldsymbol{y}) &= \int_{-\infty}^{\infty} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}x = \int_{-\infty}^{\infty} p(\boldsymbol{y}|\boldsymbol{x})\frac{1}{J}\sum_{j=1}^{J} p_j(\boldsymbol{x})\mathrm{d}x \\ &= \frac{1}{J}\sum_{j=1}^{J}\int_{-\infty}^{\infty} p(\boldsymbol{y}|\boldsymbol{x})p_j(\boldsymbol{x})\mathrm{d}x = \frac{1}{J}\sum_{j=1}^{J} p_j(\boldsymbol{y}), \end{aligned} \tag{4.22}$$

where each Gaussian (see for example [4, p.553])

$$p_j(\boldsymbol{y}) = \frac{1}{(\sqrt{2\pi})^Q|\boldsymbol{\Sigma}_y|^{1/2}}\exp(-\frac{1}{2}(\tilde{\boldsymbol{y}}-\boldsymbol{W}\boldsymbol{x}_j)^T\boldsymbol{\Sigma}_y^{-1}(\tilde{\boldsymbol{y}}-\boldsymbol{W}\boldsymbol{x}_j)), \tag{4.23}$$

has the same covariance matrix

$$\boldsymbol{\Sigma}_y = \boldsymbol{W}\boldsymbol{\Sigma}_x\boldsymbol{W}^T + \boldsymbol{\Sigma}_\epsilon.$$

[4]For notational clarity, we drop the hat notation in the following and consider the estimated distribution $\hat{p}(\boldsymbol{x})$ to be the true distribution $p(\boldsymbol{x})$.

This distribution is therefore a mixture of factor analysis models, however with fixed covariance matrices – a "kernel factor analysis" model. We see that the global density Eq. 4.23 may be completely non-Gaussian.

Posterior density

For optimization, we need to compute the posterior distribution of the subspace variable given the observation, $p(\boldsymbol{x}|\boldsymbol{y})$. From Bayes formula we have

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} = \frac{1}{p(\boldsymbol{y})}\frac{1}{J}\sum_{j=1}^{J} p(\boldsymbol{y}|\boldsymbol{x})p_j(\boldsymbol{x}), \tag{4.24}$$

which can be rewritten into a kernel density function as (App. A.1)

$$p(\boldsymbol{x}|\boldsymbol{y}) = c\frac{1}{J}\sum_{j=1}^{J} c_j \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)) = c\frac{1}{J}\sum_{j=1}^{J} c_j\Gamma_j(\boldsymbol{x}), \tag{4.25}$$

where c is a constant,

$$\boldsymbol{\Sigma} = (\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W} + \boldsymbol{\Sigma}_x^{-1})^{-1} \tag{4.26}$$

and

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}(\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}} + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j). \tag{4.27}$$

The posterior probability is therefore also a kernel estimate, but this time each Gaussian kernel has the covariance $\boldsymbol{\Sigma}$ and is centered at the mean $\boldsymbol{\mu}_j$. The mean is composed of a term that depends on the observation and a term that depends on the learning samples. The kernels are weighted by the coefficients

$$c_j = \exp(-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_x)^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_x)) \tag{4.28}$$

where

$$\boldsymbol{\Psi} = (\boldsymbol{I}_Q - \boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1})^{-1}\boldsymbol{\Sigma}_x, \tag{4.29}$$

and

$$\boldsymbol{\mu}_x = (\boldsymbol{I}_Q - \boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1})^{-1}\boldsymbol{\Sigma}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}. \tag{4.30}$$

These weights thus decay with increasing distance from $\boldsymbol{\mu}_x$, which again depends on the observation, but not on the learning samples.

With the kernels $\Gamma_j(\cdot)$, the global normalization constant c in Eq. 4.25 can be calculated by integrating the probability

$$\frac{1}{c}\int p(\boldsymbol{x}|\boldsymbol{y})\mathrm{d}x = \int \frac{1}{J}\sum_{j=1}^{J} c_j\Gamma_j(\boldsymbol{x})\mathrm{d}x = \frac{1}{J}\sum_{j=1}^{J} c_j \int \Gamma_j(\boldsymbol{x})\mathrm{d}x \tag{4.31}$$

which yields

$$c = \frac{1}{\sqrt{2\pi}^Q|\boldsymbol{\Sigma}|^{1/2}}\frac{J}{\sum_{j=1}^{J} c_j}. \tag{4.32}$$

4.5.3 Modified Mean Shift

For MAP optimization of the posterior distribution, we perform a gradient ascent on $p(\boldsymbol{x}|\boldsymbol{y})$. This leads us to the modified mean shift expression, $mms(\boldsymbol{x})$, which is obtained by taking the gradient of Eq. 4.25 as follows:

$$\begin{aligned}\nabla p(\boldsymbol{x}|\boldsymbol{y}) &= c\frac{1}{J}\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{J}(\boldsymbol{\mu}_j - \boldsymbol{x})c_j\Gamma_j(\boldsymbol{x}) \\ &= c\frac{1}{J}\boldsymbol{\Sigma}^{-1}\left[\sum_{j=1}^{J}c_j\Gamma_j(\boldsymbol{x})\right]\left[\frac{\sum_{j=1}^{J}c_j\Gamma_j(\boldsymbol{x})\boldsymbol{\mu}_j}{\sum_{j=1}^{J}c_j\Gamma_j(\boldsymbol{x})} - \boldsymbol{x}\right] \\ &= p(\boldsymbol{x}|\boldsymbol{y})\boldsymbol{\Sigma}^{-1}mms(\boldsymbol{x}).\end{aligned} \tag{4.33}$$

This modified mean shift expression is exactly the same as the mean shift expression in [40] (also Eq. 3.23 p. 56), using a Gaussian kernel to estimate the density of the shifted samples $\boldsymbol{\mu}_j$ and with a bandwidth matrix $\boldsymbol{\Sigma}$. As for the standard mean shift, we have that the modified mean shift term is proportional to the gradient. The proportional factor is again inversely proportional to the density itself, from which the desirable property of an adaptive gradient ascent follows. With Eq. 4.33, the convergence to a local maximum of an algorithm based on the modified mean shift, $mms(\boldsymbol{x})$, can be proven in the same way as for the original mean shift. This proof is shown in App. A.3.

The modified mean shift expression in (4.33) can be simplified by eliminating from the quotient factors of $c_j\Gamma_j(\boldsymbol{x})$ not depending on j. This simplification is derived in App. A.2 and is more efficient to calculate than the modified mean shift in Eq. 4.33. The final expression thus becomes:

$$mms(\boldsymbol{x}) = \underbrace{\boldsymbol{\Sigma}\left(\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}} + \frac{\sum_{j=1}^{J}\Theta_j(\boldsymbol{x})\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j}{\sum_{j=1}^{J}\Theta_j(\boldsymbol{x})}\right)}_{\boldsymbol{x}^{new}} - \boldsymbol{x}, \tag{4.34}$$

where

$$\Theta_j(\boldsymbol{x}) = \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_j)^T\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{x}-\boldsymbol{x}_j)). \tag{4.35}$$

In Eq. 4.34 the first two terms give the new location $\boldsymbol{x}^{new}$, which is iteratively recalculated until convergence. This new location, $\boldsymbol{x}^{new}$, is composed of a "subspace mean shift" term (prior information, or regularization term) and a quasi-orthogonal projection, say $\boldsymbol{x}_{\tilde{y}}$, of the mean-free observation $\tilde{\boldsymbol{y}}$ (data term), both normalized by the posterior covariance $\boldsymbol{\Sigma}$. In practice, it suffices to calculate the data term $\boldsymbol{x}_{\tilde{y}}$ once. At each iteration, we then calculate $\boldsymbol{x}^{new}$ by adding the subspace mean shift of $\boldsymbol{x}^{old}$ to $\boldsymbol{x}_{\tilde{y}}$:

$$\begin{aligned}\boldsymbol{x}^{new} =& mms(\boldsymbol{x}^{old}) - \boldsymbol{x}^{old} \\ =& \boldsymbol{\Sigma}\left(\boldsymbol{x}_{\tilde{y}} + \frac{\sum_{j=1}^{J}\Theta_j(\boldsymbol{x}^{old})\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j}{\sum_{j=1}^{J}\Theta_j(\boldsymbol{x}^{old})}\right).\end{aligned}$$

4.6 Case 5: non-Gaussian noise, non-Gaussian subspace distribution

The extension to robust noise follows the same scheme as for the robust ML/MAP estimates for uniform and Gaussian priors (Secs. 4.3 and 4.4): expand the distribution $p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$ with an

auxiliary variable and perform alternate optimization. For this we perform the multiplicative expansion of the noise distribution as in Eq. 4.13, p. 64. This leads to an adjusted noise (recall that all variance matrices are diagonal)

$$\boldsymbol{\Sigma}_{adj}^{-1} = \frac{1}{\sigma_\rho}\boldsymbol{\Sigma}_\epsilon^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}_\epsilon^{-1/2}\frac{1}{\sigma_\rho} = \frac{1}{\sigma_\rho^2}\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{B},$$

which replaces $\boldsymbol{\Sigma}_\epsilon$ in Eqs. 4.26 and 4.34:

$$mms(\boldsymbol{x}) = (\boldsymbol{W}^T\boldsymbol{\Sigma}_{adj}^{-1}\boldsymbol{W} + \boldsymbol{\Sigma}_x^{-1})^{-1}\left(\boldsymbol{W}^T\boldsymbol{\Sigma}_{adj}^{-1}\tilde{\boldsymbol{y}} + \frac{\sum_{j=1}^{J}\Theta_j(\boldsymbol{x})\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j}{\sum_{j=1}^{J}\Theta_j(\boldsymbol{x})}\right) - \boldsymbol{x}.$$

With fixed weights $\boldsymbol{b}$, we therefore optimize using the modified mean shift with an adjusted covariance matrix $\boldsymbol{\Sigma}_\epsilon \rightarrow \boldsymbol{\Sigma}_{adj}$. With fixed $\boldsymbol{x}$, we have given the minimum in $\boldsymbol{b}$ from half-quadratic theory. The final algorithm is sketched below (Alg. 1).

Algorithm 1 Robust, modified mean shift optimization

$\boldsymbol{B} \leftarrow \boldsymbol{I}_D, \boldsymbol{\Sigma}_{adj} \leftarrow \sigma_\rho^2\boldsymbol{\Sigma}_\epsilon\boldsymbol{B}^{-1}, \boldsymbol{x}^{old} \leftarrow \hat{\boldsymbol{x}}_{ML}$
repeat
  repeat
    $\boldsymbol{x}^{new} \leftarrow \boldsymbol{x}^{old} + mms(\boldsymbol{x}^{old})$
  until inner loop convergence
  $\tilde{\boldsymbol{e}} \leftarrow \frac{1}{\sigma_\rho}\boldsymbol{\Sigma}_\epsilon^{-1/2}(\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x}^{new})$
  $b_d \leftarrow \frac{\rho'(\tilde{e}_d)}{\tilde{e}_d}$
  $\boldsymbol{\Sigma}_{adj} \leftarrow \sigma_\rho^2\boldsymbol{\Sigma}_\epsilon\boldsymbol{B}^{-1}$
until outer loop convergence

We have not yet derived any convergence proof of this robust, modified mean shift algorithm. Since the outer (half-quadratic) loop converges for ARTUR with an inner loop making a quadratic optimization, one would expect this complete scheme to converge as well since the inner loop is guaranteed to converge. However, this may not necessarily be true. In practice, we have never encountered convergence problems, only division-by-zero problems (subspace mean shift term in Eq. 4.34) when we try to "mean shift" too far from the prior distribution. It is clear that the posterior density function Eq. 4.25 can have multiple maxima, so that any convergence can only be guaranteed to a local optimum.

4.7 Summary of models and algorithms

In this chapter, we have proceeded by presenting increasingly more general models in order to arrive at a comprehensive model with a non-Gaussian prior distribution combined with a non-Gaussian (robust) noise distribution. For an overview of the different declinations of the model, we have summarized the different assumptions and algorithms in Tab. 4.1. A geometrical interpretation with comparison of the principal methods is shown in Fig. 4.3.

4.8 Conclusion and future work

In this chapter we have developed an original linear generative model that is non-Gaussian. First, we described the basic model and we stated the reconstruction problem. We then

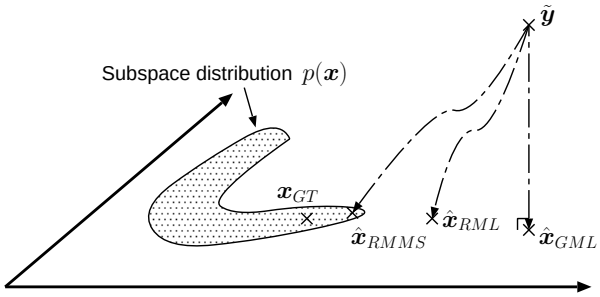


Figure 4.3: Example of a two dimensional subspace. We imagine that an image was generated from the ground truth subspace variable $\boldsymbol{x}_{GT}$ and that this image has been subject to some kind of non-Gaussian image degradation (occlusion, background clutter...) that yields the observation $\tilde{\boldsymbol{y}}$. This occluded image does not, in general, lie orthogonally to the subspace. Robust noise modeling and maximum likelihood reconstruction, tries to account for the non-Gaussian image degradation in order to correctly estimate the subspace variable that generated the original image, $\boldsymbol{x}_{RML}$. Robust MAP estimation goes one step further by additionally taking into account the non-Gaussian *a priori* distribution $p(\boldsymbol{x})$ of the subspace variable, which yields the estimate $\boldsymbol{x}_{RMMS}$.

progressed in a systematic manner to our non-Gaussian model for which an algorithm was developed that solves the associated reconstruction problem. This algorithm is based on an original extension of the mean shift procedure to what we have called the modified mean shift. A great advantage of the modified mean shift procedure is that one can extend the algorithm to account for non-Gaussian noise by using elements of half-quadratic theory. After these theoretical developments, we now continue by applying this model to reconstruct real images, both 2D images and 3D SPECT images. We show in the next chapter that the prior modeling in the subspace significantly increases recognition performance.

Let us finish this chapter with some reflexions on open paths for future investigation:

- Moghaddam and Pentland took advantage of the probabilistic PCA model to perform ML detection of faces in images [162]. This can also be done using the (W)MMS model since we have already derived the probability density function of the observation Eq. 4.22 (which can be efficiently evaluated using matrix inversion lemmas). The density of the observation cannot be obtained as easily in the case of robust noise (since many robust cost functions are not integrable, Sec. 4.3.3). The difficulty is to marginalize over $\boldsymbol{x}$: $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}x$. What we can do, is to consider all the weights $\boldsymbol{b}$ that are 0 to define the occlusion in the image - and then compute the density of the non-occluded part of the image. If we split up the observation according to the weights $\boldsymbol{b}$ found during reconstruction into a non-occluded part $\boldsymbol{y}_1$ and an occluded part $\boldsymbol{y}_2$, as $\boldsymbol{y} = (\boldsymbol{y}_1\boldsymbol{y}_2)^T$, the joint distribution of the observation is given by

$$p((\boldsymbol{y}_1^T\boldsymbol{y}_2^T)^T) = \mathcal{N}((\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2^T)^T, \begin{bmatrix} \boldsymbol{\Sigma}_{y_1} & \boldsymbol{\Sigma}_{y_{12}} \\ \boldsymbol{\Sigma}_{y_{12}} & \boldsymbol{\Sigma}_{y_2} \end{bmatrix}).$$

  The marginal distribution for $\boldsymbol{y}_1$ is then given by $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma_1})$, where the mean, $\boldsymbol{\mu}_1$, and the covariance, $\boldsymbol{\Sigma}_1$, are given by the corresponding components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_y$ [4, Th.

$p(\boldsymbol{x})$	$p(\boldsymbol{y}\|\boldsymbol{x})$	Acronym	Type	Algorithm	Section
U	GI	ML	linear	$\hat{\boldsymbol{x}} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\tilde{\boldsymbol{y}}$	4.1.1
U	GNI	WML	linear	$\hat{\boldsymbol{x}} = (\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}$	4.1.1
U	RI	RML	iterative	ARTUR or LEGEND, ML	4.3
U	RNI	RWML	iterative	ARTUR or LEGEND, WML	4.3
GI	GI	GMAP	linear	$\hat{\boldsymbol{x}} = (\sigma^2\boldsymbol{I}_Q + \boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\tilde{\boldsymbol{y}}$	4.1.1
GI	GNI	GWMAP	linear	$\hat{\boldsymbol{x}} = (\boldsymbol{I}_Q + \boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}$	4.1.1
GI	RI	RGMAP	iterative	ARTUR or LEGEND, GMAP	4.4
GI	RNI	RGWMAP	iterative	ARTUR or LEGEND, GWMAP	4.4
NP	GI	MMS	iterative	modified mean shift (MMS)	4.5.3
NP	GNI	WMMS	iterative	modified mean shift with $\boldsymbol{\Sigma}_\epsilon$	4.5.3
NP	RI	RMMS	nested iter.	MMS and ARTUR	4.6
NP	RNI	RWMMS	nested iter.	MMS and ARTUR with $\boldsymbol{\Sigma}_\epsilon$	4.6

Table 4.1: Table summarizing algorithms resulting from different hypothesis on the random variables in the linear generative model Eq. 4.1. For the subspace variable, $\boldsymbol{x}$, we have considered three different distributions (pdf): (1) uniform pdf (U) - the first four rows in the table are all maximum likelihood estimators, (2) Gaussian isotropic pdf (GI), $\mathcal{N}(0, \boldsymbol{I}_Q)$, and (3) non-parametric distribution (NP) pdf. For the noise distribution, $p(\boldsymbol{y}|\boldsymbol{x})$, we have considered (1) Gaussian isotropic $\mathcal{N}(0, \sigma^2\boldsymbol{I}_D)$, (2) Gaussian non-isotropic noise, $\mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$, (3) robust isotropic noise (RI), and (4) robust non-isotropic noise (RNI).

2.4.3, p. 31]. This marginal probability could in certain cases be used for detection (when the number of occluded pixels remains reasonable low).

- The mean shift algorithm was presented in the review article [40] as a method for localizing modes in multimodal, non-Gaussian distributions. The posterior density that we have derived for non-Gaussian subspace distributions can also have multiple maxima, both for Gaussian noise (Eq. 4.25) and for non-Gaussian noise. One way to approach the global maximum, is to perform repeated reconstructions from different initializations. For Gaussian noise, one can then evaluate Eq. 4.25 to compare different optimums. For non-Gaussian noise, one could evaluate the marginal distribution as described under the preceding point. For this last evaluation to be comparable, one can however only calculate the same marginal for all occlusions (e.g. as determined by the union of the reconstructed weights, $\boldsymbol{b}_k$, for each reconstruction k).

- A third path of further investigation is concerned with the learning of the model parameters. In the first instance, it would be interesting to develop an algorithm for the R(W)ML model. This can be done based on the EM-algorithm for PCA with missing data, derived by Roweis in [188]. The missing data could be determined from the weights, $\boldsymbol{b}$, found by the reconstruction algorithm (ARTUR or LEGEND). This would therefore be a different approach (based on half-quadratic theory) from the one presented by Skočaj *et al.* [196] that also used the EM-algorithm for PCA with missing data, however with the missing data being determined by outlier rejection and a heuristic algorithm.

- The next natural step is the derivation of a learning scheme for the final model, based

on two steps: first make an EM algorithm for estimating the (W)MMS model (non-parametric prior, Gaussian noise), followed by extension as above to robust noise.

- Finally, it would then be interesting to investigate the possibility for creating mixtures of "kernel factor analysis" models. As for many probabilistic models, this should also be possible within the EM framework.

Chapter 5

Experiments

In this chapter we present experiments we have performed and the results we obtained using three of the models developed in the last chapter. These experiments bring insight into some of the limits and possibilities of these models. The experiments are based on a simple classification scheme where we systematically add noise and occlusions to a set of test images. We then repeat the experiments for an exhaustive set of different parameter settings in order to better understand their influence. This chapter is organized as follows. We first describe the image database before we describe how the experiments were performed. The results obtained are then presented and finally enter into a detailed discussion of these results.

5.1 Image database description

We have performed our experiments using the Columbia Object Image Library (COIL-20)[1] [168]. This database is a subset of the images used by Murase and Nayar to develop a real-time object recognition system [163] based on the model described in Sec. 3.3.3 (1-D, B-spline graphical model to represent the distribution of subspace variables). We can consider this model to be optimal for the classification scheme described below in the case of non-occluded images. The models we are considering are therefore *not* optimal when there are no occlusions. Note however that the model of Murase and Nayar was specifically adapted to this particular database (developed *ad hoc*). Contrary to the model proposed here, their model may not be used for general subspace distributions. The database is considered to be a well understood database – that is, it is relatively simple to model. We chose the COIL database for our experiments because we know that the subspace distribution is non-Gaussian.

The "COIL-20-proc" database consists of images of the 20 objects that are shown in Fig. 5.1. Of each object there are 72 images, each image taken under a different observation angle (object pose). For this, the objects were placed on a turntable and with a fixed camera position, an image was acquired at every 5 degrees (thus a complete rotation of the object). An example of 6 views of object 1 is shown in Fig. 5.2. Each image has been segmented from the background and (geometrically) scaled so that every image fits into a 128×128 image matrix of 256 grayvalues.

[1] Accessible at `http://www1.cs.columbia.edu/CAVE/research/softlib/coil-20.html`

Figure 5.1: The objects of the Columbia University Image Library. Upper row, objects 1–10 and lower row, objects 11–20 (left to right).

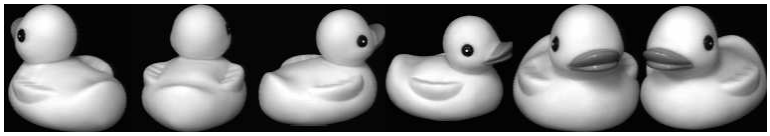


Figure 5.2: Example poses of object 1 of the COIL-20 database.

5.2 Pose estimation scenario

Using the COIL-20 database, a simple pose classification experiment was devised. The goal of this experiment was to devise a naive scheme (as opposed to finding a best method to the actual problem of estimating the pose) in which we could compare the models derived in the last chapter under controlled experimental conditions. In this experiment, each object was treated individually, i.e. the same experiment was repeated for every object of the database. In the learning stage, an eigenspace of fixed dimension was learnt from all the 72 image poses. We experimented with a varying number of eigenvectors[2]. Before learning, low-variance ($\sigma^2 = 2$) Gaussian noise was added to the learning images. The model parameters, $\boldsymbol{W}, \boldsymbol{\mu}$ and σ^2, were estimated as described in Sec. 4.1.2. For the robust algorithms, the scale parameter σ_ρ was fixed to $\sigma_\rho = 2$ for all experiments.

After this learning stage (that was repeated for every object and every choice of number of eigenvectors), we passed to the recognition stage. An overview of this stage is depicted in Fig. 5.3. Three different algorithms were used to reconstruct differently degraded images: the ML, RML and RMMS algorithms (see Tab. 4.1, p. 72). The projection associated with a particular reconstruction was classified to the pose of its nearest neighbor. We also recorded the euclidean subspace distance from the projection to the ground truth projection (of the non-degraded image). This was repeated for all 72 object poses and the results were averaged.

The images were degraded in the following manner. First, we added an occlusion of a specific type and size to the image. Second, we added Gaussian noise to the occluded image. Most experiments were performed using occlusions in the center of the image, but we also performed some limited experiments with occlusions in the lower and upper half of the image, as well as with a different background intensity. The occlusions were simply image values of the concerned region set to zero.

[2]An alternative would be to choose the number of eigenvectors based on a fixed variance proportion explained by these. One would then vary this proportion. We shall come back to this issue in the discussion Sec. 5.4.

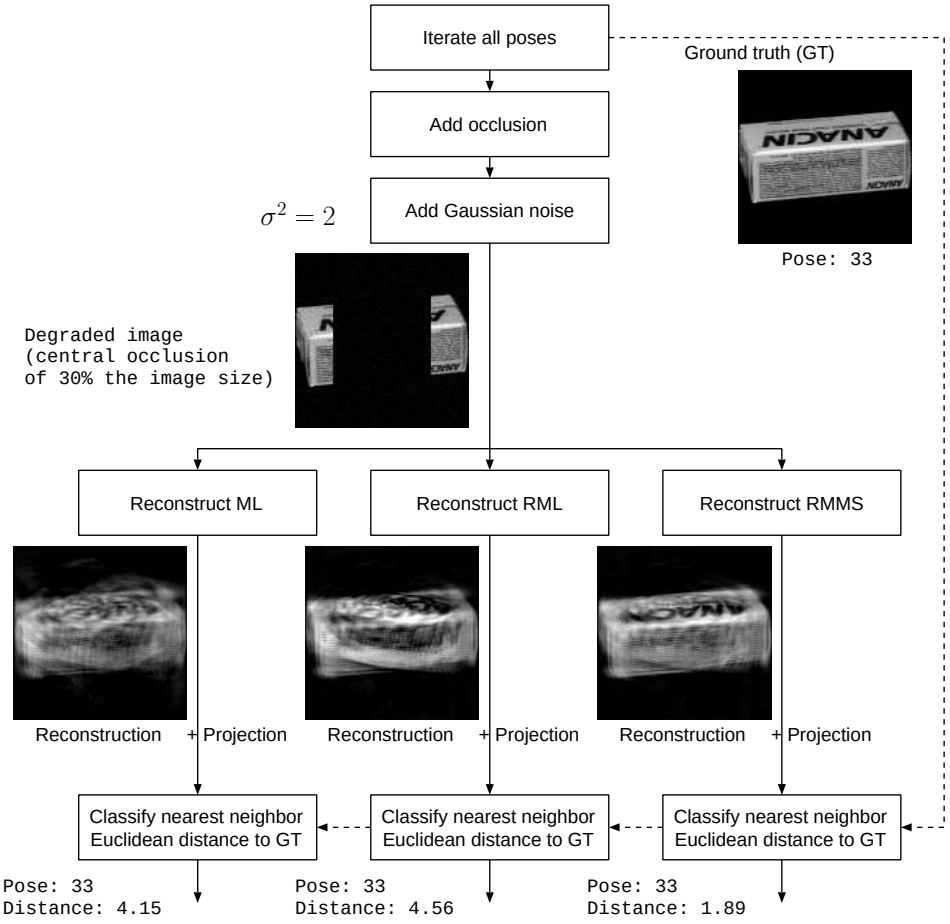


Figure 5.3: Overview of the experimental procedure. The solution to the reconstruction problem is solved using different reconstruction algorithms, which each produces a reconstructed image and a subspace projection. The subspace projection is used for nearest neighbor classification and its distance to the projection of the (non-degraded) original (ground truth) was recorded. The procedure is illustrated with an example of object 5 at pose 33. With an occlusion of 30 % size placed in the center of the image and 30 eigenvectors, all three reconstruction algorithms estimate the right pose. However, the euclidean subspace distance from the ground truth is largely inferior for the RMMS algorithm compared to the RML and ML algorithms. The reconstructed image is therefore visually closer to the ground truth image.

5.3 Experimental results

In Tab. 5.1 are shown the results obtained using 30 eigenvectors and 40 % occlusions in the middle of the images. For every object is listed the percentage of correct pose classifications, the average euclidean subspace distance to the ground truth as well as the standard deviation (stdev) of this distance. At the bottom of the table, we have given the minimum, maximum, median, and quartiles (25 % and 75 %) of each column. We see for example that for object 1 the pose was always correctly estimated with the robust algorithms, whereas the ML algorithm only makes the correct estimation for 2 poses (2.78 %). The average distance of the ML projection from the ground truth projection is 10.3 whereas the same distances for the RML and RMMS models are 0.7 and 0.6 respectively. The standard deviation of this distance is 0.47, 0.23 and 0.17 for the ML, RML and RMMS models respectively.

	Correct classifications (%)			Euclidean distance from ground truth					
				average			stdev		
Object	ML	RML	RMMS	ML	RML	RMMS	ML	RML	RMMS
1	2.78	100	100	10.51	0.7	0.6	0.47	0.23	0.17
2	37.5	94.44	95.83	7.07	2.02	1.81	0.69	1.52	1.42
3	33.33	48.61	48.61	5.54	3.65	3.49	0.9	2.17	2.29
4	51.39	75	75	5.77	3.08	2.8	0.52	3.79	3.56
5	40.28	68.06	75	5.12	4.48	3.55	0.69	1.76	1.83
6	45.83	55.56	48.61	4.58	3.59	3.49	0.88	2.61	2.63
7	26.39	80.56	84.72	12.13	3.24	2.55	0.84	2.56	2.26
8	6.94	12.5	25	7.5	8.72	6.2	0.76	3.13	3.17
9	50	100	98.61	5.57	3.3	2.63	0.78	0.57	0.51
10	51.39	52.78	61.11	5.59	5.1	4.09	0.75	2.67	2.39
11	12.5	33.33	30.56	11.3	9.43	6.99	0.86	3.61	3.01
12	5.56	75	100	11.44	4.51	1.07	0.67	3.68	0.3
13	34.72	83.33	91.67	6.91	2.87	1.58	0.52	3.56	2.22
14	18.06	84.72	98.61	14.77	4.93	1.06	0.6	3.44	1.02
15	4.17	1.39	100	38.28	61.55	0.72	0.75	1.38	0.16
16	2.78	37.5	83.33	67.82	9.15	1.88	0.8	1.47	1.83
17	6.94	1.39	100	31.49	46.29	0.52	0.86	1.21	0.15
18	45.83	56.94	76.39	6.93	6.47	3.74	0.97	3.29	2.49
19	40.28	43.06	44.44	7.17	7.06	5.76	0.83	3.47	2.99
20	15.28	25	27.78	5.43	6.42	6.04	0.76	2.53	2.69
min	2.78	1.39	25	4.58	0.7	0.52	0.47	0.23	0.15
Q-25	6.94	36.46	48.61	5.58	3.29	1.45	0.69	1.51	0.89
median	29.86	56.25	79.86	7.12	4.72	2.71	0.76	2.59	2.24
Q-75	41.67	81.25	98.61	11.61	7.47	3.82	0.84	3.45	2.64
max	51.39	100	100	67.82	61.55	6.99	0.97	3.79	3.56

Table 5.1: Object specific performance of the three models: ML, RML and RMMS (30 eigenvectors, 40 % central occlusions).

In Fig. 5.4 we have depicted box-whisker plots[3] that summarize the results obtained using 15 and 30 eigenvectors, crossed with occlusion sizes (central occlusions) of 0, 10, 20, 30 and 40 % (of the image size) for the three reconstruction algorithms examined. For every occlusion size three box-whisker plots are grouped together for the results of the ML, RML and RMMS models from the left to the right respectively. The two windows on the left show statistics of correct classifications across all objects. The two windows on the right show the statistics of the euclidean subspace distance. The upper windows were obtained using 30 eigenvectors, the lower windows using 15 eigenvectors. We see for example that with increasing occlusion size the performance of the ML model drops rapidly with respect to the robust models. For 40 % occlusion and 30 eigenvectors, we have for the RMMS algorithm a median of 75 % correct classifications. The same number for the RML and ML models are 68.06 and 34.72 % respectively.

In Fig. 5.4, the statistics used confound all objects. However, statistics on the *differences* between two methods would remove this confound. Fig. 5.5 shows histograms of object specific differences between the RML and the RMMS algorithms (i.e. difference between RML and RMMS columns in Tab. 5.1, calculated for numbers instead of percentages). For example, we can see that for three objects, the RMMS model correctly classified between 3 and 9 more poses than the RML model (third bin in top left histogram).

We also give some examples of degraded images and reconstructions obtained using the three different algorithms ML, RML and RMMS in Tab. 5.2. The experimental settings under which the images were created are listed in column one and the classifications and subspace distance from the ground truth are shown underneath each reconstructed image. The last column shows the weights ($\boldsymbol{b}$ in Sec. 4.6, p. 69) that were found by the RMMS algorithm. Different cases are shown:

- An example where the robust modeling alone completely accomodates for the occlusion (object 1).
- Failure of the RML (objects 6, 15 and 20) and the RMMS (object 6) algorithms to correctly estimate the pose (the ML algorithm mostly fails).
- An example of an object where the robust algorithms have an high failure rate, but the reconstructed images are "visually close" to the original (object 6). For this particular object (see also Tab. 5.1), it was often observed that the robust algorithms failed pose-estimation by only one (5 degrees rotation) – RML: 8 times, RMMS: 11 times – or that the reconstructed image was turned almost exactly 180 degrees around (as shown in this example for RMMS) – RML: 8 times, RMMS: 10 times.
- Correct pose classification even though the reconstructed images are visually unsatisfactory (difficult to classify pose from observed reconstruction, notice in particular object 12) – ML: objects 8, 12 and 20 – RML: objects 11 and 12.
- One of the objects for which the RML algorithm completely failed to correctly reconstruct over all poses but one, whereas the RMMS algorithm still shows 100% recognition performance.

[3]Box-whisker plots plot the minimum and maximum using a line, the quartiles using a box, inside which is drawn the median as a bar, i.e. these plots are derived from the lower 5 rows of Tab. 5.1 (and similar tables for further experimental settings).

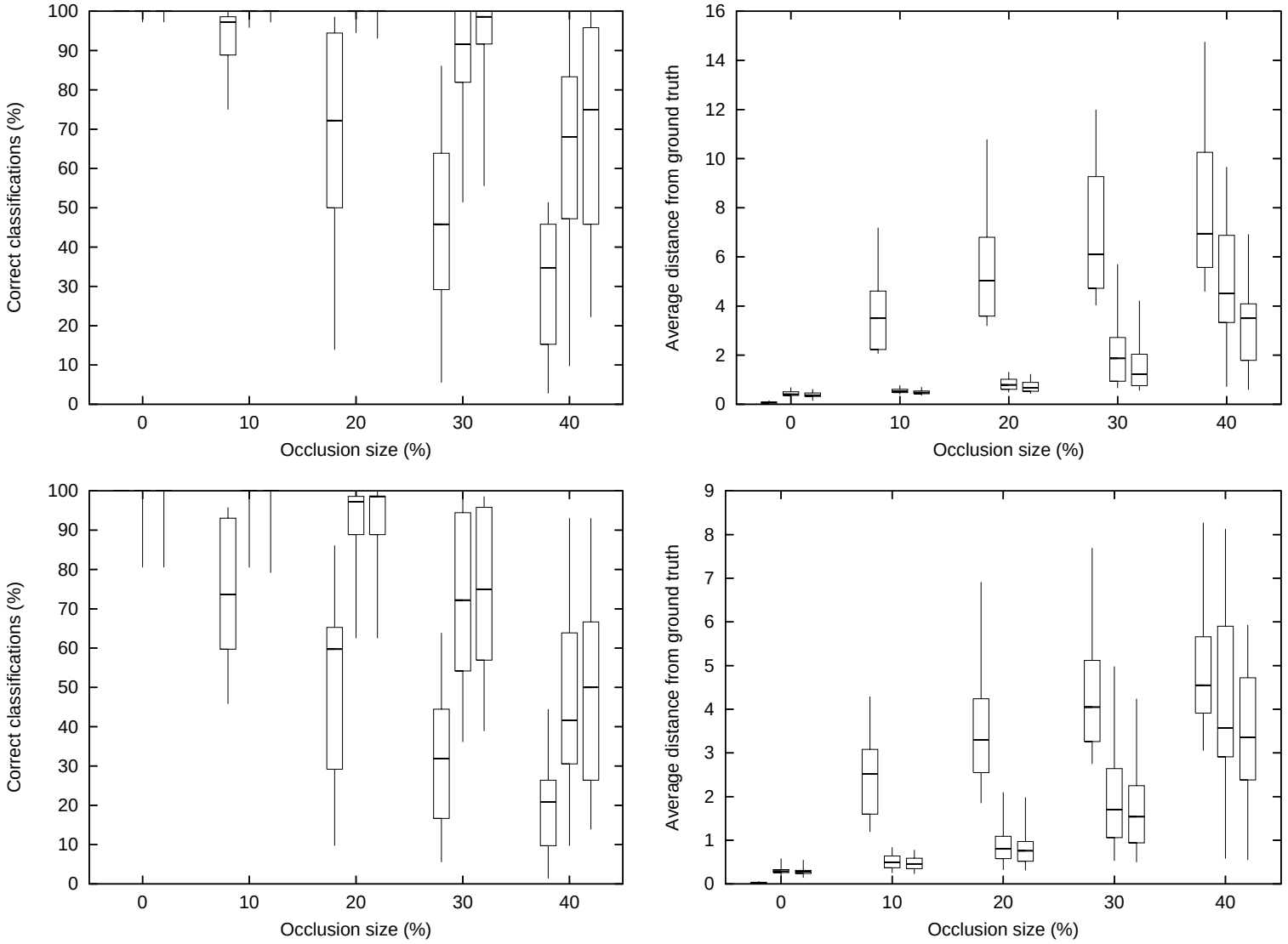


Figure 5.4: Box and whisker plots ([199]) of average performances (over all objects) using 30 (top) and 15 (bottom) eigenvectors. On the left: statistics of correct classifications. On the right: statistics of the average distance to the ground truth. For each occlusion size, the results are grouped into three (from the left): ML, RML and RMMS respectively. Note that at 0 % occlusion, the ML estimator is perfect *by definition* thus disappearing into the border of the plot.

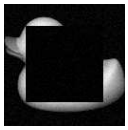
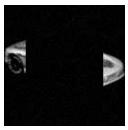
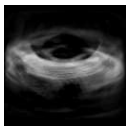
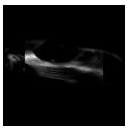
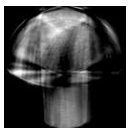
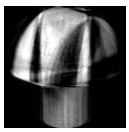
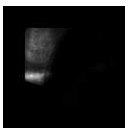
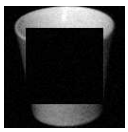
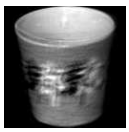
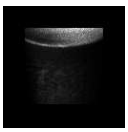
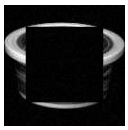
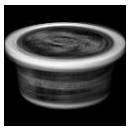
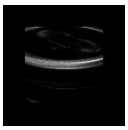
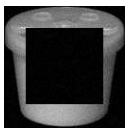
Description	Original	Degraded	ML	RML	RMMS	Weights
Object: 1 Pose: 0 Occlusion: 40% # eigenv.: 30						
Est. pose: Sub. distance:			56 10.99	0 0.58	0 0.49	
Object: 6 Pose: 8 Occlusion: 40% # eigenv.: 30						
Est. pose: Sub. distance:			66 5.05	7 2.86	40 6.49 (1.54 to NN)	
Object: 8 Pose: 42 Occlusion: 30% # eigenv.: 30						
Est. pose: Sub. distance:			42 6.37	42 4.48	42 3.23	
Object: 11 Pose: 25 Occlusion: 40% # eigenv.: 30						
Est. pose: Sub. distance:			47 10.92	25 10.13	25 3.5	
Object: 12 Pose: 34 Occlusion: 40% # eigenv.: 30						
Est. pose: Sub. distance:			34 10.82	34 8.06	34 1.23	
Object: 20 Pose: 8 Occlusion: 40% # eigenv.: 30						
Est. pose: Sub. distance:			8 3.83	9 4.56	8 2.81	
Object: 15 Pose: 0 Occlusion: 40% # eigenv.: 15						
Est. pose: Sub. distance:			62 35.16	62 61.98	0 0.95	

Table 5.2: Examples of degraded images and their reconstructions with the ML, RML and RMMS algorithms. The estimated pose is given underneath each reconstructed image as well as the euclidean subspace distance from the ground truth. The weights (intensity scaled for visualization) found by the RMMS algorithm are shown in the last column.

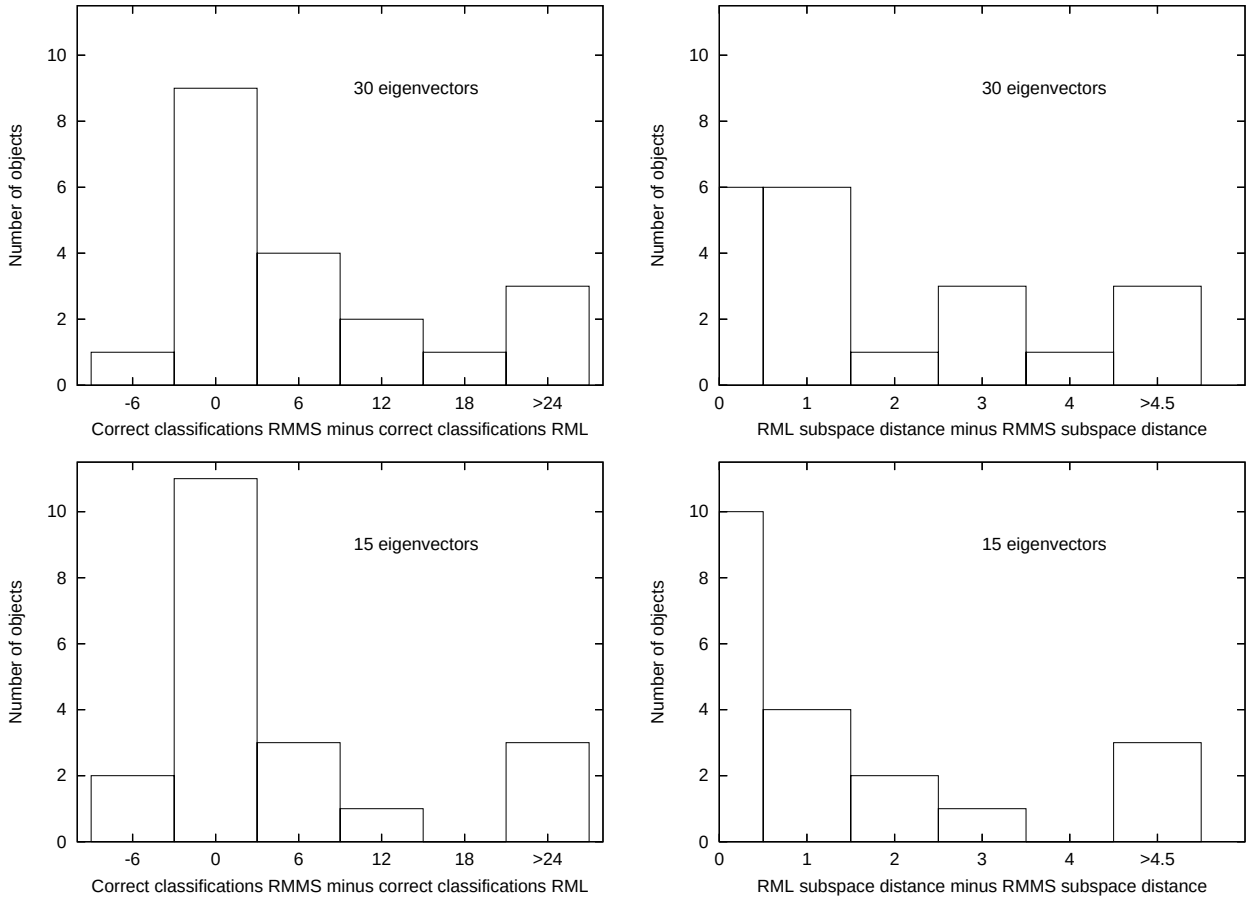


Figure 5.5: Histograms of object specific differences between the RML and the RMMS algorithms. On the left: histogram of correct classifications with the RMMS minus the RML algorithm (binsize of 6). On the right: histogram of RML subspace distances minus RMMS distances (binsize of 1, except first bin which is 0.5). The two top histograms were derived from Tab. 5.1 (30 eigenvectors and 40 % occlusions), however the number of correct classifications were used instead of the percentages. The lower histograms were obtained using 15 eigenvectors, but otherwise identical experimental settings.

We also made limited trials with other types of occlusions than the central occlusions. Notably, we made occlusions of the upper image half, the lower image half and we changed the background value. Some of these results are shown in Fig. 5.6, using the same type of presentation as in Fig. 5.4.

5.4 Discussion

The discussion is separated into three sections, one concerning the overall experimental results, one concerning modeling issues and one section on experimental issues and practical concerns.

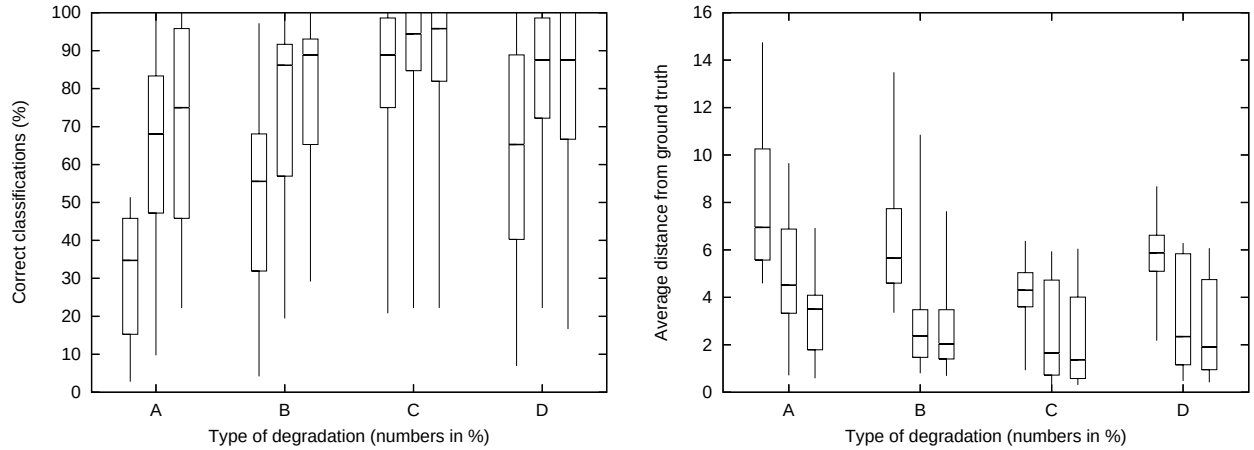


Figure 5.6: Box and whisker plots of average performances (over all objects) for different types of image degradation using 30 eigenvectors: A - center occlusion 40 %, B – lower 50 %, C – background = 127, and D – background = 127 and center occlusion 10 %. On the left: statistics of correct classifications. On the right: statistics of the average distance to the ground truth. For each type of degradation, the results are grouped into three (from the left): ML, RML and RMMS respectively.

