A Probabilistic Model to Learn, Detect, Localize and Classify Patterns in Arbitrary Images

Matthew Toews

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

May 2008

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Electrical and Computer Engineering.

2008/05/12

Abstract

This thesis presents a new, probabilistic model for describing image patterns arising from classes of visually similar objects, such as faces or brains. The model describes patterns in terms of a high level geometrical structure referred to as an object class invariant (OCI), which is invariant to nuisance parameters arising from the imaging process. The OCI itself is not directly observed from images, but can be inferred via a probabilistic model based on generic, spatially localized image features. The OCI model can be learned from a large set of natural images containing pattern instances with minimal manual supervision, in the presence of background clutter, illumination changes, partial pattern occlusion, multi-modal intra-pattern variation (e.g. faces with or without sunglasses), geometrical deformations (i.e. translations, rotations and magnifications) and viewpoint changes. In addition, it can be automatically fit to new images in similar difficult imaging conditions.

Due to the general nature of the OCI model, it has a wide range of possible applications, and its importance is demonstrated in the research fields of computer vision and medical image analysis. In computer vision, the OCI model results in the first viewpoint-invariant system for detecting, localizing and classifying object instances in terms of visual traits. Experimentation on face and motorcycle imagery demonstrates the OCI model can be used to learn, detect and localize general 3D object classes in natural imagery acquired from arbitrary viewpoints. Viewpoint-invariant OCI detection performance is shown to be superior to that of the multi-view formulation which models viewpoint information explicitly. A data-driven algorithm demonstrates the existence of stable OCIs, which can potentially be identified in a fully automatic fashion. The first results in the literature are established for sex classification of face images from arbitrary viewpoints and in the presence of occlusion. In medical image analysis, the OCI model results in the first parts-based anatomical model of the human brain, where subject images of a population are described in terms of a collage of conditionally independent local features or ’parts’. The model is the first to explicitly address the situation where one-to-one correspondence between different subjects does not exist due to natural inter-subject variability. Experimentation modeling the human brain in MR image slices demonstrates that the OCI model is capable of robustly identifying and quantifying anatomical structures in terms of their geometry, appearance, occurrence frequency and relationship to traits such as sex in a population, in cases where other models cannot cope.

Sommaire

Cette thèse présente un nouveau modèle probabiliste pour la description de formes appartenant à des images provenant de classes d'objets visuellement semblables, tels que des visages ou des cerveaux. Le modèle décrit les formes en fonction d'une structure géométrique évoluée désignée sous le nom d'invariant de classe d'objet (OCI), qui est invariante à certains paramètres nuisibles provenant du processus d'imagerie. L'OCI lui-même n'est pas directement observable à partir d'images mais peut être déduit via un modèle probabiliste basé sur des attributs génériques et spatialement localisés de l'image. Le modèle d'OCI peut être appris à partir d'un grand ensemble d'images naturelles contenant des instances de la forme d'intérêt avec une supervision manuelle minimale en présence de fouillis à l'arrière plan, de changements d'éclairement, d'occlusions partielles, de variations multi-modales intrinsèques (par exemple, des visages qui peuvent apparaître avec ou sans lunettes solaires), de transformations géométriques (i.e. translations, rotations et mises à l'échelle) et de changements de point de vue. Additionnellement, le modèle peut être automatiquement ajusté à de nouvelles images dans des conditions d'imagerie également difficiles.

De par sa nature générale, le modèle d'OCI a une grande variété d'applications potentielles et son importance est démontrée dans les domaines de la vision par ordinateur et de l'analyse d'image médicales. Dans le contexte de la vision par ordinateur, le modèle d'OCI mène au premier système intégré pour la détection, la localisation et la classification d'instances d'objets d'après leurs traits visuels de façon indépendante du point de vue. Des expériences réalisées sur des images de visages et de motocyclettes démontrent que le modèle d'OCI peut être utilisé pour apprendre, détecter et localiser des classes d'objets 3D dans des images naturelles acquises à partir de points de vue arbitraires. Il est démontré que la performance d'un détecteur d'OCI invariant aux changements de point de vue est supérieure à celle d'un détecteur basé sur une formulation multi-vues modélisant explicitement l'information relative aux points de vue. Un algorithme guidé par les données démontre l'existence d'OCIs stables qui peuvent potentiellement être identifiés de manière entièrement automatique. Les premiers résultats de la littérature sont établis pour la classification du sexe à partir d'images de visages acquises de points de vue arbitraires et en présence d'occlusions.

Dans le contexte de l'imagerie médicale, le modèle d'OCI mène au premier modèle anatomique par parties du cerveau humain, par lequel les images qui constituent les sujets d'une population sont décrites en termes d'un collage d'attributs locaux ou “parties” conditionnellement indépendants les uns des autres. Le modèle est la première approche à explicitement modéliser la situation où il n'exsite pas de correspondance bi-univoque entre des sujets différents à cause de variations inter-sujet naturelles. Des expériences modélisant le cerveau humain dans des coupes d'images de résonance magnétique démontrent que le modèle d'OCI est capable d'identifier et de quantifier de façon robuste des structures anatomiques en termes de leur géométrie, de leur apparence, de leur fréquence et de leur liaison à des traits tels que le sexe dans une population, et cela dans des cas où d'autres modèles échouent.

Acknowledgments

Thanks first and foremost to my supervisor Tal Arbel for her enthusiasm and providing a research environment in which ideas could flourish. Thanks to Louis Collins of the Montreal Neurological Institute for providing invaluable insight into the state-of-the-art in medical image analysis. Thanks to the members of my Ph.D. committee, Doina Precup for teaching me about machine learning and Jim Clark for teaching me about the world of vision. Thanks to Stephen Frey for lending his neuroanatomical expertise in validating experimentation in modeling the cortical surface. Thanks to my lab mates for stimulating discussions. Thanks to Frank Riggi my collaborator. Thanks to Rola Harmouche and Rupert Brooks for helping in ground truth labeling in experimentation involving medical imagery. Thanks to Cathy Laporte for meticulous reviews of my writing. Thanks to professors Mike Langer, Jeremy Cooperstock, Godfreid Toussaint, Vincent Hayward, Frank Ferrie, Martin Levine, Kaleem Siddiqi who all contributed to my view of the world. Thanks to my fellow students who enriched my experience at the Centre for Intelligent Machines: Prasun Lala, Peter Savadjiev, Scott McCluskey, Sandura Skaff, Carmen Au, Shufei Fan, Stephen Benoit, Phillipe Simard, Isabelle Begin, Deniz Sarikaya, Phillipe Mbonye, Albert Law, Prakash Patel, Marcel Mitran, Jerome Pasquero, Karim Abu-Mustafa, Marta Kirsten, Nils Tilton. Finally, thanks to my family and Marie-Christine Brunelle for supporting me through all of this.

Contents

List of Figures

List of Tables

List of Acronyms

AAM	Active Appearance Model
AC-PC	Anterior Commissure-Posterior Commissure
AP	Average Precision
BET	Brain Extraction Tool
DOG	Difference Of Gaussian
EM	Expectation-maximization
EMD	Earth Mover's Distance
ICA	Independent Component Analysis
ICBM	International Consortium for Brain Mapping
IID	Independent and Identically Distributed
MAP	Maximum A Posteriori
ML	Maximum Likelihood
MR	Magnetic Resonance
MNI	Montreal Neurological Institute
OCI	Object Class Invariant
PBM	Parts-based Model
PCA	Principle Component Analysis
ROC	Receiver Operating Characteristic
SIFT	Scale Invariant Feature Transform
SVM	Support Vector Machine
TRE	Target Registration Error
US	Ultrasound

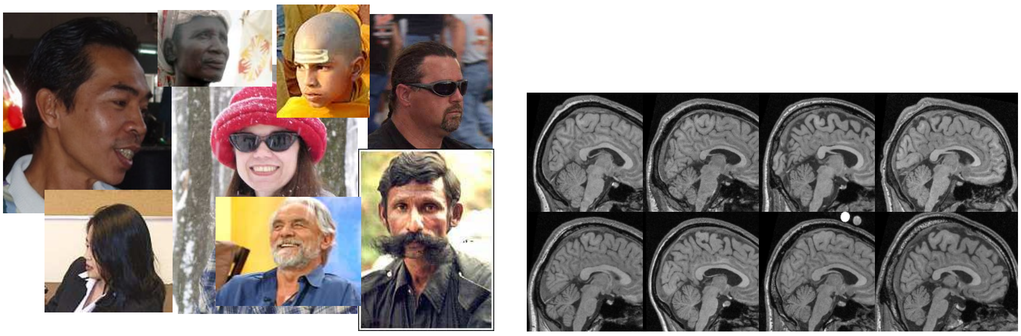
List of Symbols

m_i	An image feature i
m_i^g	The geometrical component of image feature i
m_i^a	The appearance component of image feature i
m_i^b	The occurrence component of image feature i
$\mathbf{m}$	Set of image features $\{m_i\}$
o	An object class invariant (OCI)
o^g	The geometrical component of an OCI
o^b	The occurrence component of an OCI
σ_i	The image scale of image feature i
θ_i	The image orientation of image feature i
x_i	The image location of image feature i
c	Trait variable
c^j	Trait value
f_i	Event of positive occurrence of image feature i, $m_i^{b=1}$
$\mathbf{f}$	Set of positive feature occurrences $\{f_i\}$
$Thres_i^a$	Appearance similarity threshold for image feature i
$Thres^g$	OCI geometrical consistency threshold
T_x	Location component of $Thres^g$
T_θ	Orientation component of $Thres^g$
T_σ	Scale component of $Thres^g$

Chapter 1

Introduction

Consider image patterns arising from classes of similar objects, such as the face and brain images in Figure 1.1. As humans, we have the remarkable ability to recognize these patterns despite significant differences in appearance from one pattern instance to the next, for example images of faces with or without sunglasses captured from arbitrary viewpoints. We can describe these patterns in terms of image features common to different pattern instances, for example faces in terms of facial features such as eyes and noses, and brains in terms of structures such as the corpus callosum and the cerebellum. We are able to locate new pattern instances in previously unseen images, and to classify these instances in terms visual traits, for example faces in terms of sex and brains in terms of pathology.



a) Face images

b) Brain images

Fig. 1.1 Examples of image patterns arising from two different classes of similar objects: a) faces and b) brains. Notice the variability in the appearance of face images due to lighting, viewpoint, sunglasses, and brain images due to anatomical differences between healthy subjects.

Although human observers are able to identify, describe and classify a diverse range of image patterns in similar generic ways, developing a general computational model to automatically perform these tasks is a daunting challenge. Such a model must be able to effectively describe pattern appearance despite a high degree of natural pattern variability, due to factors such as noise and unrelated clutter, partial occlusion and missing data, intensity changes, in-plane geometrical deformations such as rotations, translations, magnifications, and out-of-plane geometrical deformations due to viewpoint change. The model must generalize in order to describe a wide variety of different patterns arising from different classes of objects. The model must lend itself to computationally efficient algorithms for processing large amounts of image data. It must be possible to learn the model from natural data with minimal manual supervision or preprocessing. The model must represent and describe patterns in an intuitive manner that can be easily understood by humans, in terms of visual traits such as age, sex or pathology, or with respect to standard geometrical reference frames used in anatomical study.

Due in large part to the challenges of developing a general computational model, approaches in the literature are often tailored for specific tasks in different research fields based on specific types of image data. In the field of computer vision, for example, a primary focus is to learn the appearance of 3D object classes, such as faces, from arbitrary 2D projective imagery with minimal manual supervision, and to detect, localize and describe novel object class instances in terms visual traits. Methods to do this are typically restricted to single viewpoints of 3D object classes [1, 2] or require explicit knowledge of viewpoint [3, 4], and thus are difficult to apply to modeling general 3D object classes from arbitrary viewpoints. Methods for describing or classifying visual traits, such as sex in human faces, are typically based on frontal, non-occluded views and do not generalize easily to arbitrary viewpoints [5, 6, 7, 8, 9]. Although inseparable in a practical vision system, the tasks of detecting, localizing and classifying visual traits are typically treated in isolation, and a common framework for all three has not yet been proposed.

In the field of medical imaging, primary goals are 1) to learn an anatomical description of a population from a set of images of different subjects, MR images of the human brain for instance, 2) to quantify anatomical variability across the population and 3) to understand how anatomical structure varies with subject traits such as pathology. Modeling anatomy typically involves aligning all subjects into a common reference frame via inter-subject registration techniques [10, 11, 12], after which techniques such as morphometry [13, 14, 15, 16] are used to identify regions in which measurements such as image intensity or geometry co-vary with traits of interest. Inter-subject registration techniques typically assume the existence of one-to-one correspondence between all subjects [17, 12]. This assumption is difficult to justify in MR brain images, however, where it may be invalidated by factors such as pathology, multiple anatomical morphologies in the cortex, etc. Subsequent morphometric analysis is prone to confounding image measurements arising from different underlying anatomical structure [18, 19, 20] in regions of the brain where one-to-one inter-subject correspondence may not exist. Certain approaches focus on registering and quantifying the variability of specific structures, e.g. specific sulci in the cortex [13, 21]. However such approaches are difficult to generalize to new imaging contexts, non-cortical imagery for instance.

1.1 Outline of Approach

The goal of this thesis is to develop a general model of pattern appearance, that can be used both to detect, localize and classify patterns arising from object classes such as faces in projective imagery, and to effectively describe anatomical structure in the context of medical image analysis. Figure 1.2 details a list of the challenges faced in accomplishing this goal and the means by which they are addressed by the approach taken in this thesis. Open research challenges addressed by the modeling technique in this thesis are listed in bold typeface.

Generic, local invariant image features form the basis of the approach taken in this thesis [22, 23], and provide significant help in overcoming the challenges (1-6). They can be robustly extracted from images in a manner invariant to illumination changes (1) and in-plane geometrical deformations such as translations, rotations, magnifications, and affine shearing (2). As features are local, they can be extracted despite partial pattern occlusion (3) and clutter and background noise (4). As features are invariant to illumination and in-plane geometrical variation, they can be used to establish inter-image correspondence in a computationally efficient manner (5), without explicitly modeling or searching these variation parameters. As features are generic, they can be used to describe the appearance of a wide variety of different pattern types (6), unlike features designed for specific contexts such as eye or nose features in face images [24, 25] or sulcal features in brain images [13].

The weakness of invariant features is that they cannot generally be used to establish

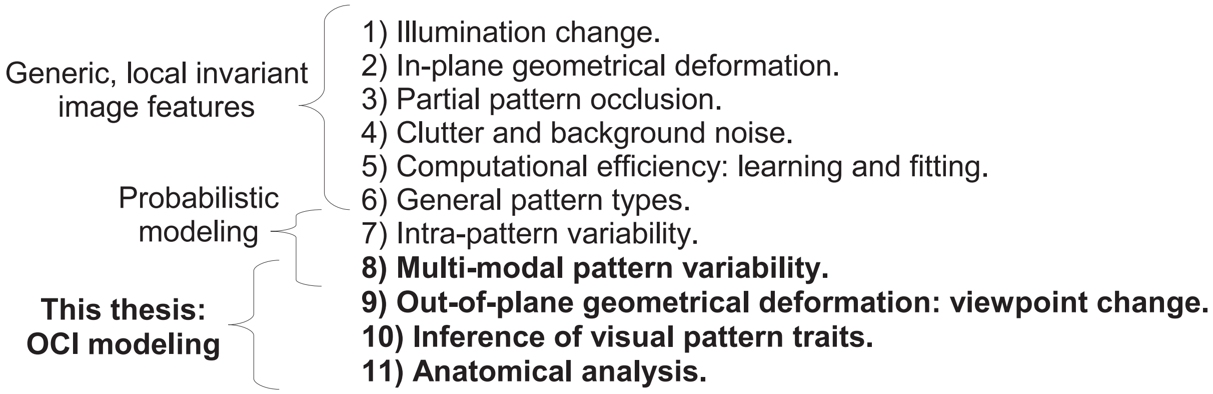

Generic, local invariant image features
Probabilistic modeling
This thesis: OCI modeling
1) Illumination change.
2) In-plane geometrical deformation.
3) Partial pattern occlusion.
4) Clutter and background noise.
5) Computational efficiency: learning and fitting.
6) General pattern types.
7) Intra-pattern variability.
8) Multi-modal pattern variability.
9) Out-of-plane geometrical deformation: viewpoint change.
10) Inference of visual pattern traits.
11) Anatomical analysis.

Fig. 1.2 A list of challenges facing a general model of image pattern appearance, and the means by which they are addressed in this thesis. Challenges 1-6 are dealt with by local invariant feature methods. Challenges 7 and to an extent 8 have been addressed in the literature by probabilistic modeling based on local features. The model proposed in this thesis builds on probabilistic modeling of local invariant image features, in order to address challenges 9-11 listed in bold font.

correspondences between different instances of the same object class, i.e. different faces or brain scans. This is due to challenges (7-8) of intra-pattern and multi-model variability. Intra-pattern variability refers to variation in the appearance or shape of the same image feature from one object instance to the next, for example natural variability of eyes or noses. Multi-modal variability is the situation where the same semantic feature or structural region exhibits multiple, distinct modes of appearance, for instance faces with or without sunglasses. Probabilistic modeling has been shown to be an effective means of describing object class appearance in the presence of intra-pattern variability (7), and to an extent, multi-modal variability. Probabilistic models based on local invariant image features [1, 2] can be used to model complex image patterns in terms of a collection of simpler, localized features or parts [26]. Parts-based modeling can be contrasted to the global modeling, i.e. simultaneously accounting for the entire spatial extent of the pattern [27, 28, 29], which has been shown to be suboptimal in terms of pattern detection performance [30].

Despite advancements in pattern appearance modeling, there remain significant challenges (9-11), i.e. coping with variation due to viewpoint change (9), inferring visual traits of patterns such as the age or sex of faces (10), and addressing anatomical description in the context of medical image analysis (11). This thesis proposes a general model of image pattern appearance designed to address these challenges, thereby making contributions to two distinct but related research fields, computer vision and medical image analysis. The following sections outline the primary theoretical contributions of this thesis and the specific contributions in the contexts of computer vision and medical image analysis.

1.1.1 Primary Contributions

The primary contributions of this thesis stem from a new probabilistic model of pattern appearance in arbitrary images. The model is generally applicable to a wide variety of image patterns, and can be efficiently learned from large sets of natural, cluttered, noisy image data with minimal manual supervision or preprocessing. In addition, the model can be efficiently and robustly fit to new images, in order to detect and localize pattern instances in the presence of similar difficult conditions. The novelty of the model is that it links generic image features to a geometrical reference frame that is 1) uniquely defined with respect to each pattern instance and 2) invariant to nuisance parameters arising from the imaging process. This geometrical structure, which is referred to as an object class invariant (OCI), represents a high-level feature that is not directly observable from image data, but can be inferred from image data via a probabilistic model relating image features to the OCI. As image features are linked to aspects of the underlying object class in a manner unaffected by context-specific nuisance parameters, the OCI model is readily generalized to new contexts. The high-level contributions of this thesis are:

1. A general probabilistic model of pattern appearance, relating invariant image features to an OCI. The OCI model can be learned from a large set of natural image data containing nuisance parameters including noise, clutter, partial pattern occlusion, multi-modal intra-pattern variation, intensity changes, and geometrical deformations arising from the imaging process, with minimal manual supervision. The OCI model can be fit to new images, in order to identify and localize pattern instances in similar imagery.

2. A general technique for learning and classifying abstract visual traits of patterns arising from object classes based on OCI model features. This technique provides a significant advancement over existing techniques as the image features required for trait classification can be robustly identified and linked to pattern instances in the presence of nuisance parameters.

3. A theoretical analysis of OCI optimality and an algorithm for determining an optimal OCI in a data-driven manner. This demonstrates that meaningful, stable OCIs exist for similar object classes, and can potentially be identified without requiring manual input.

Due to the general nature of the OCI model, it provides an effective framework for addressing open problems in a wide variety of contexts. The OCI modeling theory developed in this thesis is applied in the contexts of computer vision and medical image analysis, thereby demonstrating the general nature of the model and leading to contributions to two different but related research fields. These contributions are now described.

1.1.2 Contributions to Computer Vision

The primary contribution of this thesis to the field of computer vision is a combined solution to the tasks of learning, detecting, localizing and classifying 3D object classes, such as faces or cars, in 2D projective imagery taken from arbitrary viewpoints, as illustrated in Figure 1.3. Accomplishing this requires a mechanism for coping with appearance variation arising from viewpoint change, in addition to variability inherent to natural images including noise, unrelated clutter, partial occlusion, illumination changes, global geometrical deformations such as translations, rotations and magnifications and multimodal intra-class variation (i.e. faces with/without sunglasses, beards, etc.). Explicitly modeling these sources of variation is generally intractable and unnecessary, as they are ultimately unrelated to the tasks of interest. This thesis takes the alternative approach of formulating a model that is invariant to these parameters including viewpoint, leading to a viewpoint-invariant parts-based model of 3D object class appearance in 2D images. This is done by defining the OCI as a projective invariant, and as such, the definition of an object class instance is invariant to projective transform arising from viewpoint change. The OCI model describes appearance in terms of generic image features and a generic viewpoint-invariant reference frame, and is therefore applicable to a wide variety of object classes of arbitrary 3D shape. This differs from early modeling approaches which attempted to identify specific viewpoint-invariant properties arising from restricted classes of 3D shapes. As the variable of viewpoint is effectively marginalized from the formulation, a probabilistic model can be automatically learned with minimal manual intervention from large sets of natural, cluttered images taken from arbitrary viewpoints. The model can be used to efficiently detect and enumerate object class instances in new images taken from arbitrary viewpoints. The specific contributions of this thesis to the field of computer vision are:

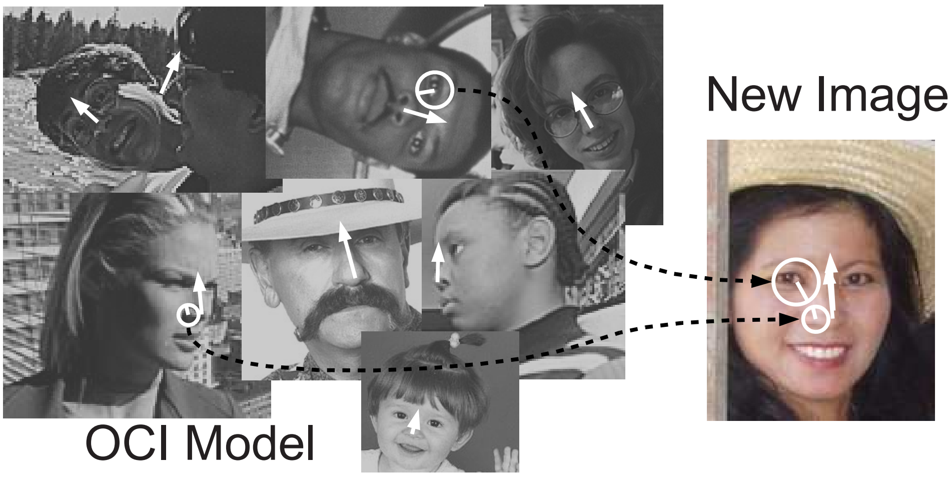


Fig. 1.3 Illustrating the computer vision contribution of this thesis: a viewpoint-invariant model of 3D object classes such as motorcycles or faces. The OCI model can be learned from natural, cluttered images taken from arbitrary viewpoints, and subsequently used to detect, localize and classify faces in terms of the visual trait of sex.

Contributions to Computer Vision

1. The first general viewpoint-invariant model of 3D object class appearance based on generic local image features. The model can be learned from natural 2D images without knowledge of viewpoint, and used to detect instances of 3D object classes in images acquired from arbitrary viewpoints [31]. The effectiveness and generality of model learning, detection and localization are experimentally demonstrated for the classes of faces and motorcycles.