5.4.1 Overall performance

Robust vs. non-robust modeling

As is seen in the box-whisker plots in Figs. 5.4 and 5.6, there is a great overall improvement in pose estimation performance using the robust algorithms over the non-robust algorithm. This is what we would expect with our experimental design which is biased towards the robust algorithms. It is however interesting that the ML reconstruction still works fairly well with about 75 % of the objects scoring over 88.9 % (correct classifications) for 10 % occlusion size and 30 eigenvectors (top left window in Fig. 5.4). However, with increasing size of the occlusions, these numbers drop substantially. It is also interesting to notice that the standard deviation of the subspace distance from the ground truth is lower for the ML algorithm than for the robust algorithms (last three columns in Tab. 5.1) – however, with the RMMS somewhat better than the RML algorithm. Of course the average euclidean distance is far lower the robust methods. We explain these results by a phenomenon we have observed. When the robust algorithms fail, they can sometimes fail "badly", probably getting caught in a local minimum. Thus, when using robust algorithms, we are playing with higher risk. A possible remedy could be to use multiple initializations for the robust algorithms (to be done).

Robust modeling with and without prior

In terms of pose classification, the difference between the robust algorithms is less clear, but shows that on average, the RMMS algorithm scores better than the RML algorithm (Fig. 5.4). Removing the confounding variance due to object specific performance, we see in Fig. 5.5 that for most objects (13 instances that form a Gaussian like histogram around zero in Fig. 5.5), there is no substantial difference in pose classification. However, for the remaining 4 objects, the difference is large (more than 9 additional errors for the RML algorithm).

In terms of the subspace distance, the difference between the RML and the RMMS algorithms becomes more clear. In particular, we see that the average subspace distance for 30

eigenvectors and 40 % occlusion size (top right window in Fig. 5.4) is quite substantial. In Fig. 5.5, we can even see that this distance is *systematically* lower for the RMMS algorithm than for the RML algorithm. Since the prior information in the RMMS model acts as a regularization term, this was indeed expected. To what degree this distance is important however, depends on the particular application at hand.

Meaning of subspace distance

Since the RMMS model shows better results concerning the subspace distance than the RML model, one might ask what this distance actually means. Recall that the learning data was whitened, so that the subspace variance of these samples is one. The average distance to the nearest neighbor was measured to about 3.7 for all objects. In this particular experiment where we searched to correctly classify poses, this distance seems to be less important as illustrated in for example the rows 4, 5 and 6 (objects 11, 12 and 20) of Tab. 5.2. Here, we see that for certain objects a subspace distance of 10 or larger does not necessarily lead to a wrong pose classification (based on the nearest neighbor classification). However, as these examples also show, it is difficult to make any general conclusions about how large this distance should be. For object 20 (second to last row), the RML algorithm leads to a wrong pose estimation at a subspace distance of 4.56, whereas just above we see that even with a subspace distance of 8.06 and 10.13 (objects 12 and 11 respectively) the pose is correctly estimated. Generally speaking however, it is clear that the reconstructed objects are visually closer to the original when this distance is small.

5.4.2 Modeling issues

Importance of model accuracy

We shall shortly present the argument that there is no simple relationship between the variance described by the model and the results we have obtained. There is however, a general improvement in the classification performance with a more accurate model in the form of an increasing number of eigenvectors as seen in the left windows[4] of Fig. 5.4. Other experiments we have performed with fewer eigenvectors also support this. We further believe that another more subtle relationship can be deduced from these results, namely that the RMMS model profits more from increased model accuracy than the RML model. This conjecture is supported by the histograms in Fig. 5.5 where we see that the histograms are skewed slightly to the right (and thus in favor of the RMMS model) for 30 eigenvectors with respect to 15 eigenvectors. This relative difference is not unambigously clear from Fig. 5.4. Another reason we believe that this supposition is correct is given by the results found for object 15 which are discussed below. An accurate model therefore seems to be important in order to obtain good results with the RMMS model.

Relation between model variance and object performance

One observation from these experiments is the large differences in the performances obtained for different objects. Are these differences mainly attributable to the objects themselves or to the fact that we have chosen a fixed number of eigenvectors across which we compare

[4]Note that we cannot make the same comparison with subspace distances since these are not directly comparable for 15 and 30 eigenvectors.

performances? Instead of fixing the number of eigenvectors, one often fixes a proportion of explained variance based on Eq. 3.5, p. 35. However, as we noticed, there is no simple relation between the performances we have observed and the variance explained by the model. This becomes clear when studying the relation between the results given in Tab. 5.1 and the proportion of variance explained, given in Fig. 5.7. For example, 30 eigenvectors describe 96.4 % of the total variance of object 8, but only 88.8 % of the total variance of object 9. However, the performances of object 9 (Tab. 5.1) are largely superior to those of object 8. There is thus no *simple* relationship between object specific variance and the observed performance (but see also the next paragraph). These differences are more related to the *difficulty* of occlusions as discussed below (Sec. 5.4.3).

Appropriateness of subspace modeling

Two small aspects are worth mentioning here. First, given the particular subspace distribution of these image sets (Fig. 3.4, p. 38) it is clear that using a kernel estimator to estimate the density results in a rather "bumpy" distribution. When using all the object images for learning – as we have done in this experiment – this estimate is probably sufficient. It is however in general a difficult distribution to estimate using a kernel estimate because the true distribution should be constant along the thread of points. This problem is similar to the problem of estimating a uniform distribution in one dimension using a kernel estimate which requires a high number of kernels to be well estimated.

The second aspect concerns the dimension of the subspace and non-parametric density estimation. The curse of dimensionality dictates an explosively high number of samples in order to correctly estimate a high dimensional density. In their article on the mean shift, Comaniciu and Meer [40] argue that the method is only appropriate up to about six dimensions. However, we are actually observing that the subspace modeling adds value even for 30 eigenvectors. Maybe one could say that some subspace estimation is better than none.

5.4.3 Experimental issues and practical concerns

Occlusion difficulty

It is clear that most of the object specific performance differences that we observe in these experiments must be attributed to what we call the "occlusion difficulty", i.e. how much does a particular object suffer from the artificial degradation it has been subjected to. For example, consider object 12 or object 20 with respect to object 1 in Tab. 5.2. Whereas it is extremely difficult, even for a human observer, to estimate the correct poses of the degraded images 12 and 20, this is a fairly simple task for object 1. Even so, it is interesting to note that there are large differences in performance on these objects (RMMS, Tab. 5.1 and Fig. 5.4): whereas the pose of object 12 was correctly estimated for all 72 poses – indeed very impressive, this number is only 20 (27.78 %) for object 20. For 30 % occlusions, the number of correct pose classifications for object 20 becomes 67 (93.06 %). It is therefore quite reasonable that the degradations we have applied to the images – even though identical – lead to quite heterogeneous results across the objects.

Computational time

The drawback of the RMMS model is that the reconstruction becomes fairly cost intensive. We have not made any theoretical analysis of the computational complexity of these algorithms,

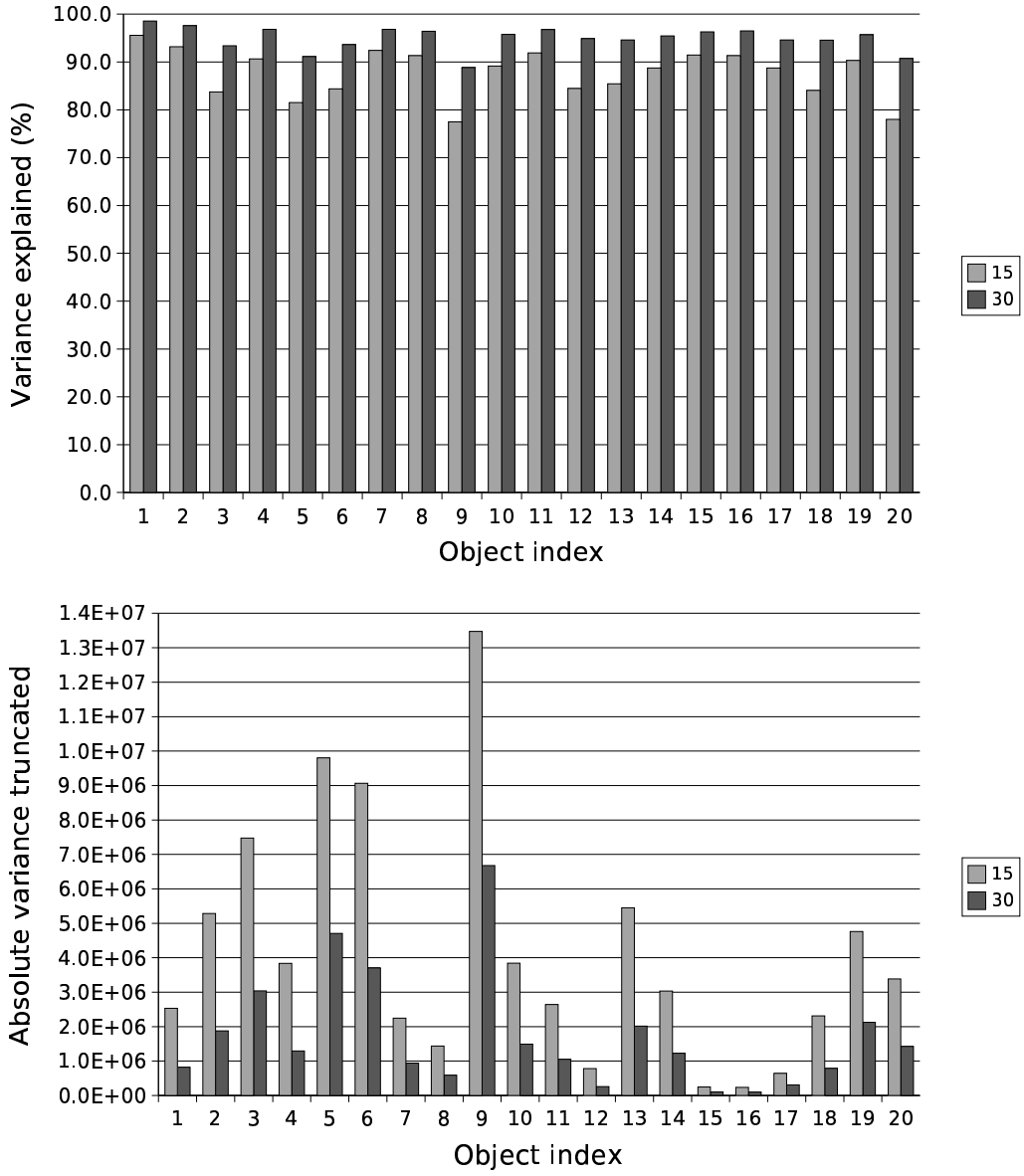


Figure 5.7: Top: percentage of variance explained by 15 and 30 eigenvectors (Eq. 3.5, p. 35) for each of the COIL-20 objects. Bottom: absolute variance that is *not* explained by 15 and 30 eigenvectors.

Occlusion size	Object	15 eigenvectors		30 eigenvectors	
		RML	RMMS	RML	RMMS
10 %	1	1.6	8.9	2.9	24
	8	1.2	9.6	3.3	29
30 %	1	1.3	11	3.6	36
	8	6.8	88	11	344
40 %	1	1.6	12	3.5	37
	8	6.0	109	11	312

Table 5.3: Example of computation times in seconds as experimentally observed for object 1 and 8 at pose 0. We see that the computation time depends strongly on the object and the occlusion. Generally, the computation time increases with increasing occlusion size and with an increasing number of eigenvectors used in the model. ML reconstruction is always fast ($\sim$ ms).

nor any convergence analysis. However, it can be seen from Eq. 4.34, p. 69, that the subspace mean shifting (prior term) depends on the number of samples and the subspace dimension. With the modified noise variance matrix due to the robust weights, there is also a dependence on the observation dimension (data term in Eq. 4.34). Furthermore, the computation depends on the type and severity of degradation of the image. In practice we observed that the reconstruction times varied greatly with the RMMS algorithm. Some examples that were obtained with our implementation are shown in Tab. 5.3 (code in C++, running on a Linux operating system over a 2.4 GHz Pentium 4 with 2GB of memory).

Other models

In these experiments we chose to concentrate on the three models ML, RML and RMMS instead of evaluating every model that was derived in Ch. 4 (and listed in Tab. 4.1, p. 72). This was done because of practical concerns (too many results can occlude principal trends), but also because we considered that these models would be most interesting to contrast in this particular experiment. The GMAP and RGMAP models were not considered because we already knew that the subspace distribution was far from Gaussian. The "Weighted" (non-isotropic) models were not considered because we have not derived the maximum likelihood estimates of the model parameters (notably the image generating matrix $\boldsymbol{W}$) for all these models yet. For the models we have used, we considered that the eigenvectors were sufficiently good estimates of $\boldsymbol{W}$ since these models are closer to the PPCA model and for this model the eigenvectors asymptotically becomes the maximum likelihood estimate of $\boldsymbol{W}$.

5.5 Conclusion

We have performed a series of experiments on the COIL-20 database. We conclude from these that a general learning approach to object modeling is indeed difficult. This is evident from the high variability of results across the different objects of the database. The results do however show that on average, the added prior modeling does significantly improve performance. In general, added modeling does in many cases lead to a more constrained model. However, in our case, because of the general form of this prior distribution, we have quite a versatile model

that does not seem to be constrained in any particular way with respect to the other models we have considered. Several paths for further investigation are open. In particular, we would like to continue investigating the results for the objects 15, 16 and 17 of the database. It would also be interesting to compare these results to the non-isotropic models developed in Ch. 4.

These experiments show that the improved modeling of images as well as the accurate reconstruction of new images under the model is important in order to obtain high recognition performances when the images are highly degraded. In the next part of this document, we shall be concerned with the comparison of images of brain perfusion of potential patients with an atlas of normal subjects. These images might potentially contain large zones of abnormal brain perfusion. As we shall use the residual between the observed image and the reconstructed image as a measure of abnormality, it is important that the reconstruction task be performed as accurately as possible. The recognition task in this medical setting can be interpreted as the problem of determining the *nearest equivalent normal image* to the image under evaluation and corresponds to solving the reconstruction problem.

Part III

Brain Perfusion Atlas: Construction and Evaluation

Chapter 6

Models and preprocessing: overview and state of the art

The goal of this work has been the creation of an atlas of brain perfusion as seen in SPECT images (Ch. 2). We chose to do this using statistical models from the domain of computer vision/pattern recognition (Part II). In this chapter, we start by defining an atlas and explaining the connection to the problem of pattern recognition. We then review statistical models used in the analysis of functional brain images. In order to situate these models, it is necessary to understand the different clinical or research questions for which answers are sought. Attempts to answer these questions lead to different *experimental designs*. An atlas can be considered to be a special case in experimental design. The notion of experimental design comes from statistical hypothesis testing and has an analogue in the formulation of statistical pattern recognition systems. Finally, we review image preprocessing techniques that are necessary in order to make statistical modeling possible: registration, segmentation and normalization.

6.1 Atlas, definition

What is a probabilistic atlas of brain perfusion? An atlas can best be described by answering the following three questions:

1. What does it describe?

2. How does it describe it?

3. For what can it be used?

For example a geographical map is a description of a part of the earth. It describes the spatial distribution of hills, forests, roads, houses etc. This is done by means of an image using specific symbols to signify different objects or elements being described. Finally the map can have a multitude of uses, like finding the shortest path between two houses.

Neurologists are interested in atlases of the brain. There exist anatomical and functional atlases. Anatomical atlases are maps describing the shape, size, spatial extent and relative locations of different brain structures. Functional atlases uses the same features to describe different physiological processes in the brain and how they are related. A functional atlas of the brain contributes to explain how the brain works, a question that has thrilled human

curiosity for at least a few thousand years[1]!

In our application, we can answer the questions posed in the introduction of this section as follows: (1) The atlas describes patterns of brain perfusion. By perfusion pattern, we understand the level of blood flow (intensity) and the spatial distribution of it in the brain. Measures of these brain perfusion patterns are obtained from SPECT images. (2) The atlas describes patterns of brain perfusion by statistically modeling the intensity values observed at each voxel. (3) The purpose of the atlas is to detect abnormal perfusion patterns in an image by assigning to each voxel a measure of normality.

We thus arrive at the following definition: A *probabilistic atlas of brain perfusion* is a statistical model that describes patterns of brain perfusion in a population and allows for inferences about new, unseen patterns relative to this population.

6.2 Overview, construction and modeling

To see how an atlas can be related to a classical pattern recognition problem, let us consider a general pattern recognition framework as depicted in Fig. 6.1. Here, we distinguish between learning and classification. In learning, we are concerned with the choice of data representation and/or to extract characteristic features of the data. This is done with the goal of finding an appropriate statistical model that can describe and/or capture salient properties of the data. The statistical model is typically parametric (in order to be compact) and its parameters are estimated from a set of learning patterns. Based on this model a decision rule is defined, which is then implemented as the classifier itself. New patterns are processed in the same manner as the learning patterns before they are classified or they may possibly be characterized in some manner. In our case this process consists of identifying regions of abnormal brain perfusion in SPECT images.

The exact preprocessing steps to be performed might vary slightly. In general, however, registration, segmentation and normalization of the images is necessary, Fig. 6.2. When comparing brain images, a first step is to align them. This is done using registration algorithms. We describe different problems and possibilities of registration in Sec. 6.4. Second, the observed gray values must be made comparable. As we have seen in Ch. 2, the total number of photons emitted may vary due to, among other things, variation in the injected dose. Such variations lead to global differences in the image intensities and must be compensated. We describe and review the methods of intensity normalization of SPECT images in Sec. 6.6. Finally, only the brain is considered for statistical modeling. This is because the activity distribution in the surrounding tissues has properties different from those of the brain (Sec. 2.6.2). Furthermore, a segmentation is sometimes necessary for the intensity normalization of the images. The segmentation of the brain does not represent a major concern, but we point out the different techniques that have been applied in Sec. 6.5.

6.2.1 Non-probabilistic approaches

The focus of this thesis has been on probabilistic atlases. Another possible, non-probabilistic approach could easily be deviced based on prototypes and nearest-neighbor comparisons (template matching). Beside the large memory requirements of this method, we see two inconveniences. First, in such an approach there is a great risk of sampling prototypes in the tails of

[1]Actually, what has fascinated the human being is the question of what a human *is*. The brain seems to be a good place to look for answers.

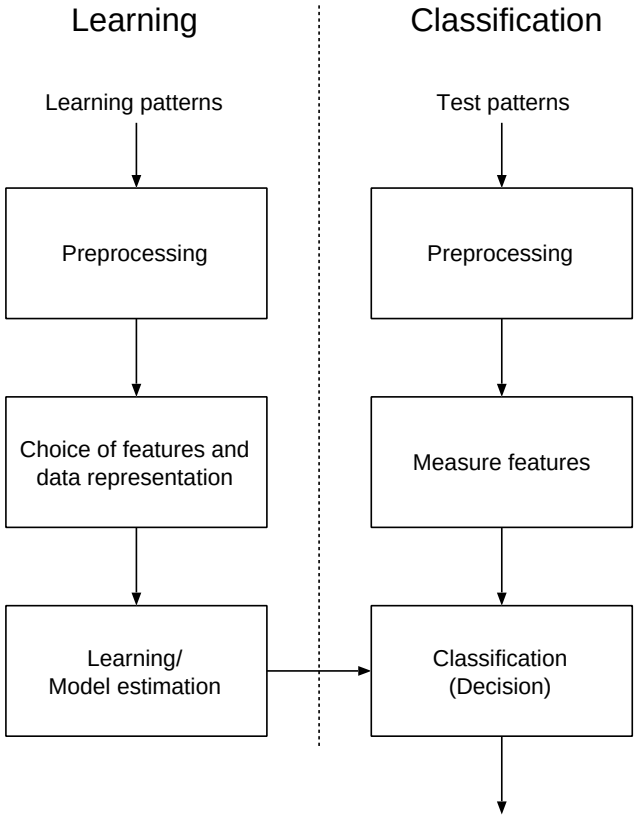


Figure 6.1: A general model for statistical pattern recognition. Similar to [110, 58].

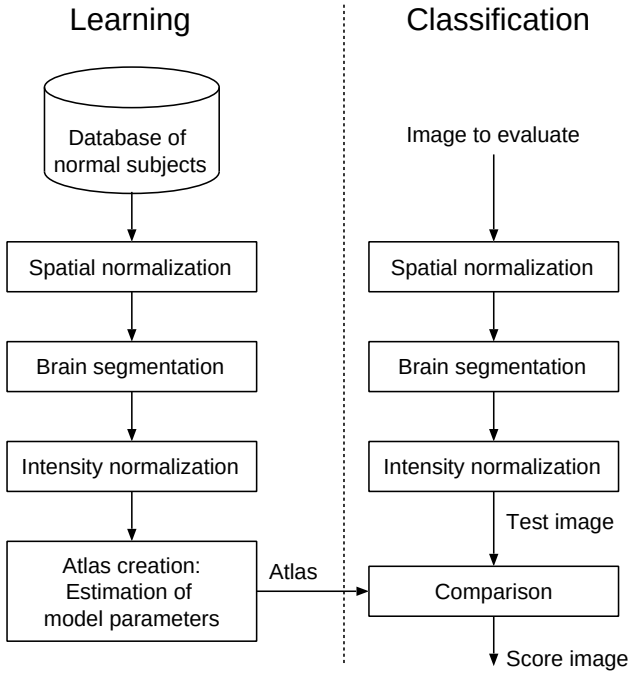


Figure 6.2: Overview of atlas creation and image comparison.

the distribution and therefore sometimes comparing an image to a prototype which is close to abnormal. This makes this approach less robust. The only way of avoiding this problem is to perform some kind of density estimation (parametric or non-parametric), which becomes prohibitive with the large dimension of the observation space due to the curse of dimensionality. We are therefore in a situation as described in Sec. 3.2.3, p. 31 where the use of a dimension reduction techniques (such as PCA) becomes a necessity. Second, we expect an approach using a dimension reduction technique to generalize better to unseen images than an approach based on prototypes as it tries to reveal "deeper causes" (structure or patterns) that can explain the observed data.

6.2.2 Pattern recognition and hypothesis-testing

We have seen how the creation and use of an atlas can be considered to be a pattern recognition problem. Most methods, however, that have been applied in functional brain studies are based on statistical hypothesis testing (e.g. whether normal perfusion depends on age, gender etc.). *How does an atlas fit into a hypothesis testing framework?*. To answer this question and in order to understand similar statistical models in the literature, it is necessary to clarify the concept of experimental design (also called study design).

6.2.3 Experimental design

In pattern recognition, we typically are *given* an application or some kind of system that we would like to reproduce. This can for example be an automatic system that recognizes a person in a picture. The input and (desired) output of the system is therefore defined. In statistics, however, we anticipate some behavior of a complex system and *design* an experiment that can highlight this behavior. The statistical analysis, of course, depends on this design. In statistical terms, we say that we make an inference about the behavior.

The goal of the design is to obtain a *valid* analysis that is as *sensitive* as possible. The validity of a statistical experiment is assured by *randomization* and *replication*. The sensitivity can be augmented by using analysis of variance models (ANOVA). For further discussion on these issues, see for example [79]. Let us however clarify some typical aspects on design types that helps in the understanding of the neuroscience literature on functional anatomy.

First, we distinguish between studies that are dynamic or not. In dynamic studies (i.e. fMRI studies or cardiologic scintigraphic studies) the temporal order of the images has a significance and eventual correlations between successive images must be accounted for. This is done in fMRI studies by taking the so-called hemodynamic response into account. In studies that are not dynamic, the analysis is invariant to the ordering of the images.

Second, we can distinguish between single-subject or multi-subject studies and, in the latter, whether one or multiple scans are obtained of each subject. In addition to adding inter-scan variability to the model (two-way ANOVA), it can be of interest to model inter-subject variability in multi-subject studies (three-way ANOVA) [225][2].

Third, in order to determine the necessary sample size, one distinguishes between two classes of inference: inference about "typical" characteristics or about "average" characteristics of a population. This leads to the notions of *fixed-* or *random*-effect models respectively [225, 65]. The former makes an inference about a specific sample (e.g. 5 out of 6 farmers owned a tractor - it is *typical* that a farmer owns a tractor), the latter makes an inference

[2]Inter-scan variability is typically modeled as global intensity differences, see Sec. 6.6.4.

about the population from which the sample has been taken (e.g. farmers in Alsace have, on *average*, more than 0.86 tractors). Making inferences or statements about the population from which the sample has been drawn, as opposed to the sample itself, is typically what interests the pattern recognition practitioner. This notion is linked to the power of a classifier to *generalize* from a *learning set* (sample) to the unseen *test set* (underlying population).

Fourth, in functional brain studies, one typically distinguishes between three types of study design [71]: *activation studies*, *parametric studies* and *factorial designs*. In activation studies, images are obtained for one or more subjects under two different conditions (i.e. activation - rest) and the activated regions are determined. Variations of this consists of analyzing several pairs of tasks conjointly, and then determining commonly activated regions. Activation studies lead to the classical analysis of variance (ANOVA) design, i.e. presence or absence of an effect. Activation studies are most often associated with single-subject, dynamic studies, but are also applied in multi-subject studies. An analogy to our atlas problematic can be made here. The normal subjects are considered as a series of baseline images and the image to test is considered as the activation image. Activations in the test-image would indicate non-normality.

In parametric studies the relationship between a physiological parameter and the degree of activation is analyzed (e.g. the dependency of regional perfusion in a specific region on the frequency of aural word presentation). Parametric studies are analysis of covariance (ANCOVA) models. Finally, in factorial designs, several factors are combined and the interactions between these are analyzed. An interaction can for example be a change in a change: a series of baseline-activation pairs before and after giving a drug to a group of subjects. These are three-way or higher ANOVA/ANCOVA models.

6.3 Related work and statistical models

In this section we present a review of different statistical-based approaches for functional brain studies. The review is done with particular attention paid to methods that can be interpreted as an atlas, like multi-subject studies and studies with none or one replication for each subject. This includes studies of normal perfusion (atlas description), implicit atlases (activation studies where a single subject is compared with a database) and explicit atlases as defined in Sec. 6.1. Some related FDG-PET studies of similar design type have also been included. Furthermore, since this work has focused on multivariate methods, we have also included multivariate activation studies that have been applied in SPECT. However, we do not present activation studies in general. These represent the large majority of functional brain studies (fMRI and PET), and are most often interpreted using voxel-wise univariate models. A unified formulation of a large number of ANOVA/ANCOVA and MANOVA type hypothesis tests can be formulated by means of the general linear model (GLM) [199]. The GLM has been implemented in the statistical parametric mapping framework (SPM) [69][3]. Voxel-based, implicit atlases and normal perfusion studies are therefore typically studied using SPM, whereas explicit atlases are typically based on commercial or proprietary software[4].

The methods are organized below according to the features and the modeling approach that are used. As features, we have (1) the voxel values (voxel-based approaches), (2) the mean

[3]SPM is a software written in MATLAB^{TM} for the statistical analysis of functional brain studies. It is freely available and a large number of "toolboxes" exist for the analysis with other methods.

[4]In a recent comparison study [129] tailored to SPECT images, SPM was found to be rather unsatisfactory in comparison with a commercial program BRASS (Nuclear Diagnostics, Sweden), VOI analysis and visual inspection.

values in a region/volume of interest (ROI/VOI approaches) or (3) the voxel values projected onto the cortical surface. The modeling approaches are all linear, but we distinguish between (A) univariate models and (B) multivariate models. The univariate models are actually multiple univariate models, one for each feature, e.g. SPM has one linear model for each voxel. These models therefore consider each feature as statistically independent of the other features. "Pure" multivariate models on their side, test all the features as a whole. In order to make localized analysis, additional analysis or response characterization is therefore necessary. We also consider as multivariate models, approaches where some kind of spatial cross-correlation has been taken into account.

6.3.1 Other reviews

Whereas this review is more focused on methodological aspects, there exist other reviews in the literature that cover other aspects. In [127], Van Laere *et al.* review eleven other studies of normal perfusion/tracer uptake in SPECT. Since the article reports normal perfusion data, the review is focused on the acquisition parameters in each study (e.g. collimator, number of camera heads etc.), as well as the selectivity of the databases representing normal subjects. We have noticed that very few such studies (one and one partially) use MRI co-registration to register the SPECT images of different subjects (see Sec. 6.4.4). The review reports the major findings in the individual studies and indicates whether ROI (VOI) analysis was employed. The statistical hypothesis/models that were used are not reported.

Gardner *et al.*, [76], have made a review of HMPAO-SPECT studies in neuropsychiatric disorders (schizophrenia and major depression). The review focused on the relating of theories concerning neural substrates in psychiatric disorders with findings in functional brain imaging (using SPECT), i.e. *what is the origin of altered radiotracer uptake and distribution in psychiatric disorders ?* Possible methods for functional brain imaging studies are briefly mentioned, but the studies referenced (nine schizophrenia and eleven major depression studies) were not linked to the methods.

Another review, made by Amen *et al.* [3], presents different findings in SPECT studies that are useful in a clinical setting. For each study they report tracer (ECD, HMPAO, Xe, I123), age distribution, database size, year of study and significant findings. The review is organized with respect to pathology: normal, brain trauma, dementia, temporal lobe epilepsy.

Finally, let us mention the review in [228, 229] of statistical methods that have been applied in fMRI and PET studies. In the former, a unified overview of univariate as well as multivariate methods is presented, whereas the latter only reviews multivariate studies, but in more depth.

6.3.2 Univariate methods

Voxel-based approaches

A list of univariate approaches based on individual modeling of each voxel is shown in Tab. 6.1 (all studies were done using SPM). In these approaches, neighboring voxels are considered to be statistically independent. This is clearly not true in low resolution SPECT images, but can serve as an approximation. However, because of the high number of statistical tests that are repeated, one typically has to correct for multiple comparisons, e.g. with a total of 10000 brain voxels, a p-level of 0.05 would yield 500 significant voxels under the null hypothesis. In order to reduce this number one uses either the Bonferroni correction [225] (used in for

example [127]) or a posterior correction based on the theory of Gaussian random fields [227][5]. These methods of post-correction are implemented in SPM, but does not seem to be employed by all researchers.

In a pattern classification setting, Laliberté *et al.* have presented a method for classifying SPECT images as normal or abnormal, based on the comparison with a normal atlas [131]. This is done using two features: voxel intensities and voxel displacements (as found by image registration, see also section 6.4.5). A univariate test is done at each voxel, one for intensity and one for displacement. Then classification is done based on the number of abnormal voxels. Using cross-validation on a database of 23 subjects with diffuse anomalies and 21 normal subjects, the best results were obtained by counting abnormal displacement voxels in the gray matter (93 % correct classification).

Region-based approaches

Until recently, region-based approaches were the approaches most widely adopted for quantitative SPECT studies. With region-based approaches we understand methods where the images are segmented into volumes/regions of interest (VOI/ROI) and further analysis is done based on the mean value of such VOIs. The advantage of these approaches is the reduced noise due to averaging. A major disadvantage of these methods is their dependence on the segmentation. An error in the segmentation of a region obviously falsifies the mean flow value of the region. A further problem is linked to the so-called partial volume effect (sample aliasing). Because of the low resolution of SPECT images, the voxel values on the border between neighboring regions have "mixed" intensity values. This is particularly annoying if the two regions in question have highly different intensities. In such cases voxels belonging to the region of high (low) intensity will have values lower (higher) than the mean of the region. In small regions, these effects can lead to large errors in the estimated flow value.

In Tab. 6.2, two studies are summarized. Both have used VOI-based and voxel-based analysis. Tab. 6.3 summarizes two studies that are only based on VOI analysis. Whereas in [203] and [109] the VOIs where manually defined (on 2D slices), the others have registered the images with a presegmented template to define the VOIs. In [127], the template partitions the image, whereas in the (PET) study of Kang *et al.* [122], the authors calculate a weighted average by multiplying the intensity value of a voxel with the probability of the voxel belonging to a particular brain structure. This probability is given in an anatomic atlas named statistical/probabilistic anatomic map (SPAM) from ICBM[6]. The method can be viewed as a sort of fuzzy partitioning.

Surface-based approaches

Another univariate method, denoted *3D-SSP*, is based on a quite different feature [157]. The method was first developed for PET, but has later been used with SPECT (ECD [5] and [^{123}I]-iodoamphetamine [102]). The method consists of segmenting the brain surface by non-linear registration of each image with a presegmented template. In addition to the presegmentated surface, vectors that point perpendicular, inward from the surface to the center of the brain, are given. The value of each surface pixel is then set to the maximum value found along this

[5]The assumption of a Gaussian random field is actually in conflict with the assumption of statistical independence between voxels.

[6]There is no reference to this atlas, which was developed at the McGill university by A. C. Evans *et al.*. The ICBM atlases can be found at `http://www.loni.ucla.edu/ICBM/ICBM_Probabilistic.html`

Author	Application	Design	Test	Regressors	Database	Tracer	Reg.	Norm.
Ebmeier et al. (1998) [60]	dementia and depression in old patients	activation, parametric	ANCOVA	confounds: age + occipital ROI. interest: MRI measures + age of onset	depressed (39), AD[a] (15), normals (11)	HMPAO	affine	ANCOVA occ. ref.
	Inference: Early- and late onset of depression, influence of MRI atrophy **Findings:** Late onset depression associated with more brain changes than early onset							
Lee et al. (2000) [135]	TL[b] epilepsy	activation, atlas	t-test	patient, ictal, interictal	epilepsy (21), normal (9)	ECD	affine	n.d.
	Inference: Lateralization **Findings:** (Single) Ictal vs. normal better than ictal vs. interictal and better than subtraction method when low threshold							
Migneco et al. (2001) [153]	apathy in AD and personality disorders (non-demented)	activation	ANOVA and conjunction	patient group and neuropsychiatric diagnosis	AD (28), PD[c] (13)	ECD	affine	whole
	Inference: Apathy related to hypoperfusion in > ant. cingulate gyrus **Findings:** Conjunction analysis showed that hypoperfusion in ant. cing. region was common for both study groups							
Chang et al. (2002) [32]	TL epilepsy	atlas of difference images	t-test	patient	epilepsy (12), normal (7)	HMPAO	affine	whole
	Inference: Localization of seizure foci **Findings:** Consistent with visual analysis. Ictal is hyper w/ time of injection (t) < 100s and hypo w/ t > 100s							
Stamatakis et al. (2002) [198]	head injury	atlas, activation	ANCOVA	age, scan time (acute and follow-up)	head injury (61=22+22+17), normals (32)	HMPAO	affine	whole
	Inference: Abnormality in SPECT and MRI **Findings:** More abnormality detected with SPECT than MRI							

Table 6.1: Voxel-based approaches based on univariate statistical methods. All these studies were performed with SPM.

[a] Alzheimer's disease
[b] Temporal Lobe
[c] Personality disorder

Author	Application	Design	Test	Regressors	Database	Tracer	Reg.	Norm.	Segm.
Imran et al. (1998) [109]	normal tracer uptake and validation of AIR[a]	activation (comparison to existing atlas)	t-test, brain overlap (co-registration w/ X-ray CT as reference)	none	normal (30)	HMPAO	affine	whole	manual, 2D
	Inference: Registration algorithm (human brain atlas vs. AIR) **Findings:** Max. anatomical variability 4.7mm. differences in rCBF as compared to HBA (which is based on another sample)								
Van Laere et al. (2001) [127]	normal tracer distribution	parametric	ANCOVA	age, gender, handedness	normal (89)	ECD	piecewise affine	whole	presegmented template
	Inference: Variability, asymmetry, age, gender **Findings:** Decline w/ age in ant. cingulate gyrus, bilateral basal ganglia, l. prefrontal, l. lat. frontal and l. sup. temporal and insular cortex Increase in asymmetry w/ age (frontal and temporal neocortex). R. par. high in women, cer. and l. ant. temp. and orbitofrontal high in men								

Table 6.2: Approaches using univariate statistical models based on voxels as well as volumes of interest (VOI).

Author	Application	Design	Test	Regressors	Database	Tracer	Reg.	Norm.	Segm.
Tanaka et al. (2000) [203]	normal patterns	parametric	ANCOVA, t-test	age, gender	normal (48)	ECD	proprietary, SPECT	whole, cerebellum	manual, 2D
	Inference: Variability, asymmetry, age, gender, >hemispheres **Findings:** Different results depending on normalization. rCBF decline w/ age in anter. + post. cingulate cortex, sup. prefrontal and parietal cortex, striatum + hippocampus								
Kang et al. (2001) [122]	TL[b] epilepsy	activation	paired t-test	lateralization	epilepsy (18), normal (22)	FDG	affine	whole	SPAM[c]
	Inference: Asymmetry in TL **Findings:** Symmetric VOIs in controls (except inf. temp. gy.), correct lateralization in 14/18, visual inspection consistent in 17/18								

Table 6.3: VOI-based approaches using univariate statistical models.

[a] Automated Image Registration, algorithm and software for registration from R. Woods [223]
[b] Temporal Lobe
[c] Statistical/probabilistic anatomic map, International Consortium of Brain Mapping (ICBM)

inward pointing vector, searching to a depth of 13.5 mm (6 pixels). The method is thus based on a geometric dimension reduction. The surface is then visualized with lateral, superior and medial views. The authors have used this data representation to create a univariate atlas, consisting of a mean and a variance for each surface voxel. To compare images with the atlas, a z-score is calculated. The atlas has been used in Alzheimer's disease [157, 5] and to predict hyperperfusion after carotid endarterectomy (deblocking arteries) [102]. Recently, in a comprehensive comparison study, involving 16 nuclear medicine physicians, the method was shown to improve the diagnosis in Alzheimer's disease [100].

6.3.3 Multivariate methods

Multivariate approaches in functional brain studies are most widely used in an *exploratory* fashion, i.e. to discover relationships between regions. A large number of such approaches have been applied in temporal studies (PET and fMRI) where such relationships are based on the idea that brain function is mediated by a network of connected regions (i.e. *functional integration*, see also Sec. 2.4). In multi-subject brain perfusion studies, there is no similar cognitive interpretation of such spatial relationships. However, spatial correlations of brain perfusion across subjects could have a physiological explanation, e.g. higher relative flow values in one region is correlated with higher flow values in another region due to a typical vascular structure. Further exploration of correlation patterns will require interpreting and analyzing these as a function of age, gender, handedness, etc. as has been done in studies of normal perfusion [179, 238, 115]. Once identified, spatial correlation patterns can be used either as *discriminators*, i.e. to distinguish (classify) two groups [159, 160, 200], or as *descriptors* of normal perfusion variations in a population [104, 105] (which is also implicitly suggested in the normal perfusion studies mentioned above).

All the models that have been used in multi-subject SPECT studies are linear and based on either PCA (SVD), Fischer discriminant analysis or partial least squares (PLS). This is in contrast to fMRI, PET and dynamic scintigraphic studies, where other linear models, such as ICA [150, 183] and factor analysis [10], as well as non-linear methods, such as nonlinear-PCA [66], have been employed.

Voxel-based approaches

In Tab. 6.4 a summary of multivariate, voxel-based methods is provided. The first author to use PCA to create a SPECT atlas, was Houston *et al.* [104]. The model is the same as the ML model in Tab. 4.1, but has an interesting twist: the *residual* is analyzed to detect voxels with abnormal flow values. We have adopted this idea, so let us explain how this is done. Let us recall the linear model from Ch. 4.1[7]:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\mu}$ is the database average, $\boldsymbol{W}$ is a matrix of eigenvectors (found by PCA), and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Psi})$ is Gaussian isotropic noise. The noise matrix $\boldsymbol{\Psi}$ is a diagonal matrix of voxel variances. Each diagonal component $d, d = 1 \ldots D$ of this matrix is estimated from the learning base residuals as follows:

$$(\boldsymbol{\Sigma}_\epsilon)^d = \left(\frac{1}{J-1-Q} \sum_j \boldsymbol{e}_j \boldsymbol{e}_j^T \right)^d, \tag{6.1}$$

[7]The formulation of the method has been adapted from that of the authors' to fit our own theoretical framework.

where $j = 1 \ldots J$ are the subject indexes and Q is the number of eigenvectors retained in the model (number of columns of the matrix $\boldsymbol{W}$. The residuals $\boldsymbol{e}_j$ are obtained from maximum likelihood (or least-squares) reconstruction of the learning base images as:

$$\boldsymbol{e}_j = \boldsymbol{y}_j - \hat{\boldsymbol{y}}_j = \boldsymbol{y}_j - \boldsymbol{\mu} - \boldsymbol{W}\hat{\boldsymbol{x}}_j, \tag{6.2}$$

with $\hat{\boldsymbol{x}}$ from Eq. 4.8 (first row in Tab. 4.1).

The atlas is thus represented by the mean, $\boldsymbol{\mu}$, the matrix of eigenvectors (eigenimages) $\boldsymbol{W}$, and the voxel variance $\boldsymbol{\Psi}$. To compare an image $\boldsymbol{y}$ with this atlas, a residual image $\boldsymbol{e}$ is first calculated, which is then compared, voxel-by-voxel, to the learning base variance as follows:

$$\boldsymbol{t} = \boldsymbol{\Sigma}_\epsilon^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{W}\hat{\boldsymbol{x}}). \tag{6.3}$$

The resulting image, $\boldsymbol{t}$, is a *score image* or *significance image* where each voxel value expresses a significance in terms of deviation from normal values.

A few remarks are worth to mention. First, this model is actually *inconsistent*. This is because the ML estimate is used to solve the reconstruction problem. This assumes isotropical noise. When constructing the score image, however, non-isotropical noise is considered. Nevertheless, we think this inconsistency is not very important in practice[8]. Second, it is also interesting to notice that the model was developed independently of the models in computer vision. Finally, we can see from the distribution of the observation when given the subspace variable, $p(\boldsymbol{y}|\boldsymbol{x})$ (given in Eq. 4.4, p. 60), that the score image is nothing else than the marginal probability of this distribution at every voxel:

$$t_d = \int \cdots \int p(\boldsymbol{y}|\boldsymbol{x}) \mathrm{d}y_1 \ldots \mathrm{d}y_{d-1} \mathrm{d}y_{d+1} \mathrm{d}y_D = p(y_d|\boldsymbol{x}).$$

In [104], this method was used to classify subjects with Alzheimer's disease, infarcts and normals. In a later study, Houston *et al.* used the same method to characterize typical patterns of abnormality in a series of pathologies, see also Tab. 6.4.

In a recent, multi-subject PET study, the authors analyzed the dependence of the eigenvectors (correlation patterns) on age in a normal population [238]. Interestingly, they found a significant correlation between age and the second principal component. Such a relation is also probable in SPECT, since univariate methods [203, 127] and multivariate, region-based methods [115, 179] have shown a similar dependence on age.

Region-based approaches

In Tab. 6.5 we give a summary of multivariate, region-based methods. The mathematical formulation is analogous to the one for voxel-based approaches: an image is considered to be a random vector, the elements being mean, regional flow values. In a multi-subject PET study Moeller *et al.* calculated correlation patterns in a population with Parkinson's disease and in normals [159]. In the normal group, representing a large age-span, a set of patterns were associated with aging: a score based on the subspace projection of each subject onto these patterns was found to correlate strongly with age. In the literature, it is predicted that Parkinsons's disease accelerate ageing. However, the scores of Parkinson's disease patients revealed a negative correlation with the ageing patterns and so the authors concluded that the disease does not accentuate normal change in metabolism due to aging. The correlation patterns here

[8] In our experiments, the differences in the ML and WML projections were small.

Author	Application	Design	Test	Regressors	Correlation	Database	Tracer	Reg.	Norm.
Houston et al. (1994) [104]	AD[a], infarcts	atlas	z-score	patient	PCA, descriptive	AD (10), infarct (12), normal (53+8)	HMPAO	affine	cerebellum
	Inference: Classification (based on the interpretation of abnormal voxels found in the gray matter) **Findings:** At optimal decision level: AD 10/10, infarct 11/12, one false-positive								
Houston et al. (1998) [105]	normals, divers, divers w/decompression illness, boxers, schizophrenics and AD	atlas	z-score	patient	PCA, descriptive	normals (50+40), divers (18), DCI[b] (50), boxers (34), schizophrenics (23), AD (21)	HMPAO	affine	cerebellum
	Inference: Typical patterns in different diseases **Findings:** Multiple small lesions are as common as single large lesions for divers w/ DCI, not for AD and schizo. Large lesions in parietal and inf. temp. regions in AD. in par. and occipital regions for divers w/ DCI and boxers. Inferior frontal region in schizophrenia								
Zuendorf et al. (2003) [238]	normal metabolism	parametric	ANCOVA, T^2 test in subspace	age	PCA, exploratory w/regressors	normals (74)	FDG	affine	whole
	Inference: Normal correlation patterns and dependence on age **Findings:** Second PC[c] is more strongly correlated w/ age than first. PC1: in and adjacent to ventricles and basal cisterns. PC2: high loadings on prefrontal, post. parietal and post. cingulate								

Table 6.4: Multivariate, voxel-based approaches.

[a] Alzheimer's disease
[b] Decompression illness
[c] Principal component

were found using the so-called scaled subprofile model (SSM). The model assumes that the observed rCBF_{ij} in subject j and region i is composed of three multiplicative components: subject-specific global flow (GSF_j), region-specific flow (group mean profile - GMP_i) and a subject- and region-specific component (SRP_{ij}) as:

$$\mathrm{rCBF}_{ij} = \mathrm{GSF}_j \times \mathrm{GMP}_i \times (1 + \mathrm{SRP}_{ij}).$$

From this equation the logarithm is taken in order to obtain an additive relationship and PCA is performed on the SRP to obtain the correlation pattern. To obtain the SRP the authors proceed by first calculating approximate values for the GSF (as global blood flow) term and the GMP (average region flow) term. These approximate values are then corrected for after reconstruction with the estimated SRP. The SSM/PCA method has been used by the same research group in a series of applications [190, 167, 29] and the reproducibility of the patterns has been assessed in [160].