2. The first combined system for detecting, localizing and classifying visual traits of 3D object class instances from arbitrary viewpoints, in the presence of noise, clutter, partial pattern occlusion, multi-modal intra-pattern variation, intensity changes, and geometrical deformations such as translations, rotations and magnifications. The system results in the first approach to detecting, localizing and classifying the sex of faces in images acquired from arbitrary viewpoints [32]. Additionally, the system provides the first automated analysis of local image cues of sex in face images acquired from arbitrary viewpoints about the head.

3. An iterative algorithm to determine an optimal viewpoint-invariant OCI for general 3D object class modeling. This algorithm was validated in the context of learning and detecting 3D faces, showing that a stable OCI consistent with the 3D geometry of the underlying object class can be estimated a data-driven manner, and used to improve face detection in terms of precision-recall [33].

1.1.3 Contributions to Medical Imaging

The primary contribution of this thesis to the field of medical imaging is a general anatomical model which describes the appearance of a population in terms of distinct, local features or parts. This represents the first general approach explicitly addressing the case where one-to-one correspondence does not exist between all subjects of a population [34, 35, 36]. By defining the OCI as invariant to the tomographic reconstruction process of an MR imaging device, for example, the OCI model can be used as a parts-based anatomical description of the human brain which can be learned from a database of images of different subjects [34, 35, 36], as illustrated in Figure 1.4. The model can be used to efficiently learn a description from large sets of images (100s of subjects). Once learned, the model can be subsequently registered to new images in a manner that is robust and stable in the presence of unexpected local deformation, unlike other approaches that attempt to model the image globally or as a whole. The OCI model provides a natural, intuitive description of anatomy in terms of localized structures, which can be easily understood by human experts. Additionally, anatomical structure characteristic of subject traits such as pathology, age or sex can be learned and identified in new subjects. The specific contributions of this thesis to the field of medical imaging are:

Contributions to Medical Imaging

1. The first generic anatomical model to specifically address the case where one-to-one correspondence between subjects of a population does not exist [34, 35, 36]. The OCI model achieves this by describing anatomy in terms of a collection of independent, spatially localized image features or 'parts', which do not occur in all subjects but rather with a particular occurrence frequency in a population. Other anatomical models typically assume the existence of one-to-one correspondence between all subjects, an assumption which is generally invalid in cases where one-to-one inter-subject correspondence is ambiguous or non-existent due to factors such pathology,

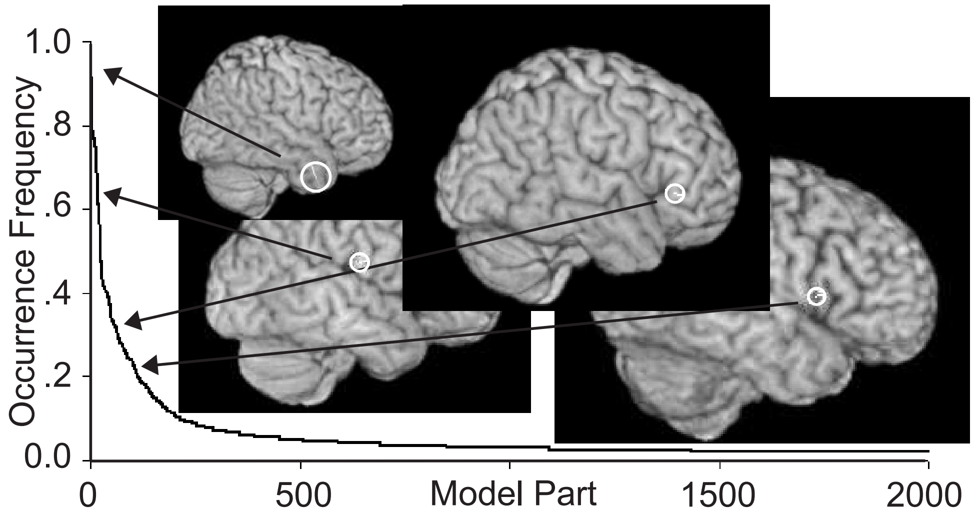


Fig. 1.4 Illustrating the medical imaging contribution of this thesis: learning a parts-based anatomical description of brain imagery, here the surface of the cerebral cortex. The appearance of the cortical surface is described as a collection of local features (white circles), which are quantified in terms of their appearance, geometry and occurrence frequency (shown on the graph) within a population.

surgical resection, multiple modes of anatomical morphology, or natural inter-subject variability. Validation based on slices of MR human brain imagery demonstrates that the OCI model can be fit to new subjects with accuracy comparable to human raters, in a manner that is significantly more robust and stable than the benchmark AAM technique [29].

2. The first method for learning unlabeled folding patterns that occur with statistical regularity in the highly variable cortical surface of the human brain. Other related methods focus on reproducing expert labelings of specific cortical structures, and thus cannot be easily applied to describing the anatomy of new cortical regions where no expert labelings exist. This method is tested using volume renderings of the cerebral cortex, and validation of automatically identified cortical structures is performed by an expert neuroanatomist.

3. A general technique for linking distinct, localized anatomical structures to subject traits. The technique can be potentially applied to identifying anatomical structure indicative of traits such as pathology, abnormality, handedness, etc. The technique is

applied to determining image features indicative of sex in sagittal slices of MR brain images as a proof of concept.

4. A technique for using the parts-based anatomical model as a basis for inter-subject registration of new subjects of a population. The technique uses the parts-based model to identify statistically regular anatomical structure shared by new subjects to be registered, thereby identifying image regions where valid inter-subject correspondence is likely to exist. This technique is demonstrated in the context of inter-subject registration of coronal MR slices of the human brain [34].

The remainder of this document is organized as follows. Chapter 2 outlines the main bodies of research literature related to this thesis, in three main sections. Section 2.1 discusses general work relating to modeling image pattern appearance from localized image features. Sections 2.2 and 2.3 then discuss literature related to the specific contributions of this thesis, i.e. modeling the appearance of 3D object classes from arbitrary viewpoints in 2D projective imagery, and modeling anatomy in medical imagery.

Chapter 3 describes the contributions relating to the OCI model of pattern appearance and specific contributions to the research fields of computer vision and medical imaging, in three main sections. Section 3.1 describes the object class invariant (OCI) model, the primary theoretical contribution of this thesis, including the model formulation, algorithms for learning the model from training images and fitting the model to new images, and using the model to learn and classifying object instances in terms of visual traits. Section 3.2 describes how OCI modeling theory can be applied to modeling the appearance 3D object classes in projective imagery taken from arbitrary viewpoints. Section 3.3 then describes how OCI modeling theory can be used as an anatomical description of medical imagery which can be automatically learned from a set of images of a population.

Experimentation demonstrating the contributions of this thesis to computer vision is presented in Chapter 4, where the OCI model is applied to viewpoint-invariant modeling of the classes of human faces and motorcycles. Section 4.1 contains three experiments involving learning, detection and localization of 3D faces from natural, cluttered imagery taken from arbitrary viewpoints. The first experiment demonstrates how an OCI model of the human face can be learned from a set of cluttered, natural images acquired from arbitrary viewpoints, and used to detect faces in similar imagery. The second experiment quantitatively compares the viewpoint-invariant OCI model presented in this thesis to the multi-view model in terms of face detection and localization performance, based on the benchmark CMU profile database [37]. Results show that the viewpoint-invariant OCI has a superior precision-recall characteristic, as the multi-view model is prone to a higher rate of false positives. The third experiment demonstrates how a stable OCI reference frame can be derived in data-driven manner from a set of 500 different face images. The data-driven OCI geometry converges to a definition that remains geometrically consistent with the 3D geometry of the head in images of arbitrary viewpoints. Modeling based on the data-driven OCI results in superior detection performance as compared to a manually selected OCI.

Section 4.2 presents experiments involving integrated detection, localization and classification of 3D faces in terms of the visual trait of sex, in images taken from arbitrary viewpoints. These experiments are performed on the public FERET database [38] and represent the first reported results on sex classification of faces from arbitrary viewpoints and in the presence of occlusion. The Bayesian classification approach proposed in this thesis is shown to outperform support vector machine (SVM) classification, another popular technique which could potentially be used. Aside from simply classifying faces in terms of visual traits, the OCI model-based classification allows one to understand the local image features which serve as visual cues of traits such as sex. Finally, Section 4.3 presents experimentation on OCI model learning, detection and localization for the class of motorcycles, based on the benchmark PASCAL database [39]. Results show that the OCI model can be learned from a database of natural imagery and used to identify motorcycles in new imagery with accuracy comparable to other state-of-the-art approaches trained and tested on the same data.

Experimentation detailing the contributions of this thesis to medical image analysis is presented in Chapter 5, applying the OCI model to describing the anatomy of the human brain in MR imagery. All experimentation is based on the international consortium of brain mapping (ICBM) 152 data set [40], consisting of brain images of 152 unique subjects. Section 5.1 outlines experimentation involving OCI modeling of normal brain anatomy, describing the brain in terms of a set of distinct, localized features or parts. The learned model can be visualized by sorting model parts according to their occurrence frequency in a population, where frequently occurring parts are indicative of stable anatomical structure shared by many brains, and infrequently occurring parts are indicative of noisy or subject-specific characteristics. The accuracy of fitting of the model to new brains is evaluated quantitatively and compared with manual model fitting by human raters, indicating that the accuracy of automatic OCI model fitting is similar to that of humans.

One of the major advantages of the parts-based model is robust, stable fitting in the presence of unexpected local perturbation, as such perturbations are treated as unrecognized anatomical structure and disregarded. A quantitative evaluation of parts-based model fitting stability in the presence of artificial local perturbations is performed, demonstrating this advantage. Furthermore, fitting stability is quantitatively compared to that of another popular appearance model, the active appearance model (AAM) [29], based a global modeling methodology common in the medical imaging literature. Results demonstrate the superior stability of the parts-based approach and the instability of the global AAM, which in unable to effectively cope with local perturbation.

The ability of the OCI model to identify anatomical characteristics reflective of subject traits is presented in Section 5.2, in this case the trait of subject sex. Results show that several anatomical structures thought to be indicative of sex in the brain are automatically identified, and new anatomical structures potentially indicative of sex are identified. In the study of the brain, it is important that a general pattern appearance model be adaptable to modeling specific regions of interest. Section 5.3 outlines the result of a study adapting the OCI model to describe the highly-variable cerebral cortex, resulting in the first technique for automatically identifying new, unlabeled cortical structures. Expert validation by a neuroanatomist on a subset of learned model parts indicates that same similar cortical structures are successfully identified by the learning process. Finally, Section 5.4 outlines a demonstration of how the parts-based OCI model can be used as a basis for the challenging task of inter-subject image registration. By identifying modeled anatomical structure shared by the subjects to be registered, registration can be avoided in image regions where correspondence may not exist.

Chapter 2

Related Work

The goal of modeling abstract image patterns is to be able to automatically learn, detect, localize and classify traits of the underlying 3D object classes from which patterns arise. There are three distinct challenges in doing this. First, modeling must successfully cope with appearance changes due to nuisances arising from the imaging process, such as illumination changes, noise, unrelated clutter, partial pattern occlusion, intra-class appearance variability, multi-modal appearance variability (i.e. faces with or without sunglasses), in-plane geometrical deformations such as translation, orientation and scale changes, in-depth deformations due to viewpoint changes, etc. Second, learning must efficiently identify patterns and quantify their variability in appearance from large sets of images with little or no manual supervision, and detection. The tasks of pattern detection, localization and classification must be similarly efficient. Third, modeling must generalize to a large number of different abstract patterns.

This chapter outlines work related to addressing these challenges and is organized as follows. Section 2.1 outlines general work relating to modeling abstract patterns of image appearance. Section 2.2 describes work relating the contribution of this thesis to the field of computer vision, learning, detecting, localizing and classifying traits of 3D object classes from arbitrary viewpoints. Section 2.3 describes work relating the contribution of this thesis to the field of medical imaging, learning a parts-based description of medical image anatomy.

2.1 Modeling Patterns in Images

Vision involves making inferences about objects or scenes in the world based on the image patterns they produce. Basic vision tasks such as detecting, localizing and classifying object categories require describing or modeling the appearance of the image patterns produced by such object categories. This section reviews general computational approaches to image pattern modeling in the literature. Just as any verbal description is based on a vocabulary of words, a general description or model of an image pattern is based on a set of image features. Section 2.1.1 reviews image features upon which general, computationally efficient models of image patterns can be constructed. Section 2.1.2 reviews mechanisms by which correspondence is achieved between features in different images arising from the same underlying objects in the world. Section 2.1.3 reviews methods for combining features into models of image patterns, and techniques for learning such models from data. As the work in this thesis follows from techniques that emerged primarily from the computer vision literature, much description refers the case of 2D projective imagery. The theory is directly applicable to 3D volumetric and 4D temporal imagery common in the field of medical imaging, and an attempt is made to generalize concepts to arbitrary image dimensions.

2.1.1 Image Features

Images are measurements of signals arising from objects in the world, which are captured by a sensor and organized in a spatial lattice or grid. Images can be formed by a variety of different processes, i.e. photons incident on a receptive field in photographic imagery, radio frequencies emitted by excited protons in magnetic resonance imagery, reflecting sound waves in ultrasound imagery, etc. Image measurements in the form of image intensities can be transformed into other measurements, i.e. spatial image derivatives [41], spectral decompositions such as Fourier frequencies [42] or principal components [27], histograms of such measurements [43].

In order to localize a pattern within an image, it must first be possible to localize the image measurements or features on which the pattern description is based. Feature localization is complicated by nuisance parameters associated with the imaging process, such changes in image intensity and pattern geometry, unrelated clutter. These parameters must be accounted for in order to detect and localize patterns, but they are ultimately unrelated to the underlying objects in the world from which features and image patterns arise. For example, a pattern can be translated, rotated and scaled to an arbitrary degree within the image plane and illuminated in a variety of different ways, while still arising from the same underlying object in the world. The goal of feature localization is to normalize image content with respect to nuisance parameters, at which point patterns can be described strictly in terms of shape and appearance variability directly associated with the underlying object. Normalization can be though of as marginalizing or explaining away nuisance parameters.

Localized features can be described as specific or generic. Specific features are those designed to identify application specific, semantically meaningful structures such as eyes or noses in face images [24, 25] or particular sulci in brain images [13]. Generic features refer to natural image structures arising in a variety of arbitrary imaging contexts, such as corners or blobs. Generic features are advantageous in that they can be used to automatically describe a wide variety of general patterns, whereas specific features must be redefined for each application domain.

Simple generic features such as edges [44, 45] are by themselves of limited use for modeling image patterns, because they can only be localized in a single spatial image dimension. They can be useful for localization when used in combination, by considering intersections of different lines arising from edges of blocks for instance, but to do so requires modeling this interaction. In early computer vision literature, a variety of approaches propose identifying image features that can be parameterized in terms of their location x in the image plane, such as corners [46], maxima of a local autocorrelation function [47], eigenvalues of local image derivatives [48, 49]. Such features are robust to linear image intensity change, as they consist of maxima of functions of image intensity, and are useful for tasks such as tracking or narrow-baseline stereo vision, where the geometrical transform of an image pattern from one image to the next can be modeled as a translation. They cannot generally be used for correspondence in cases where the geometrical transform between images involves significant changes in scale.

A general approach is to identify features in a manner that is invariant to classes of transforms of image intensity and geometry. In this way, similar image patterns can be automatically normalized with respect nuisance parameters of intensity and geometry. Once normalized, they can be efficiently matched between images without requiring an explicit search over nuisance parameters. Early invariant feature methods attempted to identify configurations of simple generic features such as edges, vertices or points arising from planar

3D shapes in the world, which could be normalized geometrically. For instance features invariant to affine transforms can be defined by a configuration of 3 co-planar points [50] or two edges and a point [51]. Such features were described in terms of geometrical invariance in [52]. The main drawback of invariant features defined by configurations of simple generic features is that there are generally many possible combinations of configurations that can be constructed and it is thus difficult to repeatably identify the same configurations in different images.

Scale-space theory showed that the notion of an image feature is intimately tied to the resolution or scale σ at which the images are processed [53, 54]. This led to the development of distinctive invariant features, which attempt to overcome the drawback of having to combine multiple features to obtain invariance, by instead identifying distinctive image regions defined by their scale in addition to their location. So-called scale-invariant features, for instance, were developed to address correspondence in the presence of changes in image scale, by localizing image features in terms of image scale σ in addition to image location [55]. Invariant features have been developed which are invariant to linear intensity variation, in addition to geometrical deformations such as translation, scale, orientation and affine transformations [22, 23], and can be robustly and repeatably identified in different images of the same natural scenes in the presence of these deformations. The advantage of distinctive invariant features is that they are explicitly associated with an image region with distinctive appearance, which can be used for the purpose of feature correspondence.

Detection of distinctive invariant feature regions is generally accomplished via a search over the geometrical parameters of the image transformation under which features are invariant. In general, localizing a distinctive image region in $\Re^N$ image space requires determining an N-dimensional location. In order to associate image content with the feature for the purpose of correspondence, a minimum of one additional scale parameter σ is required in order to define a hyperspherical image region. Scale-invariant features, for example, are invariant to similarity transforms, i.e. location, scale and orientation changes. Features must therefore be localized in terms of image location x, scale σ and additionally orientation θ. In $\Re^N$ image space, this is done by identifying extrema in an $N+1$-dimensional scale-space function $G(I, x, \sigma)$ defined over location x and scale σ in an image I:

$$\{x_i, \sigma_i\} = \underset{x_i, \sigma_i}{\text{argmax}} \{|G(I, x, \sigma)|\}. \tag{2.1}$$

The process of extrema identification can be efficiently implemented by representing $G(I, x, \sigma)$ as a scale-space image pyramid [22]. Once parameters of feature location x and scale σ have been extracted, a local orientation θ can be determined, thereby achieving invariance to orientation changes. There are generally $N(N-1)/2$ orientation parameters to be estimated, which can be determined from image gradient orientations within an image window of size proportional to σ at location x, for example as peaks in a histogram of gradient orientations. Figure 2.1 illustrates examples of scale-invariant features extracted in a sagittal MR brain image.

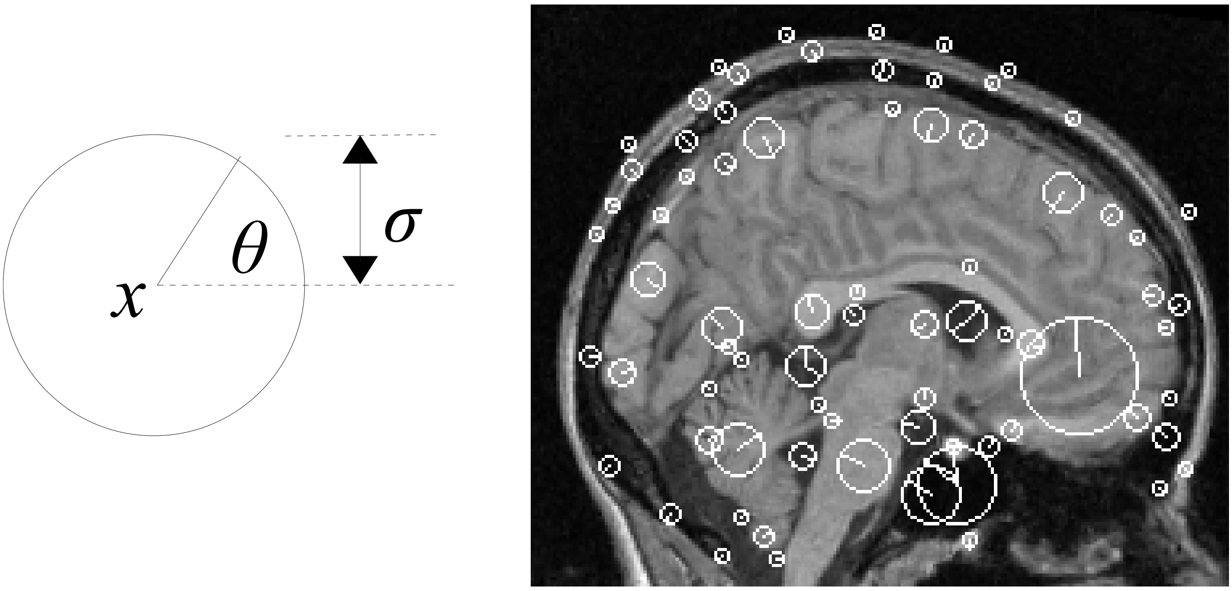


Fig. 2.1 Scale-invariant features automatically extracted in a mid-sagittal slice of a MR volume. Scale-invariant features, illustrated as circles inset with radial lines, are oriented regions in $\Re^N$ image space, defined geometrically by an N-parameter location x, a scale σ, and an $N(N-1)/2$ parameter orientation θ. Features shown here were extracted in $\Re^2$ image space using the scale-invariant feature transform (SIFT) method [22].

A variety of different distinctive invariant feature types exist. A given detection technique can be described by the geometrical transform under which it is invariant, and by the specific scale-space $G(I, x, \sigma)$ used. Detectors invariant to similarity transform [22, 56, 57, 58, 59] and affine transform (accounting for image shear) [60, 23, 61, 62, 63] are currently widespread in the literature. Note that in the case of affine-invariant features, an explicit search over additional parameters of shear is computationally intensive. As a result, affine-invariant features are often identified sub-optimally by growing covariant regions [64], potentially around extrema first detected in scale-space pyramids [61].

The scale-space function $G(I, x, \sigma)$ can be defined according to a variety of different image measurements, in order to model different image characteristics. A scale-space based on the difference-of-Gaussian (DOG) operator [65], such as used in the SIFT method [22], tends to localize blob-like image structures. Scale-spaces based on spatial or directional derivatives [41] can be used to localize edge-related image structures [57]. Informative image regions can be identified using scale-spaces constructed on information theoretic measures such as entropy [58]. Other features can be extracted based on characteristics such as object boundary contours [66], image phase [56] or color moments [59] for multi-valued image intensities. As a result, a wide variety of different invariant features can be extracted from the same images, all of which can be used to obtain complementary information regarding the image content. Figure 2.2 illustrates the process of scale-invariant feature extraction in an MR image slice. Examples of different invariant feature types identified in the same image can be seen in Figure 2.3.

Although may seem desirable to attain the highest degree of invariance in local features, the literature suggests that this is not necessarily the case. High order invariants such as perspective invariants are relatively rare and difficult to detect in general [67]. Although affine invariants have been shown to be effective at modeling patterns arising from 3D planes, empirical evidence [22] and theoretical arguments [68] suggest that affine detection techniques create an overly large equivalence class of different features, resulting in more ambiguous pattern correspondence.

There are several arguments for the use of scale invariance. Influential work has defined pattern shape as the variability remaining after geometrical normalization with respect to translation, rotation and scale [69, 70]. Hyperspherical image features associated with scale invariance are consistent with the notion of an uncommitted visual front-end exhibiting no bias as to the shape of features being extracted [71]. Experimentally, scale-invariant features can be robustly extracted and matched over a significant range of viewpoint changes, particularly for non-planar 3D objects [22].

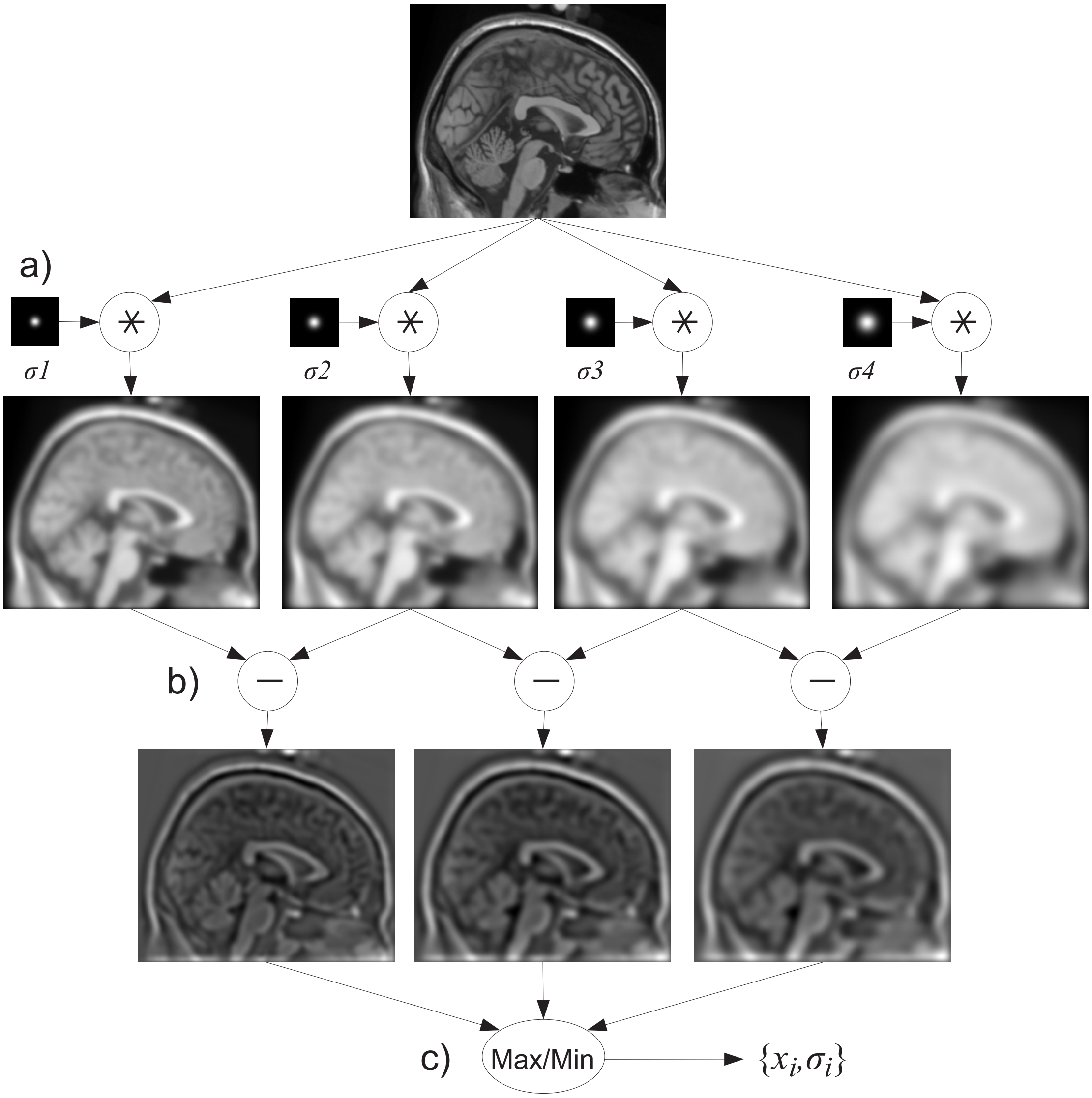
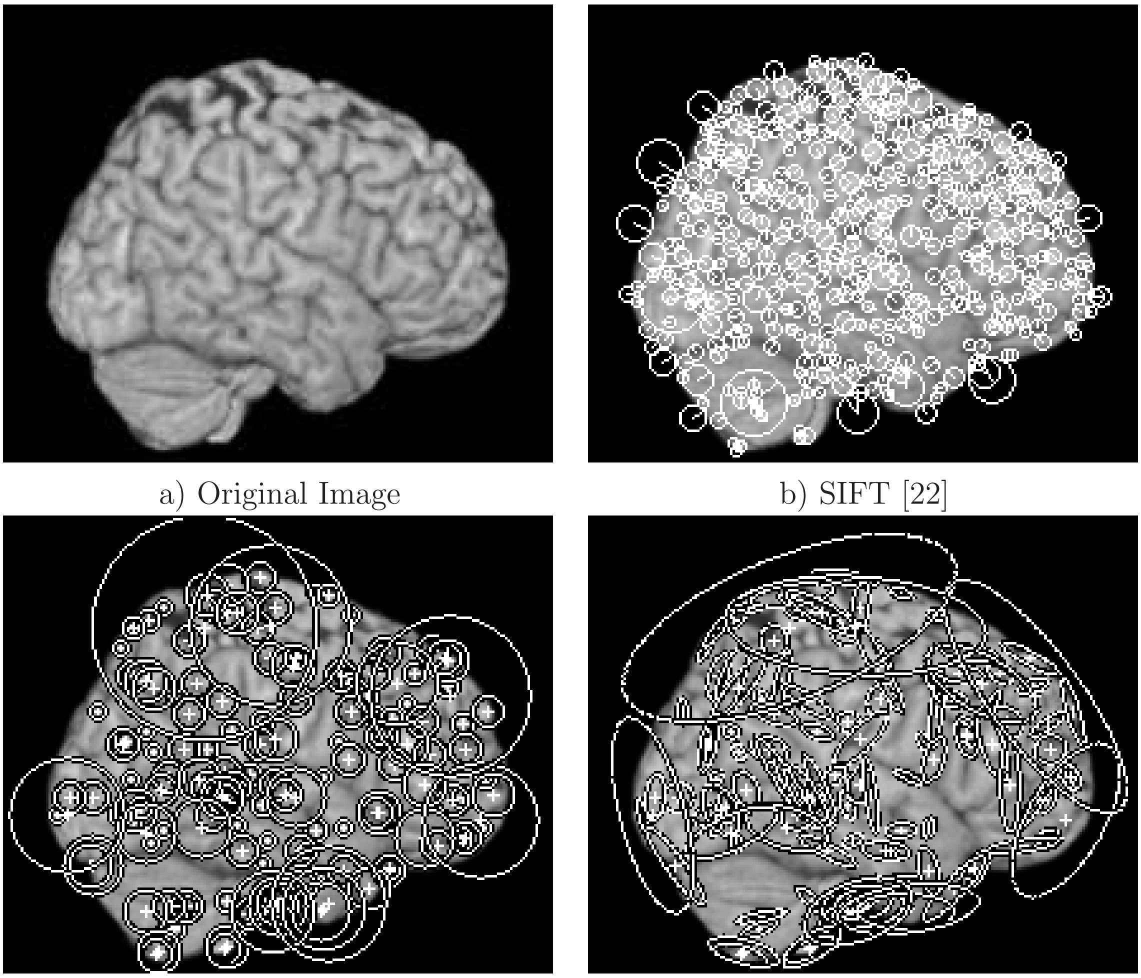


Fig. 2.2 Illustrating invariant feature extrema detection based on a difference-of-Gaussian scale-space. In step a), the original image is first convolved with Gaussian kernels of varying scale σ, generating Gaussian blurred images. Next in step b), the image difference is computed between blurred images of adjacent scales, resulting in difference-of-Gaussian images. Finally in step c), maxima and minima are detected between difference-of-Gaussian images in adjacent scales, resulting in a set of invariant feature locations and their corresponding scales $\{x_i, \sigma_i\}$.



a) Original Image

b) SIFT [22]

c) Scale-invariant Harris [57]

d) Affine-invariant Harris [61]

Fig. 2.3 Various types of invariant features extracted from the same volume rendering of the cortex shown in a). SIFT features, shown in b), are extracted from a difference-of-Gaussian scale-space and tend to correspond to blob-like features. Scale-invariant Harris features, shown in c), are based on the Harris edge criterion [49] defined by spatial image derivatives and tend to indicate edges or corners. Affine-invariant Harris features are calculated by growing elliptical affine regions around scale-invariant Harris features, capturing elongated structures. Note how in general, large-scale features correspond roughly to larger anatomical structures such as cerebral lobes, while small-scale features arise from smaller structures such as cortical sulci and gyri.

2.1.2 Feature Correspondence

Once image features are localized within images, they can be used as a basis for establishing image correspondence, i.e. identifying instances of the same image content in different images. There are three important components to computing correspondence: 1) encoding feature appearance, 2) evaluating feature similarity and 3) enforcing inter-feature geometrical constraints.

Encoding Feature Appearance: Once features have been localized, they are normalized with respect to their geometry, after which their associated image content is represented or encoded for subsequent use in feature correspondence, i.e. identifying different features arising from the same underlying image measurements. Encoding is based on image measurements associated with localized image features. For the purpose of feature correspondence, an ideal encoding both maximizes feature distinctiveness (i.e. the likelihood of obtaining correct vs. incorrect correspondences), and minimizes the computational burden of calculating correspondence or similarity.

Simple image features such as image points or line segments [67] with no associated image content are described solely in terms of their geometry. Such features are not distinctive as there is no means of distinguishing one feature from another (e.g. one line from another) without considering inter-feature relationships. For distinctive image features associated with image content, e.g. regions such as scale-invariant features, the simplest encoding is the image itself. There is a large amount of redundant information within the image, however, and encoding strategies attempt to transform raw image information into a more useful form. The manner in which image data are encoded can be described within the dichotomy of compact vs. sparse coding [72]. Compact coding includes techniques such as principal component analysis (PCA), which strive to determine a low-dimensional linear basis that captures the majority of data variance in an image set [73, 27]. The compact coding basis is effective for reconstructing images from a small number of linear coefficients with minimum squared error [73]. Sparse coding, in contrast, typically involves a high-dimensional (possibly over-complete) basis, for which all but a small percentage of dimensions are zero for any given pattern. The sparse code representation has shown to be effective for discriminating different patterns [74] and has been proposed as a model of biological vision systems [75]. Sparse codes are typically characterized by components that are spatially-localized, oriented and statistically independent [76]. These closely resemble components arising from independent component analysis (ICA) [77, 78].

In the computer vision literature, a comparison of a variety of local feature encoding techniques showed the SIFT representation of [79, 22] to be superior to other representations in terms of discrimination when matching features in different images of the same planar surfaces [80, 81] or 3D objects [82]. Although not explicitly referred to as such, the SIFT representation bears many of the hallmarks of a sparse code, consisting of a high-dimensional histogram of 128 spatially-localized and oriented derivative components. Additionally, similarity of SIFT codes can be effectively measured via the Euclidean distance metric, indicating that individual components can be modeled as statistically independent and identically distributed (IID) random variables. Interestingly, a recent publication indicated that performing PCA on SIFT vectors may improve discriminability in the context of features arising from the same scene [83], although subsequent comparisons suggest that this may not be the case in the context of general object classes [84]. Other such embeddings [85] have been recently proposed, but have not yet gained widespread acceptance. Several variations on the SIFT encoding of gradient orientation histograms have been shown to be effective for object detection [86] and general inter-image feature correspondence [87, 81].

Measuring Feature Appearance Similarity: An important consideration in the choice of feature appearance representation is the manner in which correspondence or similarity will be computed between measurements. The particular measure of similarity used depends on modeling assumptions regarding the relationship between intensities in different images [88]. Correlation is useful when measurements are linearly related. The Mahalanobis distance [89] is relevant when measurements follow a Gaussian distribution. The Euclidean distance is a special case of the Mahalanobis distance where measurements are Gaussian and IID. Correlation and Euclidean distance are equivalent in the case where measurement vectors are normalized. The earth mover's distance (EMD) [90] and mutual information (MI) [91, 92, 93] are effective for evaluating similarity between measurements viewed as discrete distributions. Measures such as the MI and the correlation ratio [94] are useful for evaluating similarity between images acquired in different imaging modalities, for example MR and ultrasound [95]. Similarity measures and their parameters can be generally derived from examples of similar and/or dissimilar features [96, 97].

Enforcing Geometrical Constraints: While measuring feature appearance similarity is an effective starting point in computing correspondence, instances of false correspondence invariably arise. Features arising from different underlying structure in the world may appear similar resulting in correspondence ambiguity, and features arising from the same underlying image content may appear different due to noise and natural variability. To cope with these situations, a variety geometrical constraints between features can be applied to reduce false correspondences and increase true correspondences. Certain constraints are based on assumptions regarding the shape of the underlying 3D world. The geometry of features arising from a 3D plane in different images is governed by a homographic transformation, which can be locally approximated as a affine transform [98]. Features in images of the same arbitrary object or scene taken from different viewpoints follow the epipolar constraint defined by a fundamental matrix (FM) [99, 98], where a point in one image is constrained to lie on an epipolar line in the next image. Computing the geometrical constructs required to enforce geometrical constraints generally requires a number of correct correspondences, and thus may be difficult or inefficient. For example, the FM required to enforce the epipolar constraint requires a minimum of 8 point-to-point correspondences between two images [100]. Invariant features offer geometrical information beyond feature locations, which can be used to reduce the number of correct correspondences required. For instance, work done in conjunction with this thesis showed that the FM can be estimated from as few as 3 affine invariant feature correspondences [101]. Instead of global constraints, topological constraints such as inter-feature proximity can be applied, for example requiring that inter-feature distances remain constant in different images [56, 102]. Where images arise from the same viewpoint of an object, features are related by a similarity transform consisting of a translation, rotation and magnification [103]. A similarity transform is an efficient means of evaluating geometrical consistency of scale-invariant feature correspondences, as it can be estimated from a single correspondence. Geometrical consistency can be enforced via robust methods such as the geometric hashing [104, 50], the Hough transform [105, 106] or the random sample consensus (RANSAC) algorithm [107].

2.1.3 Modeling Pattern Appearance

Once image features have been defined, they can be combined into a model in order to describe abstract appearance patterns. At this point, it is worth distinguishing between modeling specific objects and modeling classes of visually similar objects. Modeling specific objects is much less demanding than modeling object classes, as local feature correspondence between different images of the same textured object is often dense and unambiguous. As a result, features arising from single image templates are often sufficient for the task of object recognition, i.e. detecting and localizing specific objects [79, 108] such as the puppet or bottle in Figure 2.4. Recognition of specific 3D objects from arbitrary viewpoints can be achieved by using a small set of images acquired from different viewpoints about the object [109] or by organizing features into a 3D model [110].

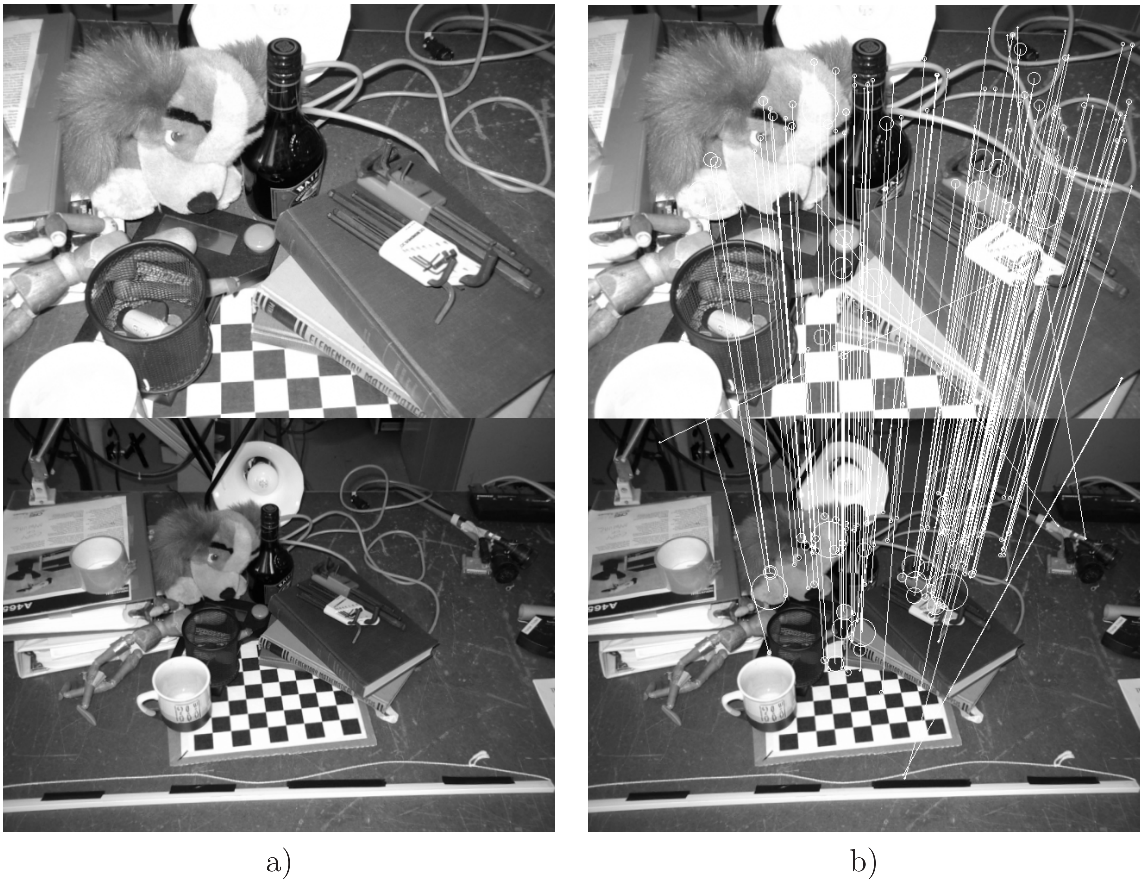


Fig. 2.4 An example of dense feature-based correspondence achievable between different views of the same textured scene. The upper and lower images in a) illustrate different views of the same scene. Note the significant change in illumination and scale. The white lines overlaying the same images in b) illustrate feature correspondences determined between the two images based on SIFT feature matching [22].

In the case of patterns arising from different objects from the same class, different faces or brains for example, direct feature correspondence is often weak or non-existent due to intra-class variation, as illustrated in Figure 2.5. In general, the range of appearance variability of such abstract patterns cannot be observed in a single image, but can be learned and summarized from a set of images spanning the range of pattern appearance variability. In order to do this, a model must first be designed in order to represent data in a useful and meaningful manner. This model must then be used in conjunction with algorithms capable of deriving this knowledge from image data. This section discusses such models and algorithms.

Modeling

The machine learning literature offers a wide range of modeling choices including decision trees, neural networks, etc. Models based on probability theory are attractive as principled mechanisms exist for parameter estimation and inference from data [111]. Probabilistic models consist of random variables whose parameters can be estimated from a data set in order to quantify data variability in a meaningful manner. Parameters of image-based models are typically based on variables of image intensity, image space, and mappings between different images. As images consist of a large number of data samples (intensities), directly modeling all samples and their interdependencies is generally intractable. Probabilistic modeling must thus resort to simplifying assumptions regarding parameters and their dependencies, which can be broadly classified according to the image scale at which modeling takes place, i.e. global vs. local models. Global models refer to those that considering the entire spatial extent of the image pattern simultaneously [27, 29, 112]. Variation in global models is typically quantified via a linear Gaussian model, consisting of additive ’modes’ of covariance about a mean. As all image data are related linearly, the modeling assumption is one of statistical dependence between spatially separated image regions, and the simplification arises from the fact that the majority of data variance can be accounted for by a small number of principal components [27]. Local models refer to those that consider appearance variation on a spatially localized scale [1]. The modeling assumption is that once aligned within a common reference frame, spatially separated image regions can be treated as statistically independent, and model simplification arises from the fact that only local interdependencies are modeled. Other models such as Markov models [113] can be considered a hybrid of global and local characteristics, where modeling is based on spatially local interactions which can propagate globally.

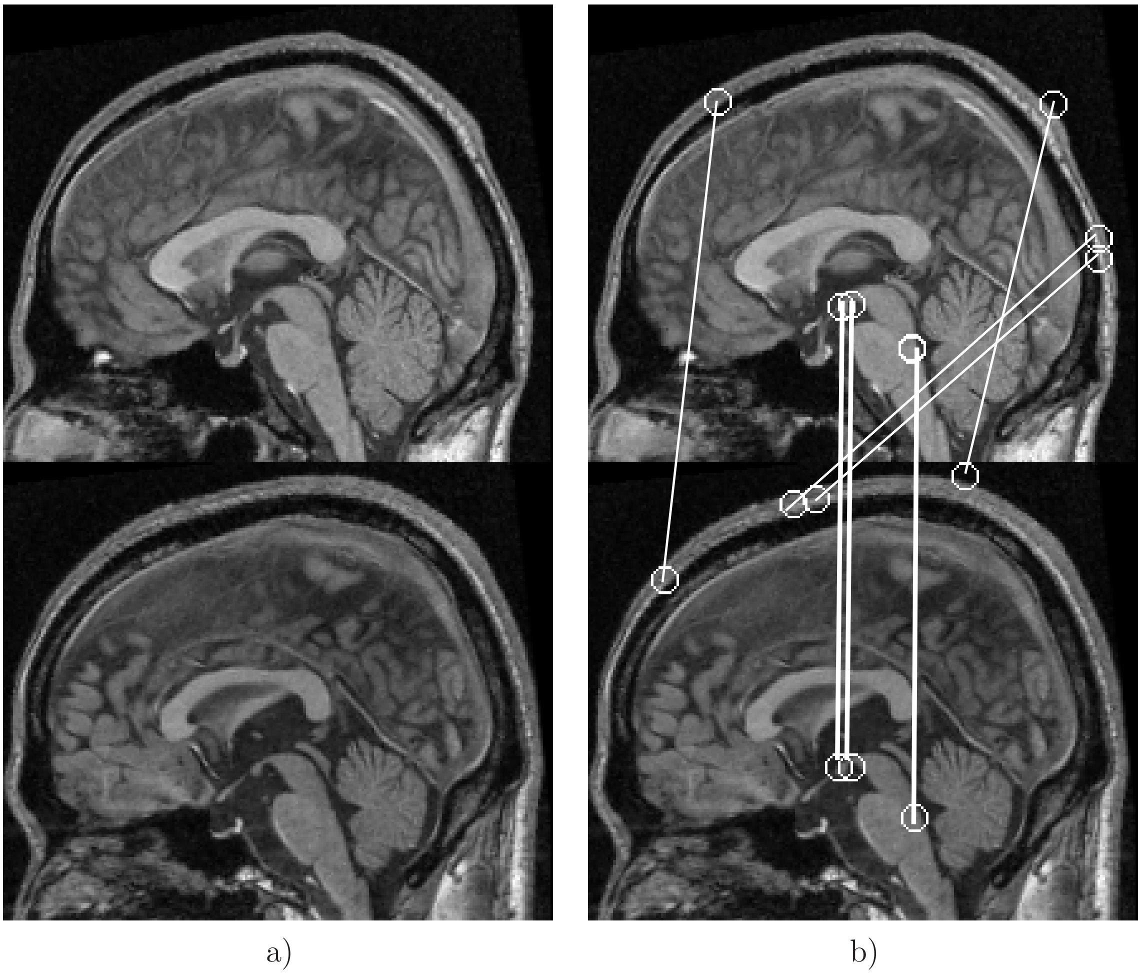


Fig. 2.5 An example of weak feature-based correspondence achievable between of different instances of the same abstract object class, here human brains. The upper and lower images in a) illustrate similar slices of different human brains. Note that similar structures exist in both images but vary significantly in shape. The white lines overlaying the same images in b) illustrate weak feature correspondences determined between the two images based on SIFT feature matching [22]. Only 3 correct correspondences are identified, along with several erroneous correspondences along the skull.

Probabilistic models of local image features as described in Section 2.1.1 are typically based on variables of feature appearance, occurrence and geometry. Central to such models is the notion of a *geometrical reference frame* that is defined in the image plane and related to the underlying object class in a geometrically meaningful way. The reference frame serves as a mechanism for spatially grouping image features arising from the same underlying object class instance, thereby providing a means of describing, identifying and localizing individual pattern instances within the image. Variables of feature appearance/occurrence can be modeled as conditionally independent given this reference frame. In this way, model features can be efficiently identified individually in images, and used to construct independent hypotheses as to their geometrical relationship within the image. Examples of geometrical reference frames include bounding boxes [114], feature constellations [1, 115], individual features [116, 117, 31] and flexible meshes [118, 3]. The geometrical reference frame is characterized by dependency assumptions regarding inter-feature geometrical relationships. A variety of relationships have been investigated, including full dependence where all features are related [119], Markov dependence where features are related to those in a local neighborhood [118, 3], Naive-Bayes dependence where features are related to a single 'parent' variable [116, 120, 121, 122] and full independence where no geometrical dependencies exist [123, 2, 124], in addition to intermediate approaches [117]. These spatial dependency assumptions are illustrated graphically in Figure 2.6.