Another multivariate technique, called partial least squares (PLS) has been employed by Leibovitch *et al.* [136] to characterize typical hypoperfusion in a group of patients that have left hemispatial neglect after stroke (they do not perceive visual stimuli from the right). The technique is quite similar to multivariate linear regression, but does not account for any autocorrelation [16]. After removing confounding data (age, gender, education and CT lesion volume), the flow data is covaried with predictor variables from neuropsychological tests. In this particular study, 81 subjects were studied, each passed 4 tests and the image data was quantified into 152 regions. Let $\boldsymbol{Y}$ be the 152×81 sample matrix of image data with one column for each subject, and let $\boldsymbol{X}$ be the 4×81 test score matrix, also with one column for each subject. Whereas PCA consists of approximating the covariance matrix $\boldsymbol{Y}\boldsymbol{Y}^T$, PLS consists of approximating the cross-correlation matrix using (sparse) singular value decomposition:

$$\boldsymbol{Y}\boldsymbol{X}^T = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T.$$

Here, $\boldsymbol{U}$ is a matrix of saliency images (equiv. to eigenimages), the diagonal matrix $\boldsymbol{S}$ contains the covariances and $\boldsymbol{V}$ the saliency test scores. They found that the relative influence of the first saliency image was 94% with an equivalent contribution from all four testscores. This image was therefore interpreted as the significant multivariate response to the (multiple) predictor variables (test scores). Furthermore, each subject's image data and test scores was projected on the saliency image and the saliency test score, and a scatterplot of these projections revealed groupings of stroke patients with and without hemispatial neglect. PLS has been criticized in [229] because the results change with scaling of the predictor variables (meaning care must be taken since there is not necessarily a well defined euclidean distance in predictor space). For fMRI and PET, multilinear models [229] and canonical correlation analysis [64] has been proposed as alternative methods. These have not, however, been applied in conjunction with SPECT images.

Other multivariate studies (Tab. 6.5) include that of Jones *et al.*, who performed SVD (PCA) on regional flow data [115]. The regions were determined by determining a polar grid of radial spokes and an external contour on image slices. In [200], Stoeckel *et al.* took a more classical pattern recognition approach to classify patterns using Fischer discriminant analysis. Finally, Pagani *et al.* performed PCA on regional flow data defined by using a software tool that is called Computerized Brain Atlas [83].

Author	Application	Design	Test	Regressors	Correlation	Database	Tracer	Reg.	Norm.
Moeller et al. (1997) [159]	PD[a]	parametric	ANCOVA	age	PCA, exploratory + discriminative w/regressors	idiopathic PD (37), normals (20)	FDG	n.d.[b]	whole
	Inference: Dependence of covariance patterns with age in patient group **Findings:** Age-metabolism relationship is progressively disrupted in PD. PD is not simply accentuation of normal ageing								
Jones et al. (1998) [115]	normal patterns	parametric	ANOVA, MANOVA	age, gender	SVD, exploratory w/regressors	normal (152)	HMPAO	n.d.	n.d.
	Inference: Age and gender specific correlations **Findings:** No sign. change in whole-brain uptake. regional declines in lat. ventr. Higher uptake in women								
Leibovitch et al. (1999) [136]	Hemispatial neglect (after stroke)	activation, classification	z-score, subspace projection	confounds: CT lesion volume, age, sex, education. covariates: neuropsychiatric test score	PLS	stroke patients (49+32)	HMPAO	n.d.	cerebellum (one hemisphere)
	Inference: Identification of brain regions with hypoperfusion which can be related to hemispatial neglect **Findings:** Right temporo-parieto-occipital junction is more important than other, earlier reported findings								
Stoeckel et al. (2001) [200]	AD[c]	atlas, classifier	Gaussian classifier	none	Fisher discriminant	AD (29), normals (50)	HMPAO	affine,	whole, top 1%
	Inference: Classification **Findings:** 90% correct classification								
Pagani et al. (2002) [179]	normal patterns	parametric	ANOVA	age, gender	PCA, exploratory w/regressors	normal (50)	HMPAO	CBA[d]	13% of voxels with highest intensity
	Inference: Age and gender specific correlations **Findings:** Higher CBF right. Decrease w/ age (in part. left) - stronger in females. rCBF decrease w/ age in vertex, l. frontotemporal and temporocingulate cortex, rel. rCBF increased								

Table 6.5: Multivariate, region-based approaches. Moeller *et al.* also perform voxel-based PCA in addition to region-based PCA.

[a] Parkinsons's disease
[b] no description
[c] Alzheimer's disease
[d] Computerized Brain Atlas, a commercialized software of segmented anatomical structures and a non-linear registration algorithm with up to 18 free parameters [83].

6.3.4 Image databases

Given the large variability of brain perfusion in a normal population, it is obviously important to develop reference (or normative) data in order to improve the understanding of normal CBF and thus the diagnosis of abnormal CBF. Several studies mentioned in the last section have addressed this question [109, 127, 203, 179]. However, given the complexity of the data processing and the analysis it is difficult to reproduce such results from paper publications. This is why several groups are constructing databases of images that are available on the Internet. The idea of these databases is to provide images and standardized algorithms for the processing of images. The two major projects for this are: Neurogenerator[9], which is a European project and ICBM[10] (International Consortium of Brain Mapping), which is a north American project.

However, neither the ICBM project, nor the Neurogenerator provide SPECT images (they are biased toward PET, fMRI, MRI, cytoarchitectural data, etc.). A more modest project that provides normal SPECT images has been initiated by the Society of Nuclear Medicine (SNM)[11]. Because of less reliable quantification in SPECT, the comparison of SPECT images across centers is, however, more difficult than for PET/fMRI [126]. As a note, the integration of neuroscience and informatics has been termed *neuroinformatics*. A list of databases in this domain can be found in the bibliography on Neuroinformatics [169].

6.3.5 Partial conclusion

We have reviewed different statistical methods used for group studies and atlas techniques based on SPECT/PET images. In order to better situate these, we structured the review based on the statistical approach (uni-/multivariate) and on the feature that is modeled (voxel/region/other). Notions about experimental design and how an atlas fits into statistical hypothesis testing is important for understanding relations and differences between the approaches. We now turn our attention from the modeling aspects to the preprocessing of the atlas.

6.4 Registration

Registration is the problem of superimposing corresponding anatomical structures from two different images of the brain. For a brain SPECT atlas, the goal is to superimpose all the database images so that the intensities observed at a given voxel can be considered to represent brain perfusion at the same anatomical location. For a complete, in-depth survey on medical image registration techniques, we point the reader to [145]. In this section we give an overview of registration techniques that have been of importance to this work. We describe the principal elements of the different registration scenarios and algorithms that have been implicated. These algorithms compose a toolbox that has been used for devising a complete registration scheme. This scheme is described in the next chapter.

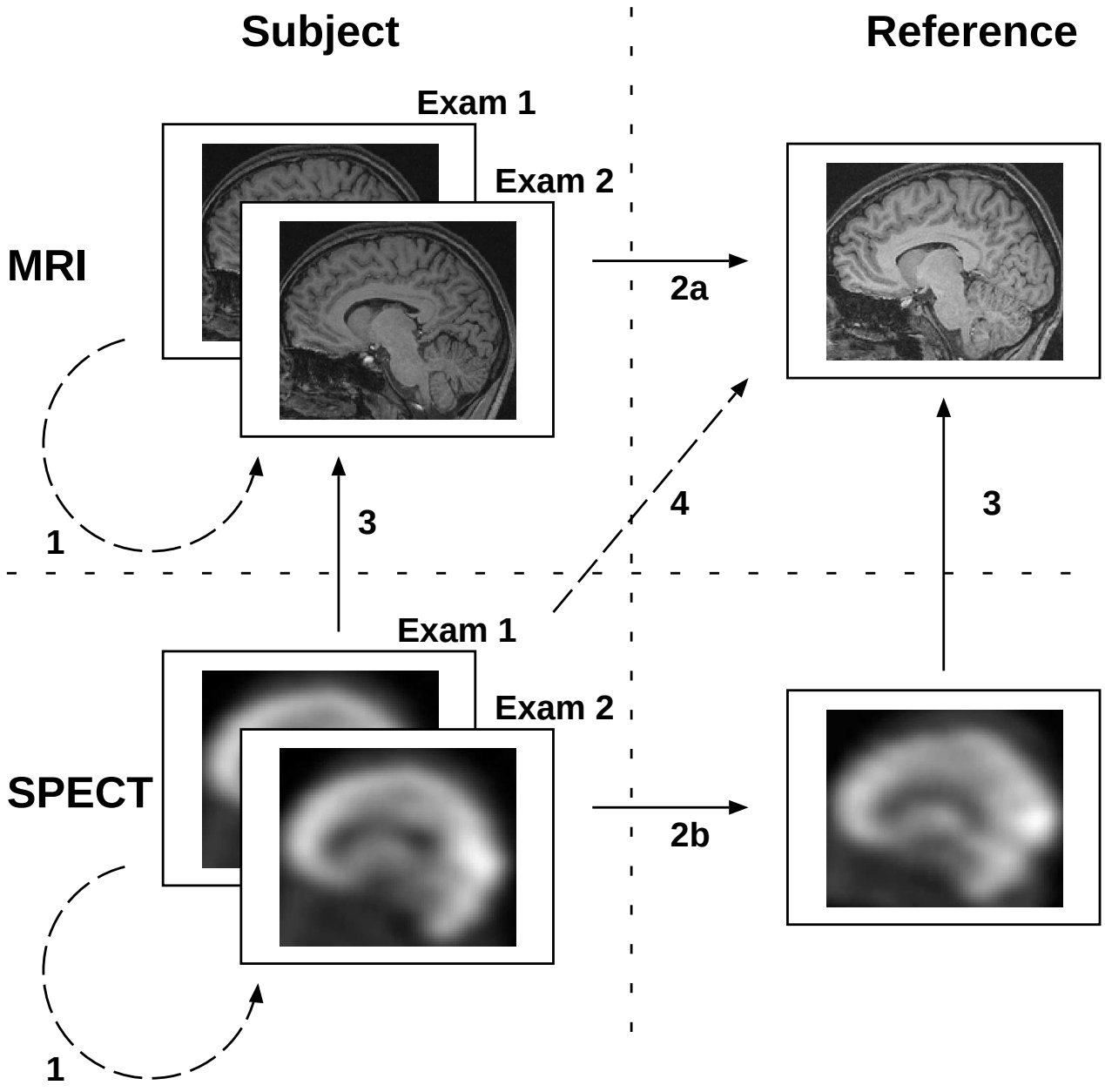


Figure 6.3: The main scenarios for registration. The reference image can either be a template, an atlas (model) or simply another subject.

6.4.1 Overview

As seen in the schema Fig. 6.3, we can distinguish four principal registration scenarios: (1) intra-subject, intra-modal, (2) inter-subject, intra-modal, (3) intra-subject, inter-modal, and finally (4) inter-subject, inter-modal registration. The first scenario occurs when several images are obtained for the same subject over time. These must then be registered in order to follow the development of a pathology such as multiple sclerosis [18]. The last scenario can occur when we have only one image of a subject (e.g. a SPECT image), in which we would like to automatically define volumes of interest (VOI). In this case, it is necessary to register the image across modalities and subjects. Scenarios 2 and 3 are described in the following paragraphs.

In general, the registration problem can be cast to an energy minimization problem. For this, one has to (1) define a similarity measure between images and (2) choose a deformation model. Since the resulting target or energy function might be complex, having many local minima, a main problem is to find a good algorithm to optimize the target function. The principal problems in the different scenarios are situated differently. We will now describe these for scenario 2 and 3.

6.4.2 Intra-subject, SPECT-MRI rigid registration

As already mentioned in Sec. 2.3, intra-subject, inter-modal registration has an important application in superimposing functional images (i.e. SPECT, PET) on anatomical images (i.e. MRI, CT). Because the functional images often have a low spatial resolution, it is difficult to localize abnormal zones anatomically in the brain. This difficulty is alleviated by bringing the functional image into correspondence with an anatomical image of the same subject.

In most applications, the brain is considered to be a rigid structure. One therefore chooses a transformation model that only consists of six free transformation parameters: three for translation and three for rotation (assuming there are no deformations caused by the imaging system). There are however situations where a 12 parameter affine[12] model, or even a deformable model are more appropriate [18, 45]. The main concern in this type of registration is the choice of a similarity measure. Since the images have different physical interpretations, one cannot directly compare voxel intensities. Simply minimizing a square error function is therefore excluded. A multitude of different similarity measures have been developed. These can be classified into three different groups: (1) landmark-based [184], (2) surface-based [95] and (3) intensity-based [96]. For the first two groups, it is fairly easy to define a similarity measure between landmarks or a surface. However, the problem's focus has been shifted to that of the automatic detection of landmarks or the segmentation of surfaces.

For fully automated registration, intensity-based measures seem to yield the best results. The most successful intensity-based measures have been the "Wood" criterion [224] (also variance of intensity ratios), cross-correlation, mutual information and normalized mutual information.

In a comparative, multi-center retrospective study, the mutual information criterion, together with multiresolution, simplex optimization yielded the best results [222]. This scheme has been implemented by other researchers in our laboratory (Nikou, Musse, Sinapin) and was accessible for this work. In particular, with this approach, no pre-segmentation of the

[9] `http://www.neurogenerator.org`

[10] `http://www.loni.ucla.edu/ICBM/index.html`

[11] `http://brainscans.indd.org/brncncl4.htm`

[12] An affine transformation maps parallel lines on parallel lines

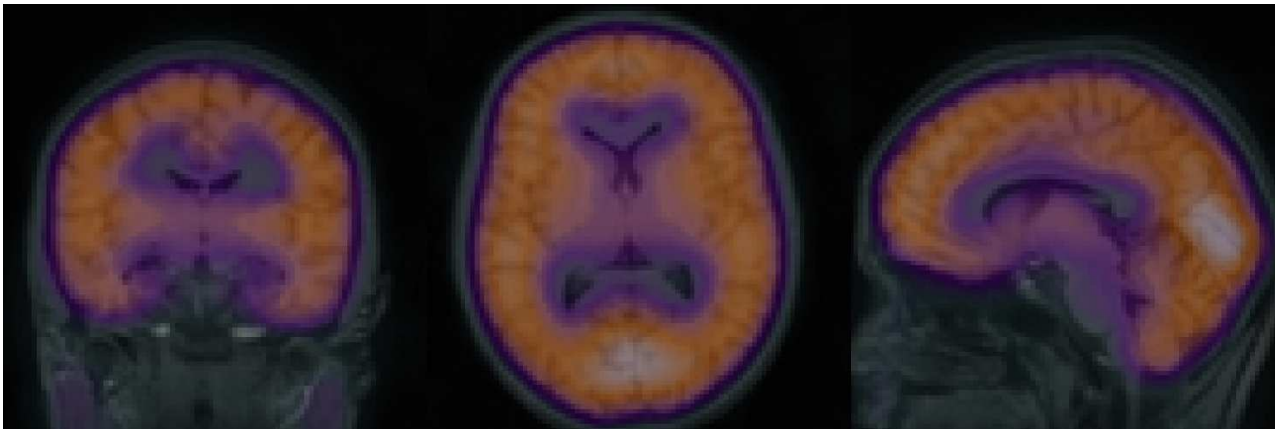


Figure 6.4: Example of a SPECT exam registered with a MRI exam of the same subject using a rigid transformation model (translation and rotation). The images are superimposed with the SPECT image displayed in colors (violet for the lowest values, over orange to white for the highest intensity values) and MRI in gray values "underneath". The registration is indeed, visually satisfying.

scalp is necessary and the capture range ("distance" of initial displacement, still yielding good results) is rather good. The problem of SPECT-MRI registration of the same subject is largely considered to be "solved" since a precision of about a few millimeters is reached and considered to be sufficient. Visual assessment of the quality of SPECT-MRI registration is nevertheless necessary. This is best done by superimposing the transformed SPECT image on the MR image, using a color palette for the SPECT image. An example is shown in Fig. 6.4.

6.4.3 Inter-subject, MRI deformable registration

Whereas for intra-subject, SPECT-MRI registration the main problem is to define a robust, sensible similarity measure, the main problem in inter-subject registration is to define a (valid) deformation model and to solve the related optimization problem. Given the high anatomical variability between subjects, an affine transformation model is not sufficient. For example, gyri and sulci are defined by how the cortex is folded, and some of these show high variability (typically those that are developed late in the fetus [214]). Ideally, one would like to have a global deformable, parametric model that could span all brain "prototypes" and at the same time be mathematically tractable (or more precisely: computationally tractable). This is the case for the cortex, which is inherently a two-dimensional surface. The best method for comparing the cortex of different brains therefore consist of segmenting the outer cortex and inflating it to a sphere [49, 61]. This method is however not adapted for the whole brain. A similar method maps the outer cortex to a sphere and permits the mapping of deeper structures as well, but necessitates much user intervention [207].

A vast literature on intensity-based methods for deformable matching of brains, witnesses the complexity of the problem [145, 208]. These range from physical models (e.g. elastic and fluid deformation models) to purely mathematical models (e.g. polynomial-, Fourier- or spline-based deformation models). In our laboratory, an algorithm based on the hierarchical spline-decomposition of the transformation has been developed [166, 172, 173, 174]. The principal properties of this algorithm shall now be described.

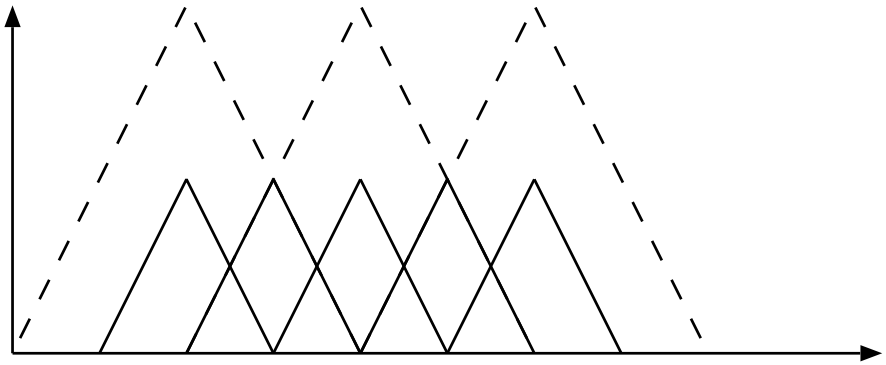


Figure 6.5: First order B-cubic spline functions in one dimension at two different scales (supports).

As for all deformable registration algorithms, the images are first brought into an initial registration using an affine transformation model. This affine transformation compensates for global differences in position, rotation, shear and size between the two brains. From this initial registration, a discrete deformation field is calculated. This field is a parametrized field of vectors that define the spatial displacement of each voxel. The parametrization is done by choosing a representation of the field using basis functions. First order B-cubic splines were chosen as basis functions (see Fig. 6.5). They were chosen (among other reasons) because they are well adapted to a multiscale approach (i.e. the B-cubic splines are scaling functions associated to wavelet representations). This means the algorithm can start by minimizing the cost function associated with a large scale decomposition of the field, i.e. large spline support and thus few parameters. The resulting minimum is used to initialize the decomposition of the field at a finer scale. Since there is an analytic relation between the field decomposition at different scales, this transition is simple. The algorithm thus calculates a deformation field at finer and finer resolutions that gradually compensates for anatomic differences. See Fig. 6.6 for an illustration and Fig. 6.7 for a synthetic example of this registration.

Whereas the main problem in inter-subject registration is to define an appropriate deformation model, the choice of similarity measure is of course also important. As an intensity-based criterion, the sum-of-square-residuals,

$$C_q(h \circ S, R) = \sum_{p_i \in \Omega} (S(h(p_i)) - R(p_i))^2, \tag{6.4}$$

has been the most popular. Here, S designates the source image, R is the reference image and h is the transformation that transforms S into correspondence with R. The sum is calculated over the support of the images, Ω. This criterion is not symmetric, that is, the transformed version of the source images is compared to the reference image[13]. Because of sampling errors and limitations of the transformation model, one does not in general obtain the inverse transformation of h when swapping reference and source image. This led the authors

[13]Actually the roles of reference and source image is inversed in practice since an inverse transformation is calculated. This is in order to avoid the so-called forward-sampling problem: Because the transformation is non-linear, the regular source grid is not necessarily transformed to a regular grid. This can lead to holes in the transformed image.

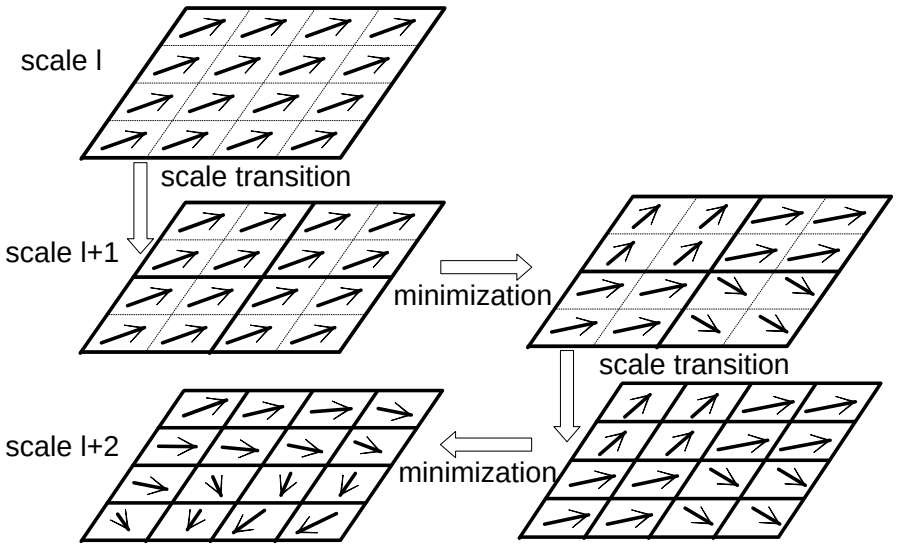


Figure 6.6: Illustration of the multiresolution approach where the deformation field is gradually calculated at finer and finer resolutions.

in [38] to propose a symmetric criterion instead:

$$C_{sym} = C_q(h \circ S, R) + C_q(S, h^{-1} \circ R) = \sum_{p_i \in \Omega} (S(h(p_i)) - R(p_i))^2 + \sum_{p_i \in \Omega} (S(p_i) - R(h^{-1}(p_i)))^2.$$

This has been taken into account in the registration algorithm that we use [173]. Furthermore, the square sum measure in Eq. 6.4 was replaced by a Lp-criterion (p=1.2) (see also Sec. 3.5.2)

$$C_{Lp}(h \circ S, R) = \sum_{p_i \in \Omega} |S(h(p_i)) - R(p_i)|^p,$$

which can be considered to be more robust to gross outliers than the square sum [173].

Other factors such as interpolation and intensity normalization also influence the registration. Undersampling of the physical reality leads to aliasing effects which might lead to violation of the assumptions on the deformation model. Physical and anatomical factors might lead to inhomogeneities in the images. These again can lead to errors in the cost function (similarity measure), so that a global minimum of the energy function does not correspond with well registered images.

Because of the complex structure of the matter let us at last mention that the validation of a registration procedure is extremely difficult [226]. This is because the ground truth is not known and the definition of what a good registration is depends on the application. Image registration has many applications, but only few new methods are taken to clinical practice where their impact on diagnostics and treatment can be measured [145]. The animated view (displaying the images in sequential order) of the transformed image and the reference image is a useful tool since human vision is extremely sensitive to movement.

6.4.4 Methods for inter-subject SPECT registration

The main problem encountered in SPECT-SPECT inter-subject registration, is the reliable estimation of the transformation parameters. This is because of the low structural information present in SPECT images. Direct registration of SPECT images is most reliable for

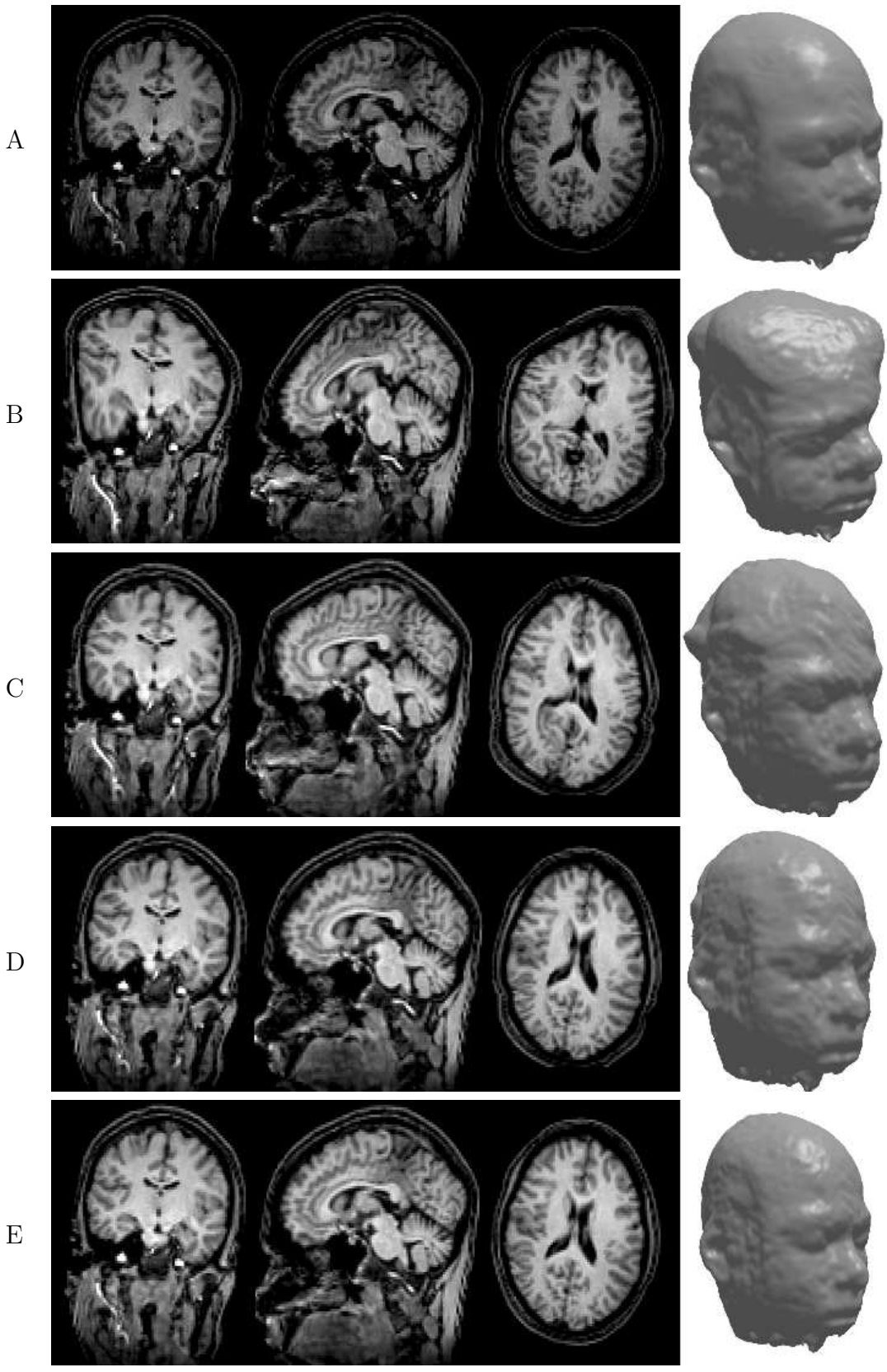


Figure 6.7: Example of non-linear, hierarchical registration at different resolutions. (A) Reference image. (B) Synthetically deformed image. (C) Registration at the second scale (81 free parameters). (D) Registration at the third scale (1029 parameters). (E) Registration at the fifth scale (89373 parameters). The algorithm consecutively corrects for deformations of finer detail.

low-parametric transformation models (i.e. rigid, affine), which are less accurate in terms of aligning anatomical structures. Even though such registration has been "validated" for statistical pooling of SPECT images [109], we have found that structural variation still remains significant. This could be assessed in the evaluation study in [221], also described in Ch. 8.

Another approach to SPECT-SPECT, inter-subject registration is to co-register SPECT images with the MRI scans of the same subject. The transformation obtained by MRI-MRI, inter-subject registration can then be applied to the co-registered SPECT images (composing step 3 and 2a in Fig. 6.3). This has the advantage that one can obtain a precise inter-subject registration using deformable transformation models that are reliably estimated from high-resolution MR images.

At last, let us mention that there also exist approaches where deformable SPECT-SPECT registration is done directly. In [131] for example, the authors have used a daemons-based approach for non-linear registration of SPECT images. The method is described in [111].

6.4.5 Integrating registration into the statistical model: modeling anatomical variance

In the scheme that we have depicted in Sec. 6.2, we have considered the registration to be a separate preprocessing step that is independent of the statistical model. This is not the only approach possible. The morphological deformation, which is necessary for the compensation of anatomical differences, contains information about a particular subject. Many neurological disorders are associated with morphological atrophy, which again leads to modifications in such deformations. An interesting approach is therefore to consider a statistical model that takes into account both morphological deformations and intensity values.

This has been done in [206, 209], where the transformation fields themselves are modeled. A field is a vectorfield of displacement vectors, defined on the same support as the image. The statistical model is simply a 3-D Gaussian distribution for each displacement vector. The distribution is estimated by registering a large number of images onto a reference image. In [131], the displacement vector is used for classification of the brain into a normal or a pathologic class. Further, statistical modeling of deformation fields is also possible. For example, Le Briquer *et al.* have explored the possibility of decomposing a group of deformation fields using PCA [24]. Rather than modeling the deformation fields themselves, Machado *et al.* have instead modeled the variance of the jacobian of the deformation fields [143]. Since the jacobian is a measure of local volumetrical change, large changes may indicate anatomical abnormality.

6.4.6 Structural approaches to inter-subject comparison

Most approaches in the neuroimaging community are based on so-called coordinate-based registration and image comparison. This is in contrast to the neuroscience community where a structural approach is taken and links are sought between architectural and cognitive models. Another approach to group analysis (atlas or group comparisons) as proposed by Mangin *et al.* [146] is therefore the so-called structural approach (see also [140]). Here, the idea is to transform the raw image into a structural representation (an approach often used in computer vision). For example, in [147] the authors use image segmentation techniques based on a priori information (atlas, learning) and the grey levels of the image in order to delineate gyri and sulci of the cortex. A structural representation is then obtained as a set of features for

each sulci (shape, position) and their interrelations. This leads to a graph representation and registration is performed as graph matching. Comparisons can then be done across graph nodes and features. An approach to SPECT atlasing based on for example VOIs could easily be devised from this kind of structural registration. With the first results appearing, [147] this kind of approach does indeed seem to be very promising. A strong argument in favor of the method, is that it facilitates the integration of neuroscientific knowledge into the analysis as shown for example in [27]. Here, models of the neurogenesis is taken into account and the development of the cortex is studied. However, as the approach is composed of a large range of techniques (image segmentation, pattern recognition, artificial intelligence techniques and others - which all must be evaluated and implemented) its application remains complex and therefore difficult.

6.5 Brain segmentation

We understand by segmentation of SPECT images to be the classification of every voxel into different classes with a physical meaning. In our model, we are interested in segmenting the brain, which is our structure of interest. This is a fairly simple task that can be done by thresholding the image. The most widely used approach in the literature, consists of fixing a relative threshold, such as 40 % of the maximum value in the image. This normally separates well the scalp and ventricles from the cortex (white and gray matter), which has higher intensity values in general. However, there exist other methods such as the one described in [154] where the authors have modeled the images using a 3D Markov random field model combined with a density mixture model. We have not considered methods for the segmentation of SPECT images into finer structures. Nevertheless we can mention that ROI/VOI segmentation of SPECT images is often done by registering with a template (atlas), see Sec. 6.3.2.

6.6 Intensity normalization of SPECT images: Existing approaches

The intensity normalization is at the heart of quantitative SPECT imaging[14]. In order to compare the activity distribution between subjects or scans, the observed counts must signify absolute measures of this activity. Reporting absolute values of rCBF also makes it possible to compare values across imaging centers. The problem of obtaining quantitative measures in emission tomography is subject to active research[15].

Among the different factors leading to count errors, it seems to be important to correct for scatter events. This can be done by modeling such errors during image reconstruction (using algebraic reconstruction algorithms). Unfortunately, the classical filtered backprojection algorithm, which is much faster, is the standard algorithm used in clinical settings, and was the only one available for the images used in this work. In this section we review the different approaches for SPECT intensity normalization used in the literature. However, there does not exist many comparative studies or studies that assesses a method's validity. As seen in Ch. 8, we found that such normalization had a high impact on the atlas that was created, but we could not assign validity to the different methods.

[14]It is more accurate to speak of semi-quantitative SPECT as discussed in [43]. The literature prefers the term quantitative however.

[15]See for example `http://www.imed.jussieu.fr/`

6.6.1 A controversial topic

Whether intensity normalization based on analysis of the images themselves is possible or not, is a topic of discussion, see [43]. Here, one author argues that the normalization of count values to a reference region yields results that are satisfactory. The opponent however, argues that such normalization may lead to random error as a cause of variation in any region of the brain. The alternative method of measuring the arterial input function during the injection of the tracer is therefore preferred. However, this measure is considered to be unpractical or invasive and is therefore rarely applied. In our study, such measures were not available so we had to consider methods for intensity normalization.

Many different strategies for intensity normalization have been proposed in the literature, but no consensus exists on which one is the most appropriate. Authors argue that one way of normalizing is better than another because it lowers the total variance (sum of voxel variances) in a set of images. This argument is not justified because we can always use a (non-linear) histogram transformation technique to obtain a global low variance, whether the transformation technique is physically justified or not. If we observe different global counts in two images, this might be caused by true global differences (which we *do not* want to compensate) and differences caused by differences in injected dose, head fraction or scanner sensitivity (which we want to compensate for by normalizing). The appropriate normalization technique in a given situation may therefore depend on the application at hand.

6.6.2 Which transfer function?

Intensity normalization can be separated into the choice of a transfer function and the estimation of this function. The transfer function maps the observed intensity values in an image to absolute values. Fig. 6.8 shows a joint histogram (scatter plot) between two images. The joint histogram is a probability function on $S = I_1 \times I_2$, where I_1 and I_2 are the gray value range of image one and image two, source and reference image, respectively. This probability function describes the frequency of co-occuring intensity values in the two images and is clearly linear[16]. Existing methods for SPECT intensity normalization therefore only consider linear transfer functions of the type:

$$y = \gamma x + \beta, \qquad x \in I_1.$$

We thus have three possible transfer functions: (1) variable scaling factor γ and $b = 0$, (2) variable constant b and $\gamma = 0$ or (3) variable γ and b. For count images, such as SPECT and PET, a proportional scaling factor, a, is considered to be the most appropriate [68]. To see this, consider the global CBF (gCBF) to be the same in the two exams: since the proportion of regional CBF remains the same, a difference in injected dose or a difference in the fraction of dose being distributed to the brain results in the same proportion between rCBF and gCBF. However, this only remains true when assuming that the uptake depends linearly on the injected product, which is not true for high perfusion rates (Sec. 2.6.2). This might explain why fitting a function with an additive constant on the joint histogram is sometimes clearly better than fitting a function without such a constant (see also Fig. 7.5 in the next chapter). A transfer function only defined by an additive constant exists in a particular case that will be described in Sec. 6.6.4.

[16] In MR images however this is rarely the case as shown in, among others, [18].

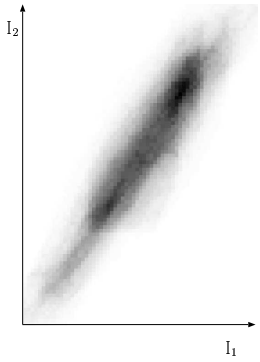


Figure 6.8: An example of a joint histogram of two registered SPECT images. The histogram is calculated as described in [18, 17], and takes into account the so-called partial-volume effect.

6.6.3 Estimating the transfer function

We first consider methods that have been used to estimate a single proportional factor for normalization. This constant is calculated by comparing the mean of a reference region to a specified value that is expected for this region. The problem consists in finding a reliable reference region that exhibits this expected value in all image acquisitions. Probably, the simplest reference region to choose is the maximum value in the image. The choice of the maximum value has the advantage that among all the values in the image, it is subject to the lowest influence of error for the estimation of γ (this is seen by setting $\gamma = v_{expected}/v_{reference}$ and deriving by $v_{reference}$). However, the maximum value as reference value leads to higher variance in the normalized images than a reference value based on the mean or median. In [185], the choice of the maximum value as a normalization constant, falsified the result in one case of epilepsy. It is therefore more widespread to calculate the mean or median in a reference region such as the whole brain, the cerebellum or more rarely, the thalamus or the basal ganglia [180]. Due to the same reasons as in Sec. 6.3.2 (partial volume effect in VOIs), the latter two are probably not ideal because of their small size and their location adjacent to regions of low activity (ventricles). Use of the cerebellum as a reference region is also error prone. This is because an alteration of the perfusion in one temporal lobe can alter the perfusion temporarily in the contralateral cerebellum (diaschisis).

For the normalization of SPECT images in Alzheimer's disease (AD) and frontal lobe dementia (FLD), Pagani *et al.* [179, 180] have proposed to use 13 % of the highest intensity values as a reference region. The choice of the 13 % was based on an analysis of the resulting size of the region in AD and FLD. The highest values were chosen based on the same reflection as for the maximum value above.

In [185], the authors compared different normalization methods for comparing ictal and interictal images in epilepsy. Their conclusion was that one scaling parameter was sufficient. The best results were obtained using a robust criterion (maximizing the number of sign changes in the two images to normalize). Linear regression on the scatterplot yielded similar results and taking the additive constant into account did not change the results. The reference for this study was visual assessment of the estimated regression lines in the scatterplot as well as diagnostic outcome.

Another method has been proposed in [21] for images of patients with epilepsy. For such images, it is characteristic to have regions of very low values or very high values. Boussion *et al.* therefore propose a strategy to automatically determine a reference region that has

values in the middle intensity range and that has a homogeneous distribution of values. This is done by first calculating the mean and variance of the whole brain and then only keeping values in the range mean plus/minus one standard deviation. This is done for both images, and the intersection of the remaining regions is chosen as the reference region. An alternative method which is based on region growing is also proposed. These methods seems to work well for the relative comparison of two images, but they do not scale naturally to the problem of quantitative intensity normalization of a complete database of subjects.

6.6.4 Integrating normalization in the statistical model: ANCOVA

As with registration, global intensity differences can also be included in the statistical model. This approach has been proposed for activation studies in the SPM framework [68]. The approach has the effect of an additive constant for intensity normalization. This constant is different for each voxel and is estimated conjointly with the other predictor variables of the model.

In SPM, each voxel is modeled separately, and an example ANCOVA model for scan $j = 1, \ldots, J$ under activation $k = 1, \ldots, K$ is:

$$y_{kj} = \alpha_k + \xi(g_{kj} - \bar{g}_\bullet) + \epsilon$$

Here, y_{kj} is the observed value at the voxel under condition k for scan j. The activation is characterized by the indicator variable α_k, which can take on the values $+1$ (activation) and -1 (no activation). Differences in global cerebral blood flow (gCBF) are accounted for by the additive normalization constant, ξ, and the difference between gCBF in scan kj and the average gCBF in all scans, $\bar{g}_\bullet$. The noise is considered Gaussian: $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The parameters of the model, α_k and ξ, are estimated using least squares estimation (linear regression). In this model the distribution of the residual follows an F-distribution and a hypothesis test based on the measured residual ("extra-sum-of-squares"), with (H_1) and without (H_0) the condition factor α_k, decides whether a voxel was activated or not. Differently formulated: the extra-sum-of-squares is a measure of how well the α_k explain the data.

The alternative to an additive normalization constant, proportional scaling before statistical analysis, leads to the model:

$$\gamma_{kj} y_{kj} = \gamma_{kj}\alpha_k + \gamma_{kj}\epsilon$$

Here, the noise variance has changed to $\gamma_{kj}^2\sigma^2$. The resulting noise is therefore dependent on the scan, with the consequence that the residual does no longer follow an F-distribution. The ANCOVA model is therefore a more rigorous approach than proportional scaling. The assumption of the ANCOVA model, however, is that the global CBF changes in an additive manner with changes in injected dose (parallel lines assumption, see Fig. 6.9). This is less appropriate for SPECT images as we have discussed in Sec. 6.6.2. One could also critizise that ξ is estimated using linear regression: the global activity, g_j, is also measured with errors, not only the rCBF at voxel d. This error is not accounted for in least squares regression (see also Sec. 7.5). In a later publication, the same authors suggest using proportional scaling for SPECT studies [1].

6.7 Conclusion

In this chapter we started out by defining what we understand by a probabilistic atlas of brain perfusion. An overview of the atlas creation process was given. We then proceeded by review-

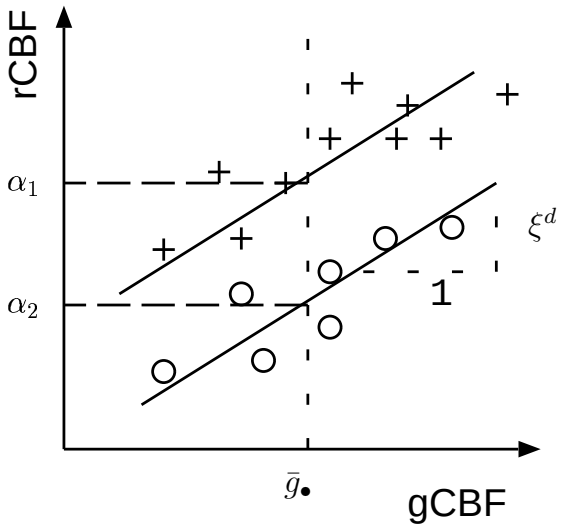


Figure 6.9: Relationship between regional CBF (rCBF) and global CBF (gCBF) at voxel d, modeled in a ANCOVA type analysis of an activation study where '+' signifies activation and 'o' rest. The average global flow over all scans and conditions is $\bar{g}_\cdot$, the slope is given by ξ. The α_q are the condition dependent flow levels after correction of differences in the global flow. In an (implicit) atlas study, only baseline images would determine the slope ξ.

ing different statistical methods that are related to this work. This review is complicated by the large spectrum of applications and methods coming from different domains of research. We have discovered an equivalent method to the appearance-based methods known from computer vision but that originated in the nuclear medicine community [104, 105]. The model can be compared to the ML-model of Ch. 4 and introduces a clever way of comparing an image to the atlas in order to obtain localized detections. The method is based on analysing the residual between the reconstructed and the observed images. We embrace this method and use it to compare images to the atlas.

Our bibliographic review further reveals that the existing atlasing methods all use affine or piecewise affine registration schemes, mostly directly registering the SPECT images from different subjects. We propose to use a multi-step procedure, passing by inter-subject, deformable MR registration as is described in the next chapter. We have not reviewed methods for image registration (which is a vast domain), but we have described the principal difficulties related to registration and we have given the principal characteristics of the algorithms that we have used.

Methods for intensity normalization on the other hand have been thoroughly discussed and reviewed. The discussion show that it is not clear which method is in general preferable. However, after a good registration of two images, the joint histogram (which actually codes the relation of the relation between many reference regions/voxels) seems to be quite linear in nature. This speaks in favour of a linear transformation function easily calculated from this histogram and is therefore our technique of choice.

The approaches we suggest in the next chapter (Ch. 7) will all be evaluated and compared to existing methods in Ch. 8.

Chapter 7

Atlas creation, our approach

In the last chapter, we defined what we understand by an atlas, we reviewed related approaches in the literature, and we reviewed techniques for registration, segmentation and intensity normalization of brain SPECT images. In this chapter we describe the material and methods that we have used and developed. We start by describing the database of normal subjects that is represented by the atlas. We also describe the patient image database which was processed in the same manner as the normal images. We then detail a registration scheme for deformable, inter-subject SPECT registration, brain segmentation, intensity normalization and the statistical models considered. The global work/data flow is depicted in Fig. 6.2. The validation of each individual processing step is rather difficult, but the complete processing chain can be evaluated as is done in the next chapter. We finish this chapter with the results of the learning process (atlas creation) and a note concerning implementation issues.

7.1 Database of normal subjects

The database of 34 normal subjects was acquired at the nuclear medicine facility attached to the institute (Institut Physique Biologique, UMR 7004, Strasbourg), 20 between april and december 2001, and another 14 until january 2004. These were volunteers without anatomical atrophies as seen in the MRI scans. Other exclusion criteria included history of neurological, psychiatric and audio/visual disorders. The project was approved by the ethical committee at the University of Strasbourg and a written consent was obtained from each subject after being informed about possible risks. An MRI scan and two SPECT scans were obtained from each subject. The MRI scans were T1-weighted, GE3D sequences with voxel-size of 1 mm obtained on a 2T Bruker scanning device. The SPECT scans were obtained using a Elsinct Helix double-headed camera with parallel collimator and filtered back-projection for reconstruction. The full width at half maximum (FWHM) was about 8 mm. All 34 images were obtained using a parallel collimator. In the research protocol, the use of a fan-beam collimator was envisaged, but had to be renounced to due to technical problems. The subjects, 12 men and 22 women, formed two age groups of 26.7 ± 6.1 and 46.9 ± 4.3 years (mean and standard deviation).