The choice of a spatial dependence model has implications regarding time and space complexity of learning and inference, and the nature of the appearance patterns being modeled. Full independence is of low complexity as it requires no geometrical modeling, however no explicit mechanism exists for grouping features arising from the same object class instance, for the purpose of localization. Full dependency is generally intractable for models composed of more than a small number of features. The Naive-Bayes model is the minimal framework for relating features geometrically and is computationally efficient. Markov-type models may be better suited for modeling flexible or articulated objects. Both offer a reasonable compromise for incorporating geometrical constraints and model complexity.

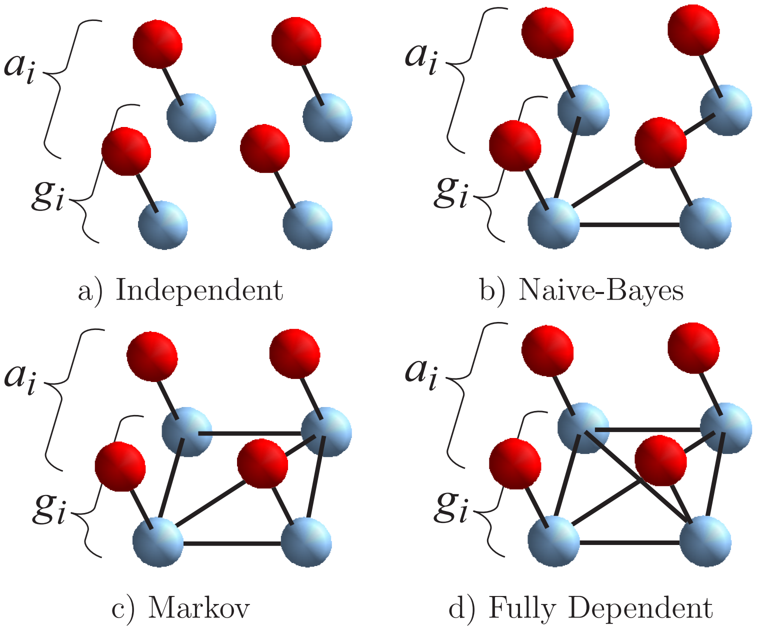


Fig. 2.6 A graphical illustration of independence assumptions regarding features in probabilistic local feature-based models. Nodes represent random variables and arcs represent statistical dependence. In each configuration, upper nodes labeled a_i colored in red illustrate variables of feature appearance, and lower nodes labeled g_i colored in blue illustrate feature geometries. Feature appearances are typically considered as conditionally dependent given feature geometries. Dependence assumptions on feature geometries range from a) independent models having no interdependencies to fully dependent models d) where all features are inter-dependent.

Learning

With assumptions regarding pattern appearance defined by a model, model parameters relating to feature appearance, occurrence and inter-feature geometry can be learned from images containing instances of the patterns to be modeled. While specific details of different learning algorithms may vary, there are general considerations regarding image feature representations and training data shared by most learning techniques.

General considerations regarding image features include the type of features to use and selecting sets of useful features. In terms of feature types, a common theme is to reduce image features into a discrete set of visual words or visual codebook. This can be done by partitioning feature appearance space, e.g. randomly [125] or via clustering techniques such as the k-means algorithm [126, 127]. The goal is to obtain a set of basis features which is general enough to account for intra-class appearance variability but informative enough for inter-class discrimination. With image features defined, algorithm-independent techniques [89] can be applied to improve the performance of modeling by selecting reduced subsets of model features which are effective for the task as hand, i.e. detection or classification. Boosting can be used to obtain a minimal set of independent model features [128], which can significantly improve the performance of detection or classification [129]. The technique of bagging can be applied in a similar manner [130]. The bootstrapping technique involves training a model, testing, then re-training the model while removing or penalizing features which lead to poor testing performance [131, 132].

General considerations regarding training data include the degree of supervision or labeling required in data preparation, the amount of data required and the variability of pattern appearance in the data. Supervised learning of object appearance models typically involves localizing, aligning and/or cropping object instances in training images prior to learning. In this way, the assumption can be made that all training image data are reflective of the pattern of interest, and that image correspondence has been achieved. Supervision is generally undesirable in that it requires either specialized equipment (e.g. a specialized image capture setup) or tedious manual labor, however the idea is that if it can be performed once off-line it will have been worth the effort if the final appearance model performs well at a particular task. For example, consider the task of detecting and localizing faces within arbitrary images, one the most prevalent computer vision applications and a good example of highly-supervised pattern learning. The detector of Jones and Viola [129] is widely recognized as one of the most effective face detectors, based on 2D Haar wavelet features [133] and the Adaboost supervised learning algorithm [128]. The detector identifies upright faces in cluttered scenes, runs in real-time, and has been extended to multi-view face detection [134]. The downside is the burden of the learning process, which requires on the order of 10,000 training faces, each of which is cropped, aligned, and sorted according to viewpoint bins, in addition to 100,000,000 non-face examples [134]. Other effective face detectors require similar learning scenarios [135, 132, 136].

Despite the effectiveness of supervised model learning approaches, they are less feasible when the goal is to learn the appearance of general object classes from natural, cluttered images. Although intensive manual labeling is possible for specific object classes such as faces, it is less realistic to manually label the large number of general object classes typically found in large image databases such as the Internet. As a result, recent work has focused on developing appearance models that can be learned from arbitrary data with little or no manual supervision [119, 137]. The so-called weakly-supervised learning scenario assumes a set of training images in which each image contains at least one instance of the pattern to be learned in an unknown geometry, and possibly a set of background images known to contain no pattern instances. Weakly-supervised learning has been accomplished via feature clustering algorithms such as the expectation-maximization algorithm [1]. Other work has focused on learning appearance models from very few (single) training images, by generalizing prior knowledge regarding pattern geometry [138] or feature appearance [139] to new patterns. An additional challenge is that of simultaneously learning multiple patterns from images containing multiple instances of different object classes [137]. Although models that lend themselves to learning in the context of minimally labeled data have not yet been shown to be effective for tasks such as face detection in general imagery [1], they are interesting in that they could someday be used to learn models of arbitrary patterns in natural imagery in a fully automatic fashion.

2.2 Computer Vision: Modeling 3D Object Class Appearance

This thesis focuses on the goal of learning, detecting, localizing and classifying general 3D object classes in terms of visual traits, all in natural cluttered images acquired from arbitrary viewpoints. Here, Section 2.2.1 reviews literature pertaining to the challenge of modeling and learning the appearance of general 3D object classes from arbitrary viewpoints, which the computer vision research community has only begun to address. Section 2.2.2 reviews literature pertaining to classifying visual traits from objects, focusing on the special case of determining the sex of face images.

2.2.1 Modeling 3D Object Class Appearance

Global modeling methodologies have been proposed for describing object appearance over changes in viewpoint, via subspace manifolds [28], view-based approaches [140] and locally linear embeddings [141]. The global modeling methodology suffers from the previously mentioned drawbacks, i.e. the inability to account for partial occlusion or local scale appearance variability. The majority of general object class appearance models based on robust local image features have focused on modeling single viewpoints. As a result, the geometrical reference frames used to relate underlying features are often single-viewpoint in nature, i.e. they do not remain geometrically consistent with the underlying 3D object class when there is a change in viewpoint. Bounding boxes change in size and shape with in-plane rotation, 2D feature configurations collapse [142] and individual features disappear with in-depth rotation. It has been observed, however, that local image features typically persist over a viewpoint range of up to 25 degrees surrounding 3D objects [22, 82], as illustrated in Figure 2.7. Given this observation, there are two mechanisms by which single-view models can be extended to modeling 3D object class from arbitrary viewpoints, via the multi-view or the viewpoint-invariant representation [143, 144].

The multi-view representation models the variable of viewpoint explicitly by maintaining a sampling of distinct single viewpoint models around the object of interest [4, 145, 132]. Detection is accomplished by fitting a new image to the nearest modeled view, the key assumption being that a reasonably similar view exists in the model. Multi-view modeling requires several important considerations. First, an adequate sampling of views around the object class of interest must be established. Over-sampling will result in a bulky model, and under-sampling will result in poor model fitting to new images falling between views. A set of views can be fixed for specific objects [4], but the same set of fixed views may not be effective or well-defined for different object classes. Certain views may be more informative [146] or representative [147] than others depending on the object class or specific class instances, which would influence sampling. Techniques have been proposed for automatically determine a set of stable views or aspects for specific objects [148, 149, 109]. It is

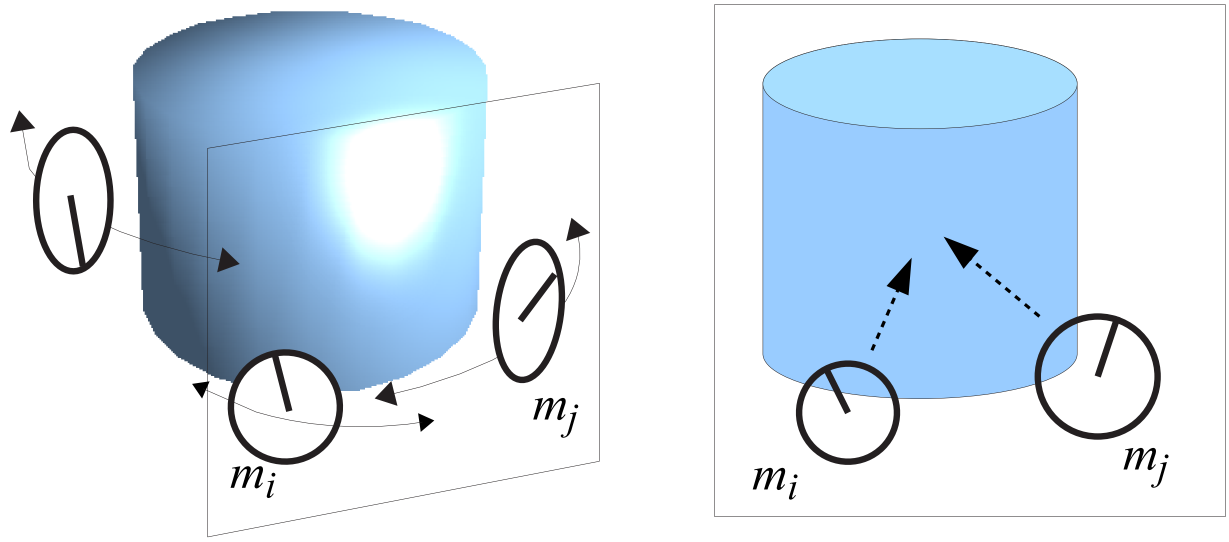


Fig. 2.7 Scale-invariant features m_i and m_j can be observed in images acquired from a range of viewpoints around a 3D object (left image). When observed in a new image from an arbitrary viewpoint (right image), features can be used to localize the object within the image plane.

not clear that these techniques are applicable to modeling general object class appearance, however, due to the variability in appearance exhibited by different object instances from the same viewpoint.

The viewpoint-invariant representation relates image features to a geometrical reference frame that is invariant to viewpoint, thereby effectively marginalizing the variable of viewpoint from the model. Detection is then accomplished by identifying the reference frame from image features in a manner independent of viewpoint. The key to the viewpoint-invariant representation is that the geometrical reference frame must represent a property of the underlying 3D object class that is invariant to perspective projection arising from viewpoint change, i.e. a viewpoint-invariant [150]. Examples include 3D object centroids which project to 2D points [151], 3D volumetric primitives such as geons [26] or spheres which project to 2D circles, 3D line segments which project to 2D lines [31, 150]. In this way, a viewpoint invariant reference frame will remain geometrically consistent with the underlying 3D object class observed in an image plane.

Early approaches to viewpoint-invariant modeling were based on identifying viewpoint-invariant properties of specific classes of 3D shapes, for example contours of parameterized volumetric shapes such as generalized cylinders [152, 153]. While feasible for modeling primitive 3D shapes, these approaches are difficult to apply to modeling general object classes in natural imagery, as viewpoint-invariant properties are rare [150] and difficult to identify in the presence of noise and clutter. The viewpoint-invariant model has not been widely investigated in the recent wave of approaches modeling object classes from local image features, with the exception of Weber et al. [142] who proposed learning a viewpoint-invariant model of 3D faces based on a single-viewpoint constellation model. The success of this approach was limited, however, as the constellation model is defined in a single viewpoint and not able to effectively cope with arbitrary viewpoints. Note that models containing no geometrical information used for image classification, e.g. the "bag-of-features" approach, provide a viewpoint-independent mechanism for modeling feature appearance [154, 2], but cannot be used directly to localize or enumerate object class instances within cluttered images. Note also that a mechanism for linking object appearance from two different viewpoints is presented in [120], but used for the task of recognizing the same object from two different viewpoints, not for general appearance modeling over an entire object class.

Although single-viewpoint models can be learned in a weakly-supervised fashion [119], learning the appearance of 3D object classes from arbitrary viewpoints is a significantly more challenging problem, as feature correspondences must be established both across viewpoints and across object class instances. As a result, multi-view approaches learning techniques require viewpoint information, either via manual labeling [3] or by capturing different images around the same object class instances [155, 156, 4], neither of which is necessarily available in arbitrary, natural image databases such as the Internet. The multi-view model of Thomas et al. [4] requires a high degree of supervision. Data are captured and sorted according to precise, regular viewpoint intervals using a special purpose studio, for each of a number of different object class instances. The multi-view models of Yan et al. [156] and Savarese and Fei-fei [155] require images captured from different viewpoints around the same object class instances. The flexible model of Kushal et al. [3] requires less supervision: object class instances are localized in the image and sorted according to coarse viewpoint bins. Due to the need to estimate parameters of viewpoint, it is difficult to learn multi-view models on databases of natural imagery where viewpoint information is not available. The advantage of the viewpoint-invariant approach is that viewpoint information is not required at any stage, reducing the number of model parameters and significantly simplifying model learning, particularly in databases of natural imagery where viewpoint information may not be present or easy to determine.

2.2.2 Visual Trait Classification: Sex from 3D Faces

Visual traits are abstract qualities of an object class identifiable from images, such as the make or model of cars, the age or sex of faces, etc. They represent a mechanism by which members of the same object class can be described or subdivided into meaningful categories. While subcategories can be defined in a hard, taxonomical fashion according to specific image features [157] or segmentations [158], visual traits such face sex are not generally determined by any individual feature but rather by their ensemble within the image. A wide range of visual traits are used by humans in order to describe objects of a class, and learning and classifying these traits is of significant interest. Facial traits such as age and sex for instance can be very useful as soft biometrics in surveillance systems, for instance. Due to the ubiquitous nature of face image analysis, one of the most common visual trait classification tasks is that of determining sex from face images. The wide range of published approaches to face sex classification can be seen as highlighting the state-of-the-art in general trait classification. Sex classification has been tackled from spatially global feature representations such as templates[159, 9], principal components [8], independent components [7] or image intensities directly [5]. Much work has contrasted the use of different machine learning techniques such as neural networks [9], support vector machines (SVMs) [8] and boosted classifiers [5]. More recently, trait classification based on local features has emerged, using localized image regions [160, 161] or Haar wavelets [6, 162].

In order to classify visual traits in a practical scenario, there must be a means of reliably identifying the image features on which classification is based (detection), associating those features with object class instances (localization), in natural, arbitrary imagery. To date, no work has addressed sex classification from arbitrary viewpoints or in the presence of occlusion. All published approaches to sex classification are based exclusively on single viewpoints, i.e. frontal faces [5, 6, 7, 8, 9]. With the exception of [162], most approaches assume that, prior to classification, faces are precisely localized and background distraction such as hair and clothing is cropped away. For example, localization is performed by manually specifying eye locations [5] or using special-purpose frontal face alignment software [6, 8], and pre-defined facial masks are subsequently applied to remove background clutter. As a result, classification error rates of 4% to 10% represent artificially low, ideal-case results, and offer little insight as to classification performance in a general vision system where object class localization is non-trivial. Indeed, recent work evaluating the effect of artificial geometrical perturbations on classification accuracy showed that accuracy drops off rapidly with even small independent perturbations in scale and orientation (e.g. a 5 degree in-plane rotation) [5]. An additional fact worth noting is that several published works reporting low error rates use different images of the same person in both classifier training and testing [6, 8]. As facial features arising from different frontal images of the same person are highly correlated, one cannot know whether the low classification error reported reflects the ability of the classifier to generalize to new, unseen faces or are simply performing classification-by-recognition.

Only a single approach has proposed a framework for general visual trait classification based on Haar wavelet image features which can be robustly detected and localized in the presence of partial occlusion [162], based on the Viola-Jones face detector [129]. The approach is single-viewpoint (frontal faces) and not invariant to in-plane rotation changes. The reported error rate of 21% reflects the increased difficulty of the combined task. This result is based on a proprietary database, however, where faces with ambiguous sex are manually removed, as are faces whose in-plane orientation is greater than 30 degrees.

2.3 Medical Imaging: Describing Anatomy

Developments in the field of medical imaging often parallel those in computer vision, as both fields deal with processing image data. The questions of interest in medical imaging often differ from those in computer vision, however, due to the nature of the images used and the clinical goals. The study of human brain imagery, for example, is the focus of intense inter-disciplinary research, bringing together neuroanatomists, neurosurgeons, engineers, computer scientists, and others. A core research goal is the development of anatomical models capable of describing brain shape, development and functionality in intuitive, meaningful terms. A primary focus lies in generating such models from images of a population, such as MR images of the human brain, in order to describe and quantify underlying anatomical structure and its variability. It is important to be able to link anatomical structure to traits such as pathology, in order to understand the relationship between the two. It is important to be able to robustly identify modeled anatomical structure in new images, despite inter-subject variability and unexpected perturbation possibly arising from natural variation or abnormality. Finally, it is important to focus on computational approaches which can be used in an automatic fashion with little or no manual input, in order to efficiently process large medical image databases. A general anatomical model with these characteristics has a wide range of potential applications, ranging from morphological study to diagnosis of pathology to understanding inter-species variation and evolution.

Although generating anatomical descriptions has traditionally been the job of trained human anatomists, approaches such as computational anatomy [17] have emerged in response to the need for automated tools capable of processing massive amounts of medical image data currently being generated. This section reviews the main bodies of literature relating to computational approaches to anatomical analysis, including inter-subject registration in Section 2.3.1, statistical modeling techniques in Section 2.3.2 and modeling of subject traits in Section 2.3.3. The use of local features in medical image modeling is outlined in Section 2.3.4 and the special case of modeling the highly variable cerebral cortex is described in Section 2.3.5.

2.3.1 Inter-subject Registration

In order to describe and measure anatomical variability between subjects in a population, images of different subjects are typically first registered into a common reference frame. This alignment task, known as inter-subject registration, represents a major challenge, as no two subjects are identical in appearance: image structures can vary significantly from one subject to the next, or simply may not exist in all subjects. A large number of techniques have been proposed for the task of inter-subject registration, the majority of which either implicitly or explicitly formulate the task as one of determining one-to-one correspondence between subjects or between a model and a subject [17, 12], driven by a measure of image similarity. Low-parameter linear registration techniques are simple and capable of determining initial coarse alignment between subjects, but do not properly account for inter-subject variation on a local scale. Deformable registration techniques involving the estimation of highly parameterized deformation fields from one image to the next have been proposed to account for variation on a local scale. Different deformable registration formulations can generally be distinguished by the manner in which the deformation field is constrained or regularized [163], examples include elastic registration [164], fluid registration [165], finite element models [166], thin plate splines [167], etc.

In the case of inter-subject registration, however, it is not clear that the estimated deformation fields represent meaningful correspondence between underlying tissues [12]. Different regularization approaches generally result in different deformation field solutions for a given pair of subjects, particularly in regions where correspondence is ambiguous, such as areas of homogenous image intensity. In general, the higher the number of deformation parameters, the more ways a deformation field can be constructed between any two subjects to obtain a near-perfect mapping in terms of pixel intensity error. This could explain the proliferation of novel highly-parameterized registration formulations in the medical imaging literature that report low error in terms pixel intensity. Recently, quantitative comparison of 6 different inter-subject registration methods showed no significant difference between high and low parameter registration approaches in terms of manually labeled ground truth [163]. In this case, the principle of Occam's razor [89] would suggest that the simpler, reduced parameter model would be more plausible.

In general, the challenge of inter-subject variability is compounded by the absence of a gold standard for validating inter-subject registration [12]. Given these difficulties, the *a priori* assumption of the existence of a one-to-one mapping between different subjects is often difficult to justify, particularly in highly variable regions such as the cortex or in case of abnormal subjects or those exhibiting pathology. Furthermore, it is reasonable to expect that models based on one-to-one correspondence perform poorly in locations where such correspondence does not exist. Although the case of smooth, diffeomorphic mappings between subjects has received a large amount of attention [168, 169, 17], the case where one-to-one correspondence does not exist is rarely modeled, with the exception of outlier detection techniques [170]. Neither of these approaches lends itself to effectively describing multiple modes of local appearance, for example a particular gyrus of the brain bearing multiple anatomical morphologies.

2.3.2 Statistical Appearance Models

Statistical appearance models have been developed in medical image analysis in order to quantify inter-subject variability. As mentioned in Section 2.1.3, statistical models of images are typically based on variables of image intensity, image space, and mappings between different images, and can be contrasted as global or local in nature. The particular manner in which variability is quantified depends on the purpose of the model. Global models, such as the multi-variate Gaussian, are often used to quantify appearance and geometry for the purpose of inter-subject registration [29] in order to generate anatomical atlases [15, 169]. Global models can be applied in a variety of ways, e.g. over combined intensity and shape [29], over deformation fields used to align different subjects [171, 12], etc. Local models have been used to quantify localized intensity or segmentation variation once subjects have been aligned into a common reference frame, a prominent example being morphometry methods [14, 15, 16]. Hybrid approaches such as Markov models have been used for medical image segmentation [113], defined by local dependencies which can propagate globally.

A major difficulty in statistically quantifying inter-subject variability is that models based either implicitly or explicitly on the assumption of one-to-one correspondence tend to break down when the assumption is invalid. Global models assuming statistical dependency between spatially distinct regions, such as the active appearance model (AAM) [29], are generally unable to properly account for deformation on a local scale. While this may be less important when considering the shape of simple isolated structures such as the hypothalamus [113], it is problematic when modeling the brain as a whole where localized variations are commonplace. While local models are able to account for spatially localized phenomena, they are prone to confounding observations of different underlying brain structures when the assumption of one-to-one correspondence is invalid, and susceptible to error introduced by the method used to obtain alignment, such as bias [169]. To minimize these problems, modeling is often based on 'normal' brains free of abnormalities [15, 17]. It has been observed, however, that even brains classified as clinically 'normal' can exhibit significant variation [171].

2.3.3 Identification of Subject Traits

A central goal of medical image modeling is to automatically identify anatomical commonalities or differences between different subject groups, for example healthy and diseased subjects. Automated segmentation and classification systems can be used to relate specific structures of interest to subject traits, e.g. particular sulci [13] to brain sex or hemispheres [21]. However, focusing on specific structures is less useful when it is not known *a priori* which image regions exhibit the strongest differences between the sub-groups, and thus which structures must be segmented. Correspondences can be identified manually by labeling and measuring structures of interest over a set of images of subjects [172] in different groups. However, manual correspondence is laborious and impractical in any medical database of non-trivial size, particularly when it is not known *a priori* which structures are relevant for the particular trait. Furthermore, it is subject to human error and results are subject to inter-rater and intra-rater variability.

General approaches to automatically learning image characteristics reflective of subject traits such as morphometry [14, 15, 16] or machine learning [173] analyze subject images which have been aligned into a common reference frame defined by a global image template or atlas. Once aligned, group differences can be reflected in differences in image intensities or deformation fields required to bring subjects into alignment. The drawback of this methodology is that, as mentioned, inter-subject alignment may be not be valid everywhere due to the non-existence of one-to-one correspondence between subjects. As a result, the outcome of subsequent morphometric analysis may be difficult to interpret, as the analysis may be confounding image content arising from different underlying anatomical tissues. This particular issue has been remarked upon by several authors [18, 19, 20].

2.3.4 Use of Local Features

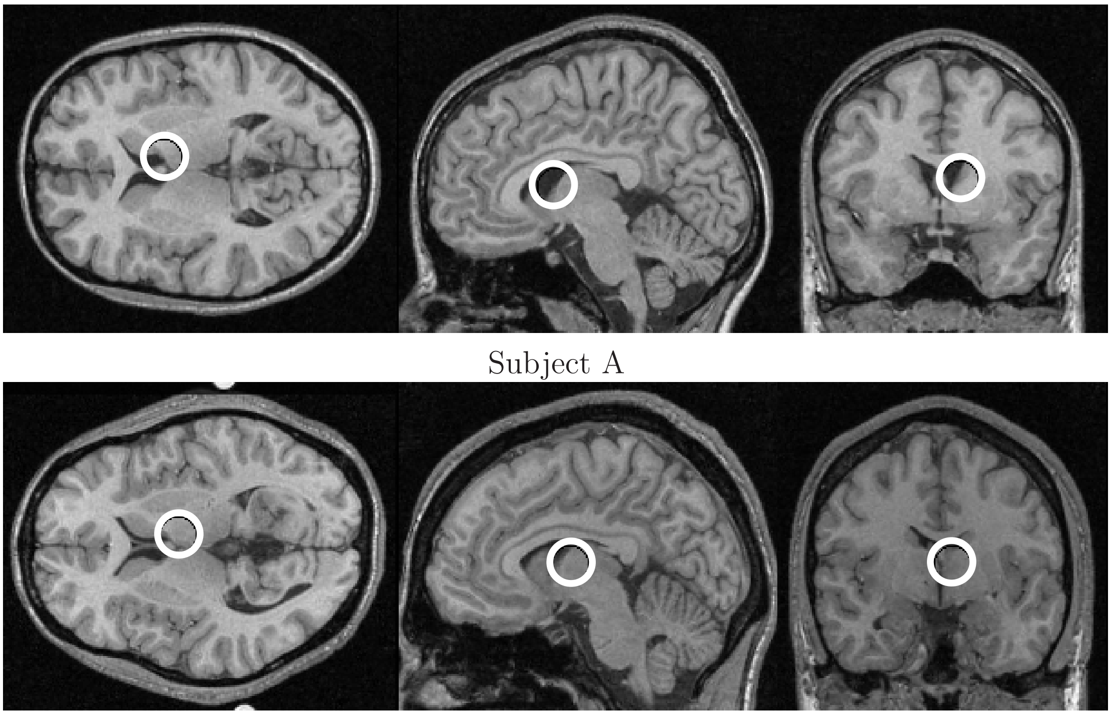
Effectively registering images of different subjects requires addressing 1) the situation where one-to-one correspondence is ambiguous or non-existent as highlighted in the previous sections and 2) the issue of computational efficiency for processing large data sets. Feature-based correspondence offers a solution to these issues by focusing on matching a set of local, informative image regions between images. As processing is based on a small set of localized features instead of entire images, correspondence can be calculated from a fraction of the total image data and can be avoided in regions where a valid correspondence may not exist.

Features used in medical imaging can be domain-specific, designed for identifying particular structures of interest such as particular brain folds or sulci [174, 175, 13, 176], or generic, identifying natural salient patterns in arbitrary imagery. The advantage of techniques based on generic features, such as invariant local features, are that they can be applied in a variety of different imaging contexts. Generic detection of point-features based on derivative operators in 3D medical imaging contexts has been proposed [177, 178] and used to register images of the same subjects [179]. Although generic scale-invariant features have been extensively used in the computer vision, their use is more recent in medical imagery where they have served primarily as a means for establishing correspondences between different images of the same subject [180, 181]. Estimating feature scale may be perceived as less important in many medical image analysis contexts, as images are often acquired under similar conditions and many do not exhibit significant scale differences from one image to the next. Although the scale component σ of invariant features is useful for establishing image-to-image correspondences in the presence of image scale change, its primary purpose is to define a characteristic scale of the underlying image content. Recently authors have begun to investigate the importance of optimal local feature scale in improving medical image registration [182]. Scale-invariant features have been developed for 3D video data [183] and 3D volumetric medical image data [184]. Extraction of feature locations and scales in N-dimensional image data is a straightforward extension of 2D techniques using scale-space pyramids, as shown in Figure 2.8. A challenge remaining is to effectively cope with the increasing number of feature orientations parameters in higher image dimensions, in order to effectively compute invariant correspondences between different subjects.

2.3.5 Modeling Cortical Appearance

Modeling the cortex from MR imagery of the human brain is a task of high interest in the medical imaging community, as cortical folding patterns are closely related to functional divisions within the brain. Describing the appearance of the cortex has proven a challenging task, even for human experts, due primarily to inter-subject variability. Particular cortical folds may vary significantly in shape from one individual to the next, possibly exhibiting multiple distinct morphologies. Although major structures such as lobes and primary folds such as the central sulcus and lateral fissure are present and readily identifiable in virtually all adult brains, secondary or tertiary folding patterns cannot typically be reliably identified in all subjects even by human experts [185]. This can be visualized in Figure 2.9.

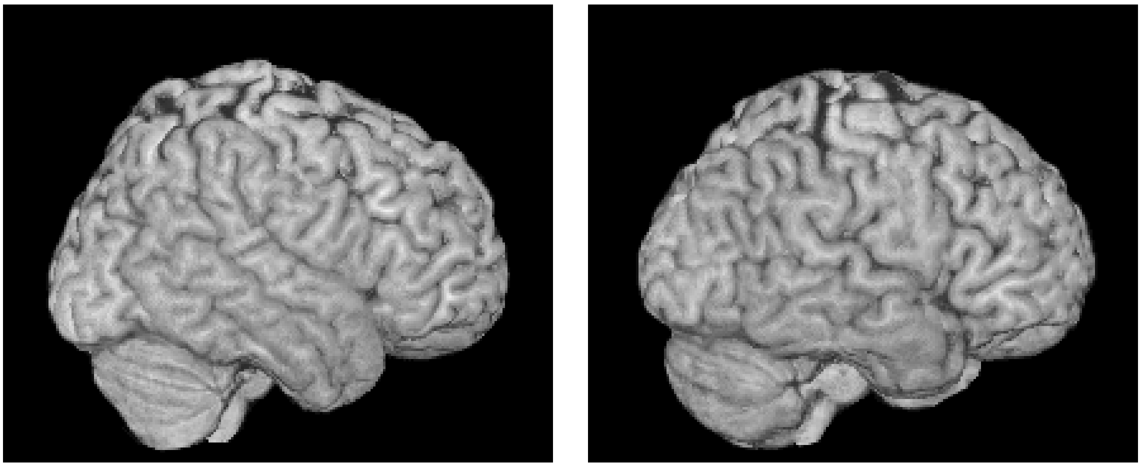
A variety of different techniques are used to describe the cortex. Manual description is widely used in the neuroanatomical community, where neuroanatomists painstakingly identify and classify folding patterns [186]. Manual description is prone to rater error and is tedious, particularly for the large data sets required to observe the full range of cortical appearance. Images of different subjects can be registered into a common reference frame or atlas and subsequently analyzed, but inter-subject cortical registration is a daunting



Subject A

Subject B

Fig. 2.8 An example of similar scale-invariant features identified in 3D brain volumes of two different subjects, A and B. The upper row of images illustrates a feature in coronal, sagittal and axial image slices of subject volume A. The lower row illustrates a feature of similar location and scale identified in a different subject. Here, features are extracted using a DOG scale-space pyramid in 3D [22], where each volume results in several thousand features.



a) b)

Fig. 2.9 Lateral surface renderings of the cortex from two different subjects. Note that although several large-scale structures such as major fissures and lobes can be identified in both brains, similar folding patterns are generally difficult to identify even by trained neuroanatomists.

task. Most registration approaches attempt to determine a one-to-one mapping between different subjects [12], whereas such a mapping does not typically exist in the cortex [163]. Image features arising from cortical sulci and gyri can be extracted and used to improve registration between different subjects [187, 13], although resolving correspondences remains difficult. Machine learning approaches have been used to reproduce expert sulcal labelings from training examples [175, 188]. These approaches are designed to detect labeled sulci known to exist in new images, however, and do not attempt to learn new, unlabeled patterns that may not exist in all subjects.

2.4 Summary

This chapter summarized the major bodies of work related to this thesis. Section 2.1 began with a general discussion of work relating to pattern appearance modeling, in particular image features and probabilistic modeling techniques. Related work was then presented in the specific computer vision and medical imaging contexts where this thesis makes contributions.

There are several distinct gaps of research which the work in this thesis aims to fill. In the computer vision literature described in Section 2.2, modeling the appearance of visual object classes for the purpose of detection and localization has become an area of intense research. Most approaches have been restricted to single views of object classes, for instance frontal views of faces, side views of cars, etc. Several approaches have begun addressing modeling appearance from arbitrary viewpoints [4, 3, 155, 156], all of which have proposed modeling viewpoint explicitly via multi-view formulations. The OCI model presented in the following section is one of the first approaches to address detection and localization of 3D object classes, and the first approach to do this in a viewpoint-invariant manner. In another vein, techniques aiming to classify faces in terms of visual traits such as sex have focused almost exclusively occlusion-free, frontal faces, which have been pre-aligned and cropped prior to classification. The OCI model presented in the following section represents the first approach to address trait classification from arbitrary viewpoints, and the first to combine detection, localization and classification into a single framework.

In the medical image analysis literature described in Section 2.3, techniques used to model anatomy over a population of different subjects, in particular that of the human brain, often assume the existence of global one-to-one correspondence between different subjects. Although this assumption permeates many aspects of anatomical modeling from inter-subject registration to subsequent morphometric study, it is often invalidated by inter-subject variability due to factors such as pathology, surgical resection, and multiple modes of morphology in regions such as the cortex. It is reasonable to expect that registration and morphometric analysis perform poorly in situations where one-to-one correspondence does not exist. The OCI model presented in the following chapter represents the first time brain anatomy has been described and analyzed in terms of a collection of conditionally independent local image features, which can be automatically identified in a robust manner in different subjects, and used as a basis for identifying anatomical characteristics reflective of subject traits such as pathology, age or sex.

Chapter 3

The Object Class Invariant Model

This chapter proposes a new, general approach to modeling image patterns arising from classes of similar objects. There are significant challenges that must be overcome in order to model image patterns in arbitrary, natural imagery. There must be a means for successfully identifying instances of the same pattern despite appearance variability due to changes in illumination, in-plane geometrical deformations such as translation, scale and orientation changes, the presence of clutter and background noise, in-depth geometrical deformations due to viewpoint changes, intra-pattern variability, in particular multi-modal intra-pattern variability. The model must lend itself to mechanisms typically used to describe image patterns, for example with respect to geometrical frames of reference used in the analysis of anatomy in medical imagery, or in terms of abstract visual traits, for example human faces in terms of sex or age. The model must be computationally efficient in order to cope with large amounts of image data, both for learning an accurate model of appearance and for identify model instances in new images. Furthermore, it must generalize to many different patterns and imaging contexts, in order to be practically useful.

The original theoretical contributions in this chapter stem from manner in which sources of pattern variability are modeled. In general, variability can be handled either explicitly by directly modeling sources of variability, or implicitly by constructing the model in a manner invariant to the sources of variability. Explicit modeling requires careful analysis in order to correctly identify, parameterize and model sources variability, which is difficult and computationally expensive when a large number of sources are to be considered. Furthermore, explicit models may not generalize between different imaging contexts, and are generally unnecessary when sources of variability can be considered as nuisance parameters, i.e. parameters which are fundamental to the imaging process but not necessarily of interest for the task at hand. To avoid these difficulties, this thesis opts for model invariance, by modeling image patterns with respect to a geometrical reference frame referred to as the object class invariant (OCI), which is defined to be invariant to parameters considered as nuisances. With an invariant model defined, remaining uncertainty in pattern appearance due to unmodeled sources of variability is then quantified probabilistically.

The remainder of this chapter is organized as follows. Section 3.1 describes the OCI model, along with issues of learning, fitting and determining an optimal OCI. Section 3.2 describes how the OCI model can be used in the field of computer vision to learn, detect localize and classify 3D object classes in terms of visual traits such as sex, in natural cluttered imagery taken from arbitrary viewpoints. Section 3.3 describes OCI modeling in the field of medical imaging, where it can be used as parts-based anatomical model, which can be learned, robustly fit to new subjects where one-to-one correspondence does not exist, and used identify anatomical structure related to traits such as sex or pathology.

3.1 OCI Modeling Theory

This section presents the theory behind modeling image patterns in terms of a geometrical reference frame, referred to as an OCI, with respect to which pattern appearance can be effectively described in terms of invariant local image features. Section 3.1.1 describes the OCI model, including its components and probabilistic formulation. Sections 3.1.2 and 3.1.3 describe how the OCI model can be learned from a set of images and fit to new images. Section 3.1.4 describes how a learned OCI model can be used as a basis for learning and classifying abstract visual traits.

3.1.1 The OCI Model of Pattern Appearance

In the OCI model, a set of local image features or parts are modeled with respect to a common reference frame or OCI, as illustrated in Figure 3.1. The OCI is defined as a geometrical reference frame which is 1) uniquely defined for each pattern instance and 2) invariant to nuisance parameters arising from the imaging process. As such, the OCI is a geometrical reference frame within which the variability of parts can be normalized and quantified. The particular OCI used is often application specific. In the context of MR brain imagery for instance, a well-known example is the midplane line defining the Talairach stereotaxic space [189], which passes from the superior aspect of the anterior commissure to the inferior aspect of the posterior commissure. In the context of 3D face modeling, the OCI could take the form of a line segment central to the face. The significance of the OCI to statistical modeling is that different features or image 'parts' can be considered as conditionally independent given knowledge of the OCI. Specifically, knowing the reference frame, parts can be automatically normalized with respect to reference frame scale, rotation and translation. At this point, the appearance variation and the remaining geometrical variation of parts can be quantified probabilistically on a local scale.

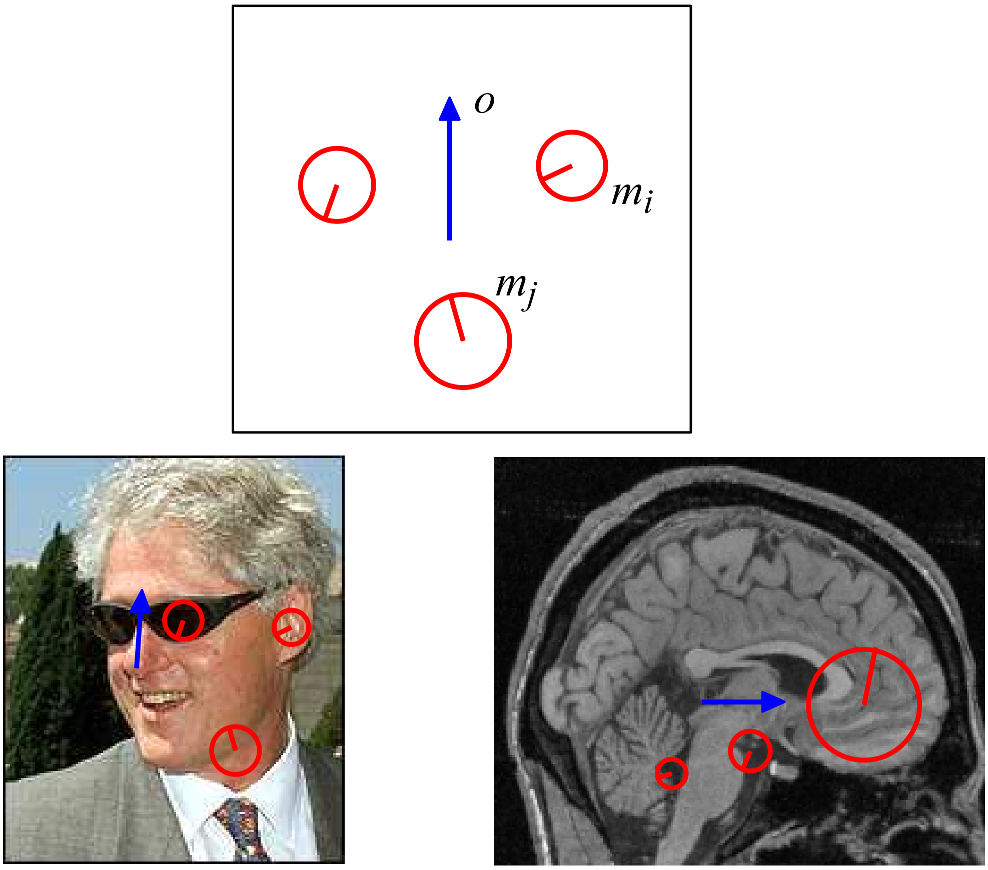


Fig. 3.1 The top diagram illustrates the components of the OCI model, local image features m_i and m_j (circles) within a geometrical reference frame o (arrow). The OCI model can be generally applied to describing a variety of patterns arising from classes of similar objects, such as faces (lower left) or brains (lower right).

Model Components

A model part represents the geometry, appearance and occurrence frequency of a scale-invariant image feature that occurs with statistical regularity in images of an object class. Within this thesis, the term 'feature' refers to a specific instance of a model part within an image, whereas a 'part' or 'model part' refers to the abstract local image pattern from which features arise or are generated. A model part is denoted as $m_i = \{m_i^b, m_i^g, m_i^a\}$ representing the occurrence, geometry and appearance of a scale-invariant feature within an image, respectively. Part occurrence m_i^b is a binary random variable representing part presence or absence in an image. Part geometry $m_i^g = \{x_i, \theta_i, \sigma_i\}$ is an oriented region in $\Re^N$ image space, represented by N-parameter location x_i, an $N(N-1)/2$ parameter orientation θ_i, and a scale σ_i. Part appearance m_i^a describes the image content at region m_i^g, and can generally be parameterized in a number of ways, such as principal components [27], histograms of gradient orientations [22], etc.

The OCI is denoted as $o = \{o^b, o^g\}$ representing the occurrence and geometry of a geometrical reference frame with respect to which the geometric variability of parts can be quantified. Note that o is identical to a model part with the exception of the appearance component. This is because o is not observed directly from the image, but must be inferred via model parts m_i. Reference frame occurrence o^b is a binary random variable indicating the presence or absence of the reference frame, i.e. whether or not an object of a particular class is present in a scene. Reference frame geometry o^g is parameterized in the same manner as model part geometry m^g. This implies that a single observed part m_i is sufficient to infer o^g in a new image, via a learned linear geometrical relationship. The following section describes the formulation of the probabilistic model for quantifying part appearance, geometry and occurrence frequency.

Probabilistic Model Formulation

Probabilistic modeling requires defining distributions over random variables and making assumptions regarding random variable independencies. The OCI model consists of a set of M model parts $\{m_1, \ldots, m_M\}$, denoted as $\mathbf{m}$, which, when observed in a new image, can be used to infer the OCI o. This can be expressed as the posterior conditional probability density function of o given $\mathbf{m}$, which, by applying Bayes' theorem, can be expressed as follows:

$$p(o|\mathbf{m}) = \frac{p(o)p(\mathbf{m}|o)}{p(\mathbf{m})}. \tag{3.1}$$

In equation (3.1), $p(o)$ is the prior probability density function over the geometry and occurrence of the OCI in the absence of model parts. $p(\mathbf{m}|o)$ is the likelihood function representing the stochastic relationship between model parts and the OCI. $p(\mathbf{m})$ is the joint probability density function of all model parts. Note that, as both $p(\mathbf{m}|o)$ and $p(\mathbf{m})$ represent joint density functions over all model parts, they are intractable to use as the number of parameters required to represent inter-feature dependencies is generally exponential in the number of model parts. Simplifying assumptions are adopted in this thesis in order to cope with tractability issues, and these are listed here. It is important to emphasize that these assumptions may be changed without changing the probabilistic formulation in equation (3.1).

Assumption 1 - Conditional Independence of Parts: Assuming that parts m_i are conditionally independent given o, the conditional probability in equation (3.1) can be further expressed as:

$$p(o|\mathbf{m}) = \frac{p(o)p(\mathbf{m}|o)}{p(\mathbf{m})} = \frac{p(o)\prod_i^M p(m_i|o)}{p(\mathbf{m})}, \tag{3.2}$$

where $p(m_i|o)$ is the likelihood function of o associated with model part m_i. The assumption of conditional independence of localized parts generally implies that knowing reference frame o, all remaining variation in part m_i can be described locally. To illustrate, consider local image features arising from face images. The assumption of conditional independence states that given that the geometry of the face is known, features arising from the eyes, for instance, will not provide information regarding features arising from the mouth. Thus, knowing that the eyes are open or closed will not provide any information as to whether the mouth is open or closed. This assumption, referred to as the 'naive Bayes assumption', has several important implications regarding modeling. It assumes the minimum degree of inter-feature statistical dependency required to relate all image features within a single geometrical reference frame. Therefore, the space and time complexity modeling (i.e. parameter estimation) are minimized. As all features are modeled locally and independently, the amount of data required to estimate model parameters is minimized. A Bayesian network diagram equivalent to the probabilistic expression in equation (3.2) is illustrated in Figure 3.2.

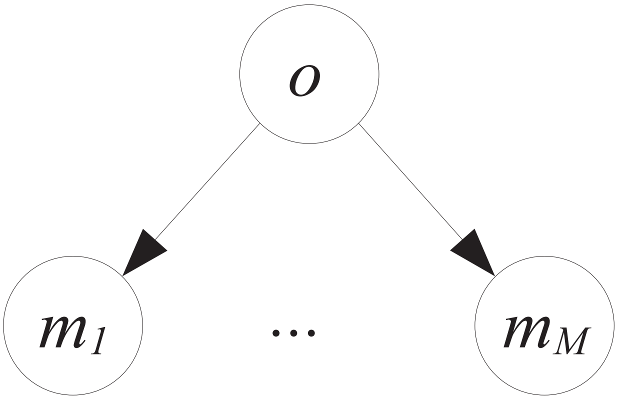


Fig. 3.2 A Bayesian network illustrating the OCI model after adopting the naive Bayes assumption. Nodes represent random variables of model parts $\mathbf{m}$ and the OCI o, and arcs represent statistical dependency. Here, model parts are conditionally independent given the OCI. The equivalent probabilistic expression is $p(o|\mathbf{m}) = \frac{p(o)\prod_i^M p(m_i|o)}{p(\mathbf{m})}$.

The assumption of conditionally independent local image regions can be contrasted with assumptions inherent to global image models, e.g. the multi-variate Gaussian, where statistical dependence is assumed between all image regions. As the number of parameters required to capture all regional interdependencies in global models is prohibitively large, the assumption of linear dependence between image regions is typically made, and techniques such as principal component analysis (PCA) are used to determine a reduced linear basis which captures the majority of the image covariance [27, 29]. The major drawback of the linear dependence assumption, however, is the inability of the global model to describe image variation on a local scale when the appearance of one image region violates the learned linear model with respect to other image regions.

Assumption 2 - Conditional Independence of Part Occurrence/Appearance and Part Geometry: After making the assumption of conditional model part independence, modeling focuses principally on the factors $p(m_i|o)$ describing the probability of individual parts m_i conditional on the OCI. By separating a model part into its components $m_i = \{m_i^a, m_i^b, m_i^g\}$, this factor can be expressed as:

$$\begin{aligned} p(m_i|o) &= p(m_i^a, m_i^b, m_i^g|o), \\ &= p(m_i^a, m_i^b|o)p(m_i^g|o), \end{aligned} \tag{3.3}$$

under the assumption that part appearance m_i^a and occurrence m_i^b are conditionally independent of part geometry m_i^g given the OCI o. This assumption states that knowing the geometry of the OCI in an image, the appearance and occurrence of a part offer no further information regarding part geometry.