7.2 Database of patients

At the institute we have access to a database of images of 154 patients with epilepsy. These come from the nuclear medicine service located at the institute. The exams were made with the same gamma camera as for the database of normal subjects, but most MR images are

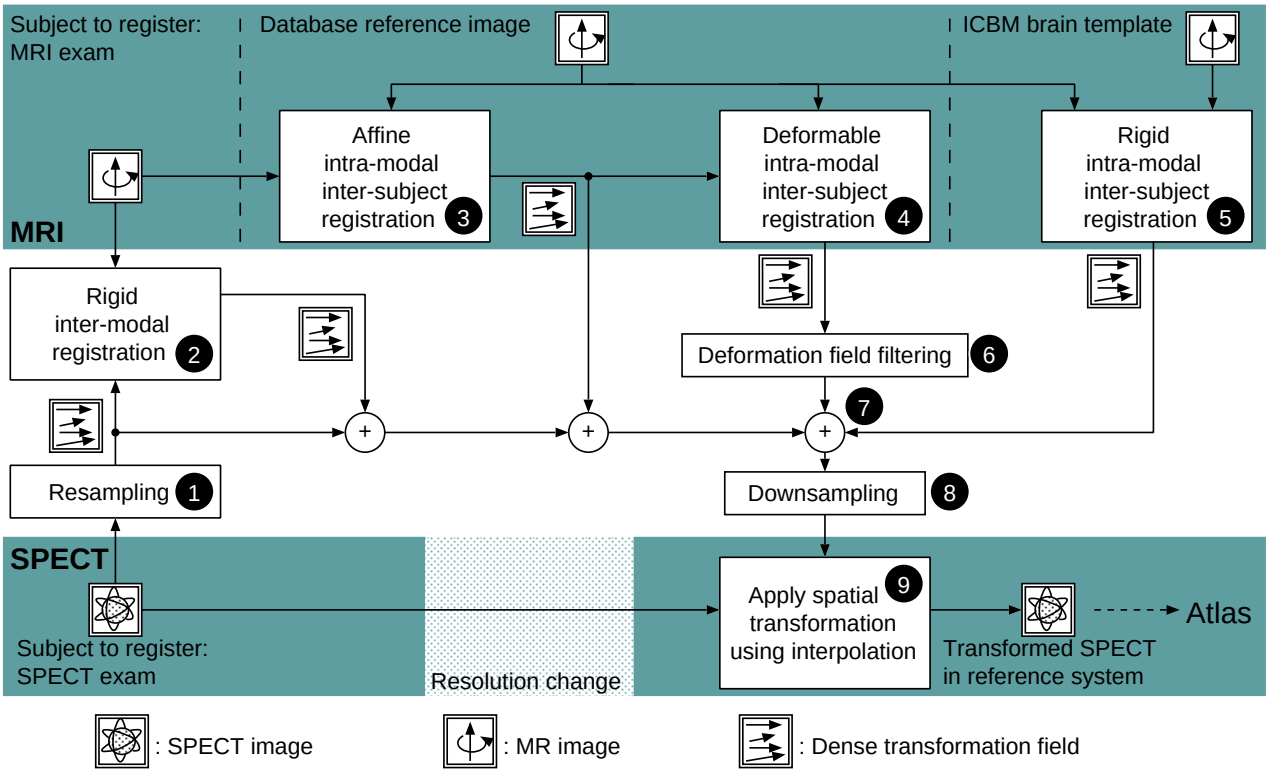


Figure 7.1: Overview of the registration scheme.

of 2 mm resolution. Some of these images show morphological anomalies in the MRI scans and some MR images have slightly different contrasts than those obtained for the normal database. These images have been analyzed by other researchers at the institute using the SISCOM technique described in a later chapter. Six images have been analyzed using the atlas. These results will be described in Ch. 9.

7.3 Spatial normalization: registration

To bring all the database images into a common reference space, we have developed a registration scheme that takes advantage of precise MRI, inter-subject registration as described in Sec. 6.4.3. The scheme is depicted in Fig. 7.1 and will be further detailed in the following. This scheme is, of course, only applicable in the case where both MRI and SPECT exams have been made of the subjects. A particularity of the scheme is the combination of all transformations into one single transformation which is then applied to the original image. This limits errors introduced by repeated interpolation of the original image. Furthermore, a filtering of the deformation field obtained by MR, inter-subject registration has been introduced. This step is further explained in Sec. 7.3.3. As to interpolation, SPECT images were interpolated using trilinear interpolation, and MR images using B-cubic spline interpolation [204]. We now describe the choice of reference space before we describe how the combination of transformations was applied to the SPECT images.

7.3.1 Choice of reference image and reference space

As our goal is to bring all images into one spatial reference space, the choice of this space merits some thought. A major consideration is the visualization of volumetric data. Visualization of 3D brain images on a 2D screen is best done by aligning the sagittal plane of the brain with one of the display axis. This way the symmetrical properties of the two brain hemispheres are highlighted in the coronal and axial views (see also Fig. 2.2). This aids the interpretation of the images significantly. The ambiguity in the remaining two axis was settled by Talairach and Tournoux who proposed the definition of a line in the sagittal plane going through the anterior commissure (AC) and posterior commissure (PC), two well defined and relatively stable locations in the brain across a larger population [202]. This line is oriented with the intersection of the axial and sagittal planes. Furthermore, Talairach and Tournoux presented a "standard" brain of a 66 year old women (on paper). A cartesian coordinate system with the origin being defined at the AC point has been superimposed on this brain. This atlas has become a *de facto* standard reference space in the neuroscience literature, denoted Talairach space. Global spatial normalization into Talairach space consists of transforming a brain by changing its position, its orientation and size so that it conforms to the Talairach brain [132]. This standard space thus permits different imaging centers to exchange coordinates of different structures in the brain (e.g. answering questions of the type "where in the brain" as opposed to "where in this brain").

Global spatial normalization is today typically done by registration with a (digital) brain image already aligned in the Talairach space using a 9 (or 12) parameter affine transformation model [132]. For this, we have used the ICBM brain template[1] which is an average of 452 brains of normal subjects that have been registered into Talairach space. The template is shown in Fig 7.2. Another average brain that is widely used as a reference brain is the average brain of the Montreal Neurological Institute (MNI), McGill University, Canada, which has been created from 152 brains[2]. Both these average brains are somewhat larger than the original Talairach brain (because of inaccuracies in the registration). The latter average template is therefore sometimes referred to as *MNI-space*[3].

Even though we do not report coordinates of our findings in this work, we appreciate the alignment with the ICBM brain atlas for its visual properties. However, because of the lack of sharpness of the averaged image, we have not used this template directly as a reference image for deformable registration. Several studies suggest that a site-specific reference (or average reference) image is preferable to an image obtained with another imaging device [81, 85]. This might be particulary true for high precision deformable registration which is less robust to differences in contrast and acquisition errors than low-parametric registration. We have therefore used as reference an (MR) image that was issued from the same imaging device as those of the database. However, an image outside of the database was chosen to avoid biased preprocessing. We chose to use this reference in its original position to avoid interpolation errors. This explains why we have two reference images in the scheme in Fig 7.1 ("Database reference image" and "ICBM brain template"). The database reference image is registered with the ICBM brain template using a rigid deformation model (rotation and translation) and a sum-of-square residuals cost function. A rigid transformation was prefered over an affine transformation because the affine would have enlarged the images to that of the average brain;

[1] From `http://www.loni.ucla.edu/ICBM/ICBM_452T1.html`

[2] This is the template brain of SPM.

[3] For a discussion on the differences and conversions between Talairach space and MNI space, see `http://www.mrc-cbu.cam.ac.uk/Imaging/Common/mnispace.shtml`

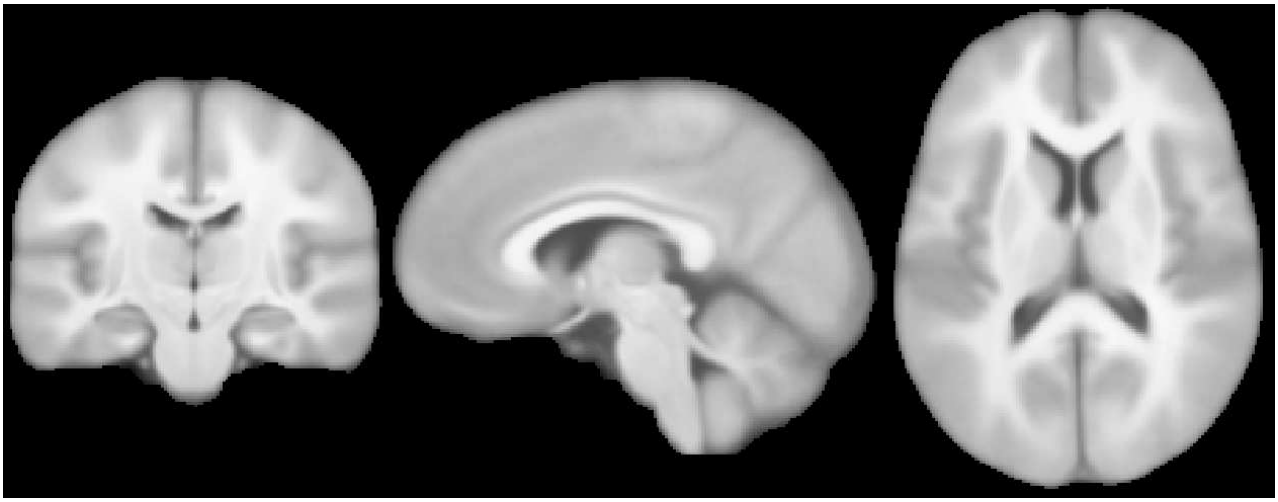


Figure 7.2: The ICBM 452 T1 brain template is an average brain of normal subjects. All subjects were registered using a fifth-order polynomial transformation model. The reference space defined by this (anatomical) atlas has been referred to as MNI-space, stereotactic space or Talairach space.

"our" coordinates are therefore not exactly MNI/ICBM-space. We shall nevertheless refer to this space as ICBM space.

7.3.2 Registration scheme

Since the algorithms we have used were described in Sec. 6.4, we only summarize the different registration steps. See Fig. 7.1 for an overview of the complete registration scheme. We start by resampling the SPECT images to 4 mm isotropic resolution (step 1). We then register each subject's SPECT image with the same subject's MR image (step 2). This is performed as described in Sec. 6.4.2 (rigid transformation and mutual information). Inter-subject, MRI registration was then performed as described in Sec. 6.4.3 with the database reference image as the reference (steps 3, 4 and 5). For images of 1 mm resolution (256^3 voxels), we used deformable matching with the finest resolution scale level 6 (which corresponds to a spline support of 2^3 voxels, 750 141 parameters). For patient images, that were of 2 mm resolution (128^3 voxels), we used deformable matching with the finest resolution scale level 5 (also spline support of 2^3 voxels, but 89 373 parameters). After each registration, the transformation and resampled images were stored to disk. All registrations were assessed by visually inspecting superimposed views (for SPECT-MRI registration), side-by-side and animated views (for MRI-MRI registration).

7.3.3 Application of transformations to SPECT images: combination and downsampling of deformation fields

In order to avoid accumulation of interpolation errors, the transformations were combined into one global transformation (step 7 in Fig. 7.1) that took the SPECT images directly from their original position (in original resolution) to the (subsampled) ICBM reference space (step 9). Care was taken to properly take the different transformations across resolution changes (step 8). For the cross-validation study described in the next chapter, this reference space was a

subsampled version (voxel resolution of 4 mm) of the original ICBM space (in order to save disk space with the large number of generated images). For the epilepsy studies described in chapter 9, a ICBM reference space with 2 mm voxel resolution was chosen.

However, we found that the direct application of the deformation field as obtained from inter-subject MRI registration was not possible. This is illustrated in Fig. 7.3 where we see an image brought into ICBM space by affine registration (without the deformable registration in Fig. 7.1) on the top and the same image brought into ICBM space by application of a non-linear deformation field with 750 141 free parameters beneath (without the deformation field filtering in Fig. 7.1). The image on the top has the appearance of a "real" SPECT image, i.e. the image smoothness is the same as that of the original SPECT image. However, in the image on the bottom we see that high frequency components have been introduced into the image. The appearance of the image has completely changed and the image no longer looks like a "valid" SPECT image. The artifacts stem from the registration of fine structures in the cortex. Since these structures are present in MR images but not in SPECT images, the deformation field transforms "invisible" structures in the SPECT image.

In order to reduce these artifical, high frequency artifacts, we experimented with different strategies (see Fig. 7.1):

1. Only using affine transformations, i.e. without deformable transformation (step 4) in Fig. 7.1. We shall refer to this strategy as the *Affine* registration strategy.

2. Deformable registration at a lower resolution, scale 3 (1 029 free parameters), without filtering of the field. We shall refer to this strategy as the *Deform3* registration strategy.

3. Deformable registration of scale 6 (750 141 free parameters) and filtering of the field before application to the SPECT image (step 6 in Fig. 7.1). We shall refer to this strategy as the *Deform6-NF* registration strategy.

4. Deformable registration of scale 6 and direct application without filtering of the field. We shall refer to this strategy as the *Deform6-F* registration strategy.

We began by evaluating the influence of these alternatives using a simple (approximative) simulation of SPECT images as depicted in Fig. 7.4. Here, we first segmented the gray matter in the MR image of one subject (using thresholding). The voxels belonging to this class were all set to 100 and subjected to a large Gaussian filter. This yielded the bottom left image in Fig. 7.4. The MR image was then transformed using a deformable registration and another SPECT image was simulated in the same manner as the first, yielding a ground truth image (bottom right image in Fig. 7.4). The first simulated SPECT image (bottom left image) was transformed using the different approaches listed above. These transformed images were then compared to the ground truth image. The comparison was done visually using an animated display. We found that the third approach yielded the best results, but we did notice that some residual difference remained after transformation. The evaluation of deformable registration techniques is a difficult subject [226], the ultimate criterion being how well a group of subjects can actually be modeled after registration. This is done in the next chapter where we evaluate the above approaches to inter-subject SPECT registration using a cross-validation scheme. This evaluation confirmed what we found visually: high resolution registration of MR images followed by a filtering of the field yielded the best results and explains step 6 in Fig. 7.1.

Whereas this is a gross simulation of a SPECT image (similar to the way PET images were simulated in [158]), there exist other ways of simulating SPECT images, either from phantoms

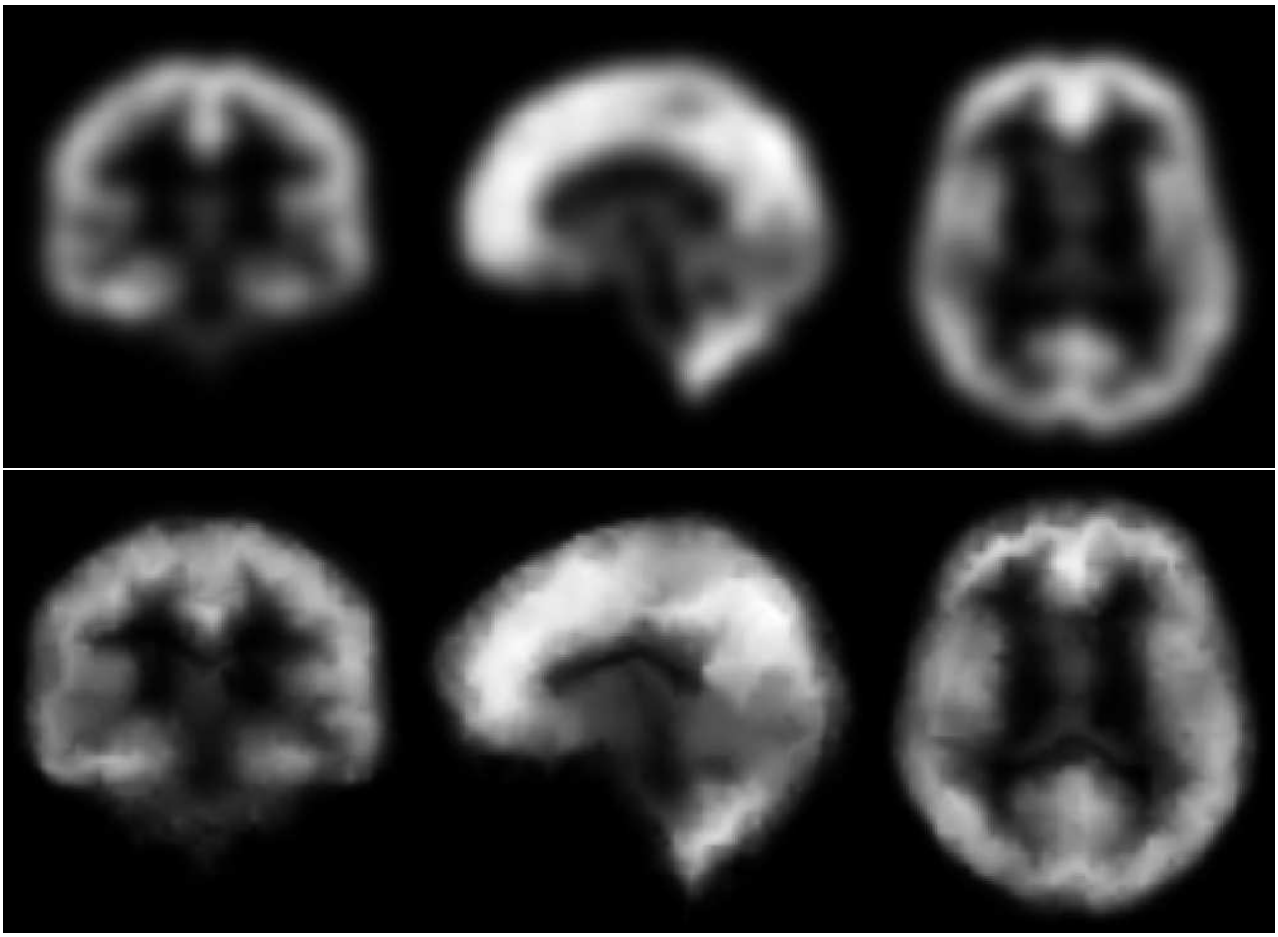


Figure 7.3: SPECT images in ICBM space. The top image was transformed using an affine registration (without the deformable registration in Fig. 7.1), the image on the bottom using a deformable transformation (without filtering in Fig. 7.1). The deformable registration is calculated from high-resolution MR images and the direct application of the resulting transformation introduces artificial, high frequency components into the image. These result from the matching of structures of finer resolution than the SPECT image. These images were transformed to 1 mm resolution for visualization (256^3 voxel images), but the same effect persisted at lower resolutions as well.

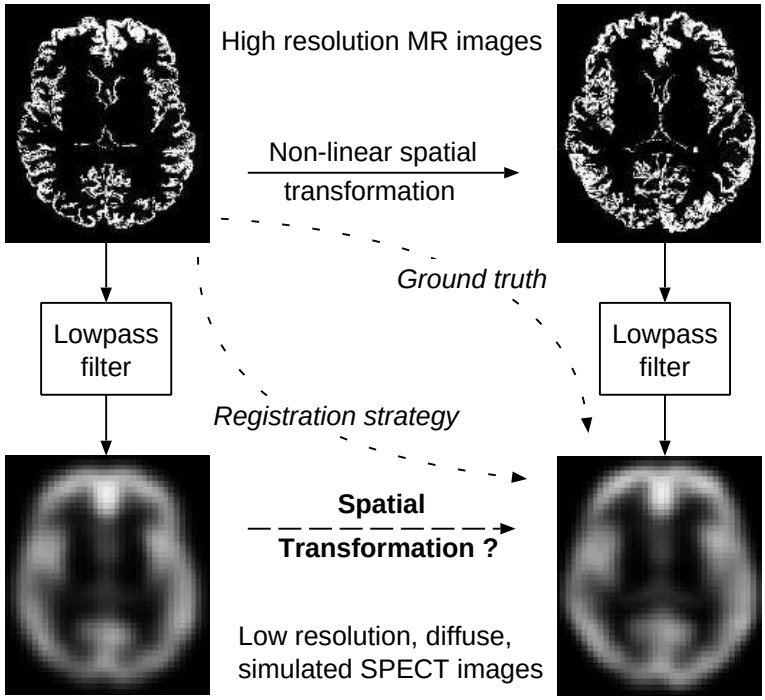


Figure 7.4: Images similar to SPECT images were created by low-pass filtering the gray matter of a high-resolution MRI image. This way we could obtain a ground-truth transformed SPECT image from the transformed MR image (bottom right image) to which we could compare differently transformed SPECT images (transforming the bottom left image). See the text for a description of the different registration schemes that were compared.

[130, 128] or from MR images [84]. This simplified simulation sufficed however to get an idea of the influence of the transformation on the SPECT images.

7.4 SPECT brain segmentation

Only the brain is considered for statistical analysis and for intensity normalization. This necessitates the determination of a brain mask. For this we use Otsu-thresholding [178] to divide the image histogram into three classes (corresponding to background, brain and scalp). Otsu thresholding is a histogram thresholding technique similar to k-means clustering. The gray value distribution of each class is modeled as a Gaussian whose parameters are found by maximizing the inter-class to intra-class variance. The thresholds are then selected so that the Bayes risk is minimized. The highest threshold determined the brain mask (because white and gray matter have the highest signal in SPECT). Otsu-thresholding is not always reliable for SPECT images as it sometimes leads to an overestimation of the brain. In these cases a manual threshold was set.

Typically there remain some "holes" in the brain mask that correspond to the ventricles. These need to be "filled". This is done by using connected component labeling on the holes and the exterior of the brain (inverse of the brain mask). Clusters of voxels not connected to the exterior are then added to the brain mask. The connected component labeling is done successively in the coronal-plane, the axial-plane and finally in the sagittal-plane. This approach was chosen because the holes were too large to be filled with morphological closing.

The resulting brain mask is sometimes slightly larger than the brain (visual assessment), leading to the modeling of voxels that do not belong to the brain. When analyzing images with the atlas we thus expect to have some false alarms on the border of the brain.

7.5 Intensity normalization by total least squares

As mentioned in the discussion on intensity normalization techniques (Sec. 6.6), the evaluation study by Pérault et al. [185] judged that a two-parameter regression function (scaling plus constant) was not necessary (scaling was sufficient). However, the fit to the scatterplot is better when adding this constant to the model (Fig. 7.5). They argue that a more complex model might be more error prone. Still, we propose to use a two-parameter function because, (1) the fit is better, (2) when the images are in good registration, the joint histogram will be quite linear and (3) we use a robust total least squares regression to estimate the function parameters. In [185] the histogram is linearly fitted using standard least squares. This means the x-values are considered to be perfectly measured - without error. Here, x are the intensity values in the reference image and y the corresponding values in the source image. If we assume the noise to be the same in both images, a more natural choice is to minimize the sum of square errors orthogonal to the regression line rather than those parallel to the y-axis. This is explained in Fig. 7.6. and leads to total-least-squares regression [51]. Total least squares is simply done by calculating the largest eigenvector on the joint histogram. Examples of fitted histograms are shown in Fig. 7.5.

7.6 Statistical models considered

The statistical models that have been considered in this work are based on the linear model in Eq. 4.1 and have been summarized in Tab. 4.1, Sec 4.7. These were all implemented and evaluated using the evaluation scheme described in the next chapter. However, their practical differences being small, we only present the results of three representative models in order to be more conclusive.

Let us recall the linear model Eq. 4.1 from Ch. 4:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}.$$

The three models, for which we present results, can then be summarized as follows:

- A univariate Gaussian model - "local" model - which is obtained by setting $\boldsymbol{W} = \boldsymbol{0}$. The noise is given by $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon)$ with $\boldsymbol{\Sigma}_\epsilon$ being diagonal.

- A multivariate Gaussian model - "global" model - which is the WML model (second row of Tab. 4.1). The noise is the same as for the local model and the subspace variable is considered to be uniformly distributed in the subspace. This model is equivalent to the ML model (first row of Tab. 4.1), also proposed by Houston *et al.* [104] and presented in Sec. 6.3.3.

- A multivariate robust, non-Gaussian model - "robust global" model - which is the original RMMS model presented in Ch. 4 (second to last row of Tab. 4.1). The noise is no longer Gaussian and the subspace variable is estimated from the training samples using a kernel density estimator.

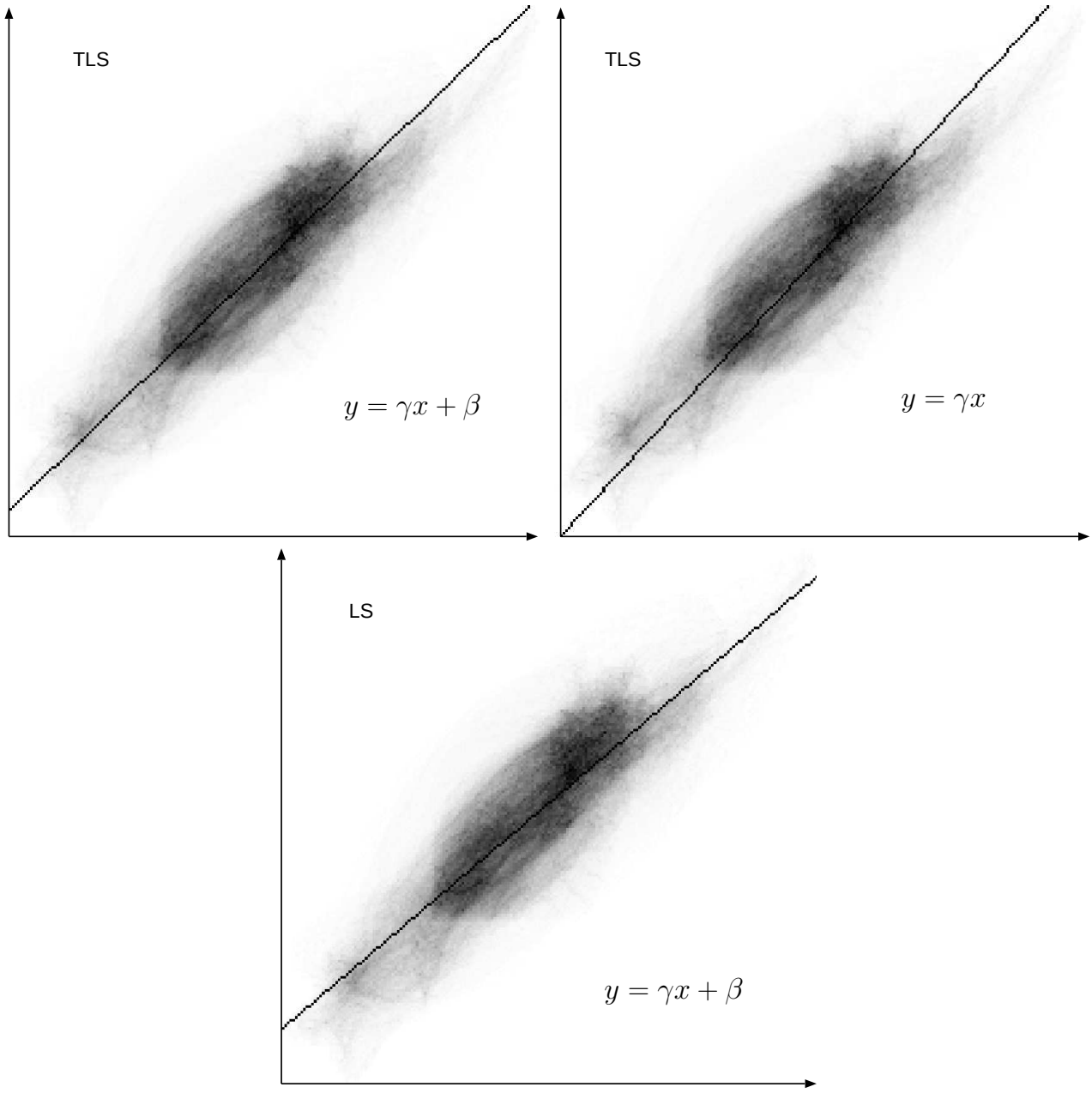


Figure 7.5: Top: Fitted transfer functions using total least squares. The function on the left has two parameters, scaling and a constant, the one on the right, only scaling. The two-parameter fit is visually more pleasing than the one-parameter fit. Bottom: Fitted two-parameter transfer function using standard linear regression. The slope is clearly too low and the constant term too high.

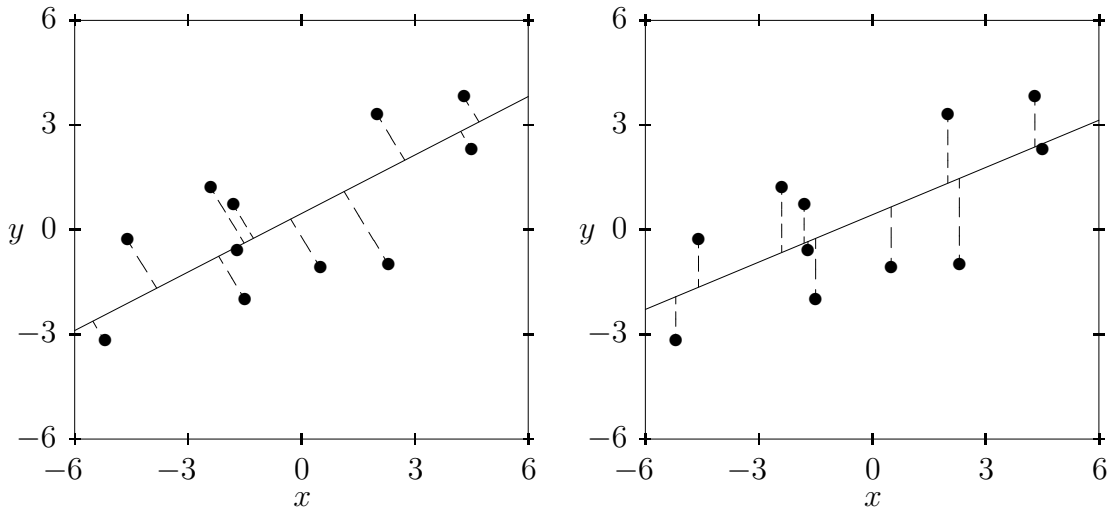


Figure 7.6: Total least squares regression (left) assumes there are errors in the measurements on the x-axis as well as on the y-axis and minimizes the euclidean distance between samples and the line to fit. Linear regression (right) minimizes the square error in y-direction.

In our experiments (Ch. 8), we observed only small differences between the isotropic and non-isotropic models, and almost no difference between the RMMS model and the RML model. This last observation is in our belief explained by the fact that the model remains quite approximate (high residuals on the learning base images). As we discussed at the end of the experimental chapter (Ch. 5), our model (RMMS) seems to improve its performance with respect to the other models when the model becomes more accurate (more eigenvectors and more learning samples). We therefore do not currently observe an improvement of using our model over the RML model in the application of a brain atlas, but we believe that such an improvement could be possible in the future when more learning images are accessible. It is also important to notice that the application of a robust model is original in the context of creating a brain atlas.

7.6.1 Comparing an image with the atlas

In order to make a localized detection of abnormally perfused brain regions we make use of the image residual after reconstruction as proposed by Houston *et al.* (see Sec. 6.3.3, p. 100). This is done by calculating the correlation-corrected $z-score$ image: $\boldsymbol{t} = \boldsymbol{\Sigma}_\epsilon^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu}-\boldsymbol{W}\hat{\boldsymbol{x}})$. Atlas creation and learning (model estimation) is described in the next section.

7.7 Atlas creation: estimation of model parameters

In this section we present the results of the learning process (model parameter estimation). Every image was brought into a $d \times 1$ vector by lexicographically ordering of the voxels inside of a common brain mask. The mask was taken as the intersection of the (segmented) brains in the learning set. This was done for practical concerns (reducing the number of voxels to about 10 %, from 64^3=262 144 voxels down to 27 422 voxels) and because we do not expect the scalp to have the same statistical properties as the brain due to a different biophysiological distribution of the tracer (Sec. 2.6.2).

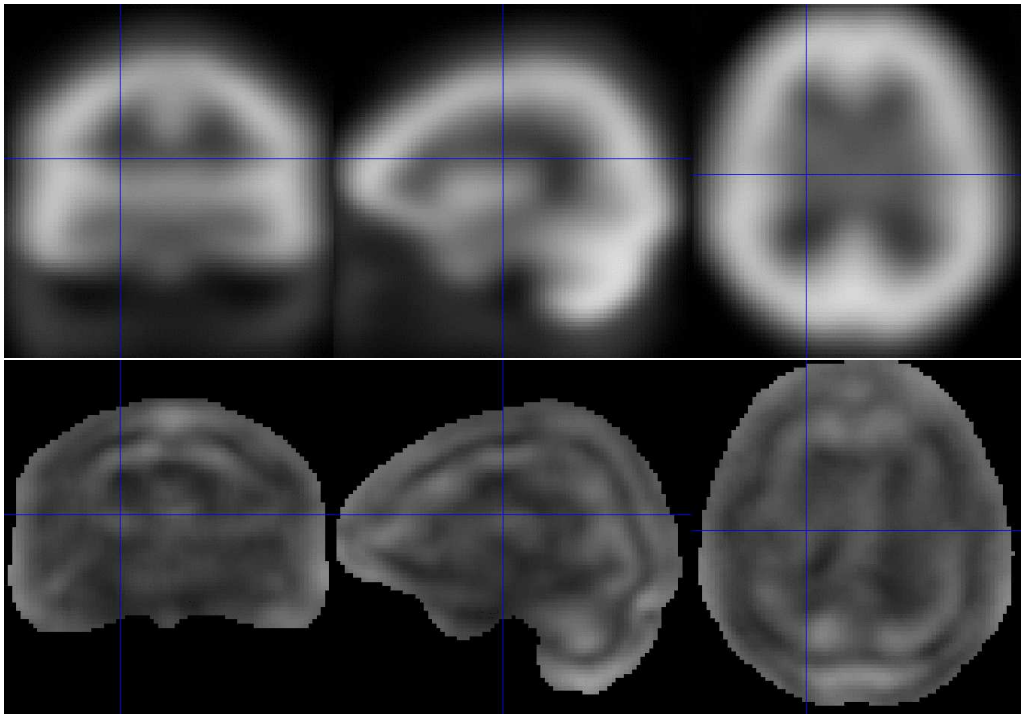


Figure 7.7: Mean and standard deviation images of the learning base (J=34). The standard deviation image was masked to enhance the contrast in image - the variance outside the brain being much larger than in the brain).

Mean and variance estimation was done using standard (unbiased) maximum likelihood estimators. Fig. 7.7 shows the mean image of the database as well as the standard deviation image. In the standard deviation image we can see high values at the inferior part of the cerebellum as well as some structure in the cortex areas. We believe that this structure comes from insufficient anatomical registration, which we have observed as "moving" structures animated sequences of the database images.

The calculations of the eigenvectors are described in App. B. Fig. 7.8 shows the eigenvalues and accumulated eigenvalues (which correspond to the relative energy located on the corresponding eigenvector/principal component). We see that most of the energy is located on the first principal component (which is shown in Fig. 7.9). Furthermore there is a steady (rather linear) decline in energy contribution up to eigenvalue number seven before the curve flattens out for higher eigenvalues. The four first eigenvectors are shown in Fig. 7.9. In the first principal component, we find highly negative values in the inferior part of the cerebellum, which is consistent with the variance image. The tendency to higher values in the superior part of the brain and lower values in the inferior part of the brain seen in this first component, might be a result of alignment (registration) problems in the axial direction. The interpretation of these images nevertheless remains difficult. The residual that remains after removing the cross-correlation of the first four principal components is shown in Fig. 7.10. We see that the distribution of residuals is more homogeneous, and as expected, the variance is lower, see Fig. 7.11.

In Fig. 7.12, we have shown two-dimensional projections of the learning base images from a five-dimensional subspace. It is difficult to say whether these datapoints are Gaussian distributed or not because of their sparsity. In some projections we can see outliers, but a

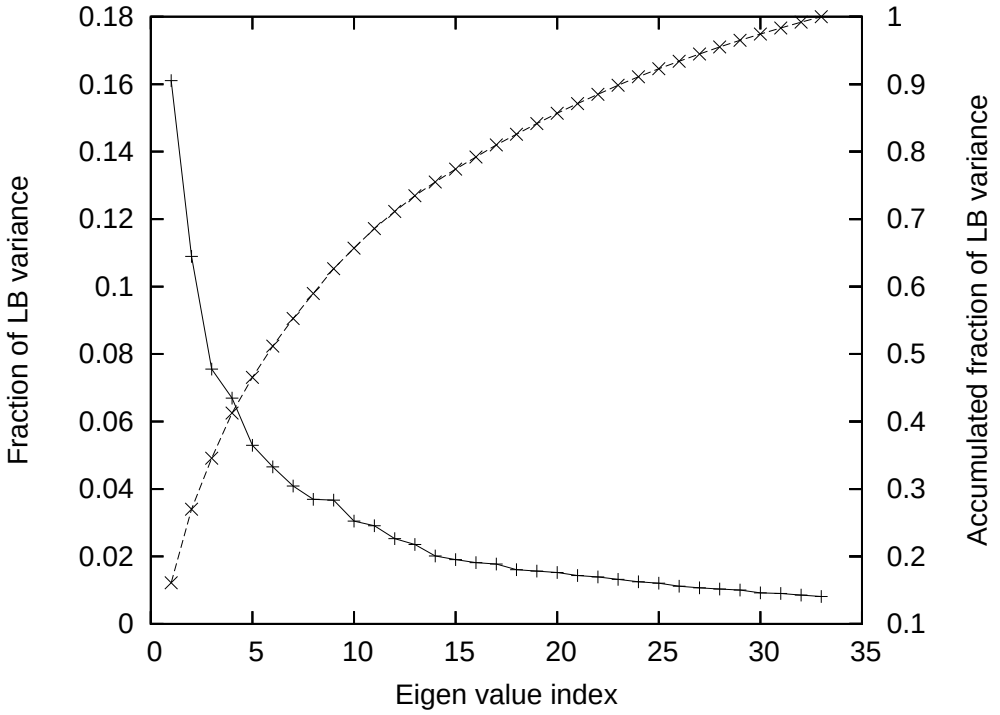


Figure 7.8: Distribution of eigenvalues of the learning base (J=34).

labeled analysis showed that none of the points (image projections) were clearly outsiders across all dimensions. For example in the plot on the top right (projections onto eigenvector one and four) we can see a point which is apart from the other points. However, this point was found in the middle of the cloud of points in other projections. We also explored the projection pursuit facilities of the data visualization tool *ggobi*[4], but we could not make out any particular (simple) structures in these projections.

7.8 Conclusion

In this chapter we have described the atlas creation process including the database of subjects, the different image processing steps, the models considered and the learning process itself. The choices made in this process were supported mainly by theoretical arguments (normalization and models) and limited simulations (registration, Fig. 7.4). In the next chapter we present an evaluation scheme that was developed to assess the chain of processing in general and to evaluate the models in particular.

[4] `http://www.ggobi.org/`

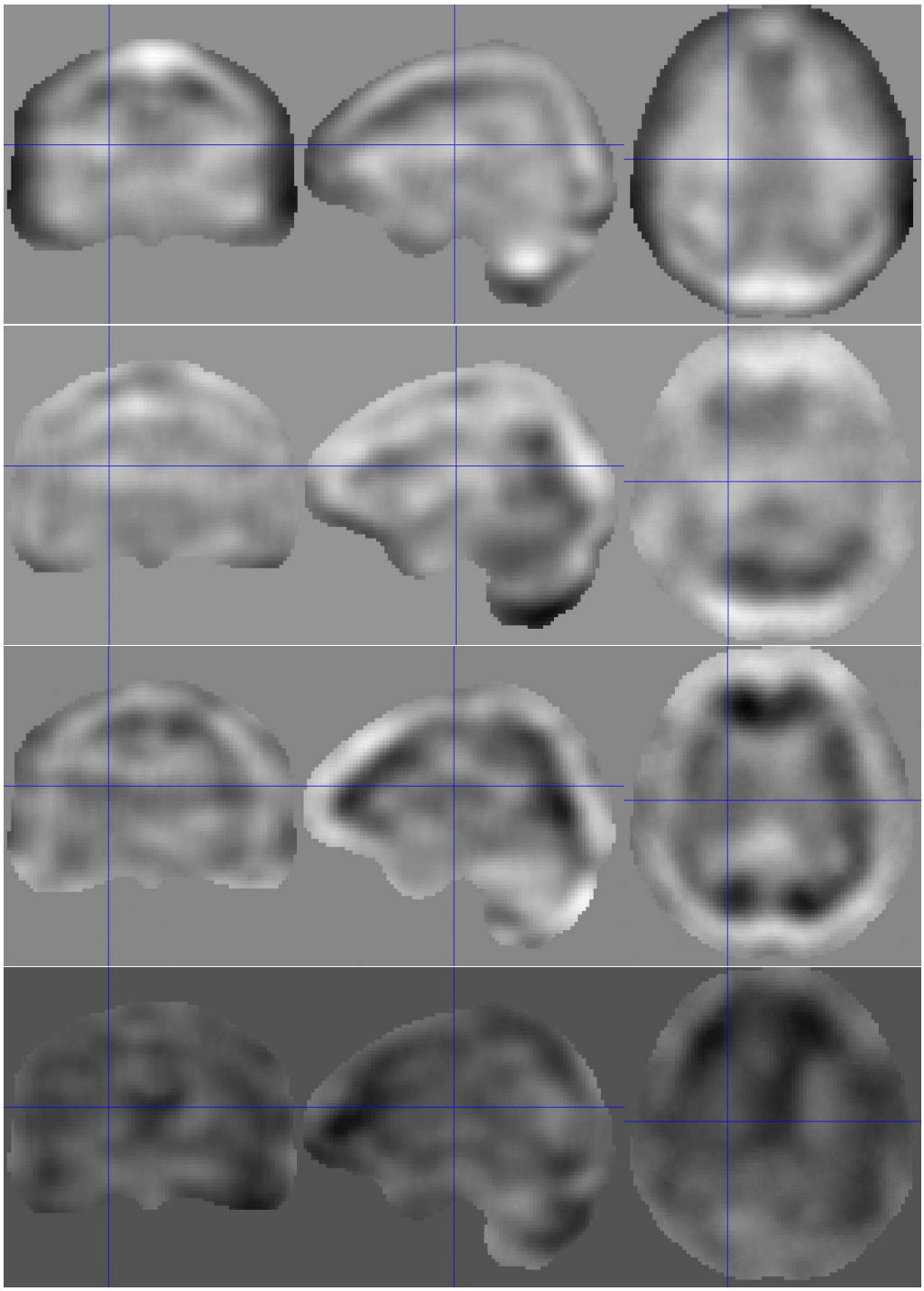


Figure 7.9: First four principal components of the learning base (J=34). A gray value display is used where the background graylevel intensity corresponds to the value zero, the minimum and maximum values of the image are mapped to the minimum and maximum values of the display for maximum display dynamic.

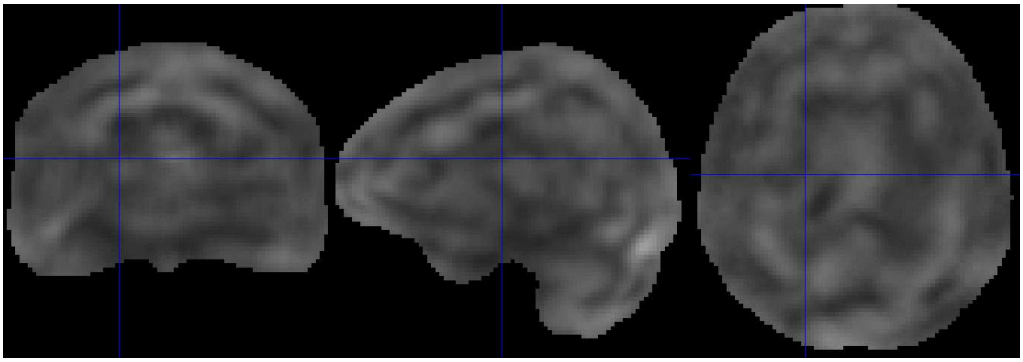


Figure 7.10: Learning base residual after removing four principal components. This is the square root of the diagonal of $\mathbf{\Sigma}_\epsilon$ in Eq. 6.1.

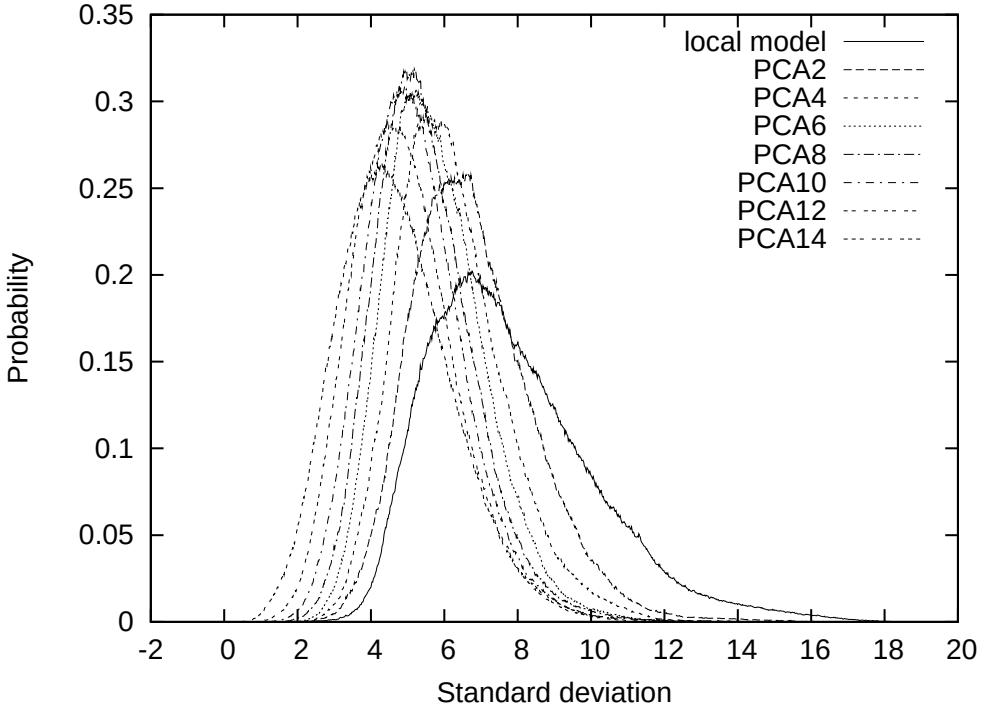


Figure 7.11: Histogram of the learning base residual images with increasing number of eigenvectors, starting with the local model (Fig. 7.7), the mean of curves move to the left with increasing number of principal components. The histogram of Fig. 7.10 is the third curve from the right (PCA4). Note that the residual for the global model was calculated using the reduced number of degrees-of-freedom as in Eq. 6.1. These curves do however hide some of the structure in the images: in the learning base residual of the model PCA14 for example, we found regions of higher variance than in the model PCA1 even though the average variance has dropped.

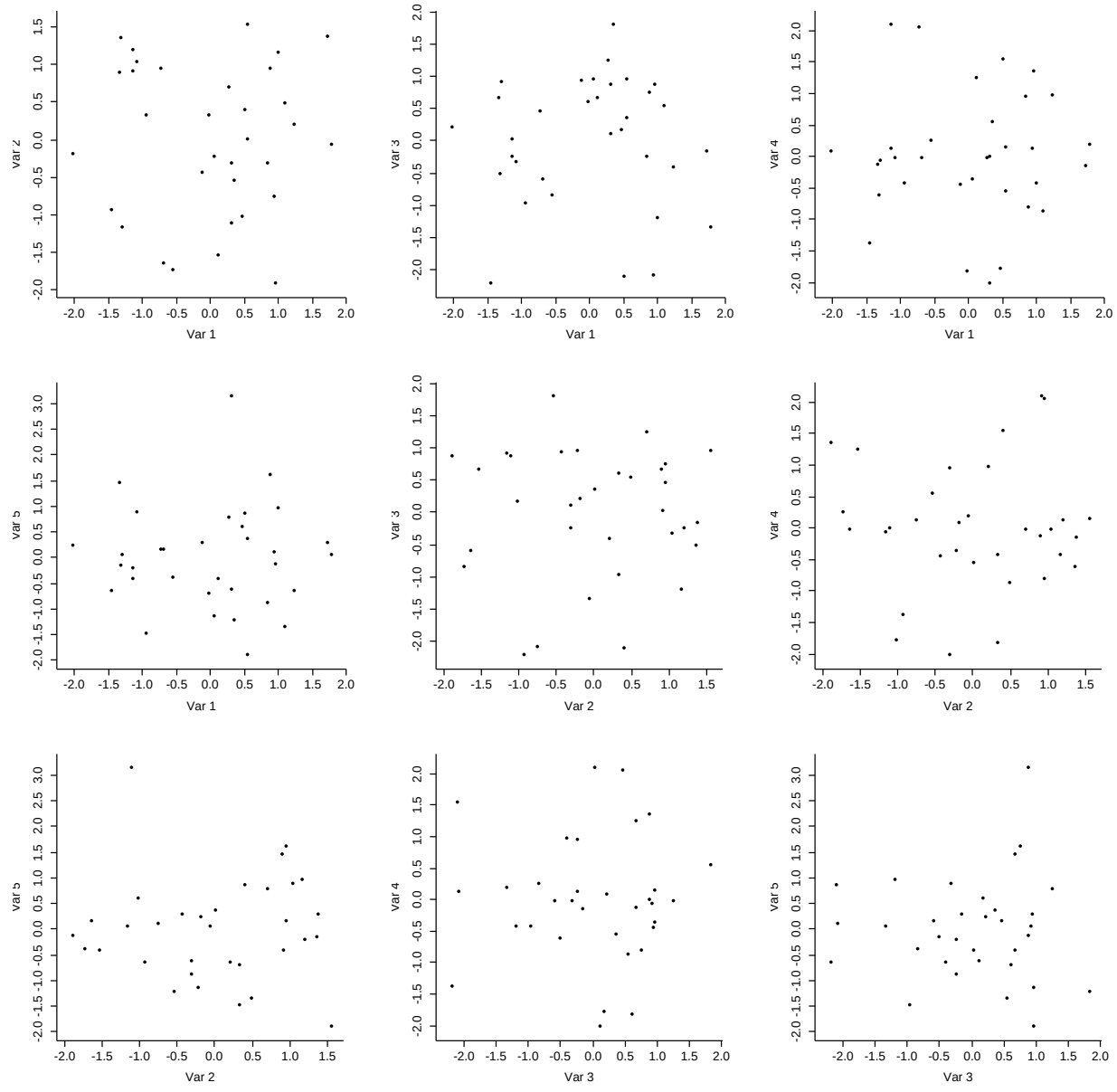


Figure 7.12: Learning base images in subspace. Projections onto two-dimensions for visualization of the learning base distribution in the space spanned by the first five principal components (denoted Var q, $q \in [1, 5]$ in this figure).

Chapter 8

Atlas evaluation

One of the main difficulties linked to the creation of a brain perfusion atlas is its assessment, i.e. *how well does the atlas describe normality and how sensitive is it at detecting abnormalities?* In this chapter we present an evaluation study based on a leave-one-out strategy combined with simulated abnormalities. This evaluation aims at making statements about both the model-fit and the detection performance of the different models considered in this work. Furthermore, it provides a powerful way of evaluating and comparing different registration schemes. With this scheme we could show that the registration scheme presented in Sec. 7.3 indeed yielded the best results. Furthermore, a slightly better performance of the standard PCA model ("global" model) over the univariate Gaussian model ("local" model) was found for small abnormalities. A superior performance of the robust PCA models (with and without subspace modeling) over the other models for large abnormalities was also found. However, there were no differences between the performances of the robust PCA model and the robust PCA model with non-parametric subspace modeling. See also [219] and [221].

8.1 The need for validation

The importance and potential benefits of validation can be seen in Fig. 1.2, p. 5. Validation can highlight intrinsic characteristics of a method, define its limits, and clarify the clinical potential of a method. Furthermore, it can help the development of better methodology by clearly defining requirements and improve the understanding of the problem domain (from a technological viewpoint). As Russell H. Taylor[1] pointed out at a recent talk in Strasbourg, industry has a lot of competence in validating methods and systems. Such competence is also necessary and useful in academia. Validation of new methods will determine whether such methods can find an application or not. A recent editorial in IEEE Transactions on Medical Imaging [112] highlighted this issue. Here, the authors brought validation of medical image analysis/synthesis research into a broader perspective. Validation is needed at several levels and there is a need to develop standard terminology, as well as common, rigorous methodologies. We shall not continue the discussion on standardization, but focus on difficulties related to the validation of the image processing methods we have used. These difficulties motivated the particular evaluation strategy that we have adopted. We continue by presenting other relevant evaluation studies before we present our study together with the results we have obtained.

[1]Professor at the John Hopkins University and director of the Computer Integrated Surgical Systems and Technology research center, Maryland, USA.

8.2 The difficulty of validation

The approval of methods for clinical use is in most countries performed by randomized and blinded multicenter clinical trials. This is costly and time consuming so that for journal publications, researchers typically show the feasability of a method or perform comparative studies. Further difficulties are linked to the complexity of a system or an algorithm to be validated (e.g. registration). This complexity can for example manifest itself as many possible parameter settings. If such parameters are correlated, an exhaustive validation across all possible parameter settings may be necessary, but impossible in practice. However, the main problem for the validation of medical image processing techniques, is the uncertainty about what is correct (and not correct). This is the problem of a missing "ground truth" or "gold standard". For example, for a brain perfusion atlas, we acquire images of subjects we assume to be normal, and who we ask to refrain from performing any particular brain task during injection. Such statements can only be true to a certain degree. Further complications for the validation, evaluation and comparison of methods are the lack of standardized databases, algorithms, software, methodology and terminology (see also [112]).

As another example let us consider the validation of registration, which is central to the atlas creation process. We mentioned in Sec. 6.4.3 that the validation of registration methods is extremely difficult. This is because we do not know the true parameters of a good registration nor if the deformation model is actually valid. As discussed in [226] a homology between different subjects may not exist and transformations can be overparametrized. Visual assessment of registration methods by superimpositioning of the registered images is not sufficient to conclude on the validity of the transformation. Simulated deformations may not be a valid assessment criterion since they are necessarily model-based (however useful in the development of the method). The authors also evoke the difficulty of defining a validation metric, even when the ground truth is known, i.e. quantifying a good registration. One of the most complete comparison/evaluation studies of inter-subject, MRI registration algorithms is the study in [94], where the authors have used several similarity measures to compare the registration quality. They do however state that the "ideal" transformation depends on the application.

8.2.1 Validating model hypotheses

Statistical tests of goodness-of-fit

Validating a model means that we want to assess a hypothesis about the model. As emphasized in [112], such hypotheses should be stated clearly and preferably be standardized (e.g. to facilitate meta-analysis). An example of such an hypothesis could in our case be: *The count values in brain SPECT images are normally distributed according to the linear model Eq. 4.1 when anatomical structures are superimposed and when global differences due to variations in the injected dose are removed.* Statistical goodness-of-fit tests have indeed been developed to assess this kind of hypothesis. We could apply multivariate tests of normality [148], but the sample size ($J = 20$) is too small to yield meaningful results [72]. Another possibility would be to test marginal probabilities by applying a Kolmogorov-Smirnov or a Lilliefors test at each voxel. This is also problematic because one would need to interpret the results (e.g. is it reasonable to obtain rejections in 10 % of all voxels, given the spatial correlations in the data?), the sample size is still small, and finally, we expect a relatively simple model like this to be rather approximate (we would therefore want to know "how approximate" the model

is). Furthermore, in order to know anything about the power of the goodness-of-fit test, one would need to formulate an alternative hypothesis (the power of a test is $1 - \beta$, where β is the probability of false negatives). Formulating a specific alternative hypothesis is difficult. Another formulation for an hypothesis could therefore be more tuned to the application of the model: *The model [...above...] is accurate enough to detect abnormal lesions of a specific type.* However, here we encounter the problem of modeling the distribution of a pathology (This is generally accepted to be a more difficult problem than to model normality, see also the discussion of generic modeling in Sec. 3.2.5).

Empirical tests of goodness-of-fit: generalization

The above considerations have led us to abandon statistical goodness-of-fit tests, mainly because of the small sample size. At the other extreme, as pointed out in [194, p.143], almost any goodness-of-fit test will fail for sample sizes that are large enough when the hypothesized distribution is only "almost true". In pattern recognition, it is rather usual to consider a distribution as approximate[2], and a more practical view on the goodness-of-fit for a model is taken: *generalization*. Generalization is a measure of how well a classifier performs on unseen data and is also useful for examining the performance of non-statistical methods. Such an evaluation is done by splitting data into training data (used for training or parameter estimation) and test data (that constitutes the unseen data). In situations where only small data sets are available, as is our case, one typically ressorts to a leave-one-out strategy. This is what we have done as described in Sec. 8.5.

8.2.2 Evaluating the model and the influence of preprocessing algorithms

It is clear that the image processing steps taken to prepare the images for statistical modeling will have an influence on its accuracy and validity. Thus, having found a way of evaluating the statistical model, one is intuitively taken to examine this influence. This is attractive because, as we have discussed in Sec. 8.2, a complex algorithm like deformable registration is difficult to validate. If one considers that the atlas model is (approximately) valid, one expects a "good" registration to yield a "good" atlas. We therefore suggest that the atlas performance is a powerful criterion for comparing different registration methods (with respect to the application of atlas or group analysis).

Let us discuss this conjecture. First, recall that the atlas describes variability in a population *after* image registration. The variation observed in a population is often considered to stem from two *different* and *independent* sources, namely (1) anatomical variability and (2) functional variability. It is likely that the two sources of variability add up rather than compensate each other. When we say an atlas is "good", we mean that the variability of the population is described with high fidelity. This fidelity is obtained by an accurate model. We believe that a good registration will improve the accuracy of the model because (1) it reduces the anatomical variability, and (2) the anatomical variability is highly non-linear, violating our linear model hypothesis. It is furthermore often hypothesized that the localization of basic brain functioning is strongly correlated with the anatomical structure of the brain (this is

[2]The Irish statistican Roy Geary (1896 - 1983) is attributed the following phrase [148, p. 279]: '"Normality is a myth; there never was, and never will be a normal distribution". This is an overstatement from the practical point of view, but it represents a safer united mental attitude...'

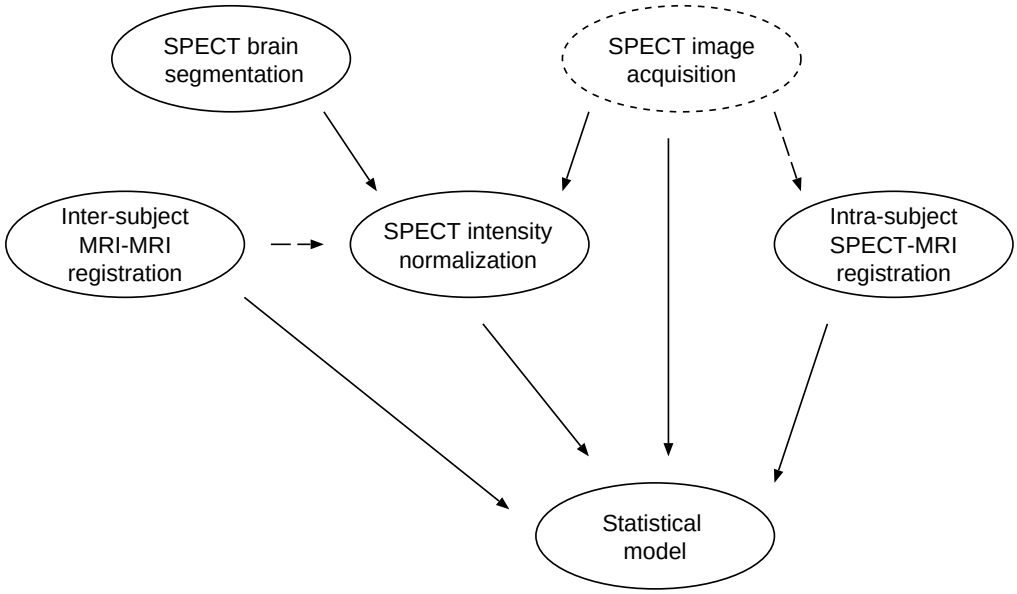


Figure 8.1: Independence diagram for different image processing units and the statistical model. The solid lines show the main influence paths between units, and dashed arrows show less important influence paths. Image acquisition and reconstruction parameters (attenuation correction, scatter correction, etc.) certainly have a significant influence on inter-subject registration, normalization and the statistical model, but are out of the scope of this thesis. Furthermore, we found that the SPECT brain segmentation had a small effect on the calculation of principal components. However, this effect was negligible for a range of visually acceptable segmentations. Inter-subject, MRI registration has an influence on the joint histogram that is used for intensity normalization (but is less important for other normalization methods). Given the simple (linear) transfer function used for normalization, we considered this influence also to be negligible. This leaves a graph where the registration algorithms and intensity normalization can be considered to be independent (given the SPECT images).

probably not so for higher cognitive brain functioning). Removing anatomical variability will therefore leave, in theory, only functional variability which is exactly what we seek to study.

In order to evaluate the influence of the parameters of a method (preprocessing algorithms in this case), it is important to understand their inter-relations and mutual dependencies. A powerful and convenient way of doing so is to consider an independence diagram[3], an example of which is shown in Fig. 8.1. We see that the different units influence each other, but we can consider the registration and intensity normalization to be relatively independent parameters. This justifies the examination of the influence of, for example, normalization under a fixed registration.

[3]Independence diagrams are a graphical way of expressing the conditional independence relationships among a set of random variables (or abstractions of these) [164, 155]. Such diagrams thus permit one to easily read conditional independences.

8.3 Evaluation studies based on simulations

Other simulation studies for assessing methods that characterize cerebral lesions in SPECT and PET images have been described. In [197], the authors decrease and increase the perfusion on a sphere in the right frontal lobe of the mean image to evaluate the capacity of SPM (statistical parametric mapping) to detect changes. We think this approach yields an optimistic estimate because an abnormality can vary in an additive manner around the *variation* of normal images, not only around their mean. Furthermore, the study is limited to only one location of the abnormality whereas we show that the detection sensitivity of the atlas is location-dependent. Another study [128] adds inclusions to a software phantom, thereby simulating a single-subject activation study, whereas we are more concerned with multiple individuals with "activations" in out-of-group subjects. In [158], the authors compare SPM and the computerized brain atlas (CBA) [83] for PET activation studies using both human volunteers and simulations. The synthesized images are also derived from a single (simulated) PET image. Our context is somewhat different from a standard activation study with multiple conditions/multiple subjects: we have several control images in rest state (learning set) and only one activation image of a subject that is not represented in the learning set. Another study also compares the effects of different registration and filtering algorithms on the detection capacity of SPM by simulating PET images [50]. These are simulated from the MR images, but this time from 16 different persons. This way natural anatomical variance is present in the database, but the functional variance still lacks.

All of these studies are biased since the resting state/inactive images of the person(s) being studied are *present* in the learning base. This is of course the way most activation studies are designed (especially in cognitive tasks), but such images are not available in real-world atlas-patient comparisons. Because of the cross-validation design of the proposed evaluation study, it is not biased in this manner. The influence of using an unbiased evaluation strategy is further discussed in Sec. 8.6.6.

8.4 Evaluation in the absence of a ground truth

Recently a class of methods have caught the attention of researchers working on the development and validation of new medical imaging technologies [101, 125, 114]. These are called *latent class analysis* methods (or are related to these). Latent class analysis is interesting because it allows us to compare different methods in the absence of a "gold standard" or "ground truth". In medical imaging this is often the case either because it is very expensive, invasive or impossible to obtain a gold standard.

In a simple binary detection scenario (pathology or not), the quality of a method is characterized by its sensitivity, specificity and the prevalence (probability of pathology). The basic idea is that applying several tests (methods) on several populations and matching the results, yields enough degrees of freedom to estimate the sensitivity and specificity of each method.

In the more complex cases of estimating covariates (continuous variables, e.g. regional cerebral blood flow), models similar to the factor analysis model (and the models used in this work) have been used [125, 114]. Here again, the models can be compared indirectly, that is without knowing the ground truth. We have not had the opportunity to study any of these methods in this work. To our knowledge there have not been any studies that have used such methods to compare voxel-based statistical atlases yet. We believe though, that latent class analysis will become more important to the image processing community in the future.

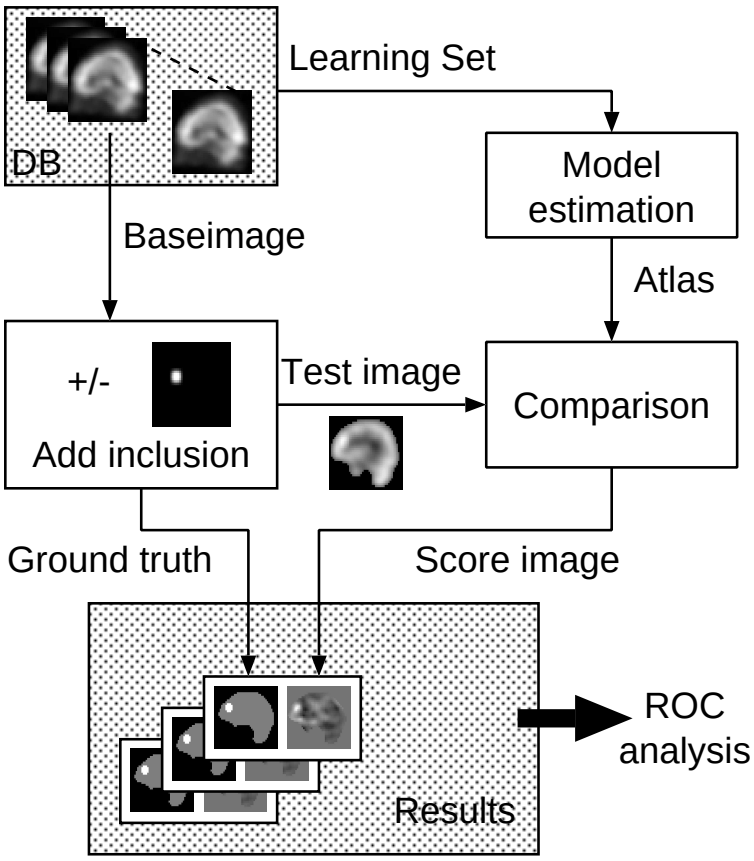


Figure 8.2: Schematic view of the leave-one-out strategy employed. At every iteration a *baseimage* is removed from the database (DB) from which a *test image* is synthesized by introducing inclusions in the baseimage (see Fig. 8.3). The atlas is constructed from the remaining images and the comparison yields the *score image*. A good atlas describes the *baseimage* well and is sensitive enough to detect the inclusions. The procedure is repeated for all images in the DB and the results are averaged to measure the overall performance. We used inclusions at six different locations, of two sizes and six intensities, so that we obtained a total set of 72 score images for every baseimage. The database contained 34 images.

8.5 Proposed evaluation scheme

An overview of the validation procedure can be seen in Fig. 8.2. The different processing steps will now be outlined before we describe the evaluation criterion.

Leave-one-out

The leave-one-out methodology is a standard method in pattern classification used to estimate the generalizing power of a method when only a small dataset is available [58]. Given a dataset of size J, our leave-one-out scheme simply consists of iterating the procedure depicted in Fig. 8.2: remove an image from the learning set, estimate the model, create a *test image*, compare this image to the estimated model. A *score image* and its corresponding *ground truth* is thus created for each image of the J dataset images. Using ROC analysis (receiver operating characteristics), the performance is then estimated by averaging the results of all these images. It is a rather time consuming technique, but the only one appropriate when data acquisition

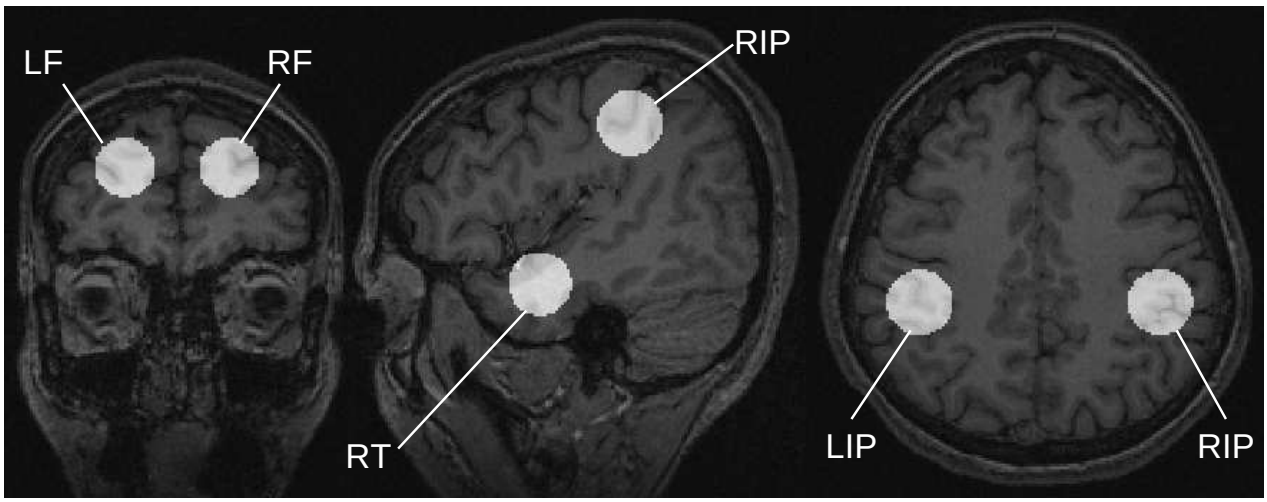


Figure 8.3: All inclusions of 20 mm diameter size superimposed on the database reference MR image (Fig. 7.1). SPECT test images were constructed with only one inclusion at a time in the leave-one-out strategy (Fig. 8.2). See Tab. 8.1 for legend.

Table 8.1: Legend of locations

Accronym	Description
LF	Left middle/superior frontopolar gyri
RF	Right middle/superior frontopolar gyri
LT	Left middle temporal gyrus
RT	Right middle temporal gyrus
LIP	Left intraparietal sulcus
RIP	Right intraparietal sulcus

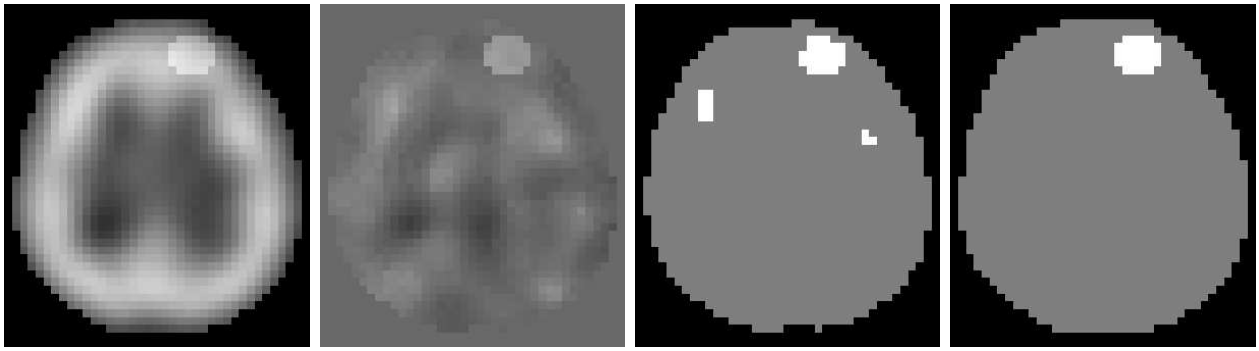
is expensive.

Synthesized images

Images with abnormal perfusion (*test images*) were synthesized by adding or subtracting *inclusions* to a *baseimage*. An inclusion is a sphere with a fixed intensity. Intensities of $\pm 15\,\%$, $\pm 25\,\%$, $\pm 35\,\%$ of average brain perfusion were used and several positions were examined as shown in Fig. 8.3 and Tab. 8.1. Two different sizes of inclusions were used, *small* inclusions of 20 mm in diameter, and *large* inclusions of 64 mm in diameter. An example *test image* is shown in Fig. 8.4a.

Performance evaluation

Performance evaluation was done by calculating the receiver operating characteristics (ROC). For this, we define the *true detection rate* (TDR) and the *false detection rate* (FDR) (or equivalently *sensitivity* and *specificity*) as probabilities [23, 58]. The *score images* are thresholded at different levels to obtain binary images from which we calculate the TDR and the FDR on a voxel basis (Fig. 8.4). By varying the threshold level we obtain a ROC curve.



(a) (b) (c) (d)

Figure 8.4: A transversal slice of (a) a *test image* with an inclusion of 25 % of average brain perfusion in the right frontal lobe (RF), (b) the *score image*, (c) the thresholded score image, which is compared to (d) the ground truth image in order to determine the FDR and TDR rates.

The FDR and TDR are calculated as follows: For a given threshold, let $\delta_k^{(j)}$ take on the value 1 for detection (value above threshold) at voxel k in subject $j = 1, \ldots, J$ and 0 for non-detection and let I designate the 3D region defined by the inclusion. We then define the TDR at this threshold as

$$\mathrm{TDR} = \frac{\sum_{j=1}^{J} \sum_{i \in I} \delta_i^{(j)}}{\sum_{j=1}^{J} \sum_{i \in I} 1} \tag{8.1}$$

and FDR as

$$\mathrm{FDR} = \frac{\sum_{j=1}^{J} \sum_{i \in B \setminus I} \delta_i^{(j)}}{\sum_{j=1}^{J} \sum_{i \in B \setminus I} 1} \ , \tag{8.2}$$

where B is the region defined by the brain (we use the same brain mask as for PCA). It is important to note the outer sum: it is not valid to calculate ROC curves for each image and sum these, each pair of FDR/TDR must be calculated for a fixed threshold across all the images.

Since the definition of the false detection rate does not make any distinction between "good" and "bad" false alarms, one might discuss whether these ROC curves are appropriate or not. In certain pathologies, the physicist is only concerned by grossly identifying an atrophical region or just lateralize abnormal brain perfusion (qualitative evaluation). However, the advantage of such a measure is that it provides an objective result.

Significance testing of differences

In order to assess the statistical significance of the difference between two curves, we summarized the ROC curves by estimating the partial area under the curve (AUC). Only the partial AUC, denoted $\mathrm{AUC}_{0.05}$, in the range of FDR $\in [0, 0.05]$ was calculated since a FDR of 0.05 already represents a large quantity of voxels (about 1000 voxels as compared to the small inclusions that are 110 voxels). Furthermore, the standard error of the $\mathrm{AUC}_{0.05}$ estimates was estimated using the jackknife ("leave-one-out") technique as described in [58, p. 473]. This error was then used to assess the significance of the difference between two curves. For this a

z-score [88]

$$\frac{AUC_1 - AUC_2}{SE}, \tag{8.3}$$

is calculated, where SE is the standard error and the subscripts denotes the two results obtained with two different methods (model, registration) or on two different locations. The z-score is compared to a normal distribution table in order to obtain a p-value. When there is reasonable evidence for postulating an alternative hypothesis where one particular method is *better* than another (e.g. robust global better than global), a one-sided value (one tail of the normal distribution) can be used. Otherwise, when the alternative hypothesis is that the two methods are *different*, a two-sided value must be used (twice the one-sided value).

Furthermore, in Eq. 8.3, the standard error depends on what is actually compared. When two methods are compared whose results stem from different sample sets, we have that $SE = \sqrt{SE_1^2 + SE_2^2}$. This is the appropriate standard error to use when comparing for example the performance on two different locations. However, when the underlying sample sets are the same, as they are when we compare models (local, global and robust global) or preprocessing steps (registration, intensity normalization), a correlation corrected standard error can be used as explained in [89]:

$$SE = \sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE2}.$$

The justification for this is analogous to the use of the paired t-test instead of an unpaired t-test and can increase the statistical power of the test considerably: fluctuation of the ROC curves (and the accuracy index AUC) will tend to fluctuate in tandem when derived from the same sample. Unfortunatly, the correlation coefficient, r, in this equation is not derived analytically, but is only tabulated in [89]. It depends on the correlation of the true negative/positive scores obtained by the two methods, as well as the AUC of the two methods. Furthermore, it is based on the assumption of a binormal distribution of the true negative and positive scores. In our experiments, the correlation coefficient was found to lie between 0.7 and 0.8, depending somewhat on the intensity of the inclusion (since this changed the AUC considerably).

8.6 Results and discussion

If not stated otherwise, all results were obtained using the deformable registration scheme (Deform6-F) described in Sec. 7.3.3 and the total least squares intensity normalization as described in Sec. 7.5. We first compare models using small inclusions (20 mm ⌀), then using large inclusions (64 mm ⌀). We then compare registration schemes and show the overall improvement of our approach over the more common atlas approach (affine registration, average normalization and local model). Finally, we describe the location specific results. Only a representative subset of hypo-inclusions are reported. The results for hyper-inclusions yielded similar results and the conclusions are similarly valid for these.

8.6.1 Comparing models

Small inclusions

In Fig. 8.5 are depicted the ROC curves for the local model and the global model using small inclusions. The curves were obtained by averaging over all locations and with a varying number of eigenvectors for the global model. The standard errors of these curves were also estimated as described in Sec. 8.5 and the significance levels of the differences between any

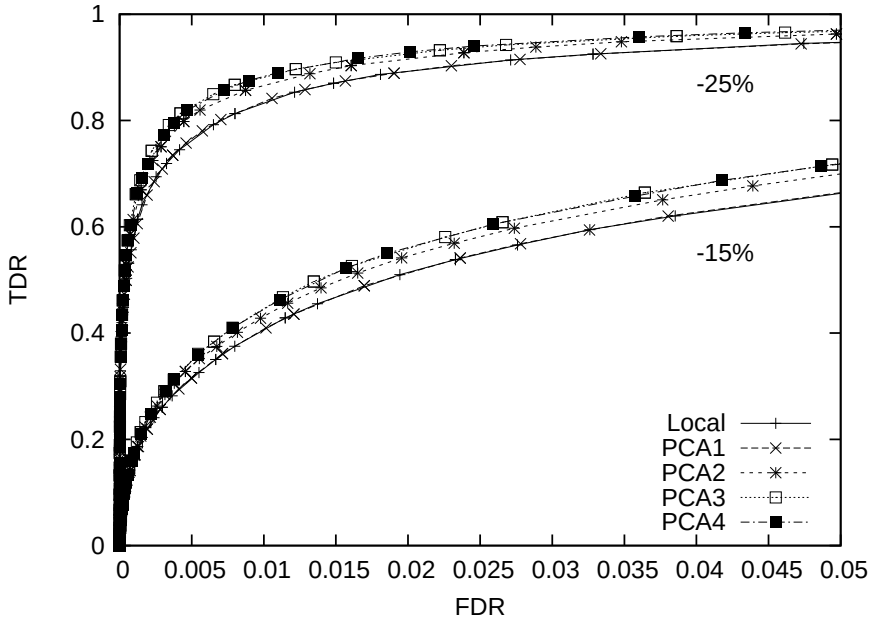


Figure 8.5: Comparison of models. *Local* is the local model, *PCAq* is the global model using q principal components. ROC curves averaged over all locations. The curves were obtained for small hypo-inclusions (20 mm ⌀) with intensities of 15 % and 25 % of average brain perfusion. Similar curves were obtained for the other intensities studied. Significance values are given in Tab. 8.2.

two curves are tabulated in Tab. 8.2 for the 25 % hypoperfusions. Here we see that the PCA2-PCA4 models are all significantly above the local and PCA1 models. The difference between the local and the PCA1 model on one hand and the difference between the PCA3 and PCA4 models are however not significant. With an odds of about 1:6 (1:7), the difference between the PCA3 (PCA4) model and the PCA2 model is not very significant.

In this paper we quantify the accuracy of different models and we show for the first time that the global model of Houston *et al.* performs significantly better than the local model as seen in Fig. 8.5 and Tab. 8.2. This is true for the model with 2, 3 and 4 eigenvectors. The appropriate choice of number of eigenvectors is however quite difficult. Too few components limit description power and too many may result in *overfitting* [58, 91] yielding eigenimages or

Model	PCA1	PCA2	PCA3	PCA4
Local	0.9972	0.0061	0.0001	0.0001
PCA1		0.0045	0.0001	4.24E-5
PCA2			0.1567	0.1398
PCA3				0.9254

Table 8.2: Two-sided p-values associated with the 25 % hypo-inclusions in Fig. 8.5. For example, one can see that the odds that the PCA3 model is different from the local model only by chance are 1 : 10000 (second row, fourth column). Since the PCA3 model shows a better ROC curve in Fig. 8.5, we therefore conclude that this model's performance is significantly better than that of the local model.

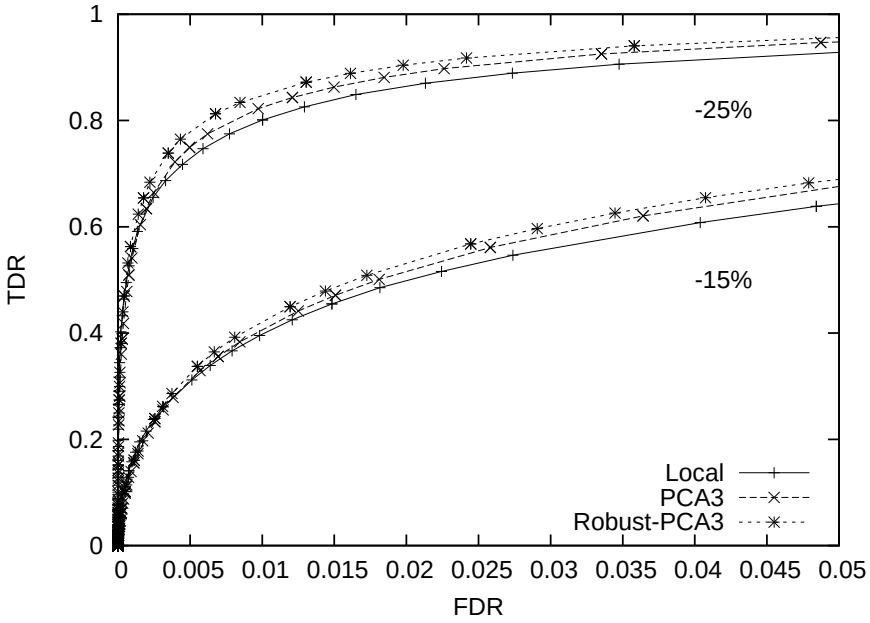


Figure 8.6: ROC curves averaged over all locations. The curves were obtained for large hypo-inclusions (64 mm ∅) with intensities of 15 % and 25 % of average brain perfusion. Note that the effective size of the inclusions were inferior to the volume of the sphere because regions outside the brain covered by the sphere was ignored.

-patterns that are specific of the sample images, *not* of the underlying distribution. Both cases lead to lower model performance. In Tab. 8.2, we see that between the models with 3 and 4 eigenvectors there is no significant difference (bottom-right entry), whereas these two are only moderately different from the PCA2 model. That covariance patterns in one or several populations carry useful information has also been shown in other studies [115, 179]. Note however, that we do not *interpret* these patterns, they merely serve as complex *descriptors* of normal variation.

Large inclusions

In Fig. 8.6 are depicted the ROC curves obtained using large inclusions. Whereas for small inclusions, the global model and the robust global model yield the same results, this is no longer the case when the size of abnormalities becomes larger. By design, the robust model (Robust-PCA3) will always be at least as good as the global model, which justifies one-sided p-values for comparing these two models. These are 0.0061 and 0.1044 for the 25 % and the 15 % hypo-perfusions respectively. The differences between the PCA3 and the local model have two-sided significance values as $p < 0.0078$ and $p < 0.0671$ for the 25 % and the 15 % hypo-perfusions respectively.

In realistic cases the zones of abnormal perfusion might be large or small. The robust model showed better results for large abnormalities than the other models (Fig. 8.6) and had equal performance for small inclusions. With increasingly good model description, due to a more representative database and better preprocessing, we expect the model to show even clearer advantages over the local and standard global model. This is observed in Fig. 8.7, which is a result from a pilot study. We synthesized test images from a baseimage which was

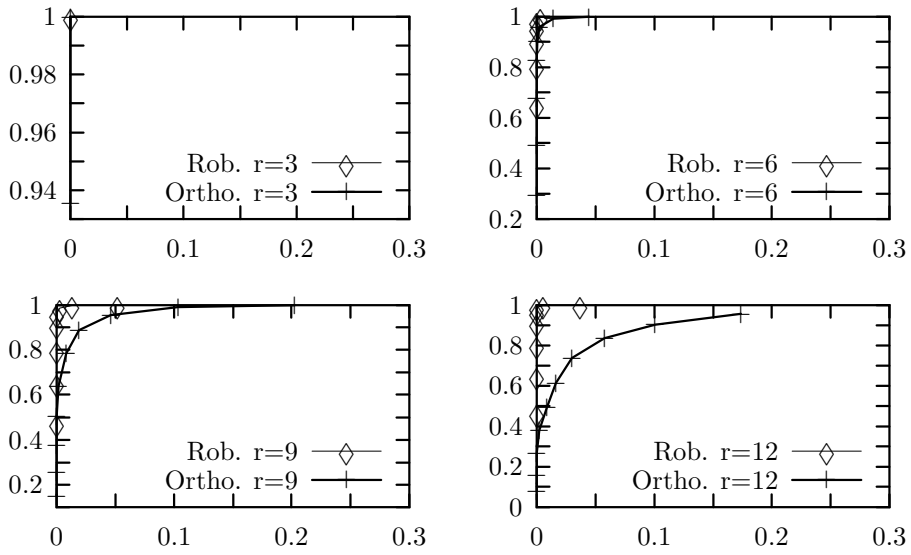


Figure 8.7: The receiver operating characteristics (ROCs) for images composed of an image from the learning base with different inclusions. Four different spherical inclusions of radius=3,6,9 and 12 voxels were added and a model with the first 12 principal components that accounts for about 93 percent of the total variance in the learning base was used. (Rob. is here the RML model, Ortho. the ML global model.)

included in the learning set. Using 12 principal components, the baseimage was thus fairly well modeled. With increasing size of the inclusions we see that the performance of the standard PCA model drops significantly. This is the same kind of improvement that we have observed in our experiments on manufactured objects, Ch. 5.

8.6.2 Comparing registration schemes

Fig. 8.8 shows the results obtained using different registration schemes. These schemes were explained in Sec. 7.3.3. All differences are significant (Deform6-F vs. Affine: $p < 0.0046$, Affine vs. Deform3 $p < 0.0033$, and Deform3 vs. Deform6-NF is negligible), again using two-sided p-values.

An influential factor on voxel-based, inter-subject studies is undeniably the spatial normalization or registration (Fig. 8.8) [225]. Since the assumption is that only similar structures should (grossly) yield comparative functional signals, these must be correctly superimposed. The leave-one-out strategy presented in this paper is clearly valid for comparing differently registered, normal images under this assumption. However, this is not true as to the algorithms capacity of registering *abnormal images* since these have been simulated after registration. The best results were obtained with the deformable registration scheme and filtering of the deformation field. In this study we have shown that a linear 12-parameter affine registration as is often used [1], is less performant than deformable co-registration. However, this approach is only possible when both SPECT and MR images of the patient are available. Finally, we notice that these conclusions are in contradiction to those found by Crivello *et al.*, where they conclude that the registration has low influence on low resolution images [44].

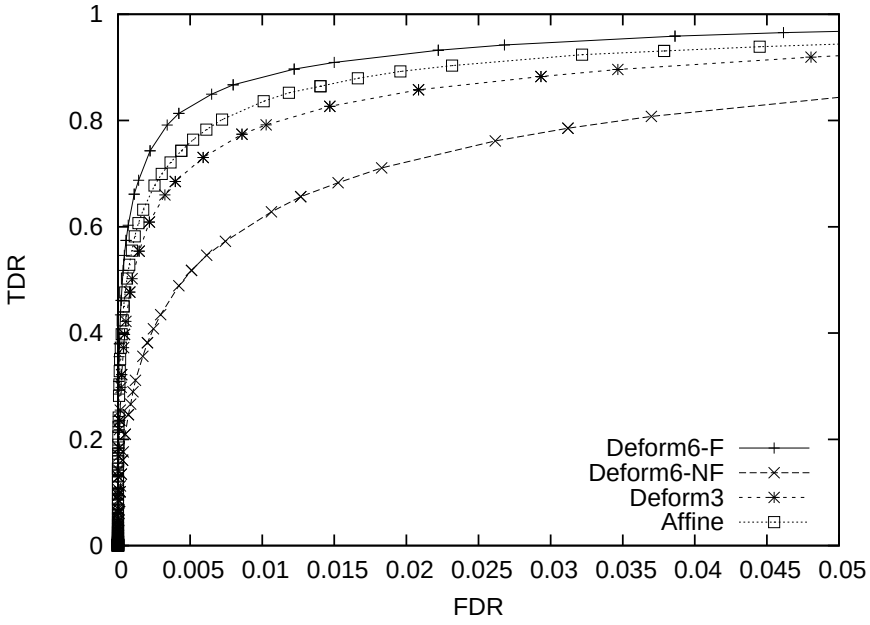


Figure 8.8: Comparison of different registration strategies for the PCA3 model. ROC curves averaged over all locations. The curves were obtained for small hypo-inclusions (20 mm ∅) with intensities of 25 % of average brain perfusion with similar curves for other intensities. See Sec. 7.3.3 for an explanation of the different strategies examined.

8.6.3 Comparing methods for intensity normalization

An example of the influence of intensity normalization is shown in Fig. 8.9. The local model and the PCA3 model are depicted in the case where the TLS normalization was used or when the global brain activity was simply normalized to have the same mean value. The difference between the two normalization techniques is highly significant for both models and all inclusion-intensities (worst case $p < 0.007$). It can be observed that the difference between the two normalization techniques for the PCA3 model is less important than the difference for the local model. This could mean that the PCA3 model is less sensitive to the intensity normalization method than the local model. However, the difference is not very significant with a two-sided p-value of 0.13. Note also that the images were renormalized after adding an inclusion.

Whereas our evaluation scheme can be used to compare different registration schemes, this is not true for comparing intensity normalization techniques. This is because intensity normalization will always lower variance, and the resulting ROC curves will show better performance. The question is whether it removes "real" variance to be studied (i.e. true differences in brain perfusion). Whether quantitative SPECT can be done using ROI methodology/normalized tissue activities without absolute blood flow measurements is still disputed, see [43] (see also Sec. 6.6.1, p. 114).

However, our evaluation scheme can be used to compare the sensitivity of different intensity normalization techniques to different types of inclusions. This can be of practical interest in order to find the most robust intensity normalization scheme for practical use. Furthermore, our scheme can also used to study interacting effects between different models and preprocessing steps (i.e. difference of a difference).

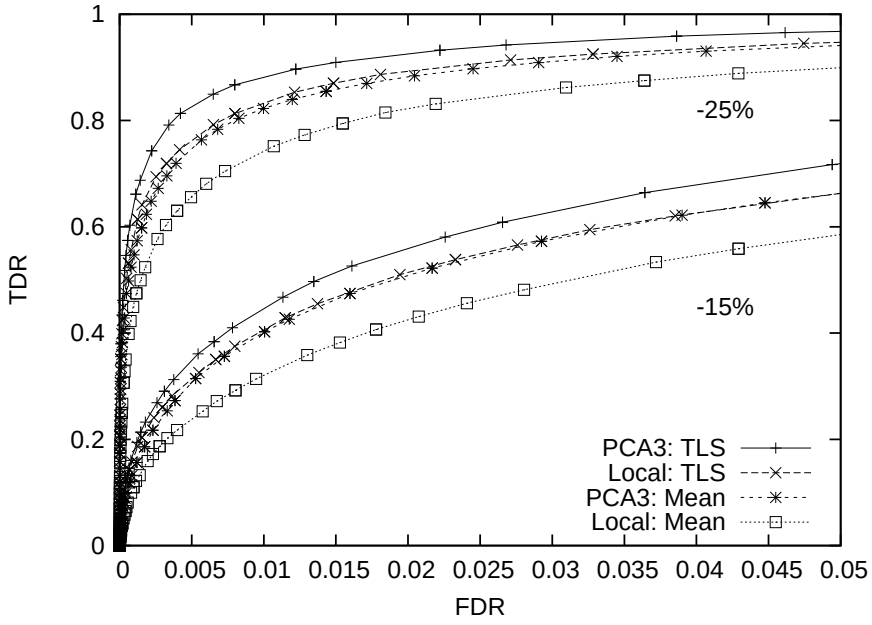


Figure 8.9: Comparison of different intensity normalization strategies (see also Sec. 7.5). ROC curves from images preprocessed with total least squares approach (TLS) and mean normalization. All locations were averaged. These curves were obtained for small hypo-inclusions (20 mm ∅) with intensities of 15 % and 25 % of average brain perfusion.