Assumption 3 - Conditional Independence of Part Appearance and OCI: The expression in equation (3.3) can be further refined by separating the OCI into its components $o = \{o^b, o^g\}$, resulting in the expression:

$$\begin{aligned} p(m_i|o) &= p(m_i^a, m_i^b|o)p(m_i^g|o), \\ &= p(m_i^a, m_i^b|o^b, o^g)p(m_i^g|o^b, o^g), \\ &= p(m_i^a|m_i^b, o^b, o^g)p(m_i^b|o^b, o^g)p(m_i^g|o^b, o^g), \\ &= p(m_i^a|m_i^b)p(m_i^b|o^b, o^g)p(m_i^g|o^b, o^g), \end{aligned} \tag{3.4}$$

under the assumption that part appearance m_i^a and OCI o are conditionally independent given part occurrence m_i^b, i.e. $p(m_i^a|m_i^b, o^b, o^g) = p(m_i^a|m_i^b)$. This assumption generally states that knowing that a model part occurs in an image, knowledge of the OCI offers no additional information regarding the appearance of the part.

Assumption 4 - Conditional Independence of Part Occurrence and OCI Geometry: A final assumption that further simplifies the expression in equation (3.4) is as follows:

$$\begin{aligned} p(m_i|o) &= p(m_i^a|m_i^b)p(m_i^b|o^b, o^g)p(m_i^g|o^b, o^g), \\ &= p(m_i^a|m_i^b)p(m_i^b|o^b)p(m_i^g|o^b, o^g), \end{aligned} \tag{3.5}$$

under the assumption that part occurrence m_i^b and OCI geometry o^g are conditionally independent given OCI occurrence o^b, i.e. $p(m_i^b|o^b, o^g) = p(m_i^b|o^b)$. This assumption generally states that knowing that the OCI occurs in an image, knowledge of the OCI geometry offers no additional information regarding whether or not a part will occur. A Bayesian network diagram equivalent to the probabilistic expression in equation (3.5) is illustrated in Figure 3.3. Incorporating Assumptions 1-4, the final expression for the posterior in equation (3.2) becomes:

$$p(o|\mathbf{m}) = \frac{p(o)\prod_i^M p(m_i|o)}{p(\mathbf{m})} = \frac{p(o)\prod_i^M p(m_i^a|m_i^b)p(m_i^b|o^b)p(m_i^g|o^b,o^g)}{p(\mathbf{m})}. \tag{3.6}$$

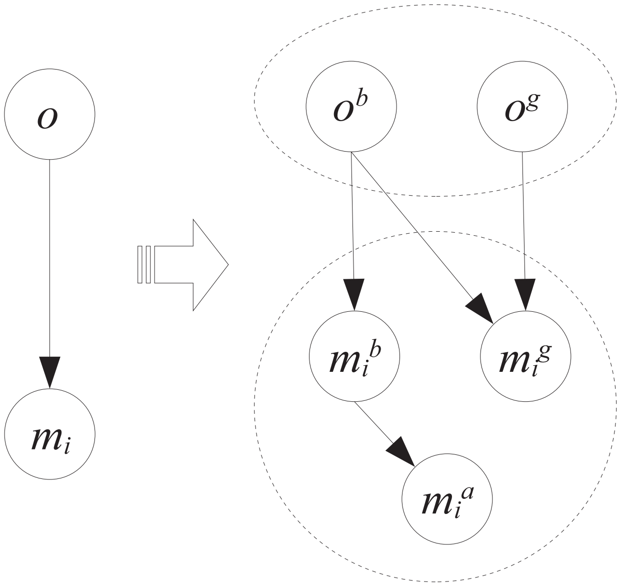


Fig. 3.3 A Bayesian network illustrating the probabilistic relationship between an individual part m_i with respect to the OCI o. The arcs represent statistical dependence assumptions made regarding OCI random variables of presence o^b and geometry o^g, and random variables of feature presence m_i^b, geometry m_i^g and appearance m_i^a. The equivalent probabilistic equation is $p(m_i|o) = p(m_i^a|m_i^b)p(m_i^b|o^b)p(m_i^g|o^b,o^g)$.

The three factors of equation (3.6) serve to describe the appearance, the occurrence and the geometry of image features in the OCI model. These are described below.

Appearance density $p(m_i^a|m_i^b)$**:** Factor $p(m_i^a|m_i^b)$ is a probability density function expressing the appearance of a part when it occurs in an image. Two density functions must be considered, corresponding to the cases $m_i^{b=1}$ and $m_i^{b=0}$. In the case of the former, $p(m_i^a|m_i^{b=1})$ is a probability density over part appearance given a valid part occurrence, which can generally be modeled in a number of ways. For example in the case of a unimodal part appearance that varies in terms of additive white noise about an average, a logical choice is a multivariate Gaussian distribution in an appearance space with parameters of mean μ_i^a and covariance matrix Σ_i^a. In general, a variety of different appearance spaces can be used, for example principal components of image intensities [27] or histograms of local image gradient orientations [22]. In the latter case, $p(m_i^a|m_i^{b=0})$ represents a probability density over part appearance in the case of part absence. This distribution can be thought of as representing all non-part appearances, and can be generally taken to be constant or uniform.

Occurrence probability $p(m_i^b|o^b)$: Term $p(m_i^b|o^b)$ is a discrete probability mass function describing the probability that a model part will occur given knowledge of the occurrence of reference frame. This function can be modeled as a discrete multinomial distribution with event count parameters $\pi_i = \{\pi_i^0, \ldots, \pi_i^3\}$ for the 4 possible combinations of binary occurrence events. Table 3.1 lists the events and their interpretations. Count π_i^0 is not of interest, as neither the part nor the reference frame are present in the image. Count π_i^1 represents the event where part m_i is observed in the absence of reference frame o. This event can be viewed as a false part occurrence or false match. Count π_i^2 represents the event where reference frame o is present but part m_i is absent. This event is the case where the part is occluded or unobservable, due to occlusion or appearance variability. Finally, count π_i^3 represents the event that both part m_i and reference frame o are present.

Table 3.1 Occurrence probability $p(m_i^b|o^b)$.

	o^b	m_i^b	**Interpretation**
π_i^0	$b = 0$	$b = 0$	No occurrence of part or object
π_i^1	$b = 0$	$b = 1$	False part occurrence
π_i^2	$b = 1$	$b = 0$	Occluded part
π_i^3	$b = 1$	$b = 1$	True part occurrence

An aspect of the occurrence probability bearing mention at this point is the likelihood ratio of true vs. false part occurrences:

$$\frac{p(m_i^{b=1}|o^{b=1})}{p(m_i^{b=1}|o^{b=0})} = \frac{\pi_i^3}{\pi_i^1}. \tag{3.7}$$

This ratio will be referred to as the *distinctiveness* of a part within this thesis, as it provides a measure of the reliability with which a part can be identified in the presence of noise and false matches [154]. The distinctiveness plays an important role both in automatic model learning and in model fitting, as described later.

Geometrical density $p(m_i^g|o^b, o^g)$**:** Factor $p(m_i^g|o^b, o^g)$ is a probability density function expressing the stochastic relationship between part geometry m_i^g and reference frame occurrence o^b and geometry o^g. It can be viewed as modeling the residual error of a geometrical transform t_i from part geometry m_i^g to reference frame geometry o^g, i.e. $t_i : m_i^g \rightarrow o^g$. This geometrical transform can be defined in a number of ways, such as the linear similarity transform described in Appendix B.

Note that in order to characterize geometrical error in terms of additive white noise, variables of scale σ are transformed logarithmically. In this way, a multiplicative error in scale magnification can be represented by addition or subtraction of the logarithm of the error. Additionally, in order to characterize location error in a manner independent of image scale, deviations in feature location x must be normalized (divided) by the scale of the reference frame. As with the appearance density, there are two distributions corresponding to cases $o^{b=1}$ and $o^{b=0}$. In the former case, $p(m_i^g|o^{b=1}, o^g)$ represents the distribution of error given a valid reference frame. Under the assumption error is distributed according to additive white noise about the actual reference frame, this distribution can be modeled as a multivariate Gaussian parameterized by a mean μ_i^g and covariance matrix Σ_i^g. In the case of $p(m_i^g|o^{b=0}, o^g)$ a valid reference frame is not present, neither o^g nor $m_i^g \rightarrow o^g$ are defined, and the error distribution can be treated as uniform or constant. Modeling valid/invalid geometrical densities can be accomplished based on a threshold defining the maximum acceptable geometrical variation for a given object class, as described in the following section.

3.1.2 OCI Model Learning

The goal of model learning is to generate a probabilistic OCI model of an appearance pattern from a set of training images containing examples of the pattern. Learning involves automatically identifying a set of M model parts $\mathbf{m}$ and estimating the parameters of their appearance, occurrence and geometrical distributions, based on the training images. Prior to learning, each training image is processed by labeling the OCI reference frame o^g and automatically extracting features $\{m_i^a, m_i^g\}$. Depending on the images, the OCI reference frame o^g can be labeled manually by defining in each training image or estimated automatically as described later. For the purpose of this thesis, an interactive software application was developed which enables a user to open images and label OCIs quickly and efficiently in 2D images, as illustrated in Figure 3.4. For the purpose of illustration, an OCI in the form of a 2D line segment in the image plane is adopted here. The particular OCI used is application specific, however, and the issue of determining an optimal OCI is addressed later.

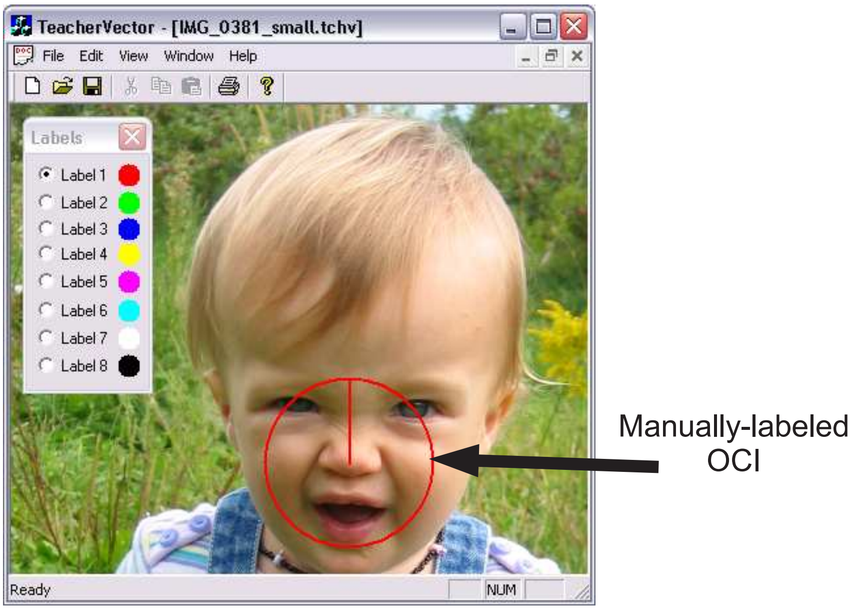


Fig. 3.4 A screen capture of an interactive application developed to allow a user to efficiently label OCIs in training images. Here, an OCI is defined for the class of face images as an oriented line segment from the tip of the nose to the forehead. The user can open an image file and specify the OCI by drawing a line segment on the image.

Parameter estimation is based on a set of data vectors of the form $\{m_i^a, m_i^g, o_i^g\}$. Here, o_i signifies the reference frame instance associated with feature m_i, i.e. the OCI in the training image from which m_i was extracted. A model part corresponds to a cluster of data vectors that are similar in terms of their appearance and their geometry relative to the reference frame, as illustrated in Figure 3.5. Parameter estimation requires identifying these clusters. In general, a variety of clustering techniques could be used for this purpose. As the data vector space contains a large, unknown number of clusters or modes, approaches such as EM (expectation-maximization) or K-means [89] are ineffective, as they require prior knowledge regarding the number of parts present, and tend to either group vectors arising from different clusters or separate vectors arising from the same cluster. This thesis adopts a different approach, instead by identifying all isolated clusters in a robust manner similar to the mean shift technique [190]. This is done by treating each feature m_i as a potential cluster center or model part, grouping features into a reduced set of parts, and finally estimating part statistics. These steps are explained here.

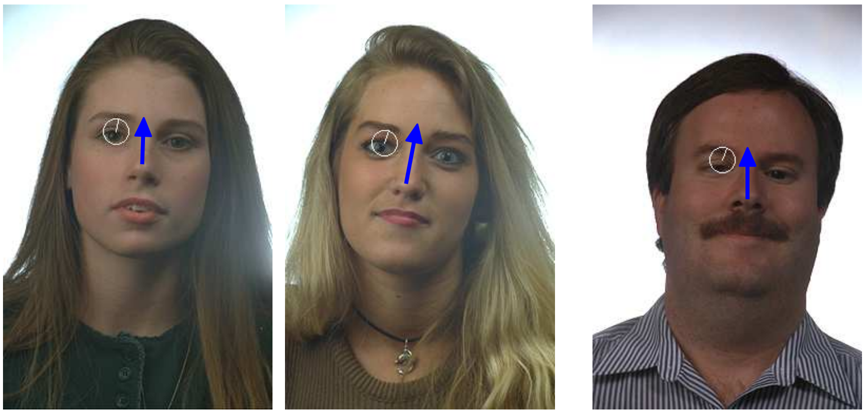


Fig. 3.5 An example of three features (white circles) which are similar in terms of appearance and geometry relative to the OCI (arrows). These features can be though of as arising from the same underlying model part. The goal of learning is to identifying clusters of such features.

Treating each extracted feature m_i as a potential model part, feature grouping proceeds by identifying two sets of features: a set of features G_i that are similar to m_i in terms of geometry, and a set of features A_i that are similar to m_i in terms of appearance, as illustrated in Figure 3.6. The goal is to identify features in the intersection of these sets $G_i \cap A_i$, which are similar both in terms of their geometry and appearance and can therefore be considered as arising from the same underlying model part m_i. The grouping process is described here.

The first step is to generate a set G_i of data vectors similar in geometry to m_i, such that the error in predicting the reference frame geometry is less than an empirically determined threshold $Thres^g$:

$$G_i = \{m_j : |t_i(m_j^g) - o_j^g| < Thres^g\}. \tag{3.8}$$

Recall that $t_i : m_i^g \rightarrow o_i^g$ is a geometrical transform between feature geometry m_i^g and reference frame geometry o_i^g. $t_i(m_j^g)$ therefore represents the geometry of reference frame o_j^g as predicted by m_i^g, o_i^g and m_j^g. $Thres^g = \{T_x, T_\theta, T_\sigma\}$ represents a scale-invariant

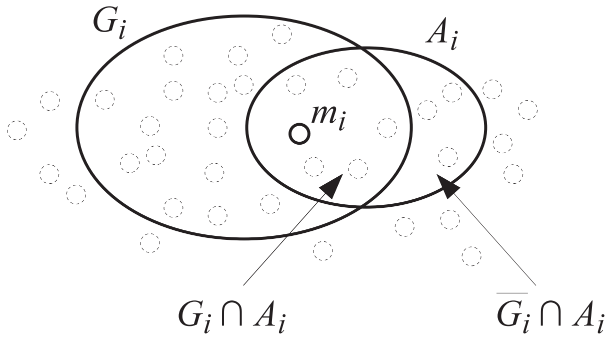


Fig. 3.6 Illustrating feature grouping. For each extracted feature m_i, sets of features G_i and A_i are identified, where features in G_i are similar to m_i in terms of geometry and features in A_i are similar to m_i in terms of appearance. The intersection of these sets $G_i \cap A_i$ defines features arising from the same model part m_i, as they are similar both in terms of geometry and appearance.

threshold on the maximum acceptable error permitted in the location, orientation and scale of the predicted reference frame geometry. These thresholds are applied independently on parameters of reference frame location, orientation and scale, and all parameters must be within their associated threshold for m_i and m_j to be considered geometrically similar. As $Thres^g$ is not related to individual features themselves but rather to their ability to predict the reference frame geometry, a single threshold is applicable to all features [1]. Figure 3.7 provides a graphical illustration of geometrical similarity of feature geometries.

The next step is to generate a set A_i of data vectors similar in appearance to m_i, such that the dissimilarity between feature appearances m_i^a and m_j^a is less than a threshold $Thres_i^a$:

$$A_i = \{m_j : dist(m_i^a, m_j^a) < Thres_i^a\}, \tag{3.9}$$

where $dist(m_i^a, m_j^a)$ is a measure of dissimilarity between m_i^a and m_j^a and $Thres_i^a$ therefore represents a threshold on the acceptable dissimilarity. A variety of dissimilarity measures could be adopted depending on the particular appearance space and modeling assumptions adopted. Examples include the Euclidean distance in a space of independent and identically distributed features, more generally the Mahalanobis distance in a space of features

[1] All experimentation in this thesis adopts a threshold of $T_x = \frac{\sigma}{2}$ pixels (where σ is the scale of o_j^g), $T_\theta = 20$ degrees and $T_\sigma = log(1.5)$ (a scaling factor range of 0.66 to 1.5 times).

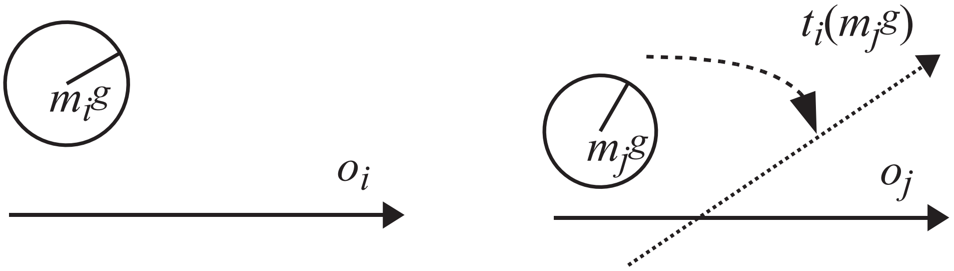


Fig. 3.7 Geometrical similarity between two features m_i and m_j. Here, feature geometries m_i^g and m_j^g differ sightly relative to their respective reference frames o_i^g and o_j^g. Features m_i and m_j are said to be geometrically similar if the geometry of reference frame o_j^g and its geometry $t_i(m_j^g)$ (straight dashed arrow) as predicted by m_i^g, o_i^g and m_j^g differ by less than a threshold $Thres^g = \{T_x, T_\theta, T_\sigma\}$ in location, orientation and scale.

exhibiting covariances [89], the mutual information [91] for measuring general statistical dependence, etc. Here, $Thres_i^a$ is automatically set to maximize the ratio of features that are similar in appearance and geometry vs. features that are similar in appearance but not in geometry:

$$Thres_i^a = \underset{Thres_i^a}{\operatorname{argmax}} \left\{ \frac{|G_i \cap A_i|}{|\bar{G}_i \cap A_i|} \right\}. \tag{3.10}$$

Note that the ratio in equation (3.10) is equivalent to the distinctiveness in equation (3.7), and $Thres_i^a$ is thus determined to maximize the likelihood of a correct match vs. an incorrect match.

Once G_i and A_i have been determined for each feature m_i, the set of features can be reduced to a small number of representative model parts. There are several potential mechanisms for doing this. Features with arbitrarily low distinctiveness can be removed [31], as they are uninformative. Features with high mutual information with other features can also be removed [31, 120], as they are redundant. In this work, a set of features R to be removed is generated, where features $m_j \in R$ are similar to, but occur less frequently than, some other feature m_i:

$$R = \{m_j : m_j \in G_i \cap A_i, |G_j \cap A_j| < |G_i \cap A_i|\}. \tag{3.11}$$

This approach has the effect of discarding redundant features, while maintaining those that are most representative of the object class as determined by their occurrence frequency. Feature removal generally reduces the entire feature set by an order of magnitude into a set of model parts. Estimation of part parameters then proceeds as follows. Variables o^b and m_i^b are determined by membership in sets G_i and A_i, respectively, allowing the estimation of event count parameters π_i. Geometry and appearance parameters $\{\mu_i^g, \Sigma_i^g\}$ and $\{\mu_i^a, \Sigma_i^a\}$ are determined from the geometries and appearances of features in set $G_i \cap A_i$.

This learning process is effective at determining a set of distinct model parts useful for the task of model-to-subject fitting, as the ratio of correct vs. incorrect matches is maximized. Model parts do not necessarily correspond to features or anatomical structures that a human might identify, such as eyes or a nose in face images, but rather natural regions identified by the particular feature detector used. Features resulting from a single anatomical structure or region may exhibit several stable modes of geometry and appearance due the particular feature detector used and to anatomical variability. In such cases, model learning generally represents each mode as a different model part. For example one model part could represent open eyes and another closed eyes. Relaxing the threshold $Thres^g$ on the permitted geometrical error results in fewer parts, each capable of describing a wider range of variability of the underlying image content, at the cost of decreased part distinctiveness. For the purpose of clarity, a high-level summary of the model learning process is provided in Algorithm 1.

Algorithm 1: Model Learning

```
1) Label reference frame geometries $o^g$ in training images.
2) Extract invariant features $\mathbf{m}$ in training images.
foreach feature $m_i$ do
    3) Identify a set of features $G_i$ similar to $m_i$ in geometry.
    4) Identify a set of features $A_i$ similar to $m_i$ in appearance.

5) Discard redundant features.
foreach feature $m_i$ do
    6) Estimate distribution parameters from $G_i \cap A_i$.
```

3.1.3 OCI Model Fitting

Once a model of an object class has been learned from a set of training images, it can be fit to new images. This is done by inferring the geometry of the reference frame o and therefore instances of the object class based on features extracted in the new image. Unlike other registration or fitting techniques based on iterative algorithms which tend to go awry when started outside a 'capture range' of the optimal solution, the OCI model can be fit globally in a robust manner. Additionally, due to the nature of scale-invariant features, the model can be automatically fit in the presence of global image translation, orientation and scale changes, in addition to inter-subject variability.

Fitting begins by first automatically matching features extracted in the new image to the learned model parts. An arbitrary feature m extracted in the new image is considered to match a model part m_i if $dist(m_i^a, m^a) < Thres_i^a$, as in equation (3.9). While the reference frame geometry in the new image is initially unknown, each image feature/model part match fixes m_i^a and m_i^g as evidence in likelihood (3.5), and infers a hypothesis as to the reference frame geometry o^g in the new image via the learned linear relationship. Dense clusters of similar hypotheses suggest the presence of a reference frame, and therefore an object class instance. Image features associated with these hypotheses indicate valid model-to-subject correspondences. A robust geometric clustering approach identical to that used in model learning can be used, where two hypotheses o_i^g and o_j^g are considered to agree if $|o_i^g - o_j^g| < Thres^g$, as in equation (3.8). Figure 3.8 illustrates the result of fitting a model learned from MR brain image slices to a new slice.