For example, we had a working hypothesis that the global models would be less sensitive to the particular intensity normalization and registration schemes employed since these models are capable of learning the associated variances from the learning images. For intensity normalization, this hypothesis is justified if the difference in AUC_z for two types of intensity normalization is significantly smaller for the PCA3 model than for the local model. In this example, for small 25 % inclusions, we found that this difference in difference has a one-sided p-value of 0.007 which supports our hypothesis. However, for the same difference (PCA3 vs. local) in registration differences (affine vs. deform6-F), the one-sided p-value of 0.15 is not significant. We will consider more intensity normalization schemes in future work in order to determine the optimal atlasing technique in terms of detection sensitivity, model fit as well as robustness.

8.6.4 Overall improvement

In Fig. 8.10 is shown the overall improvement of our atlasing approach using the robust global model, deformable registration (with filtering) and TLS normalization as compared to using the more widespread atlas method using a local model, affine registration and average normalization. The $AUC_{0.05}$ for the ROC curves in Fig. 8.10 are 0.9101 in the first case (our approach) and 0.8359 in the second case, the z-score being 4.83 ($p < 1.34\text{E-}6$).

8.6.5 Dependence on location

Fig. 8.11 shows the ROC curves obtained for the different locations used in this experiment. The variance of each region as present in the database learning images is given as standard

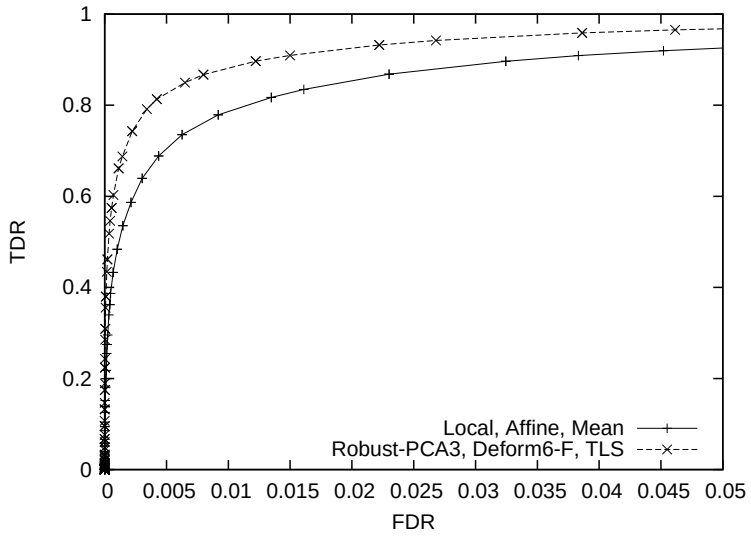


Figure 8.10: Overall improvement of our atlasing approach that consists of deformable registration, total least squares intensity normalization and a robust global model as compared to the more usual approach that consists of affine registration, mean intensity normalization and the local model. Results are averaged over all locations using small 25 % hypo-inclusions.

Location	RF	LT	RT	LIP	RIP
LF	0.0003	0.0024	0.0935	0.3036	0.2463
RF		2.62E-8	2.60E-6	0.0108	1.32E-5
LT			0.2117	0.0005	0.0759
RT				0.0155	0.6146
LIP					0.0463

Table 8.3: Two-sided p-values associated with the 25 % inclusions in Fig. 8.11.

deviations in parenthesis. The significance values corresponding to Fig. 8.11 is given in Tab. 8.3. We see that the abnormalities are more difficult to detect in the right frontal region (RF) than in the other regions ($p < 0.0108$ or less). The difference between the right and left curves is highly significant in the frontal region ($p < 0.0003$), significant in the parietal region ($p < 0.0463$) and not significant in the temporal region. Averaging the right and left regions and comparing regions, we have that the temporal region is easier for detection than the parietal region with significance $p < 0.0339$.

The asymmetry in variation found in the frontal cortex (Fig. 8.11) is in accordance with a higher uptake in this region as found in [127] as well as a known anatomical asymmetry in the frontal lobe, the right being larger than the left in general [59, p.17]. The performances obtained represent a combined measure of model-fit and detection sensitivity. The lower performance in the frontal region is therefore *not* only explained by a higher variance in this region, Fig. 8.11, but also reflects a limited capacity of the model to represent this region. For example is the average variance in the "LIP"-region (6.72^2) lower than in the "LF" region (7.98^2), even though the ROC curves show lower performance in that region (Fig. 8.11).

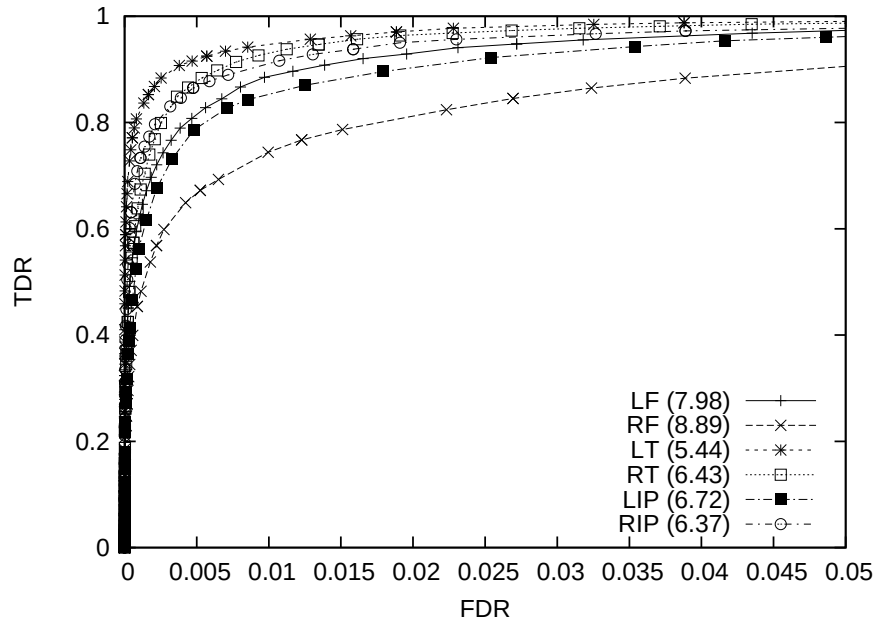


Figure 8.11: Comparison of ROC curves for each location where inclusions were added using the PCA3 model. See Fig. 8.3 for legend. The curves were obtained for small hypo-inclusions (20 mm ∅) with intensities of 25 % of average brain perfusion. The curves were similar for other intensities. In parenthesis is given the average standard deviation in the area of the inclusion.

8.6.6 Comparing validation strategies

We already mentioned in Sec. 8.3 that none of the other evaluation studies have actually tried to validate the model hypotheses (i.e. that the data is normally distributed). This is because they use a biased strategy where the baseimage is included in the learning set. These studies therefore yield optimistic performance estimates of the methods/models they validate. To see the influence of the choice of validation strategy, we repeated the validation study, but this time by leaving the baseimage in the learning set. The testimages were created in the same manner as in the leave-one-out scheme, with the same locations and intensities. The results are shown in Fig. 8.12. We see that the biased estimates are clearly superior to the unbiased estimates. It is clear that using the biased estimation method, we can obtain quite impressive results by increasing the number of eigenvectors. This is shown in Fig. 8.13 and is the reason why we believe this model will be more performant than the local model for a representative database of learning images that is more complete.

8.7 Conclusion

We do not expect a probabilistic, computerized atlas to replace visual inspection of SPECT scans. However, it can add value to SPECT examinations by permitting the comparison of an image with normative data as we suggests in the next chapter. For this to be possible, complex image processing must be undertaken, whereby the impact and consequences of such processing must be evaluated and well understood. Furthermore, the assumptions underlying the statistical models and tests must be validated. In this chapter, we have attempted to

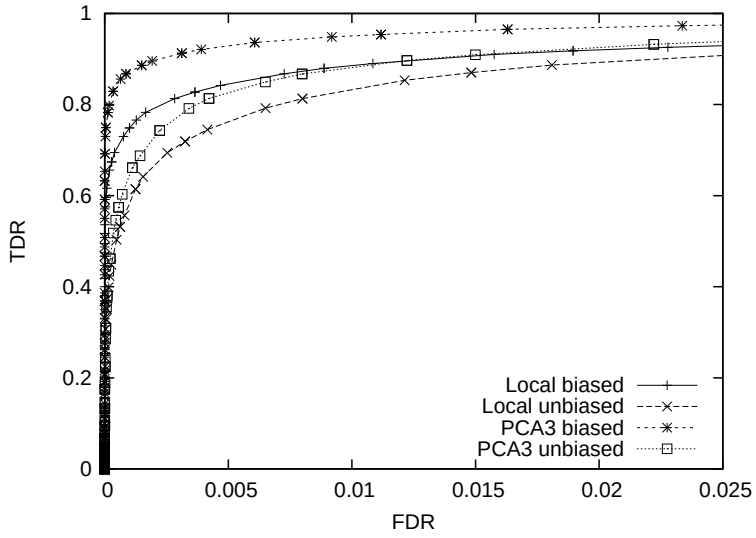


Figure 8.12: Comparison of different validation strategies. The two upper curves are the performances of the global and local models using a biased validation strategy where the baseimage was part of the training set. The two lower curves are the performances using the leave-one-out strategy described in Sec. 8.5. All locations were averaged. The curves were obtained for small hypo-inclusions (20 mm ⌀) with intensities of 25 % of average brain perfusion.

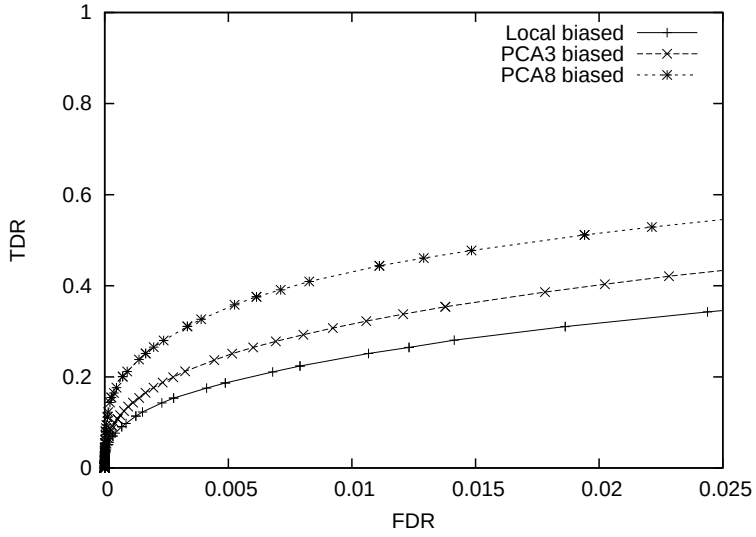


Figure 8.13: With the biased validation strategy it is clear that we obtain increasingly better ROC curves with a higher number of principal components in the model (this is only true up to a certain limit). Here are shown three curves obtained for the local model, the global model with three and eight principal components respectively. Again, all locations were averaged. The curves were obtained for small hypo-inclusions (20 mm ⌀) with intensities of 10 % of average brain perfusion.

obtain quantitative measures in this direction. Among the models studied, the robust (non-linear) global model shows the best overall characteristics: sensitivity, specificity and model validity. This kind of model also compensates to some degree errors in registration and intensity normalization. However, we do not make any statements about the clinical outcome of these models or preprocessing algorithms.

Chapter 9

Clinical Application: Epilepsy

In the evaluation study presented in the last chapter, the capacity of the atlas and the differences between atlas models were evaluated in an exploratory setting, i.e. without a priori information or hypotheses about the abnormality that is searched. In practice, this is not how the physician works. He, or she, is interested in understanding more about the impact of an atlas in the routine analysis of SPECT images as well as its possible added value for diagnostics.

In this chapter we present some preliminary results of comparing images of patients with epilepsy to the atlas. We begin by giving a brief introduction to epilepsy and the particular status of SPECT imaging for determining seizure foci. We proceed by discussing the added value that we expect an supplementary analysis using the atlas could bring. We then finally present some results that are currently motivating further atlas-based projects.

9.1 The pathology

Brief description

Epilepsy is a brain disorder where patients have a tendency to experience recurrent seizures. An epileptic seizure happens when normal brain activity is disrupted and a number of neurons start firing signals in an uncontrolled manner. This may cause temporary changes in the person's personality, mood, memory, sensations, movement or consciousness. Seizures can be classified into partial seizures or generalized seizures[1]. For partial seizures, a distinction is made into simple and complex partial seizures, depending on whether a person's consciousness is impaired or not. A partial seizure may spread to involve the whole of the brain and thus turns into a (secondarily) generalized seizure. A generalized seizure begins with a widespread discharge that involves both sides of the brain at once.

Most people have seizures in their life, sometimes without even recognizing it. However, epilepsy is not diagnostized before at least two seizures have been observed. In only three out of ten cases of epilepsy is it possible to find a cause. Causes include head injury, stroke, infections, brain tumors and others. In summary, one can say that epilepsy is a complex neurological disorder in which many pathophysiological aspects are not yet well understood.

[1]The classification of seizures is actually more complicated. This is only a typical, simplified characterization.

Treatment

About 80 % of all cases of epilepsy are successfully treated with medicaments. The medicaments aim at increasing the level of inhibitory neurotransmitters or at decreasing the amount of excitatory ones. Further treatments consists of a specialized (Ketonic) diet for children or a vagus nerve stimulation (with an implant). However, in a small number of serious cases, when other treatments fail, surgery is undergone. In partial epilepsy, the seizure center is removed from the brain. This can be done if the center is not situated in any essential part of the brain (i.e. affecting speech-, memory-, audiovisual- or language cortex). Because of the high plasticity of the brain, the functions that were present in the removed region are taken over by other regions.

Four types of surgery are possible: (1) Temporal Lobectomy - here a part of or the whole temporal lobe is removed. About 70 % of the patients have marked improvement or are cured of their seizures. (2) Extratemporal Resections - seizure centers in an area other than the temporal lobe can sometimes be removed. Fifty percent of the patients have a marked reduction or elimination of their seizures. (3) Corpus Callosotomy - the interconnection between the two hemispheres is removed. This is only done in a very few patients having serious attacks and where many seizure centers are present. (4) Hemispherectomy - if one hemisphere of the brain is abnormal and causes seizures, it is completely removed in rare cases.

9.2 Medical imaging and epilepsy

From the above, it is clear that accurate localization of the seizure foci and brain functions is of utter importance for epileptic surgery. For this, several tests are available: neuropsychologic evaluation, Electroencephalogram (EEG), invasive EEG (also called stereotactic depth EEG, SEEG), Long Term Monitoring (LTM) with EEG and video surveillance, MR imaging to find morphological atrophies, PET imaging for brain function and SPECT imaging of seizure onset among others. Typically, several tests are performed since they may yield complementary information.

High-resolution MRI has emerged as the best diagnostic tool for identification of epileptogenic lesion. In 20-40 % of patients with intractable epilepsy [205], no lesions are detected on MRI. In these cases, invasive EEG studies are very useful. Invasive EEG studies have the disadvantage of restrictive vision due to limited sampling and may not distinguish between distantly propagated seizures and initial discharge of the ictal onset zone. Occasionally, complications like hemorrhage and infections may also occur. Another aspect is that in developing countries, facilities for invasive studies are not commonly available.

9.2.1 SPECT in epilepsy

Among the techniques for localizing seizure foci, SPECT imaging is in a particular position because it is possible to image the seizure onset. For this the radiotracer is injected into the patient at the moment of onset. Because the tracer is rapidly distributed to the brain (fixates) and remains trapped a few hours, it is possible to record images when the seizure is over and the patient is calm. This is called an *ictal image*. A second *interictal image* is acquired of the patient, either before or between seizures. These two images are then compared in order to define the seizure focus (or foci) and eventually study the path of discharge.

9.3 Computer-aided evaluation and SISCOM

In clinical practice, it is still quite common that an expert compares the two SPECT images (ictal and interictal) side-by-side. The images are typically displayed slice-by-slice in three different views (coronal, axial and sagittal) and the expert has the possibiliy to scale the intensities (interactive intensity normalization). Furthermore, access to an interactive ROI (region of interest) method is provided, where the expert can delineate regions (typically using simple geometrical forms such as ellipses) in the slices and compare mean values of the ROIs.

To further improve the sensitivity of the detection of seizure foci, several automatic computer aided methods have been developed:

- Automatic intensity normalization.
- Intra-modal registration (using a rigid transformation model) that makes it possible to create subtraction images that highlight differences in the two images.
- Inter-modality registration with the patients MRI scan in order to improve the localization of the foci with respect to anatomical structures.

A method that combines these steps is known as SISCOM (Subtraction Ictal SPECT with CO-registered MRI) [237, 175, 176] and will be further discussed in the next section.

9.3.1 Intensity normalization revisited

As for intensity normalization, we have already discussed this issue at length in both Sec. 6.6, p. 113 and in Sec. 7.5, p. 126. Because of the sometimes large modifications observed in ictal SPECT images with respect to normal images, several authors argue that specially adapted intensity normalization methods are necessary [22]. During a seizure, the blood flow can become quite high and the uptake of tracer no longer depends linearly on the blood flow. Hyperperfused areas thus appear less intensely in the SPECT images as they actually are. Traditional intensity normalization techniques normalize the observed counts to the whole brain uptake or to that of a (fixed) reference region. This is no longer useful for ictal images since the whole brain perfusion pattern can be completely altered. One technique that has been specifically developed for intensity normalization of ictal and interictal SPECT images has been presented in [21] (also described in Sec. 6.6.3, p. 115). Here, the reference region is automatically determined based on a criterion of homogeneity[2]. However, the technique is not easily extensible to our case where we would like to compare ictal and interictal images to a database of normal images.

Another comparative study [185] recommends simple linear scaling based on the evaluation of the joint histogram (scatterplot) of the two images. Large zones of abnormal perfusion could in this case lead to joint histograms where the distribution is non-linear, or equivalently, the regression line shows outlier concentrations. A possible solution in this case could be to use robust regression techniques such as those discussed in Sec. 3.5, p. 3.5 (or the one proposed in [185]). However, for this outlier effect to be of any impact, the number of abnormal voxels must be quite large. We have in practice never seen clear outliers in the joint histogram. What we have observed is on the contrary that the joint histogram between the ictal image and the reference image is often more "smeared out" than the joint histogram between the

[2] Actually, two different criteria are proposed for finding this region.

interictal image and the reference image, see Fig. 9.1. This is probably the result of many micro alteration of blood flow, both hypo and hyper. Because such smearing has less influence on the estimate of the regression line than a large outlier concentration, we believe that the total least squares regression on the joint histogram is sufficiently robust also for the ictal image. The only problem that we still think may be present is an underestimation of high blood flow values because the radioactive fixation may not necessarily linearly follow the blood flow when these are high above what is normal.

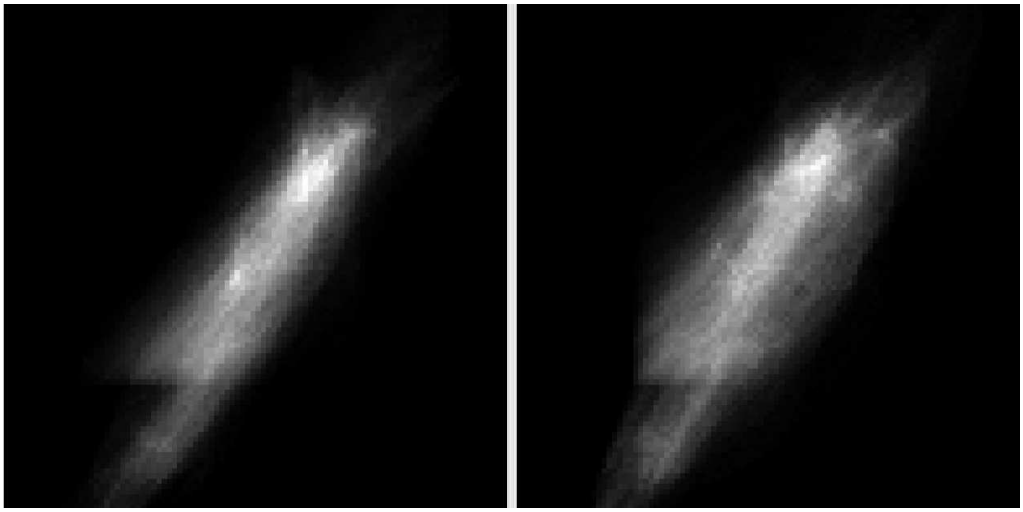


Figure 9.1: Example joint histogram between a patient with epilepsy and the database reference image. On the left: interictal image. On the right: ictal image. The histogram with the ictal image is more smeared out than that for the interictal image.

9.3.2 Difficulties with SISCOM

The usefulness of the SISCOM technique is due to the availability of both the interictal image and the ictal image. The interictal image thus serves as a reference image. In the interictal image, one expects to find regions of hypoperfusion where the seizure foci is located, in the ictal, zones of hyperperfusion. In fact, when PET exams are made, in which case one can only acquire an interictal image, zones of hypoperfusion are indicative of epileptic foci. However, these assumptions are not as simple in practice as discussed in [134] and illustrated in Fig. 9.2. True interictal images are difficult to obtain, because the pathologic brain region is sometimes hyperperfused several days after a seizure. A possible cause for this may be that the region needs to recover before blood flow again decreases to the patient's normal level. Likewise, if the injection of the radiotracer does not comply perfectly with the seizure onset, the seizure center might be exhausted and a zone of hypoperfusion may actually be observed (see Fig. 9.2). In this case, the hyperperfusions that are found correspond to the path of seizure discharge. Both these cases complicate the evaluation of the images and can lead to large errors. Semi-quantitative analysis of the interictal and ictal images where the blood levels are compared to what can be considered the patient's normal level could therefore add important information to the evaluation of SISCOM studies. Let us finally mention that it is often necessary to acquire several ictal images of the same patient because the epileptic focus (foci) cannot be accurately determined from the first ictal image. Improved analysis based on

an atlas can therefore lead to more cost-effective diagnostics and less exposition of the patient to radiation.

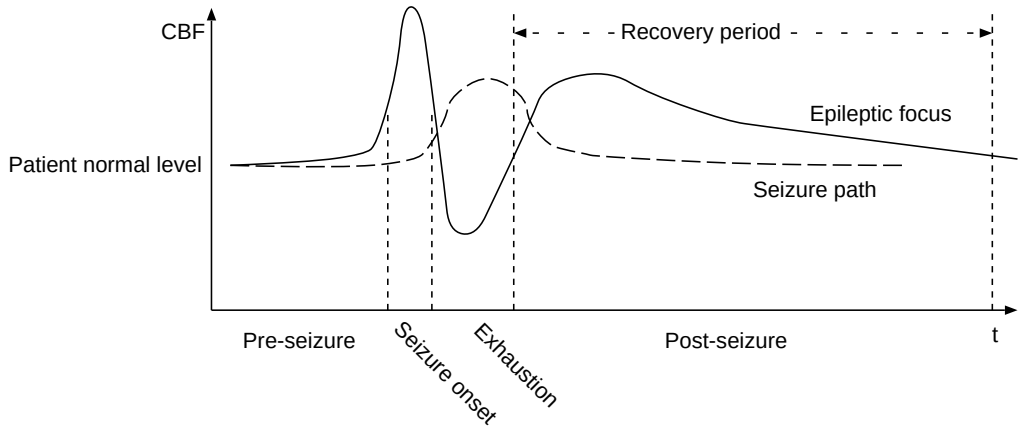


Figure 9.2: Illustration of momentaneous cerebral blood flow (CBF) at the epileptic focus and in the seizure path as a function of time (t). Since this curve cannot be exactly measured, it is only a model, based on [134]. At the epileptic focus, we have a large increase in blood flow at the seizure onset as a result of the explosive increase in neuronal activity. This brutal discharge leads to an exhaustion of neurotransmitting chemicals and to a period of hypoperfusion that precedes a period of recovery in which, according to the authors, hyperperfusion can be observed upto several days after the seizure. In the path of the seizure (where the activity propagates), increased activity leads to a period of "well-behaved" hyperperfusion. These curves illustrate the importance of obtaining *true* ictal and interictal images as otherwise the results of comparison can be completeley reversed.

9.4 Added value of an atlas

We have on several occasions discussed difficulties linked to the interpretation of SPECT images. SPECT images are often difficult to interpret, even for an expert. This is especially true when only small differences are present in the image with respect to what is normal. Let us summarize the advantages that we expect of an atlas and make some reflexions on its practical application to epilepsy. Since the modifications of cerebral blood flow (CBF) in the ictal image are often quite large, they are mostly well detected in a SISCOM analysis. However, several problems can be encountered:

- Difficulties of obtaining true ictal and interictal images (Sec. 9.3.2).
- Questions concerning the normality of the interictal image.

Additional comparison with an atlas of normal perfusion can help clarify these figures and cases of doubt. In some special cases this comparison can also reveal subtle differences that are not visible in the SPECT images. Since in most cases where an atlas evaluation is available, SISCOM analysis is also available, the additional value of an atlas is therefore mainly a complementary source of information to the clinician. Of course, if the atlas-based interpretation is well validated and successful, the interictal exam could be left out altogether. This

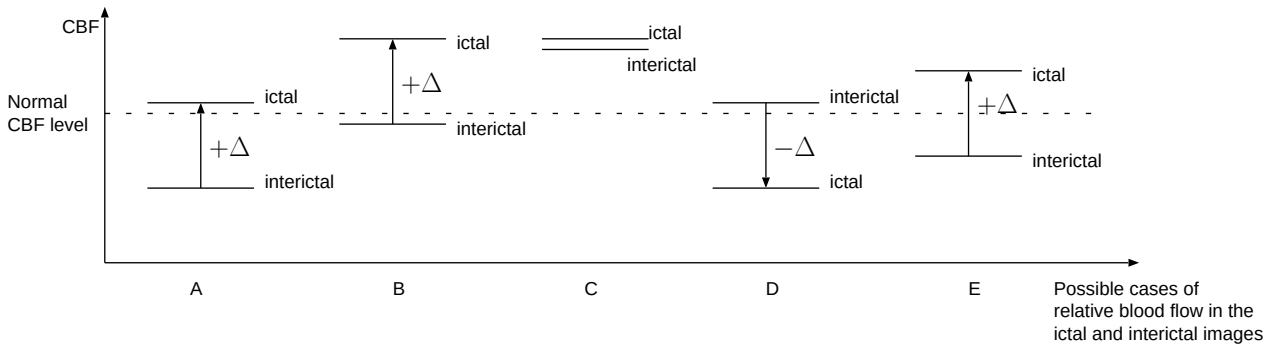


Figure 9.3: Example figures of relative and absolute blood flow levels that can be observed in a pair of (true) ictal and interictal images. (A) an epileptogenic zone where the interictal blood flow level is below that of normal and where the seizure augments this level to a normal level. (B) the ideal case where the interictal image shows a normal blood flow level and the ictal an increased level. This could be both in the epileptogenic zone or in the seizure path. (C) A less usual figure where a non-epileptogenic zone shows interictal hyperperfusion that remains relatively constant during the epileptic seizure. (D) An exhausted epileptogenic zone where a negative difference between the ictal and interictal image is observed. (E) Significant difference in ictal and interictal blood flow that is not detected by atlas comparison as both values are in the range of what is considered to be normal.

case is of great importance in other pathologies where a reference image is not available such as dementias and others.

9.4.1 Cost of an atlas

In order to make an analysis of the true added value of an atlas, one is obliged to ask at what additional cost this is available. The main cost is of course the acquisition of a database of normal subjects as well as the developments of the software and validation tools. However, once the atlas is established there are only two aspects that determines the cost of using an atlas: (1) material cost in form of computers and display interfaces, and (2) cost linked to the added complexity of evaluating the results, notably time of interpretation and training of the clinician. The material costs are relatively small. If the facility already possesses a computer for performing SISCOM analysis, the only additional computational step to perform is the non-linear registration of the patient's MR image with the reference (intensity normalization and brain segmentation is negligible in terms of computational cost).

The second point is probably more important. In a clinical routine situation, the physician only has a limited time for evaluating the images. The simultaneous multiplanar visualization of 5 to 6 images as we have used (see screenshots in the next section, Figs. 9.4 and 9.5), is indeed quite laborious and demands much expertise. However, with an adapted visual interface and approriate training, the additional time of comparing the results of the SISCOM analysis with the atlas could be kept to a minimum.

9.4.2 Evaluation revisited

Ideally, to validate the added value of an atlas, one would need to perform a large-scale study (several readers and patients). One of three possibilities could be envisaged:

1. Blindfolded (no patient record information available to the reader) interpretation solely based on the atlas must be better than random guessing. This evaluation could make statements to whether the atlas provides any useful information at all. In the latter case, any further study would be futile. However, it does not provide information as to whether atlas analysis brings additional information to SISCOM analysis and this kind of evaluation is extremely difficult to perform for the physician.

2. Blindfolded interpretation based on SISCOM with and without the atlas would therefore be more appropriate since a possible additional value of the atlas could be measured. However, it would not necessarily answer the question whether an atlas brings added value in a clinical setting where the clinician works in a hypothesis-based manner (as opposed to taking an exploratory approach).

3. In a real clinical setting, the patient record is available to the clinician. This would answer the real question whether atlas analysis brings any additional information in a clinical setting. It is however difficult to design an unbiased experiment in this case.

All these studies do need some form for ground truth. Unfortunately, we have not performed any complete study that can truly respond to the exigences of validation as listed above. The results we show in the next section are therefore to be taken as preliminary. A more elaborate study of the department's database of epileptic patients is in planning as well as a project for studying migraine patients.

9.5 Results on real images

Six subjects with temporal lobe epilepsy were preprocessed and compared to the atlas. The atlas was created from 20 images as described in Ch. 7 (the last 14 images for atlas creation were acquired at a later stage). The analysis was made by an experienced clinician working in the nuclear medicine service at the institute. The comparison was made using the three models that were compared in Ch. 8: the local model, the standard global model (PCA3, ML model in the notation of Ch. 4), and the robust global model (RMMS model with three eigenvectors). The deformable registration scheme explained in Ch. 7 (Sec. 7.3, p. 120) was used and the images were intensity normalized using total least squares regression as explained in Sec. 7.5, p. 126. A SISCOM difference image was also created. We first discuss the differences observed between SISCOM and the RMMS model, before we discuss differences observed between the different atlas models (and that were smaller). Two cases have been shown in Figs. 9.4 and 9.5. These will be used to illustrate the following analysis and discussion.

9.5.1 Similarities and dissimilarities between SISCOM and atlas

The nuclear medicine physician globally judged the results of the ictal-atlas comparison to be similar to the results of the SISCOM analysis. The same lateralizations, main foci and paths of discharge were found with both methods. However, some dissimilarities were also observed. In several cases, we observed significant increases of perfusion in the occipital lobe and the cerebellum in SISCOM that were less pronounced or not significant when comparing to the atlas. This is explained by two factors: (1) normal variation in these regions are somewhat higher than the average variation elsewhere in the brain, and (2) sometimes these zones showed

up as hypoperfusions in the interictal image (as compared to the atlas). This can for example be seen in Fig. 9.4.

Most dissimilarities between the SISCOM and the ictal-atlas comparison images were well explained by analysis of the interictal-atlas comparison image, see also schema in Fig. 9.3. Some examples of this are shown as annotations in Figs. 9.4 and 9.5: Regions detected in SISCOM that were not detected in the ictal-atlas comparison were explained by hypoperfusions in the interictal-atlas comparison. Likewise, detections in the ictal-atlas comparison that were not detected in the SISCOM image were explained by hyperperfusions in the interictal-atlas comparison. Hyperperfusions in the interictal image was indeed unexpected by the clinician and could be subject to further investigation.

In Fig. 9.5, a hyperperfusion is found in the left hippocampus (blue cursor) by the ictal-atlas comparison that is undetected in the SISCOM image. This difference is also difficult to observe in the original ictal and interictal images. This detection is interesting because the hippocampus is often implicated in epileptic seizures. Note also that this hyperperfusion is only partially explained by the interictal image.

Another difference between SISCOM and the ictal-atlas comparison that was noted, was that the patterns in the latter case often had a different characteristic, typically more focused, than in the former. As an example, consider the sagittal view (top, right image in the middle column, top row) in Fig. 9.4 of the right temporal lobe, where the pattern is quite different from the SISCOM pattern (left column, top row). Whether these patterns are truly more indicative of the path of discharge must however be further investigated. The patterns observed when comparing to the atlas being more characteristic, the clinician also perceived these as more noisy. We have shown the z-score (significance) images using the three different atlas models for another patient in Fig. 9.6 together with the SISCOM difference image. Here we see that the SISCOM image is indeed more smoothed out than the significance images which explains the perceived noisiness.

9.5.2 Similarities and dissimilarities between the atlas models

The atlasing methods (models) yield quite similar results as seen in Fig. 9.6. Whereas the differences between the local model (top left) and the global models (top right and bottom left) are perceivable, one has to create the difference image between the robust model (bottom left) and the standard global model (top right) to see any differences[3]. This difference image is shown in Fig. 9.7 using the same display dynamic (contrast) as in Fig. 9.6. A careful analysis shows that most of the high valued regions obtained by the global model are further accentuated by the robust model. This is what one would expect: high values are counted as outliers, or almost outliers, and are therefore neglected when solving the reconstruction problem. It is however doubtful that the different models in this example would lead to different diagnostic outcomes. As we showed in the simulation studies in Ch. 8, the robust model yielded a marginal, but systematic improvement over the standard global model, thus making it a better choice as the default model. This even more so, since the additional computational cost of using this model is small (around a few minutes for large subspaces).

[3]A better approach is to visualize an animation of the significance images obtained by the two methods.

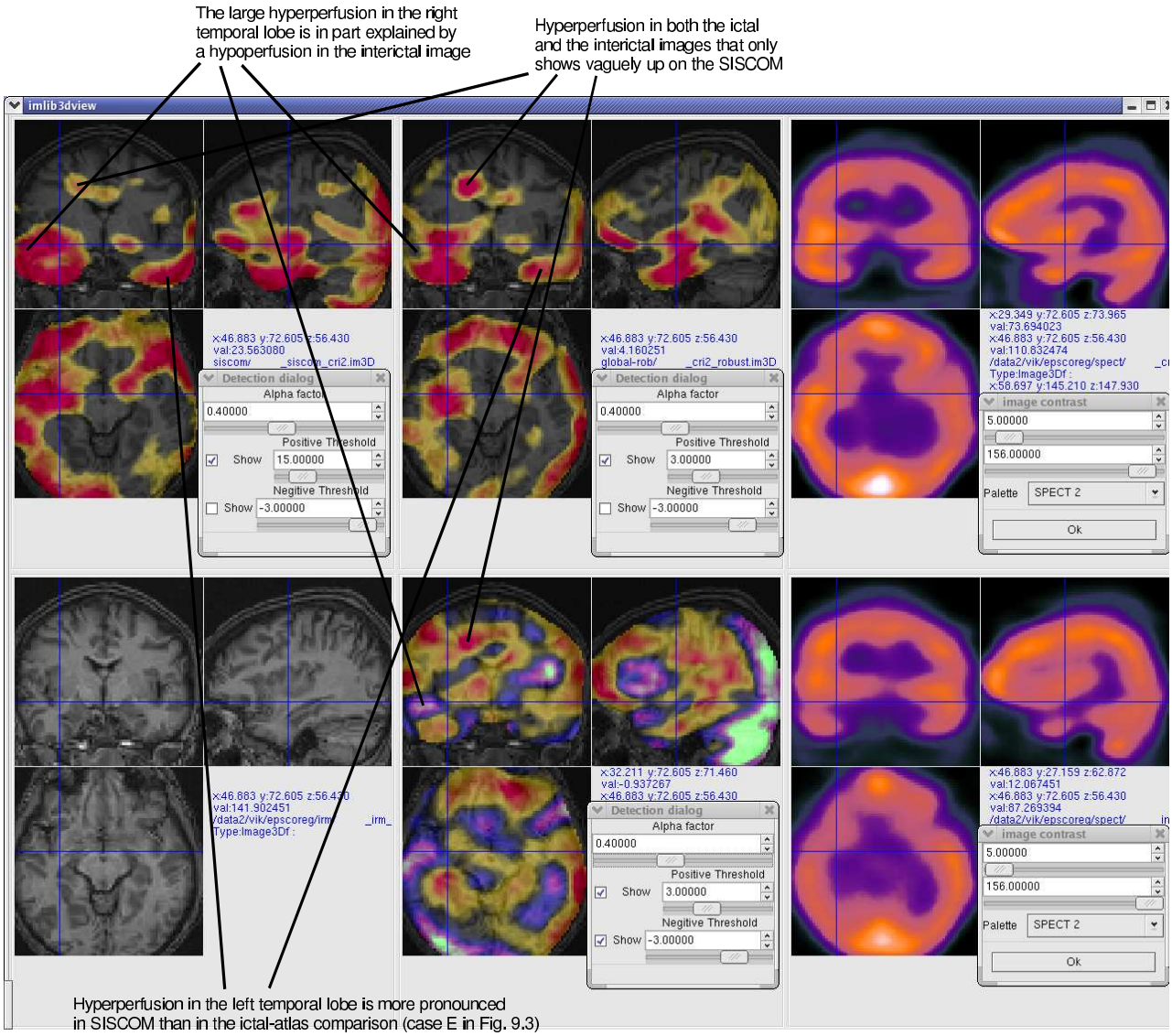


Figure 9.4: Example of a patient with bilateral temporal lobe epilepsy that is predominant in the right lobe. Right column: top – ictal image, bottom – interictal image. Middle column: top – ictal image compared with atlas, bottom – interictal image compared with atlas. Left column: top – relative difference between the ictal and interictal images (SISCOM), bottom – patient MRI. This simultaneous view of multiple results and images makes it possible to extract complementary information from the original images, the SISCOM technique and from the atlas. This is useful for comparing different techniques and for verifying the relations explained in Fig. 9.3. Some of these are annotated in the figure.

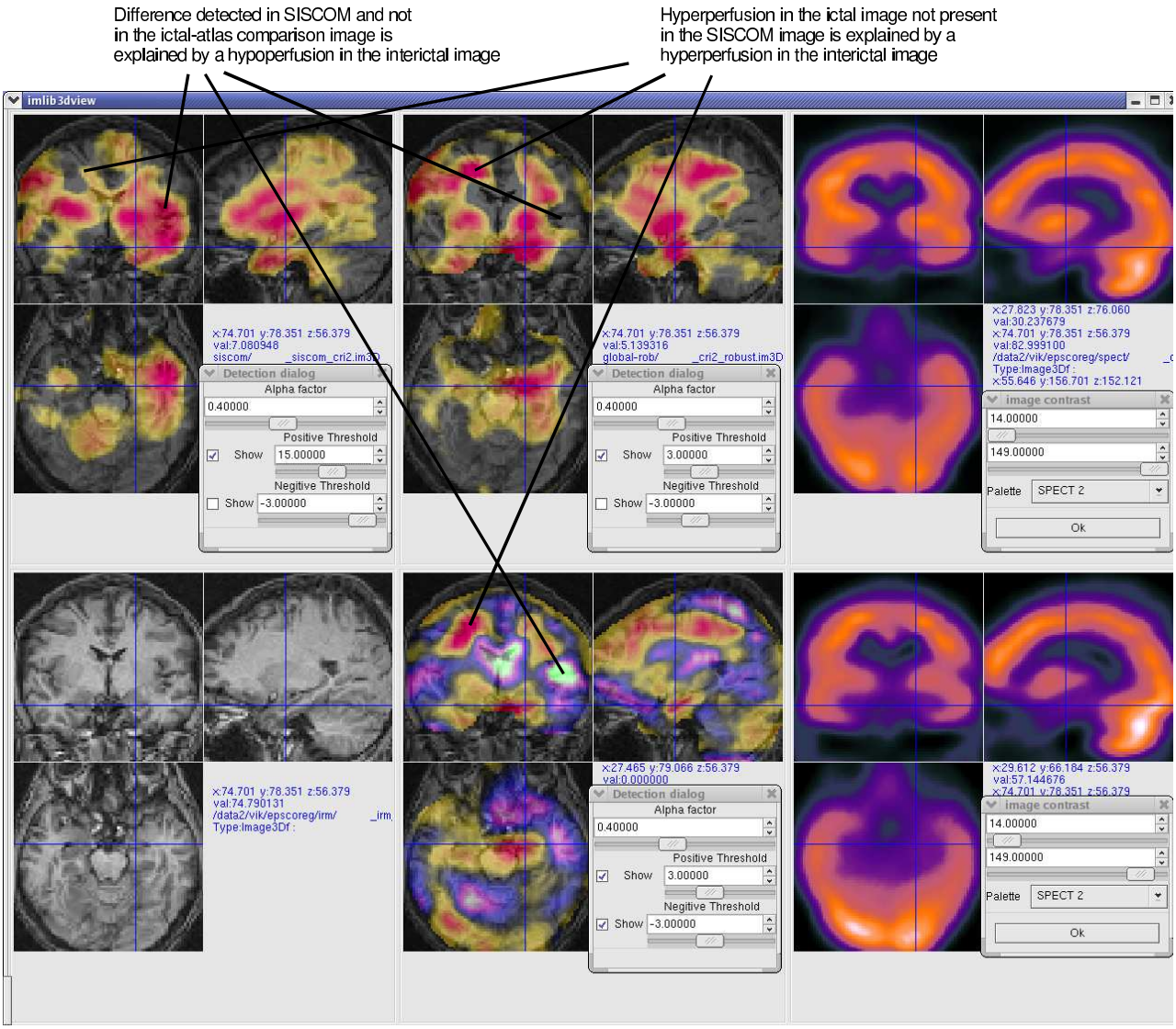


Figure 9.5: Another patient in the same kind of display as in Fig. 9.4. Right column: top – ictal image, bottom – interictal image. Middle column: top – ictal image compared with atlas, bottom – interictal image compared with atlas. Left column: top – relative difference between the ictal and interictal images (SISCOM), bottom – patient MRI. The blue cursor is placed on the left hippocampus where a hyperperfusion with respect to the atlas is detected that is partially explained by a hyperperfusion in the interictal image, and therefore goes undetected by in the SISCOM image. Note also that this kind of abnormality with respect to normality is difficult to see directly in the original ictal and interictal images.

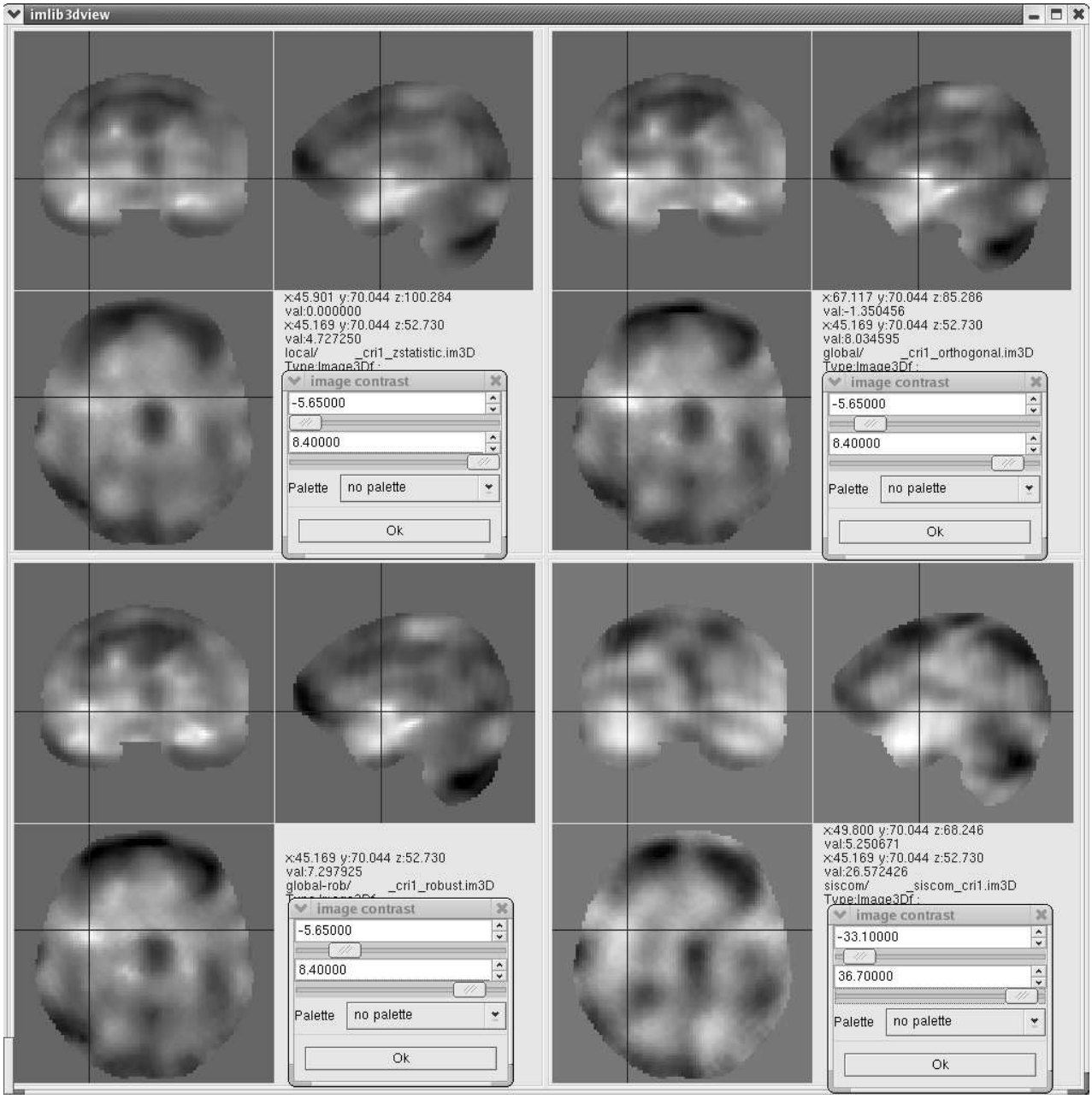


Figure 9.6: Examples of significance images of another patient with bilateral temporal lobe epilepsy. Top-left: ictal image compared voxel-by-voxel to the atlas (local model). Top-right: ictal image compared to the atlas (global model, PCA3). Bottom-left: ictal image compared to the atlas (robust global model, PCA3). Bottom-right: difference of ictal and interictal images (SISCOM). Generally, we see that SISCOM yields a much smoother significance image (bright for hyper perfusion and dark for hypoperfusion) than the atlas methods. The atlas methods show more focused abnormalities than the SISCOM image. As in this example, the local significance images are in general smoother than the global significance images. However, at the position of the cursor (cross), the global models signal more significant abnormality than the local model. Between the global methods there is only a small difference, shown in the difference image Fig. 9.7.