Once hypothesis clusters o^g have been identified in the new image, each cluster is evaluated to determine whether it is the result of a true OCI instance or a random noisy match. A true OCI instance is denoted as $o = \{o^g, o^{b=1}\}$, and a random noisy OCI is denoted as $\overline{o} = \{o^g, o^{b=0}\}$. These two hypotheses can be compared via a Bayes decision ratio:

$$\gamma(o, \mathbf{m}) = \frac{p(o|\mathbf{m})}{p(\overline{o}|\mathbf{m})} = \frac{p(o)}{p(\overline{o})} \prod_{i=1}^{M} \frac{p(m_i|o)}{p(m_i|\overline{o})}, \tag{3.12}$$

where large $\gamma(o, \mathbf{m})$ indicates the presence of a model. Here, $\frac{p(o)}{p(\overline{o})}$ is a constant representing the expected ratio of true to false model instances, and can be used as a threshold to control the false positive rate for in the context of object detection. The value of $\gamma(o, \mathbf{m})$ is, in large part, determined by the part distinctiveness defined in equation (3.7), as highly distinctive

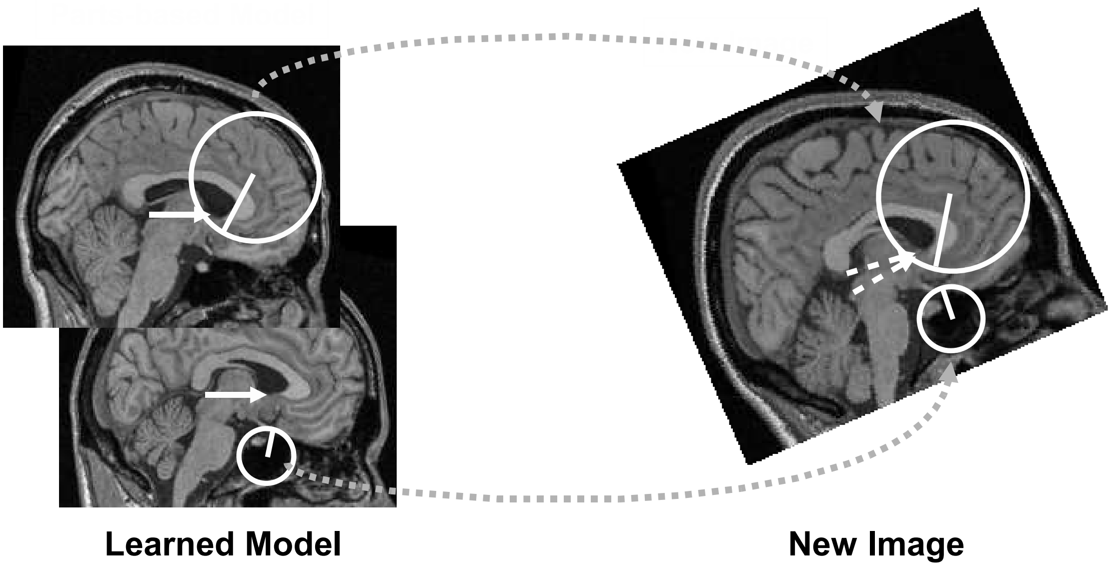


Fig. 3.8 Fitting the parts-based model to a new image, based on sagittal slices of MR brain imagery. Fitting begins by matching features in a model learned from a training dataset (left) to those extracted in a new image (right). Each model-to-image correspondence produces a hypothesis as to the reference frame geometry o^g in the new image (dashed white arrows). Clusters of geometrically similar hypotheses indicate the presence of a valid reference frame instance. Note that the new image has been rotated and magnified to illustrate the invariance of model fitting to location, orientation and scale changes.

parts will carry greater weight in determining the model fit. Note that the term $p(\mathbf{m})$ is present in both the numerator and denominator in equation (3.12) and is thus canceled. For the purpose of clarity, a high-level summary of the model fitting process is provided in Algorithm 2.

```
Algorithm 2: Model Fitting
  1) Extract invariant features $m$ in new image.
  2) Match image features to learned model parts $m_i$.
  3) Estimate reference frame hypotheses $o^g$.
  4) Cluster reference frame hypotheses.
  5) Evaluate the Bayes decision ratio of hypothesis clusters.
```

3.1.4 Learning and Classifying Visual Traits

Image patterns arising from a class of similar objects can often be described in terms of abstract visual traits. For example, face images can be described by traits such as age (young or old) or sex (male or female). Brain images can be described in terms of health (normal, diseased, or distinct disease stage). A practical model of pattern appearance should lend itself to describing and classifying patterns in terms of visual traits, and to explaining which image characteristics are indicative of traits. Doing so requires effectively detecting and localizing the image features on which trait classification is based, which is generally non-trivial in natural, arbitrary imagery. The tasks of detection, localization and trait classification are thus inextricably linked in a realistic scenario. This section proposes the OCI model as a basis for learning and representing visual traits of appearance patterns, by linking individual model parts or features directly to visual traits via their co-occurrence statistics. In this way, a single model can be used as a basis for the tasks of detection, localization and trait classification in arbitrary imagery.

Recall that the model fitting process described in Section 3.1.3 identifies an OCI instance along with a set of model feature occurrences $\mathbf{m}$ in a new image. The methodology here is to use this set of model features to classify the object class instance in terms of visual traits. This is done as follows. Let $f_i = m_i^{b=1}$ be the event of positive occurrence of model part m_i, and let $\mathbf{f} = \{f_i\}$ denote the set of all model parts positively identified with the

OCI instance to be classified. Let $c = \{c_1, \ldots, c_K\}$ denote the visual trait variable defined over K trait values of interest, e.g. $sex : \{c_1 = female, c_2 = male\}$. The notation $\bar{c_j}$ is used to indicate any trait value except c_j. A Bayesian classifier $\psi(c)$ can then be used to express the most probable trait classification given set $\mathbf{f}$:

$$\psi(c, \mathbf{f}) = \frac{p(c|\mathbf{f})}{p(\bar{c}|\mathbf{f})} = \frac{p(c)}{p(\bar{c})} \prod_{f_i \in \mathbf{f}} \frac{p(f_i|c)}{p(f_i|\bar{c})},$$

or equivalently:

$$\log \psi(c, \mathbf{f}) = \log \frac{p(c)}{p(\bar{c})} + \sum_{f_i \in \mathbf{f}} \log \frac{p(f_i|c)}{p(f_i|\bar{c})}, \tag{3.13}$$

under the assumption that model features f are conditionally independent given trait c. Note that the classifier in equation (3.13) has the computational form of a standard linear classifier [89]. The optimal Bayes classification is to choose trait value c^* maximizing $log\ \psi(c, \mathbf{f})$:

$$c^* = \underset{c}{\operatorname{argmax}} \{\log \psi(c, \mathbf{f})\}. \tag{3.14}$$

Classifier training requires estimating factors $\frac{p(c)}{p(\bar{c})}$ and $\frac{p(f_i|c)}{p(f_i|\bar{c})}$ in equation (3.13) from a set of training images. This can be done in a supervised manner, based on an OCI model learned from a set of training images labeled according to their trait values. Factor $\frac{p(f_i|c)}{p(f_i|\bar{c})}$ is the likelihood ratio of trait value presence vs. absence coinciding with feature observation f_i. Features that are important to classification, or highly informative with regard to a particular trait value c_j, have high likelihood ratios. The focus of the approach here is to use these likelihood ratios to quantify the association of model features with visual traits. Factor $\frac{p(c)}{p(\bar{c})}$ is the prior ratio of trait value presence vs. absence (e.g. male vs. not male), and is used to control classifier bias toward different trait values. The manner in which each of these factors is estimated is now described.

Estimating $\log \frac{p(f_i|c)}{p(f_i|\bar{c})}$**:** The likelihood ratios are estimated via a supervised learning process, based on observed model feature occurrences f_i and trait labels c_j for each training image. Discrete class-conditional likelihoods $p(f_i|c_j)$ are represented as binomial distributions, parameterized by event counts [191]. $p(f_i|c_j)$ is estimated from $p(c_j)$ and $p(f_i, c_j)$, the probability of observed joint events (f_i, c_j), using the definition of conditional probability:

$$p(f_i|c_j) = \frac{p(f_i, c_j)}{p(c_j)}. \tag{3.15}$$

Factor $p(c_j)$ is important in correcting bias in the training set, i.e. when the numbers of different trait value labels in the training data are unequal. In this situation, joint events are normalized by the prior probabilities of trait values to avoid biasing classification in favor of more common trait values.

The most straightforward manner of estimating $p(f_i, c_j)$ is via maximum likelihood (ML) estimation, by counting the joint events (f_i, c_j) and normalizing with respect to their sum. ML estimation is unstable in the case of sparse data, leading to noisy or undefined parameter estimates. This is particularly true in models consisting of many local features, where feature occurrences are typically rare events. Bayesian maximum a posteriori (MAP) estimation can be used to cope with data sparsity, and involves regularizing estimates using a Dirichlet hyperparameter distribution [191]. In practice, Dirichlet regularization involves pre-populating event count parameters with samples following a prior distribution embodying assumptions regarding the expected sample distribution. Where no relevant prior knowledge exists, a uniform or maximum entropy prior can be used [192]. Although both ML and MAP estimates converge as the number of data samples increases, MAP estimation using a uniform prior tends towards stable, conservative parameter estimates when the number of data samples is low. Work done in conjunction with this thesis showed how MAP estimation with a uniform Dirichlet prior resulted in stable parameter estimates in the presence of sparse data [95]. The final estimator used is:

$$p(f_i|c_j) \propto \frac{k_{i,j}}{p(c_j)} + d_{i,j}, \tag{3.16}$$

where $k_{i,j}$ is the frequency of the joint occurrence event (f_i, c_j), $p(c_j)$ is the frequency of trait value c_j in the training data and $d_{i,j}$ is the Dirichlet regularization parameter used to pre-populate event counts. In the case of a uniform prior, $d_{i,j}$ is constant for all i, j. The proportionality constant for the likelihood in equation (3.16) can be obtained by normalizing over values of f_i, but is not required for the likelihood ratios used for classification.

Estimating $\log \frac{p(c)}{p(\bar{c})}$**:** Although individual likelihood ratios have been corrected for training set bias by the estimator in (3.16), the Bayesian classifier in equation (3.13) will still exhibit bias due to the fact that the number of features and their corresponding likelihood ratios associated with different traits are generally unequal. Given the set $\mathbf{f}$ of features to classify, log $\psi(c, \mathbf{f})$ will be higher *a priori* for traits associated with a larger number of features or with features bearing higher likelihood ratios. This bias can be controlled by setting $\log \frac{p(c_j)}{p(\bar{c_j})}$ for each trait value c_j such that the expected value of log $\psi(c_j, \mathbf{f})$ based on set $\mathbf{f}$ is zero:

$$E[\ \log\ \psi(c_j, \mathbf{f})\ ] = E\left[\ \log \frac{p(c_j)}{p(\bar{c_j})} + \sum_{f \in \mathbf{f}} \log \frac{p(f|c_j)}{p(f|\bar{c_j})}\ \right] = 0, \qquad (3.17)$$

and thus:

$$\log \frac{p(c_j)}{p(\bar{c_j})} = -E\left[\ \sum_{f \in \mathbf{f}} \log \frac{p(f|c_j)}{p(f|\bar{c_j})}\ \right]. \qquad (3.18)$$

The right hand side of equation (3.18) represents the expected sum of log likelihood ratios for an arbitrary set $\mathbf{f}$ of positive model part occurrences. The random variable required to evaluate the expectation is f, which can take on M discrete values $f : \{f_1, \ldots, f_M\}$ with individual conditional probabilities $p(f_i|o^{b=1})$. Thus,

$$\begin{aligned} \log \frac{p(c_j)}{p(\bar{c_j})} &= -E\left[\ \sum_{f \in \mathbf{f}} \log \frac{p(f|c_j)}{p(f|\bar{c_j})}\ \right], \\ &= -|\mathbf{f}|\ E\left[\ \log \frac{p(f|c_j)}{p(f|\bar{c_j})}\ \right], \\ &= -|\mathbf{f}|\ \sum_{i}^{M} p(f_i|o^{b=1}) \log \frac{p(f_i|c_j)}{p(f_i|\bar{c_j})}, \qquad (3.19) \end{aligned}$$

where $|\mathbf{f}|$ is the cardinality of set $\mathbf{f}$, i.e. the number of features on which classification is based. Note that in the 3rd line of equation (3.19), $c_j \cap \bar{c_j} = o^{b=1}$, as c_j and $\bar{c_j}$ partition the event of positive OCI occurrence $o^{b=1}$. The term $\log \frac{p(c_j)}{p(\bar{c_j})}$ is thus the product of: 1) the expected likelihood ratio for trait c_j and 2) the number of features associated with a detected object class instance to be classified. Note that this term is not dependent on the specific features to be classified, and can be calculated off-line prior to classification. The final form of the classifier in equation (3.13) used is thus:

$$\log \psi(c, \mathbf{f}) = \sum_{f \in \mathbf{f}} \log \frac{p(f|c)}{p(f|\bar{c})} - |\mathbf{f}| \sum_{i}^{M} p(f_i|o^{b=1}) \log \frac{p(f_i|c_j)}{p(f_i|\bar{c_j})}. \tag{3.20}$$

3.1.5 Summary of OCI Modeling Theory

In this section, the general theory of the OCI model was presented, including the probabilistic formulation, algorithms for learning the model parameters from training images and fitting the model to new images, and finally the use of the model in classifying object instances in terms of visual traits. The following two sections show how the model can be applied to representing the appearance of 3D object classes from arbitrary viewpoints, and to modeling anatomy in medical imagery.

3.2 Computer Vision: OCI Modeling of 3D Object Classes

The first major application of the OCI theory presented in this thesis is to model appearance patterns in 2D projective imagery arising from underlying 3D object classes. The current state-of-the-art in computer vision is still far from matching the ability of the human being to learn, identify and describe patterns in images. For example, consider an intelligent vision system that must identify all males in a crowded scene such as illustrated in Figure 3.9 A). Face instances must first be identified from thousands of image features, despite nuisance parameters such as illumination and in-plane geometrical variations, partial occlusions, unrelated image clutter, and in-depth appearance variation due to viewpoint changes, as illustrated in Figure 3.9 B) and C). These face instances are then classified according to their associated features as illustrated in Figure 3.9 D). The primary challenge faced by such a system is to reliably detect and localize face instances, along with their associated image features required for trait classification, in images acquired from arbitrary viewpoints.

Explicitly modeling all sources of nuisance variability in pattern appearance is generally intractable. Variability of object class appearance due to viewpoint changes can be modeled explicitly via multi-view formulations [4, 3], however explicit knowledge of viewpoint is not required for the tasks of object class detection and localization. Furthermore, viewpoint information may not be available in training information or may be difficult

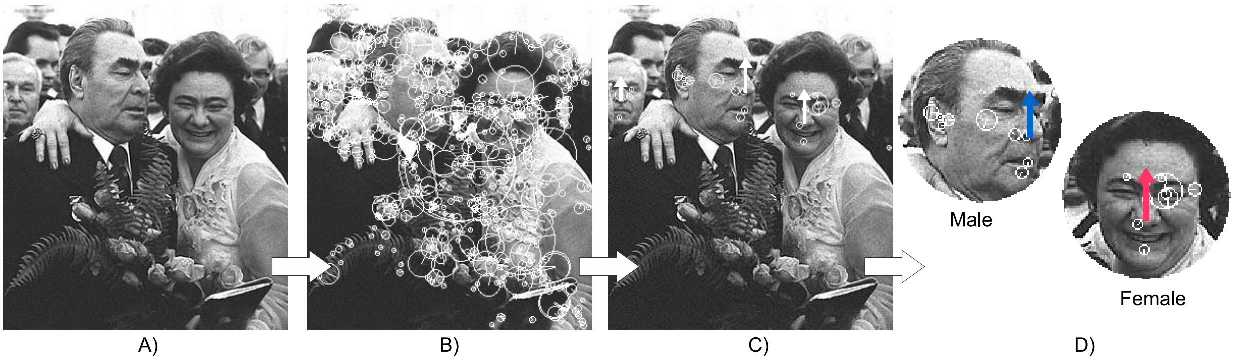


Fig. 3.9 The general OCI framework for combined detection, localization and trait classification from arbitrary viewpoints, applied to the class of 3D faces. All three tasks are embedded in a viewpoint-invariant model derived from scale-invariant image features. Image A) shows an example of a cluttered scene. In B), scale-invariant features (white circles) are extracted from the cluttered scene. Next in C), the viewpoint-invariant model is used to detect and localize face instances (small white arrows) and associated features. Finally in D), a Bayesian classifier is used to determine the sex of face instances from associated features. The image shown is from the CMU face database [37], and the probabilistic framework used is learned from 500 color FERET [38] face images taken at arbitrary viewpoints.

to estimate automatically. The OCI model presented in this section is the first approach to take the alternative approach of modeling 3D object class appearance variability in a manner invariant to viewpoint. This section describes how the OCI model can be used as a viewpoint-invariant model of 3D object classes, and used to learn, detect, localize and classify objects in natural imagery taken from arbitrary viewpoints.

3.2.1 Modeling 3D Object Class Appearance

This thesis proposes using the OCI model to describe the appearance of 3D object classes in 2D projective imagery. In this context, the OCI represents a high-level geometrical feature of the 3D object class that is 1) uniquely defined for each object class instance and 2) invariant to the nuisance parameters arising from the image formation process. As the image formation process involves the projection of light reflecting from objects in the 3D world onto a 2D image plane, the OCI must be a property of the underlying object class which is invariant to perspective transform. In this way, the OCI maintains a consistent geometrical interpretation in the image plane with respect to the underlying object class, and can be used to localize object class instances within the image.

An OCI model of a 3D object class, for example 3D human faces as illustrated in Figure 3.10, requires first defining an appropriate OCI reference frame. An example used throughout this thesis is an OCI in the form of a 3D line segment, such as the line from the tip of the nose to the forehead in Figure 3.10. This OCI projects to a line segment in 2D images of faces acquired from arbitrary viewpoints. It can therefore be used to learn a model of faces from natural, cluttered images taken from arbitrary viewpoints based automatically extracted local invariant features, via the learning process described in Section 3.1.2. Once learned, the OCI model can be used to detect new faces in images taken from arbitrary viewpoints by the model fitting process described in Section 3.1.3.

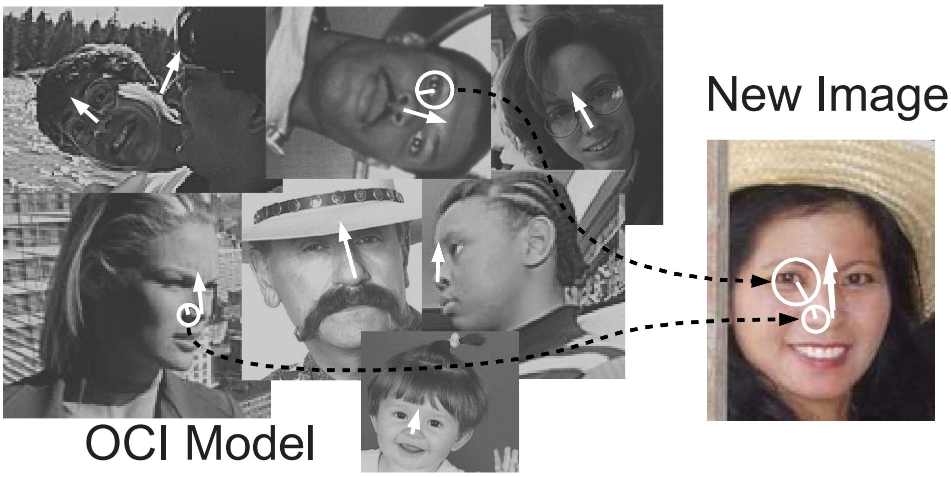


Fig. 3.10 Illustrating a viewpoint-invariant OCI model of faces, based on an OCI in the form of a line segment. Here the OCI (white arrows) has been specified as a 3D line segment from the nose to the forehead. This OCI projects to a line segment in images taken from arbitrary viewpoints, and can be used to learn a model of faces from natural, cluttered images taken from arbitrary viewpoints (left images) based on automatically extracted local invariant features (white circles). Once learned, the OCI model can be used to detect and localize new faces in images taken from arbitrary viewpoints (right image).

It is important to note that while an OCI in the form of a 3D line segment projects to 2D line segment in the image plane in arbitrary views, its geometry (location, magnitude and orientation) only remains consistent with that of the underlying 3D object in the image when viewed from a coronal plane (i.e. a plane perpendicular to the 3D line) as illustrated in Figure 3.11 a). The majority of images arising from object classes such as cars or faces typically observed from views approximately perpendicular to the vertical axis can be modeled effectively using this OCI definition. Overhead or underhead views pose a difficultly in that the line segment magnitude vanishes, and other OCI definitions are necessary to handle entire viewing ranges. An OCI in the form of a 3D sphere projects to a circle in the image plane from an arbitrary viewpoint, as illustrated in Figure 3.11 b), maintaining a location and a scale indicative of the location and size of the underlying 3D object within the image. A spherical OCI bears no information regarding object orientation, however. Another possibility would be to used a collection of perpendicular 3D line segments. Figure 3.12 shows examples of linear and spherical OCI definitions relative to the object classes of 3D faces and chairs.

3.2.2 Multi-view vs. Viewpoint Invariant Modeling

The OCI model describes pattern appearance terms of local image features relatively to a viewpoint-invariant OCI reference frame. As a result, the OCI model could be used as a basis for either a multi-view representation, where a 3D object class is modeled by a set of distinct views, or a viewpoint-invariant representation where all views are combined into a single model. This begs the question as to which representation should be used. The multi-view model is arguably more prevalent in the literature [3, 4, 151] than the viewpoint-invariant model, possibly due to the simplicity of modeling individual views: individual view models do not require viewpoint-invariant geometrical reference frames and can generally make use of a wide variety of single viewpoint techniques for model learning and fitting. The multi-view representation has several drawbacks when compared to the viewpoint-invariant representation. First and foremost, the multi-view representation requires modeling the variable of viewpoint while the viewpoint-invariant modeling representation does not. When learning models from natural images acquired from arbitrary viewpoints with no viewpoint information, e.g. images from the internet, learning a viewpoint-invariant model is facilitated by the fact that there are fewer model parameters to estimate. Learning multi-view models from such data typically involves estimating additional viewpoint variables, which is difficult to do automatically and laborious do manually, i.e. by sorting images according to viewpoint [3].