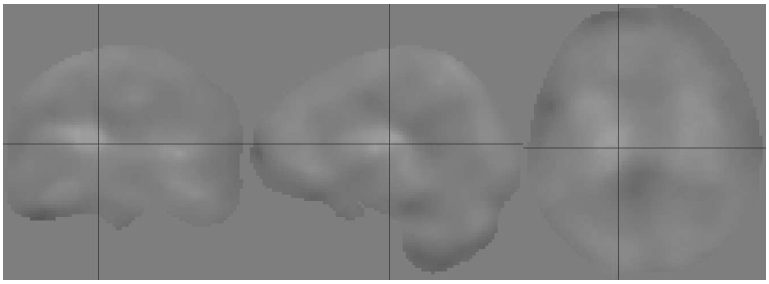


Figure 9.7: Difference between the significance image obtained by the robust global model (bottom-left) and the global model (top-right) in Fig. 9.6 using the same display contrast. This is the image where we observed the largest difference between these models, ranging from -4.1 to 2.6.

9.5.3 Affine versus deformable registration

In Fig. 9.8 is shown an example of an ictal-atlas comparison using an affine transformation model and a deformable deformation model for image registration (see also Sec. 7.3, p. 120). Even though the results often bear resemblance, there are some differences when using these two approaches. In this example, an important spot of hyperperfusion was not detected (or only weakly detected) when using the affine registration scheme instead of the deformable scheme. We have also observed that the path of discharge is quite differently depicted in the ictal-atlas (affine) comparison as to the SISCOM image, whereas the ictal-atlas (deformable) comparison is quite similar. We also know from the evaluation study in Ch. 8 that the atlas that has been constructed using the deformable registration scheme is more reliable.

9.6 Conclusion and future work

In this chapter we have presented some preliminary results on real images of patients with epilepsy. These illustrate some of the potentials that the use of an atlas have for aiding the diagnosis. In particular, the type of detection observed around the hippocampus in Fig. 9.5 seems to be of much interest to the medical expert. The application to images of epilepsy was chosen because the atlasing method can be compared to the existing SISCOM technique. However, its application is of more importance in other pathologies where no reference images are available and no other quantitative method exists.

In order to fully validate our approach, further systematic studies must be conducted. For this, we are planning a systematic study of the database of treated epilepsy patients that is accessible at the institute. As we have observed in the experimental studies of 2-D images (Ch. 5), we think the atlas will profit more from a larger database of normal subjects (the atlas that was used for these preliminary results contained only 20 images) so that the description of variations of normal perfusion is improved.

The medical experts show great interest in our atlasing system, not only for the application in epilepsy, but also for other pathologies. However, there is still some work to be done before a medical expert can use the system without any technical assistance. In the meantime, we provide a temporary solution where some functionality is made accessible through training and assistance.

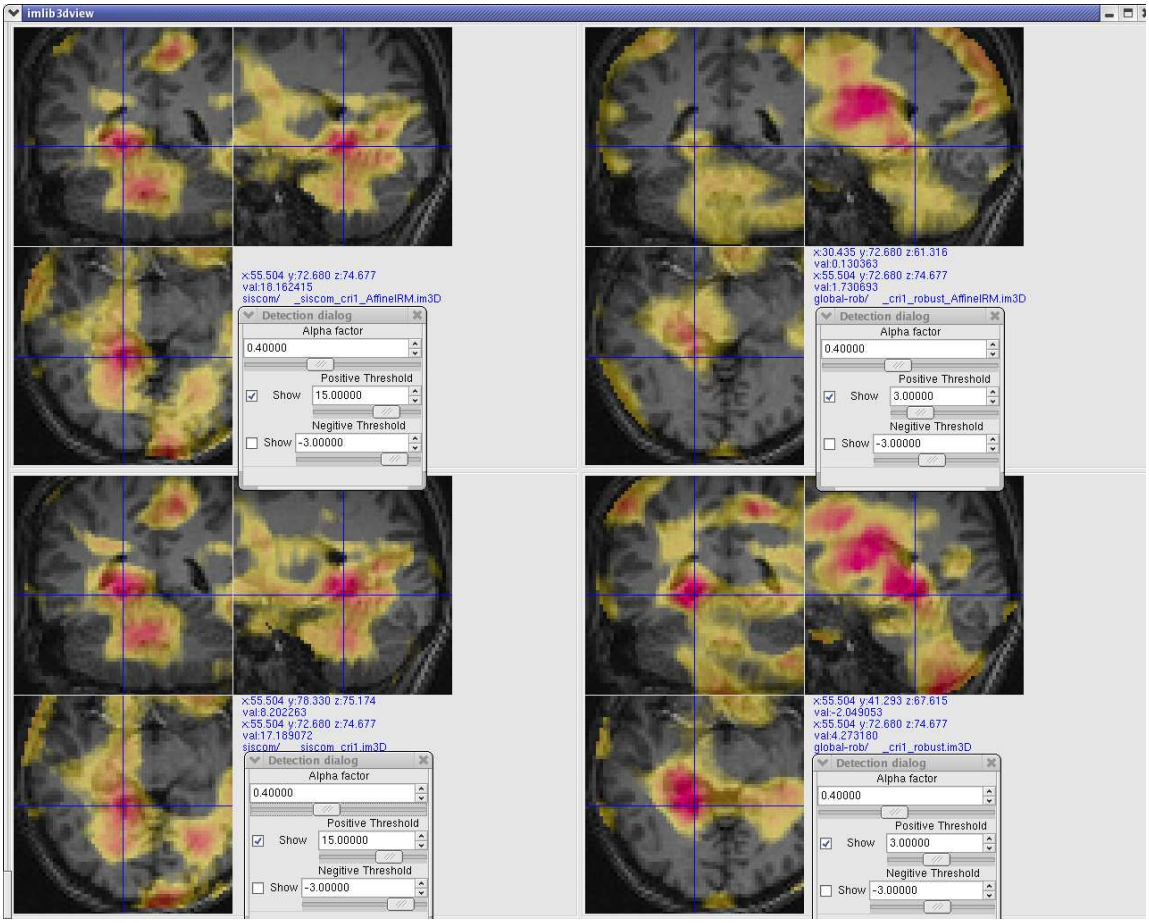


Figure 9.8: Another patient with extra-temporal epilepsy. In the left column are shown the SISCOM images, in the right column, the results of the ictal-atlas comparison. In the top row an affine deformation model was used for MRI image registration, in the lower, a deformable model as described in Sec. 7.3, p. 120. The two SISCOM images are somewhat different because of the different reference spaces. The atlas comparison results are similar, but the comparison using only the affine registration almost misses an important spot of hyperperfusion that is clearly detected in the three other images.

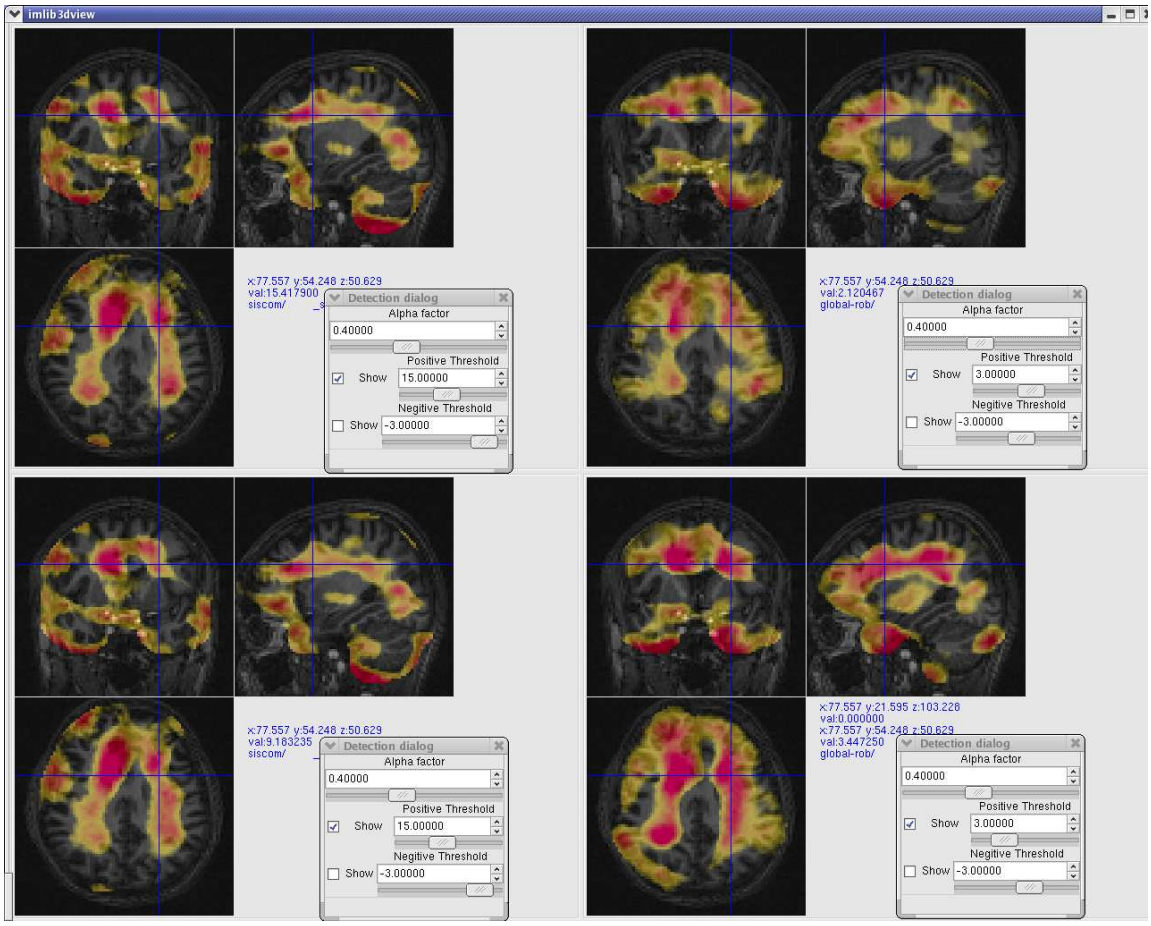


Figure 9.9: The same comparison as in Fig. 9.8 for another patient with right temporal lobe epilepsy. In the left column are shown the SISCOM images, in the right column, the results of the ictal-atlas comparison. In the top row an affine deformation model was used for MRI image registration, in the lower, a deformable model as described in Sec. 7.3, p. 120. Two bilateral paths that are clearly depicted in the SISCOM images, deviates in the affine atlas comparison, whereas the deformable atlas comparison accentuates these. Both atlases also show small regions of hyperperfusion in the temporal lobes (the deformable somewhat more than the affine) that are almost inexistent in the SISCOM images. In a second ictal scan of this patient, a strong hyperperfusion was found in the right temporal lobe with small spread to the contralateral temporal lobe and strong spread to the frontal and superior parts of the brain.

Chapter 10

Conclusions and future work

In this final chapter, we summarize this thesis and its contributions. To conclude, some ideas for future work are presented.

10.1 Summary and discussion

This thesis and the contributions of this thesis are divided into two parts. One concerns the theoretical developments of a new probabilistic model. The second concerns the creation and evaluation of an atlas of brain perfusion.

We began Part II by studying different appearance-based models used in computer vision. In particular, we studied in detail global, linear models that rely on PCA for dimension reduction. These have been applied with much success in the modeling of images of both objects and faces. In Sec. 3.3.5, p. 39, we established the relationship between the popular model of Moghaddam and Pentland [162] and the theoretical development of the PPCA model by Tipping and Bishop [211]. Even though the applications (face recognition and computer vision) for these models are far from our problem of creating an atlas of brain perfusion, they are attractive to us because (1) of their success, (2) of their simplicity, and (3) they are learning methods.

In Ch. 4, we then presented an original, non-Gaussian appearance model for unsupervised learning. The model is based on a linear factor analysis model, but has a non-Gaussian subspace distribution and a non-Gaussian (robust) noise distribution. The first important problem to solve for any such model is the reconstruction problem (also called the inference problem), which was defined in Sec. 4.1.1, p. 61. We solved this problem by developing the *robust modified mean shift algorithm*, [220]. The algorithm is based on half-quadratic theory and an extension of the mean shift procedure. In experiments performed on a standard computer vision database, we showed this model's superiority to other well known models. These experiments also permitted us to better understand the behavior of our approach in different situations. In particular, in Sec. 5.4.2, p. 84 we drew the conclusion that for the model to be effective, it must be "sufficiently accurate". This is in general obtained by using enough training images, and a sufficient number of eigenvectors. This gain must be compared to the increased computational cost of the model. Indeed, depending on the application, one must clarify whether robust modeling is indeed necessary. Our model is particularly well adapted to difficult situations with much image degradation as the experiments show. Another difficulty might arise from the multimodal nature of the posterior density function which makes it only

possible to guarantee that the algorithm can find local minima. In some situations it might therefore be necessary to use multiple initializations in order to find a satisfactory solution. Finally, note that this model is not limited to the modeling of images. It is indeed formulated in a very general manner and could be applied to pattern recognition problems in general. This opens for many interesting perspectives for future work.

In Part III, we returned to the problem of creating an atlas of brain perfusion that originated this work. We began this part with a comprehensive study of statistical models used to create atlases of functional brain images both for research and in a clinical setting. Presenting these in a unified way is difficult because of the many different applications (image modalities, pathologies, study designs) and the many possible ways of preprocessing the images before statistical modeling. This is further complicated by the rich statistical and medical nomenclature. Nevertheless, we proposed a review where the methods were classified based on (1) whether the statistical models used are multivariate or univariate in nature, and (2) what feature is modeled. To our knowledge, no such review exist in the literature.

We take particular notice that Houston *et al.* [104] independently developed a model that is very similar to the classical linear PCA-based appearance models used in computer vision. They also found a clever way of making localized detections for this kind of global models. This method is important since otherwise only a global statement about an image into a class (pathology or not) would be possible. This local detection is simply done by weighting the residual obtained from the image reconstruction problem by the non-isotropic variance of the learning base images. We formalized this model and situated it with respect to the models presented in Ch. 4.

In Ch. 6 we continued by presenting different techniques for SPECT and MRI image processing: registration, brain segmentation and intensity normalization. We presented the main characteristics of the algorithms for deformable, inter-subject, MRI-MRI registration and for rigid, intra-subject, SPECT-MRI registration that have been developed prior to, and during this work by other researchers at our laboratory. These could therefore in Sec. 7.3, p. 120 be used to develop a registration scheme that is specifically adapted to our needs of registering SPECT images of multiple subjects for the creation of an atlas. This scheme has the particularities that deformable registration was used (generally, only affine or piecewise affine is used), and that a step of transformation field filtering was introduced. Deformable registration is still a domain of intense research that is still far from being solved. The scheme that we propose seems to be the best possible with respect to atlas creation (and with the techniques to which we have access). However, we believe there is still margin of possible improvement concerning inter-subject registration.

In Sec. 7.5, p. 126, an obvious extension to the standard linear regression solution for joint histogram intensity normalization was proposed. Total least squares was used to model errors in both images as opposed to standard linear regression where only the source (and not the reference) image is considered to contain errors. We have not yet performed systematic comparison between the LS and the TLS normalization. We only argument for the TLS method with theoretical arguments. Such comparison is indeed difficult as the intensity normalization issue is associated with much uncertainty and dispute. One possible study would however be to evaluate the robustness of the different registration schemes using our evaluation scheme (Ch. 8).

In Ch. 8, we presented an original and comprehensive evaluation study that was used to compare and evaluate several important aspects concerning atlas creation: statistical models,

heterogeneity of detection across different regions in the brain, registration schemes and finally intensity normalization. The results were described in Sec. 8.6, p. 143, see also [219, 221]. The evaluation study aims at making statements about the validity of assumptions underlying the statistical models and about the sensitivity of the model at detecting abnormalities. This has not been done before.

Finally, in Ch. 9, we have shown some preliminary results using the atlas to aid the interpretation of images of patients with epilepsy. The results showed good coherence with the results obtained by the more usual SISCOM technique, but also some differences. These two techniques seem to yield complementary information about resting state brain perfusion and changes therein induced by epileptic seizures. However, further analyses and tests must be performed. In a first time, the atlas will serve the clinician as a supplemental tool with limited confidence level.

Several software libraries were developed and co-developed during this thesis. The libraries were developed using advanced features of the C++ programming language (*generic programming* by means of templates, [218], object-oriented design) and modern software engineering techniques (*design patterns* [75], unit tests[1]). In particular, the author contributed to the open source, software package *ImLib3D*[2] which was designed and conceptualized by M. Bosc. This work gave rise to a conference publication [19]. Finally, the libraries have been implemented as a module of the in-house, proprietary software *Medimax* through which the probabilistic atlas is accessible to physicians and researchers at IPB (Institut Physique Biologique).

A weakness of this thesis is that we have not explicitly modeled sources of variation that stem from the image acquisition process, nor have we explored/evaluated the intra-individual variation and dependance on age and gender. The contribution of these error sources to the total image variation need to be better understood. Since we have a database with two images of each subject, this latter kind of variation would be possible to study.

10.2 Future work

Every thesis tries to shed some light on a specific problem. The contributions of this thesis have been summarized above. However, it seems to the author of this document that the number of resolved questions remains largely inferior to the number of unresolved questions to which this work gives rise. This is no burden – rather an opportunity. However, only a limited number of problems should be attacked at a time, and it is probably wise to choose those that are the most promising. Paths for future investigation can be divided into (1) those concerning our original model, both at the theoretical and practical level, and (2) those concerning the application of the atlas to routine clinical examinations and medical research.

10.2.1 Model

Some paths for future work concerning our original model were already detailed in Sec. 4.8, p. 70. These are summarized and some are added.

[1] `http://www.extremeprogramming.org/`
[2] `http://imlib3d.sourceforge.net`

Theoretical

The most interesting problem to solve from a theoretical viewpoint is of course the system identification (or model estimation) problem for our model. We see a possibility for this by means of the EM-algorithm as mentioned in Sec. 4.8. If we succeed in this, we would indeed have a full-blown method for unsupervised dimension reduction with possibly many applications in perspective. Furthermore, it would be desirable to derive a convergence proof of the RMMS algorithm. Also, a better understanding of the convergence properties of this algorithm is necessary.

Generalized linear models are often mentioned in the context of global linear models. It would be interesting to investigate whether it is possible to derive a *generalized modified mean shift* for such models. Finally, there are several links to be established to recent probabilistic models. Palmer [182, 181] for example shows that there are links between half-quadratic theory and ICA. Saul and Roweis [191] have recently proposed a mixture model of local linear models with a new learning concept that bears similarities to our model.

Practical

There is a need to continue experimentation with multiple initialization and marginal probability for detection. We showed in Sec. 4.8 how one can calculate the marginal probability of the non-occluded part of the image. This makes it possible to use multiple initializations for the RMMS algorithm. This could solve the problem where the robust algorithm sometimes fails with a large margin as observed in our experiments.

10.2.2 Atlas, model

As we have found during the experiments in Ch. 5 and in the pilot study briefly mentioned in Sec. 8.6.1, p. 145 (Fig. 8.7, p. 146), the performance of the models we have developed increases with more model accuracy. A more accurate model is obtained with a larger (and therefore more representative) database of normal images and with more eigenvectors in the model. This is therefore important work to do in the near future. Currently the atlas is constructed from 20 subjects. Another ten images of normal subjects have recently been acquired and will therefore be included in the atlas shortly. However, a total of 30 images is probably still not sufficient. This number should probably be in the hundreds. Because of the expenses of acquisition, it is important that image centers share images. Comparing images that come from different gamma cameras may however pose problems. In this case, inter-center image variability should be modeled, or alternatively, adapted normalization (spatial and intensity) procedures should be developed.

Modeling global activity, age, gender and handedness

An aspect that distinguishes the atlas application from a typical computer vision application is that we possess additional knowledge such as age, gender and handedness of the subjects. Since these are factors that are known to influence normal blood flow [127], these should be included in the model. This necessitates however, even more reference images. From a mathematical viewpoint, this extension is straightforward and is shown for the modeling of global activity. The extension to age, gender and handedness follows the same scheme.

Instead of considering the intensity normalization as a separate preprocessing step before modeling, an alternative would be to take the global activity into account in the model as in

the SPM ANCOVA model, Sec. 6.6.4, p. 116. This can be done by considering the following MANCOVA model[3] for scan j:

$$\boldsymbol{y}_j = \boldsymbol{W}\boldsymbol{x}_j + \boldsymbol{\mu} + \boldsymbol{\xi}(g_j - \bar{g}_\cdot) + \boldsymbol{\epsilon}_j \tag{10.1}$$

The model is the same as the global linear model in Eq. 4.1, p. 60, with an additional D-dimensional constant $\boldsymbol{\xi}$, the global activity in image j, g_j, and the average global activity in the learning base images, $\bar{g}_\cdot$.

Under this model and for Gaussian isotropic noise we have that:

$$p(\boldsymbol{y}_j|\boldsymbol{\mu}, \boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\xi}(g_j - \bar{g}_\cdot), \sigma^2 I + \boldsymbol{W}\boldsymbol{W}^T)$$

From this, we can derive the maximum likelihood estimators of $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$ as

$$\hat{\boldsymbol{\xi}} = \frac{\sum_j (g_j - \bar{g}_\cdot)\boldsymbol{y}_j}{\sum_j (g_j - \bar{g}_\cdot)^2}$$

and

$$\tilde{\boldsymbol{\mu}} = \frac{1}{J}\sum_j \left(\boldsymbol{y}_j - \hat{\boldsymbol{\xi}}(g_j - \bar{g}_\cdot)\right).$$

Note, that the mean, $\tilde{\boldsymbol{\mu}}$, now is the mean of the intensity-corrected data. Least squares projection is done as usual

$$\hat{\boldsymbol{x}}_j = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T(\boldsymbol{y}_j - \tilde{\boldsymbol{\mu}})$$

whether j is an image of the learning base or not. On the mean- and global activity-free data, PCA can now be performed as usual. In this MANCOVA model, $\boldsymbol{\xi}$ corrects for global intensity changes by adding a proportion of the global intensity difference of an image to each voxel intensity. Since it is estimated, we therefore lose one degree of freedom when estimating the voxel variance estimation (see Eq. 6.1, p. 100).

Multi-modal, integrated and disease-specific atlases

For future work, we think that incorporating more sources of variability in the probabilistic atlas model will be necessary. Conjointly modeling different image modalities, image acquisition, image reconstruction and image processing (registration, intensity normalization and segmentation) could benefit the task of interpreting and analyzing these images. The image reconstruction and processing tools that today work indepently of each other, could also mutually improve from such modeling. Likewise, it may also be necessary to develop disease-specific atlases in order to improve the utility and performance of quantitative, computer aided diagnosis.

10.2.3 Atlas, application

It is desirable to evaluate the influence of the atlas in a clinical setting. For this, we envisage a retrospective study of epilepsy patients that have undergone surgery. For these patients, there is a well established ground truth that could be used as a reference. However, since the differences in models are small, very large cohorts are necessary to show any significant

[3] A MANCOVA model was also proposed by Friston *et al.* [70]. However, this model is different in that the dimensionality of the data is reduced before global intensity changes and age are modeled.

difference. Furthermore, we are in the planning stage of a study of patients with migraine. The hypothesis here is that a situation of "aura" that certain patients experience before the migraine, is explained by an increased activity in the brain stem. An injection would therefore be made at the moment these patients experience this aura to measure the blood flow at this moment.

Part IV

Appendix

Appendix A

Modified Mean Shift

In this chapter we have collected the calculations that were left out from the derivation of the modified mean shift in the main text (Ch. 4). We first show how to calculate the MAP kernel estimate from which we obtain the more "interpretable" form of the modified mean shift (Eq. 4.33, p. 69):

$$mms(\boldsymbol{x}) = \left[\frac{\sum_{j=1}^{J} c_j \Gamma_j(\boldsymbol{x}) \boldsymbol{\mu}_j}{\sum_{j=1}^{J} c_j \Gamma_j(\boldsymbol{x})} - \boldsymbol{x} \right]. \tag{A.1}$$

We then proceed to describe how to obtain the simplified Modified Mean Shift term of Eq. 4.34, p. 69. This is first done by simplifying/rewriting Eq. A.1, then an alternative approach is derived. Finally, in Sec. A.3, we bring the convergence proof for the modified mean shift. Note that we have not derived a convergence proof for robust modified mean shift optimization.

A.1 Modified Kernel Estimate

We shall show how to obtain the posterior density (Eq. 4.25, p. 68):

$$p(\boldsymbol{x}|\boldsymbol{y}) = c \frac{1}{J} \sum_{j=1}^{J} c_j \Gamma_j(\boldsymbol{x}), \tag{A.2}$$

which we denote the posterior kernel estimate, from (Eq. 4.24, p. 68)

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{p(\boldsymbol{y})} \frac{1}{J} \sum_{j=1}^{J} p(\boldsymbol{y}|\boldsymbol{x}) p_j(\boldsymbol{x}). \tag{A.3}$$

The calculations are not complicated, but a bit tedious. Intuitively, we see that since each of the two factors of the product $p(\boldsymbol{y}|\boldsymbol{x})p_j(\boldsymbol{x})$ are Gaussians, the product itself is also Gaussian. We therefore only have to find the quadratic form of the exponential of this product. From Eq. A.3 we have

$$p(\boldsymbol{y}|\boldsymbol{x})p_j(\boldsymbol{x}) = k_1 k_2 \exp(-\frac{1}{2} \left[(\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x})^T \boldsymbol{\Sigma}_\epsilon^{-1} (\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{x}_j)^T \boldsymbol{\Sigma}_x^{-1} (\boldsymbol{x} - \boldsymbol{x}_j) \right]), \tag{A.4}$$

where k_1 and k_2 are constants. Since there are only quadratic terms in the brackets of this exponential, we can introduce a mean, $\boldsymbol{\mu}_j$, and a symmetric covariance matrix, $\boldsymbol{\Sigma}$, so that

$$(\boldsymbol{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - 2\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j. \tag{A.5}$$

Writing out the brackets of Eq. A.4 yields

$$\boldsymbol{x}^T(\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}+\boldsymbol{\Sigma}_x^{-1})\boldsymbol{x}-2(\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}+\boldsymbol{x}_j^T\boldsymbol{\Sigma}_x^{-1})\boldsymbol{x}+\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}+\boldsymbol{x}_j^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j \tag{A.6}$$

and by comparing terms with Eq. A.5 we directly obtain

$$\boldsymbol{\Sigma}^{-1}=\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}+\boldsymbol{\Sigma}_x^{-1}, \tag{A.7}$$

which yields (4.26) (Note that $\boldsymbol{\Sigma}$ actually is symmetric as $\boldsymbol{\Sigma}_\epsilon$ and $\boldsymbol{\Sigma}_x$ are both symmetric). Furthermore, we have (by comparing terms in Eqs. A.6 and A.5)

$$\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}=\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}+\boldsymbol{x}_j^T\boldsymbol{\Sigma}_x^{-1}. \tag{A.8}$$

By taking the transpose

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j=\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}+\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j \tag{A.9}$$

and multiplicate from the left with $\boldsymbol{\Sigma}$, we obtain (4.27).

It now remains to calculate the weighting coefficients c_j. With Eqs. A.7 and A.8, A.6 becomes

$$\underbrace{(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)}_{\propto-\log(\Gamma_j)}\underbrace{-\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j+\boldsymbol{x}_j^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j+\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}}_{constant-2\log c_j}. \tag{A.10}$$

In this equation only terms depending on j are needed to calculate c_j, leaving only the two middle terms. All constant terms can flow into the global normalization constant c in Eq. 4.25, which is more conveniently determined by integration (Eq. 4.31, p. 68).

As above, we introduce two constants, the symmetrical matrix $\boldsymbol{\Psi}$ and $\boldsymbol{\mu}_x$, we rewrite the two terms in the middle of Eq. A.10 on the quadratic form (the same as Eq. 4.28)

$$(\boldsymbol{x}_j-\boldsymbol{\mu}_x)^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_x)=\boldsymbol{x}_j^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_j-2\boldsymbol{\mu}_x^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_j+\boldsymbol{\mu}_x^T\boldsymbol{\Psi}^{-1}\boldsymbol{\mu}_x, \tag{A.11}$$

and compare terms. For this, we need the expression $\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j$ (from Eq. A.8)

$$\begin{aligned}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j &= (\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}+\boldsymbol{x}_j^T\boldsymbol{\Sigma}_x^{-1})\boldsymbol{\Sigma}(\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}+\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j) &(A.12)\\ &= \tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}}+2\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j+\boldsymbol{x}_j^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j, &(A.13)\end{aligned}$$

with which, the last three terms in Eq. A.10 becomes

$$\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}(\boldsymbol{I}-\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1})\tilde{\boldsymbol{y}}-2\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j+\boldsymbol{x}_j^T\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{I}-\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1})\boldsymbol{x}_j. \tag{A.14}$$

Comparing with Eq. A.11 and only considering terms depending on j, yields

$$\boldsymbol{\Psi}^{-1}=\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{I}-\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1}), \tag{A.15}$$

or equivalently

$$\boldsymbol{\Psi}=(\boldsymbol{I}-\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1})^{-1}\boldsymbol{\Sigma}_x \tag{A.16}$$

which is the same expression as Eq. 4.29 For $\boldsymbol{\mu}_x$ we have

$$\boldsymbol{\mu}_x^T\boldsymbol{\Psi}^{-1}=\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1} \tag{A.17}$$

from which Eq. 4.30 follows. Finally, in Eq. A.11, we assumed that $\boldsymbol{\Psi}$ was symmetric. This is verified as follows:

$$\begin{aligned}\boldsymbol{\Psi}^T &= \boldsymbol{\Sigma}_x(\boldsymbol{I}-\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma})^{-1}\\ &= (\boldsymbol{\Sigma}_x^{-1})^{-1}(\boldsymbol{I}-\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma})^{-1}\\ &= ((\boldsymbol{I}-\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma})\boldsymbol{\Sigma}_x^{-1})^{-1}\\ &= (\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{I}-\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1}))^{-1}\\ &= (\boldsymbol{I}-\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1})^{-1}\boldsymbol{\Sigma}_x\\ &= \boldsymbol{\Psi}.\end{aligned} \tag{A.18}$$

A.2 Simplified Modified Mean Shift

We can obtain an expression for the modified mean shift $mms(\boldsymbol{x})$ that is computationally less expensive than the expression in Eq. A.1, and that is based on the standard mean shift term. For this we need to rewrite the expression $c_j\Gamma_j(\cdot)$ in Eq. A.1, so that we only keep all that is dependent on j. The rest can be eliminated from the fraction. We begin by writing

$$c_j\Gamma_j(\boldsymbol{x}) = \exp\left(-\frac{1}{2}\left[(\boldsymbol{x}_j-\boldsymbol{\mu}_x)^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_x)+(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right]\right). \tag{A.19}$$

Writing out the exponent yields the expression

$$\boldsymbol{x}_j^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_j - 2\boldsymbol{\mu}_x^T\boldsymbol{\Psi}^{-1}\boldsymbol{x}_j + \boldsymbol{\mu}_x^T\boldsymbol{\Psi}^{-1}\boldsymbol{\mu}_x + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j \tag{A.20}$$

from which we keep only the terms depending on j. Thus, substituting for $\boldsymbol{\mu}_j$ (from Eqs. A.8 and A.12) and leaving out the terms dependent on $\boldsymbol{x}$, yields

$$\boldsymbol{x}_j^T(\boldsymbol{\Psi}^{-1}+\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1})\boldsymbol{x}_j + 2(\tilde{\boldsymbol{y}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1} - \boldsymbol{x}^T\boldsymbol{\Sigma}_x^{-1} - \boldsymbol{\mu}_x^T\boldsymbol{\Psi}^{-1})\boldsymbol{x}_j, \tag{A.21}$$

which, with Eqs. A.15 and A.17, simplifies to

$$\boldsymbol{x}_j^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j - 2\boldsymbol{x}^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j. \tag{A.22}$$

We thus finally obtain

$$c_j\Gamma_j(\boldsymbol{x}) \propto \Theta_j(\boldsymbol{x}) = \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_j)^T\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{x}-\boldsymbol{x}_j)). \tag{A.23}$$

With $\Theta_j(\cdot)$ and $\boldsymbol{\mu}_j$, we can now rewrite the $mms(\boldsymbol{x})$ expression in Eq. A.1 as

$$\begin{aligned} mms(\boldsymbol{x}) &= \frac{\sum_{j=1}^J \Theta_j(\boldsymbol{x})\boldsymbol{\mu}_j}{\sum_{j=1}^J \Theta_j(\boldsymbol{x})} - \boldsymbol{x} \\ &= \frac{\sum_{j=1}^J \Theta_j(\boldsymbol{x})\boldsymbol{\Sigma}\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}} + \sum_{j=1}^J \Theta_j(\boldsymbol{x})\boldsymbol{\Sigma}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x}_j}{\sum_{j=1}^J \Theta_j(\boldsymbol{x})} - \boldsymbol{x}, \\ &= \Sigma\left(\boldsymbol{W}^T\boldsymbol{\Sigma}_\epsilon^{-1}\tilde{\boldsymbol{y}} + \boldsymbol{\Sigma}_x^{-1}\frac{\sum_{j=1}^J \Theta_j(\boldsymbol{x})\boldsymbol{x}_j}{\sum_{j=1}^J \Theta_j(\boldsymbol{x})}\right) - \boldsymbol{x} \end{aligned} \tag{A.24}$$

which is the same as Eq. 4.34, p. 69.

A.2.1 An alternative way of obtaining the simplified modified mean shift

The modified mean shift and the simplified version of the Modified Mean Shift obtained by rewriting the posterior probability Eq. A.3 into Eq. A.2, differentiating and eliminating terms from the quotient. An alternative way of obtaining the same expression as (A.24), is by deriving the composite posterior probability (A.3)

$$\arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}), \tag{A.25}$$

directly to find $\nabla p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$.

With

$$\nabla p_j(\boldsymbol{x}) = p_j(\boldsymbol{x}){\boldsymbol{\Sigma_x}}^{-1}(\boldsymbol{x} - \boldsymbol{x}_j) \tag{A.26}$$

and

$$\nabla p(\boldsymbol{y}|\boldsymbol{x}) = -p(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{W}^T{\boldsymbol{\Sigma_\epsilon}}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x}) \tag{A.27}$$

we have

$$\begin{aligned}
\nabla p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) &= p(\boldsymbol{x})\nabla p(\boldsymbol{y}|\boldsymbol{x}) + p(\boldsymbol{y}|\boldsymbol{x})\nabla p(\boldsymbol{x}) \\
&= p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\boldsymbol{W}^T{\boldsymbol{\Sigma_\epsilon}}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x}) - p(\boldsymbol{y}|\boldsymbol{x})\frac{1}{J}\sum_{j=1}^{J} p_j(\boldsymbol{x}){\boldsymbol{\Sigma_x}}^{-1}(\boldsymbol{x} - \boldsymbol{x}_j) \\
&= p(\boldsymbol{y}|\boldsymbol{x}) \left[p(\boldsymbol{x})\boldsymbol{W}^T{\boldsymbol{\Sigma_\epsilon}}^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{W}\boldsymbol{x}) + {\boldsymbol{\Sigma_x}}^{-1} \underbrace{\left[\frac{1}{J}\sum_{j=1}^{J} p_j(\boldsymbol{x})\right]}_{p(\boldsymbol{x})} \left[\frac{\sum_{j=1}^{J} p_j(\boldsymbol{x})\boldsymbol{x}_j}{\sum_{j=1}^{J} p_j(\boldsymbol{x})} - \boldsymbol{x}\right] \right] \\
&= p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \left[\boldsymbol{W}^T{\boldsymbol{\Sigma_\epsilon}}^{-1}\tilde{\boldsymbol{y}} + {\boldsymbol{\Sigma_x}}^{-1}\frac{\sum_{j=1}^{J} p_j(\boldsymbol{x})\boldsymbol{x}_j}{\sum_{j=1}^{J} p_j(\boldsymbol{x})} - \underbrace{(\boldsymbol{W}^T{\boldsymbol{\Sigma_\epsilon}}^{-1}\boldsymbol{W} + {\boldsymbol{\Sigma_x}}^{-1})}_{\boldsymbol{\Sigma}^{-1}}\boldsymbol{x} \right] \\
&= p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\boldsymbol{\Sigma}^{-1} \left[\boldsymbol{\Sigma}\left(\boldsymbol{W}^T{\boldsymbol{\Sigma_\epsilon}}^{-1}\tilde{\boldsymbol{y}} + {\boldsymbol{\Sigma_x}}^{-1}\frac{\sum_{j=1}^{J} p_j(\boldsymbol{x})\boldsymbol{x}_j}{\sum_{j=1}^{J} p_j(\boldsymbol{x})} \right) - \boldsymbol{x} \right],
\end{aligned} \tag{A.28}$$

which is the same as Eq. 4.34, p. 69.

A.3 Convergence Proof of the Modified Mean Shift

With the modified mean shift expression in Eq. A.1, the convergence of an optimization algorithm based on the modified mean shift can be proved in an analog manner to the original mean shift convergence proof [40][1]. In the following, we make use of the following notations and relations:

- The iteration index, i.
- The positive definite quadratic form:

  $$q_j(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu_j})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu_j}),$$

  where $\boldsymbol{\Sigma}$ is a non-singular, symmetric matrix with positive eigenvalues.
- The (normal) kernel:

  $$k(x) = \exp(-\frac{1}{2}x).$$

[1] This is actually the main advantage of Eq. A.1 over the simplified modified mean shift expression, Eq. A.28.

Since this kernel is convex, and since $k'(x) = -\frac{1}{2}k(x)$, the following relation holds:

$$k(x_2) - k(x_1) \geq \frac{1}{2}k(x_1)(x_1 - x_2), \tag{A.29}$$

for $x_1 \neq x_2$ and $x_1, x_2 \geq 0$.

- The updated subspace variable estimate is then given by (rewriting Eq. A.1):

$$\boldsymbol{x}_{i+1} = \frac{\sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_i))\boldsymbol{\mu}_j}{\sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_i))} = mms(\boldsymbol{x}_i) + \boldsymbol{x}_i, \tag{A.30}$$

which yields a sequence of posterior probabilities, $\{p(\boldsymbol{x}_i|\boldsymbol{y})\}_{i=1,2\ldots}$.

We can now prove the following theorem:

Theorem 1 *With the normal kernel, $k(x)$, the sequences $\{\boldsymbol{x}_i\}_{i=1,2\ldots}$ and $\{p(\boldsymbol{x}_i|\boldsymbol{y})\}_{i=1,2\ldots}$ converge, and $\{p(\boldsymbol{x}_i|\boldsymbol{y})\}_{i=1,2\ldots}$ is monotonically increasing.*

Proof

Since J is finite, the sequence $\{p(\boldsymbol{x}_i|\boldsymbol{y})\}_{i=1,2\ldots}$ is bounded (recall $p(\boldsymbol{x}|\boldsymbol{y})$ from Eq. A.2). It is therefore sufficient to show that $\{p(\boldsymbol{x}_i|\boldsymbol{y})\}_{i=1,2\ldots}$ is strictly monotonic increasing, i.e. if $\boldsymbol{x}_i \neq \boldsymbol{x}_{i+1}$, then

$$p(\boldsymbol{x}_i|\boldsymbol{y}) < p(\boldsymbol{x}_{i+1}|\boldsymbol{y}), \quad i = 1, 2 \ldots.$$

From Eq. A.2, using the above notation, we have

$$p(\boldsymbol{x}_{i+1}|\boldsymbol{y}) - p(\boldsymbol{x}_i|\boldsymbol{y}) = \frac{c}{J}\sum_{j=1}^{J} c_j \big[k(q_j(\boldsymbol{x}_{i+1})) - k(q_j(\boldsymbol{x}_i))\big]. \tag{A.31}$$

With Eq. A.29, we can rewrite this into

$$p(\boldsymbol{x}_{i+1}|\boldsymbol{y}) - p(\boldsymbol{x}_i|\boldsymbol{y}) \geq \frac{c}{2J}\sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_i))\big[q_j(\boldsymbol{x}_i) - q_j(\boldsymbol{x}_{i+1})\big]$$

$$= \frac{c}{2J}\sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_i))\big[\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j - \boldsymbol{x}_{i+1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_{i+1} + 2\boldsymbol{x}_{i+1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\big],$$

which with Eq. A.30 becomes,

$$\begin{aligned} p(\boldsymbol{x}_{i+1}|\boldsymbol{y}) - p(\boldsymbol{x}_i|\boldsymbol{y}) &\geq \frac{c}{2J}[\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_{i+1} - \boldsymbol{x}_{i+1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_{i+1} + \\ &\qquad + 2\boldsymbol{x}_{i+1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_{i+1}] \cdot \sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_i)) \\ &= \frac{c}{2J}\underbrace{(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)}_{\geq 0}\underbrace{\sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_i))}_{>0}. \end{aligned} \tag{A.32}$$

With the last product being strictly positive (recall that $\boldsymbol{x}_i \neq \boldsymbol{x}_{i+1}$ and $\boldsymbol{\Sigma}$ positive definite), we have that the sequence $\{p(\boldsymbol{x}_i|\boldsymbol{y})\}_{i=1,2\ldots}$ is monotonically increasing.

To show the convergence in the euclidean space of the sequence $\{\boldsymbol{x}_i\}_{i=1,2\ldots}$, it is sufficient to show that $||\boldsymbol{x}_{i+m} - \boldsymbol{x}_i||^2$ is bounded for $m \to \infty$. Summing successive iterations from Eq. A.32 yields:

$$\begin{aligned} p(\boldsymbol{x}_{i+m}|\boldsymbol{y}) - p(\boldsymbol{x}_i|\boldsymbol{y}) \geq \frac{c}{2J}(\boldsymbol{x}_{i+m} - \boldsymbol{x}_{i+m-1})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{i+m} - \boldsymbol{x}_{i+m-1})\sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_{i+m-1})) + \ldots \\ + \frac{c}{2J}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)\sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_i)) \\ \geq \frac{c}{2J}\big[(\boldsymbol{x}_{i+m} - \boldsymbol{x}_{i+m-1})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{i+m} - \boldsymbol{x}_{i+m-1}) + \ldots \\ + \frac{c}{2J}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)\big]M \\ \geq \frac{c}{2J}(\boldsymbol{x}_{i+m} - \boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{i+m} - \boldsymbol{x}_i)M, \end{aligned}$$

where M is the the minimum (always strictly positive) of the sum $\sum_{j=1}^{J} c_j k(q_j(\boldsymbol{x}_i))$ for all $\{\boldsymbol{x}_i\}_{i=1,2\ldots}$. Since $\{p(\boldsymbol{x}_i|\boldsymbol{y})\}_{i=1,2\ldots}$ is convergent, it follows that $(\boldsymbol{x}_{i+m} - \boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{i+m} - \boldsymbol{x}_i)$ is bounded. In particular, we have that for $\boldsymbol{\Sigma} = \boldsymbol{I}$, the euclidean distance, $||\boldsymbol{x}_{i+m} - \boldsymbol{x}_i||^2$, is bounded.

Appendix B

Small sample size and covariance matrix decomposition

How to calculate the principal components, or equivalently approximating the covariance matrix of high dimensional random variables, is repeatedly described in journal papers despite the fact that it has been known for quite some time. Fukunaga already described how to do this 14 years ago [72]. We recall this description and show that this method is actually the singular value decomposition of the sample matrix (which is today the more usual way to go).

From J observations, $\boldsymbol{y}_j$, we estimate the mean and the covariance matrix of the *random vector* $\boldsymbol{y}$

$$\hat{\boldsymbol{\mu}} = \frac{1}{J}\sum_j^J \boldsymbol{y}_j, \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{J-1}\sum_j^J (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}})^T.$$

Since $\hat{\boldsymbol{\Sigma}}$ is a function of J or less linearly independent vectors, its rank can only be J or less. Furthermore, because of the estimated mean vector $\hat{\boldsymbol{\mu}}$ we "loose" a degree of freedom and the maximum possible rank is $(J-1)$. If $J < D$, where D is the dimension of the observation space, $\hat{\boldsymbol{\Sigma}}$ is singular. Furthermore, since the covariance matrix is of the size $D \times D$ it becomes computationally unhandly, if not untractable for large D ($D >> J$).

Some carefulness is therefore necessary to solve the eigenproblem. Fukunaga [72] proposes to invert the roles of variables and samples. Let $\boldsymbol{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_J]$ be the *sample matrix* and $\bar{\boldsymbol{Y}}$ the mean free sample matrix. Then we can write $\hat{\boldsymbol{\Sigma}} = \frac{1}{J-1}\bar{\boldsymbol{Y}}\bar{\boldsymbol{Y}}^T$. Instead of using $\hat{\boldsymbol{\Sigma}}$, we can calculate the eigenvectors, $\boldsymbol{v}_q$, and eigenvalues, λ_j, of $\frac{1}{J-1}\bar{\boldsymbol{Y}}^T\bar{\boldsymbol{Y}}$, which is only $J \times J$

$$\frac{1}{J-1}(\bar{\boldsymbol{Y}}^T\bar{\boldsymbol{Y}})\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda} = [\boldsymbol{v}_1 \dots \boldsymbol{v}_J] \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_{J-1} \end{bmatrix},$$

where $\boldsymbol{V}$ is the matrix of eigenvectors. Multiplying from the left with $\bar{\boldsymbol{Y}}$ yields

$$\frac{1}{J-1}(\bar{\boldsymbol{Y}}\bar{\boldsymbol{Y}}^T)(\bar{\boldsymbol{Y}}\boldsymbol{V}) = (\bar{\boldsymbol{Y}}\boldsymbol{V})\boldsymbol{\Lambda}.$$

Thus, $(\bar{\boldsymbol{Y}}\boldsymbol{V})$ and $\boldsymbol{\Lambda}$ are the $(J-1)$ eigenvectors and eigenvalues of $\hat{\boldsymbol{\Sigma}}$. The other $(D-J+1)$ eigenvalues are all zero and their eigenvectors are indefinite.

The matrix $(\bar{\boldsymbol{Y}}\boldsymbol{V})$ represents orthogonal vectors. To obtain orthonormal ones, we have to divide each column vector of $(\bar{\boldsymbol{Y}}\boldsymbol{V})$ by $((J-1)\lambda_j)^{1/2}$ yielding:

$$\boldsymbol{U} = \frac{1}{\sqrt{J-1}}\bar{\boldsymbol{Y}}\boldsymbol{V}\boldsymbol{\Lambda}^{-1/2}. \tag{B.1}$$

We can validate this as follows:

$$\boldsymbol{U}^T\boldsymbol{U} = \frac{1}{J-1}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{V}^T\bar{\boldsymbol{Y}}^T\bar{\boldsymbol{Y}}\boldsymbol{V}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{I}.$$

Rewriting Eq. B.1, we can express $\bar{\boldsymbol{Y}}$ as

$$\frac{1}{\sqrt{J-1}}\bar{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}^T$$

which is exactly the SVD of $\frac{1}{\sqrt{J-1}}\bar{\boldsymbol{Y}}$ (or more precisely what Golub *et al.* call *thin* SVD) [82]. We have found that the SVD algorithm, based on the one-sided Jacobi orthogonalization ([74]), gives the most precise results.

Appendix C

Peer-reviewed publications by the author

International journals

T. Vik, F. Heitz, and J.-P. Armspach. On the modeling, construction and evaluation of a probabilistic atlas of brain perfusion. *NeuroImage*, 2004, 14 pages, accepted for publication.

International conferences with proceedings

T. Vik, F. Heitz, and J.-P. Armspach. Statistical atlas-based detection of abnormalities in brain perfusion: Comparing models and estimating detection performance. In R. E. Ellis and T. M. Peters, editors, *Int. Conf. on Medical Image Computing & Computer Assisted Intervention 2003*, Lecture Notes in Computer Science 2879, pages 838-845, Toronto, Canada, November 2003.

M. Bosc, T. Vik, J.-P. Armspach, and F. Heitz. ImLib3D: An efficient, open source, medical image processing framework in C++. In R. E. Ellis and T. M. Peters, editors, *Int. Conf. on Medical Image Computing & Computer Assisted Intervention 2003*, Lecture Notes in Computer Science 2879, pages 981-982, Toronto, Canada, November 2003.

T. Vik, F. Heitz, and P. Charbonnier. Mean shift-based bayesian image reconstruction into visual subspace. In *Proceedings of the 2003 International Conference on Image Processing (ICIP 2003)*, Barcelona, Spain, September 2003.

Bibliography

[1] P. Acton and K. Friston. Statistical parametric mapping in functional neuroimaging: beyond PET and fMRI activation studies. *European Journal of Nuclear Medicine*, 25(7):663–667, July 1998. Editorial.

[2] D. G. Amen. *Images Into Human Behavior: A Brain SPECT Atlas*. Mindworks Press, nd. `http://www.brainplace.com/bp/atlas/default.asp`.

[3] D. G. Amen, J. C. Wu, and B. Carmichael. The clinical use of brain SPECT imaging in neuropsychiatry. *Alasbimn Journal*, 5(19), January 2003.

[4] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 1984.

[5] P. Bartenstein, S. Minoshima, C. Hirsch, K. Buch, F. Willoch, D. Mösch, D. Schad, M. Schwaiger, and A. Kurz. Quantitative assessment of cerebral blood flow in patients with Alzheimer's disease by SPECT. *The Journal of Nuclear Medicine*, 38(7):1095–1101, July 1997.

[6] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *IEEE Trans. Neural Networks*, 13:1450–1464, 2002.

[7] C. F. Beckmann and S. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imag.*, 23(2):137–152, February 2004.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(7):711–720, July 1997.

[9] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions. *International Journal of Computer Vision*, 28(3):245–260, March 1998.

[10] H. Benali, I. Buvat, F. Frouin, J. P. Bazin, and R. Di Paola. A statistical model for the determination of the optimal metric in factor analysis of medical image sequences (FAMIS). *Phys. Med. Biol.*, 38:1065–1080, 1993.

[11] C. Bishop and M. Tipping. A hierarchical latent variable model for data visualization. *IEEE Trans. Pattern Anal. Machine Intell.*, 20(3):281–293, 1998.

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

[13] M. J. Black and A. D. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[14] M. J. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1):57–91, 1996.

[15] D.A. Boas, D.H. Brooks, E.L. Miller, C.A. DiMarzio, M. Kilmer, and R.J. Gaudette. Imaging the body with diffuse optical tomography. *IEEE Signal Processing Magazine*, pages 57–75, November 2001.

[16] M. Borga, T. Landelius, and H. Knutsson. A unified approach to PCA, PLS, MLR and CCA. Report LiTH-ISY-R-1992, ISY, Linköping University, Sweden, SE-581 83 Linköping, Sweden, November 1997.

[17] M. Bosc. *Contribution à la détection de changements dans des séquences IRM 3D multimodales*. PhD thesis, Université Louis Pasteur I, Strasbourg, 2003.

[18] M. Bosc, F. Heitz, J.-P. Armspach, I. Namer, D. Gounot, and L. Rumbach. Automatic change detection in multi-modal serial MRI: Application to multiple sclerosis lesion evolution. *NeuroImage*, 20(2):643–656, October 2003.

[19] M. Bosc, T. Vik, J.-P. Armspach, and F. Heitz. ImLib3D: An efficient, open source, medical image processing framework in C++. In R. E. Ellis and T. M. Peters, editors, *MICCAI 2003*, LNCS 2879, pages 981–982, November 2003.

[20] P. Bourdieu. *La distinction : Critique sociale du jugement*. Les Editions de Minuit, 1979.

[21] N. Boussion, C. Houzard, K. Ostrowsky, P. Ryvlin, F. Maugière, and L. Cinotti. Automated detection of local normalization areas for ictal-interictal subtraction brain SPECT. *The Journal of Nuclear Medicine*, 43:1419–1425, 2002.

[22] N. Boussion, P. Ryvlin, J. Isnard, C. Houzard, F. Maugière, and L. Cinotti. Towards an optimal reference region in single photon emission tomography difference images in epilepsy. *European Journal of Nuclear Medicine*, 27:155–160, 2000.

[23] K. W. Bowyer. Validation of medical image analysis techniques. In M. Sonka and J. M. Fitzpatrick, editors, *Handbook of Medical Imaging*, volume 2, pages 567–607. SPIE Press, 2000.

[24] L. Le Briquer and J. C. Gee. Design of a statistical model of brain shape. In J. S. Duncan and G. R. Gindi, editors, *XVth Int. Conf. on Information Processing in Medical Imaging*, pages 9–13. Springer-Verlag, Heidelberg, 1997.

[25] P. P. Bruyant. Analytic and iterative reconstruction algorithms in SPECT. *The Journal of Nuclear Medicine*, 43(10):1343–1358, October 2002.

[26] I. Buvat. Toward a standardization of the description and validation of monte carlo simulation codes in spect and pet. *The Journal of Nuclear Medicine*, 43:207P, 2002.

[27] A. Cachia, J.-F. Mangin, D. Rivière, F. Kherif, N. Boddaert, A. Andrade, D. Papadopoulos-Orfanos, J.-B. Poline, I. Bloch, M. Zilbovicius, P. Sonigo, F. Brunelle, and J. Régis. A primal sketch of the cortex mean curvature: a morphogenesis based approach to study the variability of the folding patterns. *IEEE Trans. Med. Imag.*, 22(6):754–765, 2003.

[28] R. Cappelli, D. Maio, and D. Maltoni. Multispace KL for pattern representation and classification. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(9), September 2001.

[29] M. Carbon, M. Ghilardi, A. Feigin, M. Fukuda, G. Silvestri, M. Mentis, C. Ghez, J. Moeller, and D. Eidelberg. Learning networks in health and Parkinsons's disease: Reproducibility and treatment effects. *Human Brain Mapping*, 11:197–211, 2003.

[30] M. Carreira-Perpiñán. A review of dimension reduction techniques. Technical report, Dept. of Computer Science, University of Sheffield, January 1997.

[31] B. Chalmond and S. C. Girard. Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Trans. Pattern Anal. Machine Intell.*, 21(5):422–432, May 1999.

[32] D. Chang, I. Zubal, C. Gottschalk, A. Necochea, R. Stokking, C. Studholme, M. Corsi, J. Slawski, S. Spencer, and H. Blumenfeld. Comparison of statistical parametric mapping and SPECT difference imaging in patients with temporal lobe epilepsy. *Epilepsia*, 43(1):68–74, 2002.

[33] K.-Y. Chang and J. Ghosh. A unified model for probabilistic principal surfaces. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(1):22–41, January 2001.

[34] P. Charbonnier. *Reconstruction d'image : régularisation avec prise en compte des discontinuités.* PhD thesis, Université de Nice-Sophia Antipolis, France, 1994.

[35] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Two deterministic half quadratic regularization algorithms for computed imaging. In *IEEE International Conference on Image Processing*, pages 168–172, Austin, USA, 1994.

[36] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6(2):298–311, 1997.

[37] Z.-H. Cho, J. P. Jones, and M. Singh. *Foundations of Medical Imaging.* John Wiley & Sons, 1993.

[38] G. Christensen and H. Johnson. Consistent image registration. *IEEE Trans. Med. Imag.*, 20(7):568–582, July 2001.

[39] D. Comaniciu. An algorithm for data-driven bandwidth selection. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(2):1–8, February 2003.

[40] D. Comaniciu and P. Meer. Mean Shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(5):603–619, May 2002.

[41] P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36(3):287–314, April 1994.

[42] DuPont Merck Pharmaceutical Company. Neurolite®: Kit for the preparation of Technetium Tc99m BICISATE. Product Monograph., November 1993. Summarizes studies undergone in the approval process of bicisate.

[43] D. C. Costa and K. Schmidt. Can ROI methodology/normalised tissue activities be used instead of absolute blood flow measurements in the brain? For and Against. *European Journal of Nuclear Medicine*, 29(7):948–956, 2002.

[44] F. Crivello, T. Schormann N. Tzourio-Mazoyer, P. E. Roland, K. Zilles, and B. M. Mazoyer. Comparison of spatial normalization procedures and their impact on functional maps. *Human Brain Mapping*, 16:228–250, 2002.

[45] M. Dahlbom and S. Huang. Physical and biological bases of spatial distortions in positron emission tomography images. In I. N. Bankmann, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 439–447. Academic Press, 2000.

[46] R. Dahyot. *Appearance based road scene video analysis for the management of the road network (in french).* PhD thesis, Université Louis Pasteur - Strasbourg I, France, 2001.

[47] R. Dahyot, P. Charbonnier, and F. Heitz. Robust visual recognition of colour images. In *IEEE Int. Conf. Computer Vision and Pattern Recognition*, volume 1, pages 685–690, June 2000. CVPR 2000, Hilton Head Island, USA.

[48] R. Dahyot, P. Charbonnier, and F. Heitz. Robust bayesian detection using appearance-based models. *Pattern Analysis and Applications*, 2004. Submitted, 1st revision.

[49] A. Dale, B. Fischl, and M. Sereno. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, 9:179–194, 1999.

[50] C. Davatzikos, H. H. Li, E. Herskovits, and S. M. Resnick. Accuracy and sensitivity of detection of activation foci in the brain via statistical parametric mapping: A study using a PET simulator. *NeuroImage*, 13:176–184, 2001.

[51] P. de Groen. An introduction to total least squares. *Nieuw Archief voor Wiskunde*, 4(14):237–253, 1996.

[52] F. de la Torre and M. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1/2/3):117–142, 2003.

[53] F. de la Torre and M. J. Black. Robust principal component analysis for computer vision. In *Proceedings of the International Conference on Computer Vision*, July 2001. Vancouver, Canada.

[54] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. Neural Networks*, 8(1):148–154, January 1997.

[55] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[56] B. Draper, K. Baek, M. Bartlett, and J. Beveridge. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding*, 91(1–2):115–137, July 2003.

[57] C. Drexler, F. Mattern, and J. Denzler. Appearance based generic object modeling and recognition using probabilistic principal component analysis. In *DAGM-Symposium*, pages 100–108, 2002.

[58] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, 2nd edition, 2001.

[59] H. M. Duvernoy. *The Human Brain: Surface, Blood Supply, and Three-Dimensional Sectional Anatomy*. Springer-Verlag, second edition, 1999.

[60] K. P. Ebmeier, M. F. Glabus, N. Prentice, A. Ryman, and G. M. Goodwin. A voxel-based analysis of cerebral perfusion in dementia and depression of old age. *NeuroImage*, 7:199–208, 1998.

[61] B. Fischl, M. Sereno, and A. Dale. Cortical surface-based analysis. II. Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9:195–207, 1999.

[62] B. J. Frey, A. Colmenarez, and T. S. Huang. Mixtures of local linear subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 1998*. IEEE Computer Society Press: Los Alamitos, CA, 1998.

[63] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Computers*, C-23:881–889, 1974.

[64] O. Friman, M. Borga, P. Lundberg, and H. Knutsson. Exploratory fMRI analysis by autocorrelation maximization. *NeuroImage*, 16(2):454–464, June 2002.

[65] K. Friston, A. Holmes, and K. Worsley. Comments and controversies: How many subjects constitute a study? *NeuroImage*, 10:1–5, 1999.

[66] K. Friston, J. Phillips, D. Chawla, and C. Büchel. Revealing interactions among brain systems with nonlinear PCA. *Human Brain Mapping*, 8:92–97, 1999.

[67] K. J. Friston et al. SPM course notes 1997, chapter 1, Data analysis: Basic concepts and overview. `http://www.fil.ion.ucl.ac.uk/spm/course/notes02/overview/Refs.htm`, 1997.

[68] K. J. Friston, C. D. Frith, P. F. Liddle, R. J. Dolan, A. A. Lammertsmaa, and R. S. J. Frackowiak. The relationship between global and local changes in PET scans. *J Cereb Blood Flow Metab*, 1990.

[69] K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

[70] K. J. Friston, J.-P. Poline, S. Strother, A. P. Holmes, C. D. Frith, and R. S. J. Frackowiak. A multivariate analysis of PET activation studies. *Human Brain Mapping*, 4:140–151, 1996.

[71] K. J. Friston, C. J. Price, and C. Büchel. Designing activation experiments. In B. Gulyas and H.W. Miller-Gartner, editors, *Positron Emission Tomography: A Critical Assessment of Recent Trends*. Kluwer Academic, 1998.

[72] K. Fukunaga. *Statistical Pattern Recognition*. Academic Press, 2 edition, 1990.

[73] K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory*, 21:32–40, 1975.

[74] M. Galassi et al. GNU Scientific Library (GSL). World Wide Web. A collection of routines for numerical computing. Available at http://sources.redhat.com/gsl/.

[75] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns. Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1995.

[76] A. Gardner and M. Pagani. A review: SPECT in neuropsychiatric disorders: neurobiological background, methodology, findings and future perspectives. *Alasbimn Journal*, 5(21), July 2003.

[77] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Machine Intell.*, 14(3):367–383, March 1992.

[78] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization and FFT's. *IEEE Trans. Image Processing*, 4(7):932–946, 1995.

[79] P. G.Hoel. *Introduction to Mathematical Statistics*. John Wiley & Sons, 1954.

[80] F. Girosi, T. Poggio, and B. Caprile. Extensions of a theory of networks for approximation and learning: Outliers and negative examples. In R. Lippmann, J. Moody, and D. Touretzkey, editors, *Proceedings Neural Information Processing Society Conference*. Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[81] J. D. Gispert, J. Pascau, S. Reig, R. Martínez-Lázaro, V. Molina, P. García-Barreno, and M. Desco. Influence of the normalization template on the outcome of statistical parametric mapping of PET scans. *NeuroImage*, 19:601–612, 2003.

[82] G. H. Golub and C. F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

[83] T. Greitz, C. Bohm, S. Holte, and L. Eriksson. A computerized brain atlas: Construction, anatomical content, and some applications. *J Comput Assist Tomogr*, 15(1):26–38, 1991.

[84] C. Grova, A. Biraben, J.-M. Scarabin, P. Jannin, I. Buvat, H. Benali, and B. Gibaud. A methodology to validate MRI/SPECT registration methods using realistic simulated SPECT data. In W. Niessen and M. Viergever, editors, *MICCAI 2001*, LNCS 2208, pages 275–282, 2001.

[85] A. Guimond, J. Meunier, and J.-P. Thirion. Average brain models: A convergence study. Technical Report 3731, Institut National de Recherche en Informatique et en Automatique, Sophia-Antipolis, France, July 1999.

[86] P. Hall and K.-C. Li. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21(2):867–889, 1993.

[87] R. Hamdan. *Détection, suivi et reconnaissance des formes et du mouvement par modèles probabilistes d'apparence*. PhD thesis, Université Louis Pasteur - Strasbourg I, France, 2001.

[88] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, April 1982.

[89] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843, September 1983.

[90] P. Hannequin and J. Mas. Statistical and heuristic image noise extraction (SHINE): a new method for processing Poisson noise in scintigraphic images. *Phys. Med. Biol.*, 47:4329–4344, 2002.

[91] L. Hansen, J. Larsen, F. Nielsen, S. Strother, E. Rostrup, R. Savoy, N. Lange, J. Sidtis, C. Svarer, and O. Paulson. Generalizable patterns in neuroimaging: How many principal components. *NeuroImage*, 9:534–544, 1999.

[92] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[93] P. Hellier. *Recalage non rigide en imagerie cérébrale : méthodes et validation.* PhD thesis, Université de Rennes I, 2000.

[94] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D.L. Collins, A. Evans, G. Malandain, N. Ayache, G.E. Christensen, and H.J. Johnson. Retrospective evaluation of inter-subject brain registration. *IEEE Transactions on Medical Imaging*, 22(9):1120–1130, 2003.

[95] M. Van Herk. Image registration using chamfer matching. In I. N. Bankmann, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 515–527. Academic Press, 2000.

[96] D. Hill and D. Hawkes. Across-modality registration using intensity-based cost functions. In I. N. Bankmann, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 537–553. Academic Press, 2000.

[97] G. E. Hinton, P. Dayan, and M. Revow. Modelling the manifolds of images of handwritten digits. *IEEE trans. on Neural Networks*, 8(1):65–74, 1997.

[98] H. Hoffmann and J.-P. Krischewski. *Gegenstandskatalog 2: Radiologie.* Georg Thieme Verlag, 12 edition, 1999.

[99] B. L. Holman, P. K. Chandak, and B. Garada. Atlas of brain perfusion SPECT. Internet: `http://brighamrad.harvard.edu/education/online/BrainSPECT/index.html`, june 1998. Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston.

[100] N. Honda, K. Machida, T. Matsumoto, et al. Three-dimensional stereotactic surface projection of brain perfusion SPECT improves diagnosis of Alzheimer's disease. *Annals of Nuclear Medecine*, 17(8):641–648, 2003.

[101] J. Hoppin, M. Kupinski, G. Kastis, E. Clarkson, and H. Barrett. Objective comparison of quantitative imaging modalities without the use of a gold standard. *IEEE Trans. Med. Imag.*, 21(5):441–449, May 2002.

[102] K. Hosoda, T. Kawaguchi, K. Ishii, S. Minoshima, Y. Shibata, M. Iwakura, S. Ishiguro, and E. Kohmura. Prediction of hyperperfusion after carotid endarterectomy by brain SPECT analysis with semiquantitative statistical mapping method. *Stroke*, 34:1187–1193, 2003.

[103] G. N. Hounsfield. A method of and apparatus for examination of a body by radiation such as X-ray or gamma radiation, London (1972). British Patent No: 1283915.

[104] A. S. Houston, P. M. Kemp, and M. A. Macleod. A method for assessing the significance of abnormalities in HMPAO brain SPECT images. *The Journal of Nuclear Medicine*, 35(2):239–244, February 1994.

[105] A. S. Houston, P. M. Kemp, M. A. Macleod, T. James, R. Francis, H. A. Colohan, and H. P. Matthews. Use of significance image to determine patterns of cortical blood flow abnormality in pathological and at-risk groups. *The Journal of Nuclear Medicine*, 39(3):425–430, March 1998.

[106] P.J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[107] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[108] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13((4-5)):411–430, 2000.

[109] M. B. Imran, R. Kawashima, K. Sato, S. Kinomura, H. Ito, M. Koyama, R. Goto, S. Ono, S. Yoshioka, and H. Fukuda. Mean regional cerebral blood flow images of normal subjects using Technetium-99m-HMPAO by automated image registration. *The Journal of Nuclear Medicine*, 39(1):203–207, January 1998.

[110] A. K. Jain, R. P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(1):4–37, January 2000.

[111] C. Janicki, J. Meunier, J.-P. Soucy, B. Imbert, and A. Guimond. A 3-D non-linear registration algorithm for functional SPECT brain imaging using optical flow. In *Proceedings of the 44th Annual Meeting*. Society of Nuclear Medicine, June 1997.

[112] P. Jannin, J. M. Fitzpatrick, D. J. Hawkes, X. Pennec, R. Shahidi, and M. W. Vannier. Validation of medical image processing in image-guided therapy. *IEEE Trans. Med. Imag.*, 21(12):1445–1449, December 2002. Editorial.

[113] M. Jogan and A. Leonardis. Parametric eigenspace representations of panoramic images. In *International Conference on Advanced Robotics 2001 - Omnidirectional Vision Applied to robotic orientation and nondestructive testing (NDT)*, pages 31–36, Budapest, Hungary, August 2001. IEEE Computer Society.

[114] W. Johnson, J. Gastwirth, and L. Pearson. Screening without a "Gold Standard": The Hui-Walter paradigm revisited. *Am J Epidemiol*, 153(9):921–924, 2001.

[115] K. Jones, K. Johnson, J. Becker, P. Spiers, M. Albert, and B. Holman. Use of singular value decomposition to characterize age and gender differences in SPECT cerebral perfusion. *The Journal of Nuclear Medicine*, 39(6):965–973, June 1998.

[116] T. Jones. Quantification of brain function with PET. *NeuroImage*, 2(2):S1–S5, June 1995.

[117] C. Jutten and J. Hérault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[118] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrica*, 23:187–200, 1958.

[119] N. Kambhatla and T. K. Leen. Fast non-linear dimension reduction. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing systems 6*, San Fransisco, CA, 1994. Morgan Kaufmann Publishers.

[120] N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural Computation*, 9:1493–1516, 1997.

[121] H. Kamiya and S. Eguchi. A class of robust principal component vectors. *Journal of Multivariate Analysis*, 77:239–269, 2001.

[122] K. W. Kang, D. S. Lee, J. H. Cho, et al. Quantification of F-18 FDG PET images in temporal lobe epilepsy patients using probabilistic brain atlas. *NeuroImage*, 14:1–6, 2001.

[123] T. Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[124] P. M. Koulibaly. *Régularisation et corrections physiques en tomographie d'émission*. PhD thesis, Université de Nice-Sophia Antipolis, 1996.

[125] M. Kupinski, J. Hoppin, E. Clarkson, H. Barrett, and G. Kastis. Estimation in medical imaging without a gold standard. *Acad Radiol*, 9(3):290–297, March 2002.

[126] K. Van Laere, M. Koole, J. Versijpt, S. Vandenberghe, B. Vandenberghe, B. Brans, Y. D'Asseler, O. De Winter, and R. A. Dierckx. Transfer of normal ^{99m}Tc-ECD brain SPET databases between different gamma cameras. *European Journal of Nuclear Medicine*, 28:435–449, 2001.

[127] K. Van Laere, J. Versijpt, K. Audenaert, M. Koole, I. Goethals, E. Achten, and R. Dierckx. 99mTc-ECD brain perfusion SPET: variability, asymmetry and effects of age and gender in healthy adults. *European Journal of Nuclear Medicine*, 28(7):873–887, July 2001.

[128] K. Van Laere, J. Versijpt, M. Koole, S. Vandenberghe, P. Lahorte, I. Lemahieu, and R. A. Dierckx. Experimental performance assessment of SPM for SPECT neuroactivation studies using a subresolution sandwich phantom design. *NeuroImage*, 16:200–216, 2002.

[129] K. Van Laere, J. Warwick, J. Versijpt, I. Goethals, K. Audenaert, B. Heerden, and R. Dierckx. 99mTc-ECD brain perfusion SPET: variability, asymmetry and effects of age and gender in healthy adults. *The Journal of Nuclear Medicine*, 43(4):458–469, April 2002.

[130] P. Lahorte, S. Vandenberghe, K. Van Laere, K. Audenaert, I. Lemahieu, and R. A. Dierckx. Assessing the performance of SPM analyses of SPECT neuroactivation studies. *NeuroImage*, 12:757–764, 2000.

[131] J.-F. Laliberté, J. Meunier, M. Mignotte, and J.-P. Soucy. Detection of abnormal diffuse perfusion in SPECT using a normal brain atlas. In *SPIE Conference on Medical Imaging, 5032-04*, February 2003.

[132] J. Lancaster and P. Fox. Talairach space as a tool for intersubject standardization in the brain. In I. N. Bankman, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 555–567. Academic Press, 2000.

[133] M. LeBlanc and R. Tibshirani. Adaptive principal surfaces. *Journal of the American Statistical Association*, 89(425):53–64, 1994.

[134] D. S. Lee, S. K. Lee, and M. C. Lee. Functional neuroimaging in epilepsy: FDG PET and ictal SPECT. *J Korean Med Sci*, 16:689–96, 2001.

[135] J. D. Lee, H.-J. Kim, et al. Evaluation of ictal brain SPET using statistical parametric mapping in temporal lobe epilepsy. *European Journal of Nuclear Medicine*, 27(11):1658–1665, November 2000.

[136] F. Leibovitch, S. Black, C. Caldwell, A. McIntosh, L. Ehrlich, and J. Szalai. Brain SPECT imaging and left hemispatial neglect covaried using partial least squares: The sunnybrook stroke study. *Human Brain Mapping*, 7:244–253, 1999.

[137] M. Lennon. *Méthodes d'analyse d'images hyperspectrales*. PhD thesis, Université de Rennes I, 2002.

[138] A. Leonardis and H. Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78:99–118, 2000.

[139] A. Leonardis, H. Bischof, and J. Maver. Multiple eigenspaces. *Pattern Recognition*, 35:2613–2627, 2002.

[140] A. Leow, C. L. Yu, S. J. Lee, S. C. Huang, H. Protas, R. Nicolson, K. M. Hayashi, A.W. Toga, and P. M. Thompson. Brain structural mapping using a novel hybrid implicit/explicit framework based on the level-set method. *NeuroImage*, 2004. In Press.

[141] J. Léveillé, G. Demonceau, M. D. Roo, P. Rigo, et al. Characterization of Technetium-99m-L,L-ECD for brain perfusion imaging, part 2: Biodistribution and brain imaging in humans. *The Journal of Nuclear Medicine*, 30(11):1902–1910, November 1989.

[142] K. Lorenz. *Das sogenannte Böse*. Deutsche Taschenbuch Verlag, 1998.

[143] A. Machado and J. Gee. Atlas warping for brain morphometry. In *Proceedings of the SPIE Medical Imaging 1998: Image Processing, San Diego, Bellingham*, 1998.

[144] D. J. C. MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research, Section A*, 354(1):73–80, 1995.

[145] J. Maintz and M. Viergever. A survey of medical image registration. *Med. Image Anal.*, 2(1):1–37, 1998.

[146] J.-F. Mangin, D. Rivière, O. Coulon, C. Poupon, A. Cachia, Y. Cointepas, J.-B. Poline, D. Le Bihan, J. Régis, and D. Papadopoulos-Orfanos. Coordinate-based versus structural approaches to brain image analysis. *Artificial Intelligence in Medicine*, 30:177–197, 2004.

[147] J.-F. Mangin, D. Rivière, A. Cachia, Y. Cointepas E. Duchesnay, D. Papadopoulos-Orfanos, D. L. Collins, A. C. Evans, and J. Régis. Object-based morphometry of the cerebral cortex. *IEEE Trans. Med. Imag.*, 23(8):968–982, August 2004.

[148] K.V. Mardia. Tests of univariate and multivariate normality. In P.R. Krishnaiah, editor, *Handbook of Statistics 1: Analysis of Variance*, pages 279–320. North-Holland, 1980.

[149] A. Martínez and A. Kak. PCA versus LDA. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(2):228–233, 2001.

[150] M. McKeown, S. Makeig, G. Brown, T.-B. Jung, S. Kindermann, A. Bell, and T. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6:1–6, 1998.

[151] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[152] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6(1):59–70, 1991.

[153] O. Migneco, M. Benoit, P. Koulibaly, I. Dygai, et al. Perfusion brain SPECT and statistical parametric mapping analysis indicate that apathy is a Cingulate syndrome: A study in Alzheimer's disease and nondemented patients. *NeuroImage*, 13:896–902, May 2001.

[154] M. Mignotte, J. Meunier, J.-P. Soucy, and C. Janicki. Segmentation and classification of brain SPECT images using 3D Markov random field and density mixture estimations. In *Concepts and Applications of Systemics and Informatics*, volume X, pages 239–244, Orlando, Florida, USA, July 2001. 5th World Multi-Conference on Systemics, Cybernetics and Informatics, SCI'01.

[155] T. Minka. Independence diagrams. Tutorial notes. `http://www.stat.cmu.edu/~minka/papers/diagrams.html`.

[156] T. P. Minka. Automatic choice of dimensionality for PCA. Technical report, MIT, Media Laboratory Perceptual Computing Section, May 2000.

[157] S. Minoshima, K. Frey, R. Koeppe, N. Foster, and D. Kuhl. A diagnostic approach in Alzheimer's disease using three-dimensional stereotactic surface projections of Fluorine-18-FDG PET. *The Journal of Nuclear Medicine*, 36(7):1238–1248, July 1995.

[158] J. Missimer, U. Knorr, R. P. Maguire, H. Herzog, R. J. Seitz, L. Tellman, and K. L. Leenders. On two methods of statistical image analysis. *Human Brain Mapping*, 8:245–258, 1999.

[159] J. R. Moeller and D. Eidelberg. Divergent expression of regional metabolic topograhpies in Parkinson's disease and normal ageing. *Brain*, 120:2197–2206, 1997.

[160] J. R. Moeller, T. Nakamura, M. J. Mentis, V. Dhawan, P. Spetsiers, A. Antonini, J. Missimer, K. L. Leenders, and D. Eidelberg. Reproducibility of regional metabolic covariance patterns: Comparison of four populations. *The Journal of Nuclear Medicine*, 40(8):1264–1269, August 1999.

[161] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(6):780–788, June 2002.

[162] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(7):696–710, July 1997.

[163] H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[164] K. Murphy. An introduction to graphical models. Tutorial notes. `http://www.ai.mit.edu/~murphyk/Bayes/bnintro.html`.

[165] O. Musse. *Contribution à la mise en correspondance non rigide d'images médicales : une approche paramétrique hiérarchique sous contraintes topologiques. Application au recalage déformable du cerveau en imagerie IRM.* PhD thesis, Université Louis Pasteur, Strasbourg I, 2000.

[166] O. Musse, F. Heitz, and J.P. Armspach. Topology preserving deformable image matching using constrained hierarchical parametric models. *IEEE Trans. Im. Proc.*, 10(7):1081–1093, 2001.

[167] T. Nakamura, M. Ghilardi, M. Mentis, V. Dhawan, M. Fukuda, A. Hacking, J. Moeller, C. Ghez, and D. Eidelberg. Functional networks in motor sequence learning: Abnormal topographies in Parkinson's disease. *Human Brain Mapping*, 12:42–60, 2001.

[168] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). Technical report, Columbia University, 1996.

[169] F. Nielsen. Bibliography on neuroinformatics. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, 2004.

[170] M. Nikolova and M. K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. Technical Report 15, CMLA, ENS de Cachan, France, 2003.

[171] C. Nikou. *Contribution au recalage d'images médicales multimodales : approches par fonctions de similarité robustes et modèles déformables sous contraintes statistiques.* PhD thesis, Université Louis Pasteur, Strasbourg I, 1999.

[172] V. Noblet, C. Heinrich, F. Heitz, and J.-P. Armspach. A topology preserving method for 3-D non-rigid brain image registration. In R. E. Ellis and T. M. Peters, editors, *MICCAI 2003*, LNCS 2879, pages 977–978, November 2003.

[173] V. Noblet, C. Heinrich, F. Heitz, and J.-P. Armspach. Topology preserving non-rigid registration method using a symmetric similarity function - Application to 3-D brain images. In *8th European Conference on Computer Vision (ECCV), Prague*, April 2004.

[174] V. Noblet, C. Heinrich, F. Heitz, and J.-P. Armspach. 3–D deformable image registration: a topology preserving scheme based on hierarchical deformation models and interval analysis optimization. *IEEE Trans. Image Processing*, 2005. in press.

[175] T. O'Brien, E. So, B. Mullan, M. Hauser, B. Brinkmann, N. Bohnen, D. Hanson, G. Cascino, G. Cascino, and F. Sharbrough. Subtraction SPECT co-registered to MRI improves postictal SPECT localization of seizure foci. *Neurology*, 50(2):445–54, February 1998.

[176] T. J. O'Brien, E. L. So, B. P. Mullan, M. F. Hauser, B. H. Brinkmann, C. R. Jack, G. D. Cascino, F. B. Meyer, and F. W. Sharbrough. Subtraction SPECT co-registered to MRI improves postictal SPECT localization of seizure foci. *Neurology*, 52(1):137–46, January 1999.

[177] Society of Nuclear Medicine. Society of nuclear medicine procedure guideline for brain perfusion single photon emission computed tomography (SPECT) using Tc-99m radiopharmaceuticals, 1999. version 2.0, approved February 7.

[178] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9:62–66, March 1979.

[179] M. Pagani, D. Salmaso, C. Jonsson, R. Hatherly, H. Jacobsson, S. A. Larsson, and A. Wägner. Regional cerebral blood flow as assessed by principal component analysis and ^{99m}Tc-HMPAO SPET in healthy subjects at rest: normal distribution and effect of age and gender. *European Journal of Nuclear Medicine*, 29(1):67–75, January 2002.

[180] Marco Pagani. *Advances in Brain SPECT. Methodological and Human Investigations.* PhD thesis, Departments of Radiology and Hospital Physics, Section for Nuclear Medicine, Karolinska Hospital and Karolinska Institue, Sweden, 2000.

[181] J. A. Palmer. Theory and algorithms for linear and nonlinear component analysis. Technical report, ECE department, University of California, 2003.

[182] J. A. Palmer and K. Kreutz-Delgado. A general framework for component estimation. In *Proceedings of the 4th International Symposium on Independent Component Analysis*, 2003.

[183] H.-J. Park, J.-J. Kim, T. Youn, D. S. Lee, M. C. Lee, and J. S. Kwon. Independent component model for cognitive functions of multiple subjects using [^{15}O]H_2O PET images. *Human Brain Mapping*, 18:284–295, 2003.

[184] X. Pennec, N. Ayache, and J.-P. Thirion. Landmark-based registration using features identified through differential geometry. In I. N. Bankmann, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 499–513. Academic Press, 2000.

[185] C. Pérault et al. Computer-aided intrapatient comparison of brain SPECT images: The gray-level normalization issue applied to children with epilepsy. *The Journal of Nuclear Medicine*, 43:715–724, 2002.

[186] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(10):1090–1104, October 2000.

[187] R. Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability in images of a lambertian object. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(10), October 2002.

[188] S. Roweis. EM algorithms for PCA and SPCA. In *Neural Information Processing Systems 10 (NIPS'97)*, pages 626–632, 1997.

[189] S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

[190] H. Sackheim, I. Prohovnik, J. Moeller, R. Mayeux, Y. Stern, and D. Devanand. Regional cerebral blood flow in mood disorders. II. Comparison of major depression and Alzheimer's disease. *The Journal of Nuclear Medicine*, 34(7):1090–1101, July 1993.

[191] L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. of Machine Learning Research*, 4:119–155, 2003.

[192] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. Technical Report 44, Max-Planck-Institut für biologische Kybernetik, December 1996.

[193] R. N . Shepard and J. D. Carroll. Parametric representation of nonlinear data structures. In P. R. Krishnaiah, editor, *Proc. Int'l Symp. Multivariate Analysis*, pages 561–592. Academic-Press, 1965.

[194] D. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, second edition, 2000.

[195] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, 1987.

[196] D. Skočaj, H. Bischof, and A. Leonardis. A robust PCA algorithm for building representations from panoramic images. In *Europ. Conf. Comp. Vision (ECCV), LNCS 2353*, volume IV, pages 171–178, 2002.

[197] E. A. Stamatakis, M. F. Glabus, D. J. Wyper, A. Barnes, and J. T. L. Wilson. Validation of statistical parametric mapping (SPM) in assessing cerebral lesions: A simulation study. *NeuroImage*, 10:397–407, 1999.

[198] E. A. Stamatakis, J. T. L. Wilson, D. M. Hadley, and D. J. Wyper. SPECT imaging in head injury interpreted with statistical parametric mapping. *The Journal of Nuclear Medicine*, 43(4):476–483, 2002.

[199] Statsoft. *Electronic Statistics Textbook*. StatSoft, 2004. http://www.statsoft.com/textbook/stathome.html.

[200] J. Stoeckel, G. Malandain, O. Migneco, P. M. Koulibaly, P. Robert, N. Ayache, and J. Darcourt. Classification of SPECT images of normal subjects versus images of Alzheimer's disease patients. In W.J. Niessen and M.A. Viergever, editors, *4th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI'01)*, volume 2208 of *LNCS*, pages 666–674, Utrecht, The Netherlands, October 2001.

[201] J. F. M. Svensén. *GTM: The Generative Topographic Mapping*. PhD thesis, Aston University, April 1998.

[202] J. Talairach and P. Tournoux. *Co-Planar Stereotactic Atlas of the Human Brain*. Thieme Verlag, Stuttgart/New York, 1988.

[203] F. Tanaka, D. Vines, T. Tsuchida, M. Freedman, and M. Ichise. Normal patterns on ^{99m}Tc-ECD brain SPECT scans in adults. *The Journal of Nuclear Medicine*, 41(9):1456–1464, September 2000.

[204] P. Thévenaz, T. Blu, and M. Unser. Image interpolation and resampling. In I. N. Bankmann, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 393–420. Academic Press, 2000.

[205] R. Thomas, M. Bhatia, C. S. Bal, S. B. Gaikwad, V.P. Singh, and S. Jain. Correlations of ictal EEG and SPECT studies in patients of intractable epilepsy with normal MRI. *Neurology India*, 50:440–443, December 2002.

[206] P. Thompson, D. MacDonald, M. Mega, C. Holmes, A. Evans, and A. Toga. Detection and mapping of abnormal brain structure with a probabilistic atlas of cortical surfaces. *J Comput Assist Tomogr*, 21(4):567–581, 1997.

[207] P. Thompson and A. Toga. A surface-based technique for warping three-dimensional images of the brain. *IEEE Trans. Med. Imag.*, 15(4):402–416, August 1996.

[208] P. Thompson and A. Toga. Warping strategies for intersubject registration. In I. N. Bankmann, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 569–601. Academic Press, 2000.

[209] P. M. Thompson and A. W. Toga. Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations. *Medical Image Analysis*, 1(4):271–294, 1997.

[210] R. Tibshirani. Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.

[211] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999. Also available as technical report from Aston University.

[212] M. E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999. Also available as technical report from Aston University.

[213] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[214] N. Tzourio-Mazoyer, F. Crivello, M. Joliot, and B. Mazoyer. Biological underpinnings of anatomic consistency and variability in the human brain. In I. N. Bankman, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 449–463. Academic Press, 2000.

[215] M. Uenohara and T. Kanade. Optimal approximation of uniformly rotated images: Relationship between Karhunen-Loeve expansion and discrete cosine transform. *IEEE Trans. Im. Proc.*, 7(1):116–119, January 1998. *There is a glitch in a formula.*

[216] S. Vallabhajosula, R. E. Zimmerman, M. Picard, et al. Technetium-99m ECD: A new brain imaging agent: In vivo kinetics and biodistribution studies in normal human subjects. *The Journal of Nuclear Medicine*, 30(5):599–604, May 1989.

[217] V. Vapnik. *Statistical Learning Theory.* Springer-Verlag, New York, 1995.

[218] T. Veldhuizen. Techniques for scientific c++. Technical report, Indiana University Computer Science Technical Report 542, 2000.

[219] T. Vik, F. Heitz, and J.-P. Armspach. Statistical atlas-based detection of abnormalities in brain perfusion: Comparing models and estimating detection performance. In R. E. Ellis and T. M. Peters, editors, *MICCAI 2003*, LNCS 2879, pages 838–845, November 2003.

[220] T. Vik, F. Heitz, and P. Charbonnier. Mean shift-based bayesian image reconstruction into visual subspace. In *Proceedings of the 2003 IEEE International Conference on Image Processing (ICIP 2003)*, Barcelona, Spain, 2003.

[221] T. Vik, F. Heitz, I. Namer, and J.-P. Armspach. On the modeling, construction and evaluation of a probabilistic atlas of brain perfusion. *NeuroImage*, 2004. 15 pages, in press.

[222] J. West, J. Fitzpatrick, M. Wang, B. Dawant, C. Maurer Jr., R. Kessler, and R. Maciunas. Comparison and evaluation of retrospective intermodality brain image registration techniques. *J Comput Assist Tomogr*, 21:554–566, 1997.

[223] R. Woods, S. Cherry, and J. Mazziotta. Rapid automated algorithm for aligning and reslicing PET images. *J Comput Assist Tomogr*, 16(4):620–633, 1992.

[224] R. Woods, J. Mazziotta, and S. Cherry. MRI-PET registration with automated algorithm. *J Comput Assist Tomogr*, 17:536–546, 1993.

[225] R. P. Woods. Modeling for intergroup comparisons of imaging data. *NeuroImage*, 4:S84–S94, 1996.

[226] R. P. Woods. Validation of registration accuracy. In I. N. Bankmann, editor, *Handbook of Medical Imaging, Processing and Analysis*, pages 491–497. Academic Press, 2000.

[227] K. Worsley, S. Marrett, P. Neelin, A. Vandal, K. Friston, and A. Evans. A unified statistical approach for determining significant signals in location and scale space images of cerebral activation. *NeuroImage*, 4:58–73, 1996.

[228] K. J. Worsley. An overview and some new developments in the statistical analysis of PET and fMRI data. *NeuroImage*, 5:254–258, 1997.

[229] K. J. Worsley, J-B. Poline, K. J. Friston, and A. C. Evans. Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage*, 6:305–319, 1997.

[230] L. Xu and A. L. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Trans. Neural Networks*, 6(1):131–143, 1995.

[231] C. Yang, R. Duraiswami, and D. DeMenthon L. Davis. Mean-shift analysis using quasi-Newton methods. In *Proceedings of the 2003 IEEE International Conference on Image Processing (ICIP 2003)*, Barcelona, Spain, 2003.

[232] M.-H. Yang, N. Ahuja, and D. Kriegman. Face recognition using kernel eigenfaces. In *Proceedings of the 2000 IEEE International Conference on Image Processing (ICIP 2000)*, volume 1, pages 37–40, Vancouver, Canada, September 2000.

[233] M.-H. Yang, D. Kriegman, and N. Ahuja. Face detection using multimodal density models. *Computer Vision and Image Understanding*, 84:264–284, 2001.

[234] Z. Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. *International Journal of Image and Vision Computing*, 15(1):59–76, January 1997. Also as technical report 2676 from INRIA, Sophia-Antipolis (2676).

[235] L. Zhao and Y.-H. Yang. Mosaic image method: A local and global method. *Pattern Recognition*, 32:1421–1433, 1999.

[236] X. Zhou, B. Moghaddam, and T. Huang. ICA-based probabilistic local appearance models. Technical report, MERL - A Mitsubishi Electric Research Laboratory, July 2001.

[237] I. Zubal, S. Spencer, K. Imam, J. Seibyl, E. Smith, G. Wisniewski, and P. Hoffer. Difference images calculated from ictal and interictal technetium-99m-HMPAO SPECT scans of epilepsy. *The Journal of Nuclear Medicine*, 36(4):684–9, 1995.

[238] G. Zuendorf, N. Kerrouche, K. Herholz, and J.-C. Baron. Efficient principal component analysis for multivariate 3D voxel-based mapping of brain functional imaging data sets as applied to FDG-PET and normal ageing. *Human Brain Mapping*, 18:13–21, 2003.