Assuming that the variable of viewpoint can be learned, an additional difficulty with the multi-view model is as follows: local features arising from images of a 3D object class

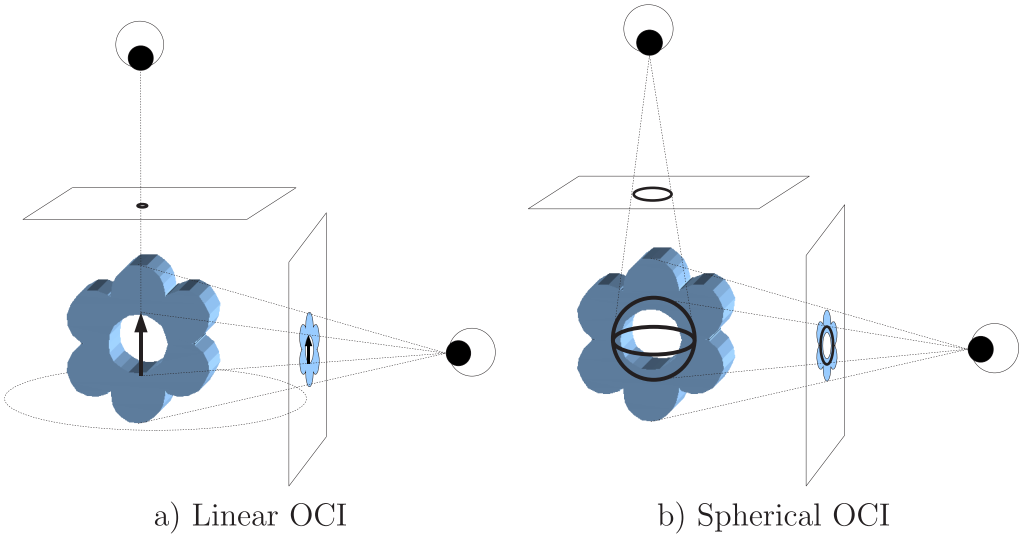
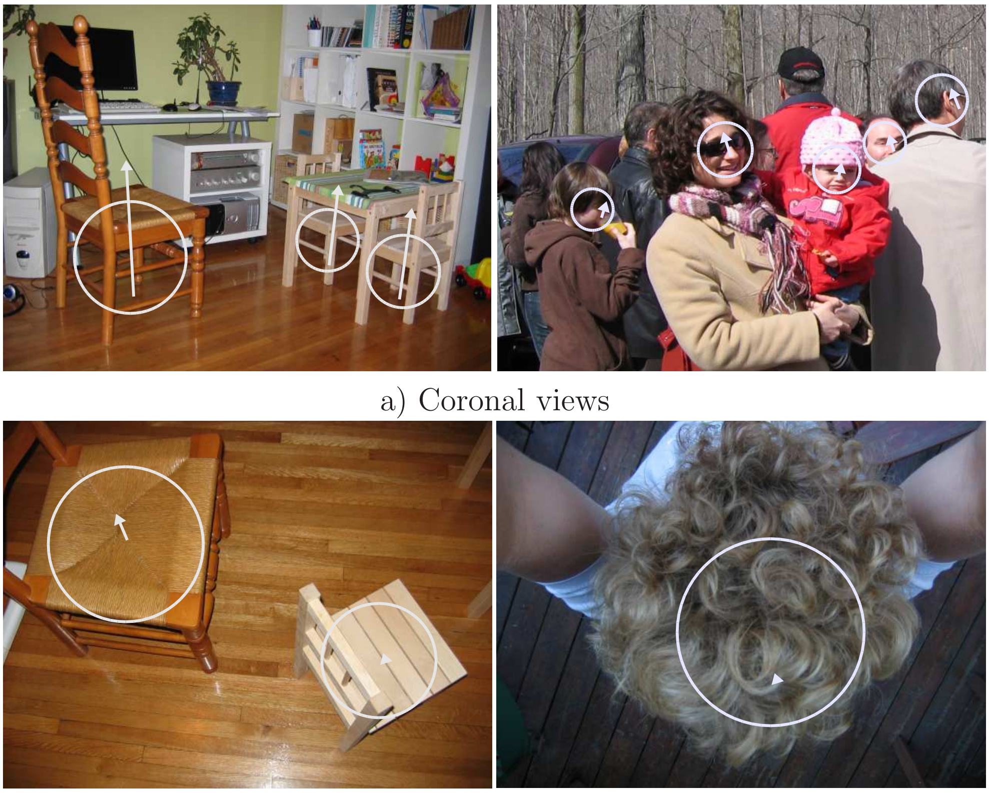


Fig. 3.11 The projective geometry of OCIs in the form of a) a 3D line segment and b) a 3D sphere. Here, an object in the 3D world is represented by the extruded flower shape central to both diagrams a) and b). The bold arrow overlaying the object in diagram a) represents an OCI in the form of a 3D line segment. The bold circles overlaying the object in diagram b) represent an OCI in the form of a 3D sphere. The parallelograms above and to the side of the objects represent image planes onto which the object projects when viewed from above or to the side. Lines from the object through the image planes to the eyes indicate rays of light which reflect off the object to form the images seen by the eyes at each viewpoint. In diagram a), the 3D linear OCI projects to a 2D line segment in the image plane, and maintains a location, orientation and magnitude indicative of the location, orientation and size of the underlying 3D object in the image from all coronal (side) views around the object. The linear OCI projects to a point in overhead views. In diagram b), the 3D spherical OCI projects to a circle in the image plane, and maintains a location and magnitude indicative of the location and size of the underlying 3D object in the image from all viewpoints.

typically persist over a range of viewpoint, as illustrated in Figure 3.13. The particular range depends on the specific image feature. When an image feature is visible in more than one view of the multi-view model, which is often the case, the same image feature is linked to distinctly different reference frames in different view models. When a multi-view model is used for detection, the same image features will thus produce strong, distinctly different hypotheses as to the geometry of the reference frame in the image. Indeed, a major focus of multi-view modeling is developing strategies to cope with false detections or imprecise

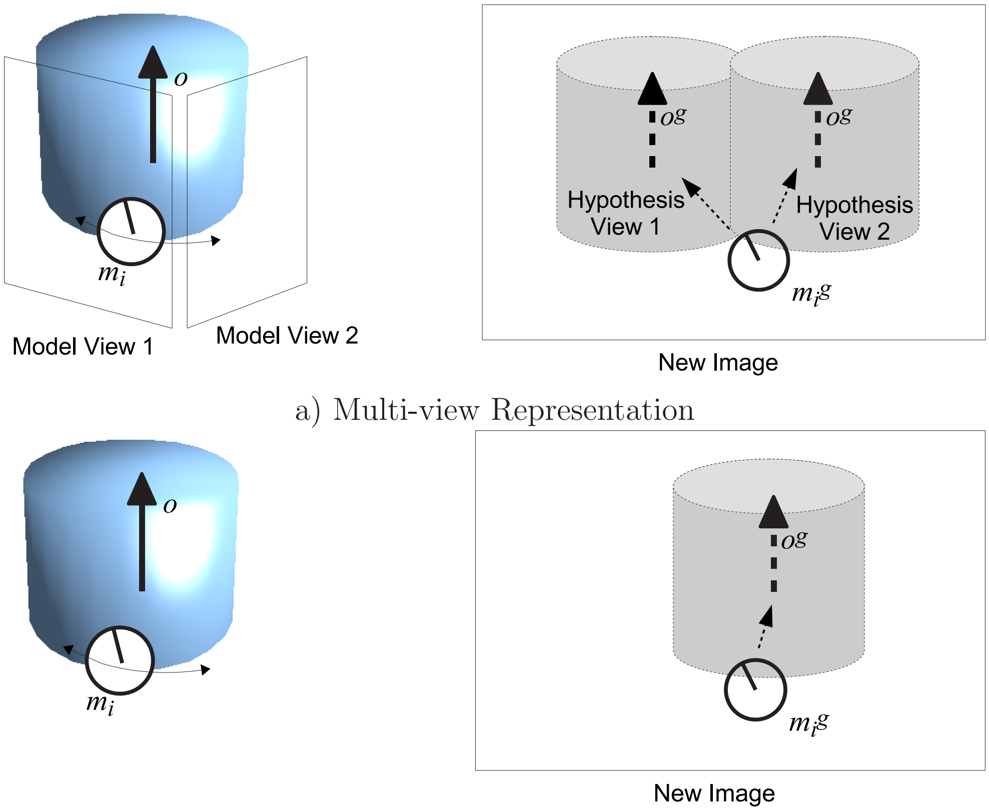


a) Coronal views

b) Overhead views

Fig. 3.12 OCI definitions in coronal and overhead views of chair and face object classes. The OCI can be defined as a 3D line segment or a sphere relative to 3D chairs (left images) and faces (right images), which project to 2D line segments (white arrows) and 2D circles (white circles) respectively in the image plane. In typical coronal views a), both OCI definitions can be used to model 3D object classes, as they maintain a constant magnitude/radius in the image plane with respect size of the underlying object class. In overhead views b), the line segment OCI definition is not effective as its 2D magnitude vanishes.

localization resulting from the same image features supporting different hypotheses as to the object class in the image [4]. Additionally, special learning algorithms have been proposed to effectively share features across views [145]. The viewpoint-invariant OCI model presented in this thesis avoids this difficulty as follows: the appearances of individual image features are learned over the range of viewpoints in which they best predict the reference frame geometry, and not according to a fixed sampling of viewpoints. For this reason, the hypothesis is that a viewpoint-invariant representation will generally result in fewer false positive detections, and thus improved detection performance. This hypothesis is borne out later in experimentation.



a) Multi-view Representation

b) Viewpoint-invariant Representation

Fig. 3.13 Object class detection via multi-view and viewpoint invariant representations. The multi-view representation in a) generally relates a single image feature m_i (circle) to multiple reference frames (large arrows) associated with different model views around the object (upper left). When m_i is observed in a new image (upper right), it will support multiple, distinct hypotheses as to the object geometry o^g in the image, as different views model different geometrical feature-to-OCI relationships (thin dashed arrows). The viewpoint-invariant representation in b) relates the image feature m_i directly to a viewpoint invariant reference frame (lower left). When observed in a new image, m_i supports a single hypothesis as the object geometry o^g in the image (lower right).

3.2.3 An Optimal Viewpoint Invariant Reference Frame

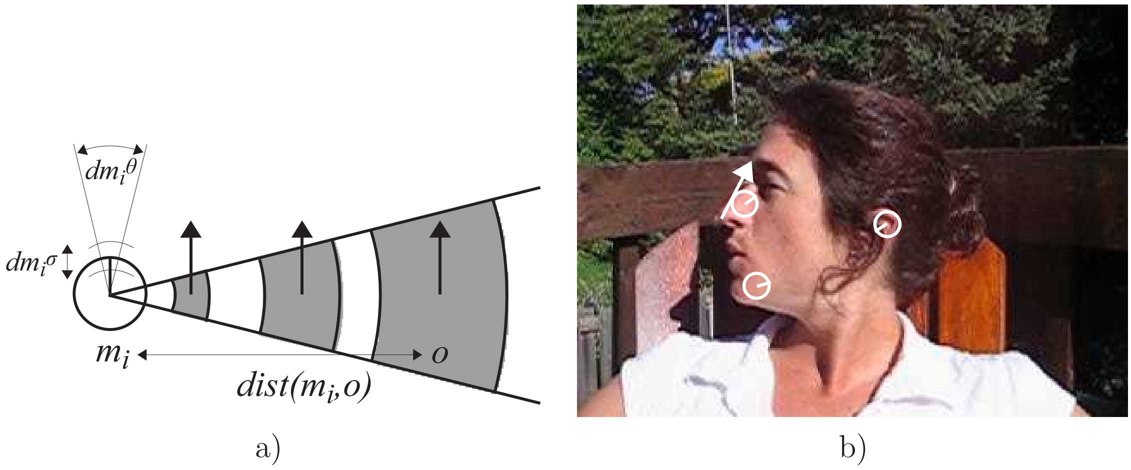
Up to this point, the definition of the OCI has been deliberately left general, as a geometrical structure that is 1) uniquely defined with respect to the underlying appearance pattern and 2) invariant to nuisance parameters arising from the imaging process. A variety of different definitions for the OCI reference frame are possible for a given object class. The definition as a 3D line segment from the nose to the forehead used to model faces has an intuitive interpretation, and facilitates learning from arbitrary, natural imagery. When labeling arbitrary images of faces, for instance, one could even guess the location of the nose from rear views of a head. It is also defined according to a natural plane of facial symmetry, and thus the symmetry of the face can be used to effectively double the amount of training data as mirrored images can be considered as valid training images [73].

Although an OCI can be defined manually according to intuition, it may not result in an OCI model that is optimal for tasks such as detection or classification. Ideally, an optimal OCI would be derived from training images in an iterative, data-driven manner, with minimal manual supervision. To investigate an optimal OCI, consider the error in predicting the OCI geometry from an image feature in model fitting for the task of detection. The error increases with the distance separating the feature and the OCI in the image, as illustrated in Figure 3.14. An optimal reference frame should therefore minimize the expected distance between image features observed from arbitrary viewpoints and the reference frame.

OCI model learning can be extended in order to determine an optimal viewpoint invariant in an iterative, data-driven fashion. Starting from an initial, coarse OCI labeling as before, model parts $\mathbf{m}$ are learned by maximizing their likelihoods based on reference frames labels on a set of training images $\{o_j\}$ via the learning processes described in Section 3.1.2:

$$m_{i_t} = \underset{m_i}{\operatorname{argmax}} \{ \; p(m_i | \{o_{j_t}\}) \; \} \, , \tag{3.21}$$

where m_{i_t} represents model feature i at iteration t, and o_{j_t} represents the $j'th$ OCI training instance at iteration t. OCI labels $\{o_j\}$ can then be re-estimated from $\mathbf{m_t} = \{m_{i_t}\}$, the set of learned model parts at time t, by fitting the model to the training images as described



a)

b)

Fig. 3.14 Figure a) illustrates how the error in predicting the location of the reference frame o (black arrows) from an image feature m_i (black circle) increases with the distance $dist(m_i, o)$ between the two. In a), the three grey arc regions illustrate the localization error for three different values of $dist(m_i, o)$, which is a function of errors in feature scale dm_i^σ and orientation dm_i^θ and $dist(m_i, o)$. To illustrate, consider Figure b) showing an OCI defined along the nose (white arrow) and three features arising from the face (white circles). Features near the OCI in the image plane such as the nose feature are generally able to predict the OCI geometry with less error than those that are far such as the ear feature.

in Section 3.1.3:

$$o_{j_{t+1}} = \underset{o_j}{\operatorname{argmax}} \{ \ \gamma(o_j, \mathbf{m_t}) \ \} \, . \tag{3.22}$$

where $o_{j_{t+1}}$ is the $j'th$ OCI instance at time $t+1$. The hypothesis is that this iterative learning process will converge to a stable OCI definition that: 1) is centrally located with respect to image feature in order to minimize OCI localization error, 2) remains geometrically consistent with the underlying 3D object class, and 3) results in improved detection performance. In this process, initial OCI labeling still constitutes a degree of manual supervision. A partially supervised learning approach could be adopted, where OCI labels are specified for a small portion of image data, and then propagated to the rest of training data via iterative learning. Alternatively, a fully unsupervised learning approach could be constructed by determining OCI labels via initial, coarse feature correspondences between training images. This would open the door to automatically learning pattern appearance models from arbitrary imagery with no manual input whatsoever.

3.2.4 Classification based on Visual Traits

The goal in this thesis is not only to detect and localize instances of 3D object classes from arbitrary viewpoints in natural scenes, but also to describe and classify instances in terms of visual traits. Consider the example of a surveillance system that must automatically identify all male faces in a cluttered scene. Such a system must first detect and localize face instances based on image features, then use the same image features to classify the faces in terms of traits such as sex. Techniques addressing trait classification, sex classification from faces in particular, have focused exclusively on unoccluded frontal faces. Underlying image measurements used are typically global and/or single viewpoint in nature and, as such, cannot be used in realistic scenarios where object class instances are occluded or viewed from arbitrary viewpoints. To date, no work has yet proposed trait classification from arbitrary viewpoints.

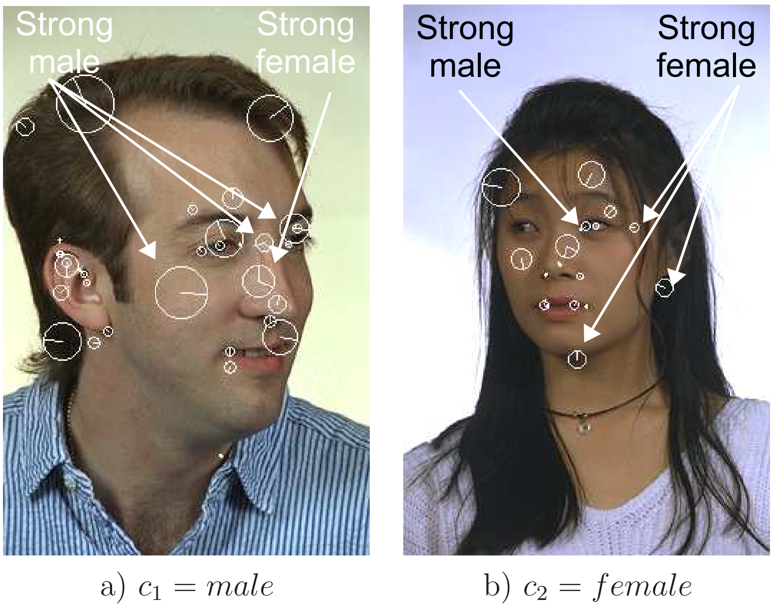
This thesis proposes using local invariant image features of a learned viewpoint-invariant OCI model in conjunction with a Bayesian classifier for learning and inferring visual traits from arbitrary images. To do this, an object class instance o^* along with an associated set of model feature occurrences $\mathbf{m}$ are first identified in an image by maximizing a Bayes decision ratio as described in Section 3.1.3:

$$o^* = \underset{o}{\operatorname{argmax}} \{ \ \gamma(o, \mathbf{m}) \ \} \, . \tag{3.23}$$

The set of positively occurring model features, $\mathbf{f} = \{m_i^{b=1}\}$, is then used as a basis for determining the trait value of OCI instance o^* using the Bayes classifier as described in Section 3.1.4:

$$c^* = \underset{c}{\operatorname{argmax}} \{ \ \log \psi(c, \mathbf{f}) \ \} . \tag{3.24}$$

Using this methodology, the same image features used to detect and localize general 3D object class instances from arbitrary viewpoints can be used as a basis for subsequent trait classification. As individual modeled features are linked directly with traits of interest, they can also be considered as local visual cues indicative of traits, and provide insight into the image features operative in inferring traits, as illustrated in Figure 3.15. The classification framework is generally applicable to modeling a variety of different traits, such as sex, age or ethnicity in the case of faces, or make and model in the case of cars. Modeling of continuous valued traits such as age can be achieved by variable quantization or by using likelihoods based on continuous random variables.



a) $c_1 = male$ b) $c_2 = female$

Fig. 3.15 Local feature occurrences f_i (white circles) indicative of the visual trait of sex $c : \{c_1 = male, c_2 = female\}$ in individual faces. A given face instance consists of a set of local features, a subset of which are reflective of either gender, and it is their ensemble which determines the final decision regarding sex. Here, a feature f_i is described as strongly male or female if its likelihood ratio $\frac{p(f_i|c)}{p(f_i|\bar{c})}$ of co-occurring with the indicated sex in training images is greater than 2:1.

3.3 Medical Imaging: OCI Modeling of Anatomical Variability

The previous section described how the OCI model can be used to represent classes of similar 3D objects in 2D projective imagery, primarily for the tasks of detecting, localizing and classifying object instances in cluttered imagery. This section describes the second major application of the OCI theory presented in this thesis, modeling anatomical appearance in medical imagery. In the analysis of medical imagery, for instance in the case of MR images of the human brain, the task of interest is not to detect brains in cluttered imagery, but rather to quantitatively describe the variability of brain structure across different subjects of a population. This task is difficult because of the fact that no two subjects are identical. Anatomical structure or tissue may exhibit significant appearance variation from one subject to the next, or may simply not exist in all subjects, in the case of pathology for example. This phenomenon is referred to as inter-subject appearance variability, or the manner in which images of different subjects vary within a population. Effectively quantifying inter-subject variability is of great importance to the medical imaging community, as it lies at the heart of understanding how anatomical structure varies within a population. This leads to various open research questions: What image structures are common within a population, what structures are rare, and how do they vary in appearance and geometry? In what ways is an abnormal subject similar to or different from the population?

The OCI model presented in this thesis specifically addresses the challenge of quantifying inter-subject variability in the case where one-to-one correspondence between all subjects in a population does not exist. The model represents MR brain images as a collection of 'parts', which are defined as spatially localized image regions. Each model part consists of an appearance, a geometrical description and an occurrence frequency, all of which are quantified statistically. The strength of the parts-based model is that brain appearance can be modeled locally in terms of multiple, distinct modes of appearance. In this way, the model explicitly accounts for occlusion and inter-subject variation on a local scale, as model parts are not expected to (and typically do not) occur in all images.

Medical imaging devices such as MR scanners are typically designed to produce images based on a tomographic reconstruction process. The resulting images consist of a lattice where inter-intensity distances are proportional to real-world distances. In this scenario, the geometry of the imaging process can be modeled as an orthographic projection of the world onto the image lattice, and the OCI reference frame must be invariant to orthographic projection. MR images of the human brain are often analyzed within precisely such a reference frame: the Talairach stereotactic reference frame [189]. In keeping with this established methodology, this thesis adopts the Talairach reference frame as the OCI for modeling slices of the entire brain. In particular, the OCI takes the form of the AC-PC line which passes from the superior aspect of the anterior commissure to the inferior aspect of the posterior commissure, as illustrated in Figure 3.16. It is important to note that other OCI definitions exist and could potentially prove useful. In the future, an optimal OCI could be determined in a data-driven manner as described in Section 3.2.3. Alternatively, OCIs could be chosen according in order model specific anatomical structures or regions of interest, for instance the cortical surface as discussed later in the following section.

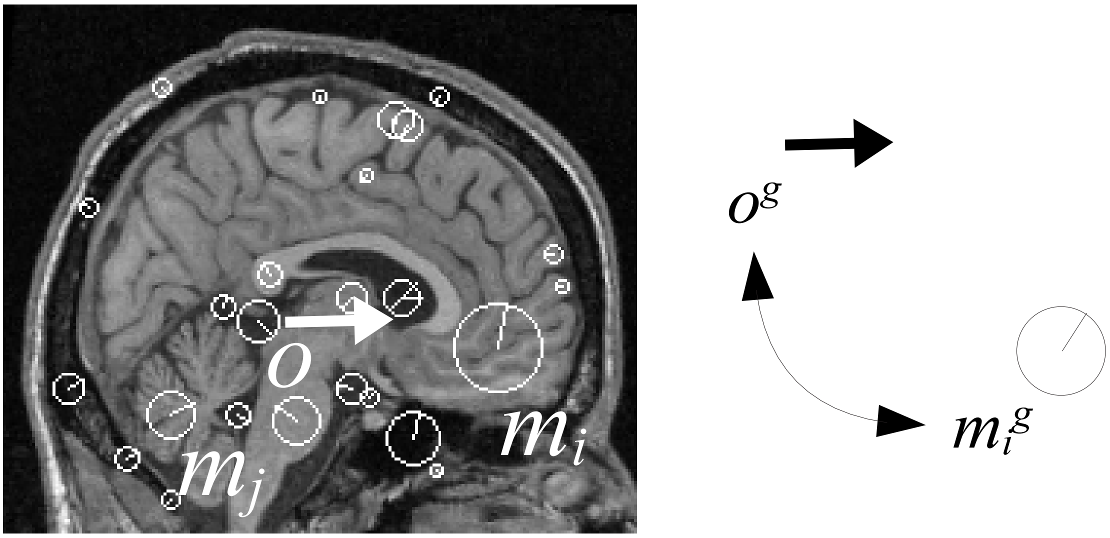


Fig. 3.16 Model parts and a reference frame in a sagittal slice of a T1-weighted MR brain image. In the left image, reference frame o is illustrated as a white arrow, and represents the projection of the Talairach AC-PC line onto the slice. Reference frame and model part geometry o^g and m_i^g are related via an invertible linear transform $t_i : m_i^g \rightarrow o^g$, as illustrated by the diagram on the right, and thus a single observed part is sufficient to infer the reference frame geometry.

The parts-based OCI model represents several important advancements with respect to current statistical appearance models in the medical imaging literature. The OCI model can be constructed over a large set of training images via the automatic machine learning procedure described in Section 3.1.2. The procedure automatically identifies a set of image parts **m** reflective of the underlying population anatomy which occur with statistical regularity in a population. Once identified, this set of localized model parts serves as a natural and intuitive basis for describing and communicating the variability of anatomical structure, in comparison with other representations such as modes of global covariance which are arguably less intuitive. All subjects of a population can be modeled simultaneously without making *a priori* classifications as to which subjects are 'normal'. This is because image structure common to multiple subjects is automatically recognized, allowing irregular structure due to subject-specific or abnormal characteristics such as pathology to be disregarded.

The OCI model can be robustly fit to new subject images via the process described in Section 3.1.3, in the presence of inter-subject variation and abnormality, along with global image translation, rotation and scale changes. While the fitting process seeks to identify an OCI reference frame hypothesis which maximizes the Bayesian decision ratio $\gamma(o_j, \mathbf{m})$, it is the set of model parts $\mathbf{m}$ identified in the new subject image which indicates how the anatomy of the subject relates to the population. Fitting is globally optimal, and does not relying on iterative search techniques prone to getting trapped in suboptimal local minima when poorly initialized. Additionally, model fitting is stable in the presence of local image perturbations in the sense that a local perturbation results in a local change in the fitting solution. This contrasts with global modeling approaches where a local perturbation will generally give rise to a global change in the fitting solution. This is demonstrated later in experimentation.

3.3.1 Modeling Cortical Appearance

The previous sections described how the OCI model can be used to describe the anatomy of a population from a set of medical images with respect to a standard geometrical reference frame, for example MR images of the human brain within the Talairach stereotactic coordinate system. In many instances, it may be of interest to focus modeling on specialized anatomical structures or regions of the brain. Modeling the anatomy of the cerebral cortex, for instance, is a task of special interest as cortical folding patterns tend to define functional regions of the brain. The goal of cortical modeling is to generate a description of how cortical folding patterns vary across subjects of a population. Cortical modeling techniques generally attempt reproduce expert sulcal labelings from training examples [175, 188]. These approaches are designed for detecting labeled sulci in new images, however, and do not attempt to learn new, unlabeled patterns. This thesis proposes using a parts-based OCI model to automatically learn a description of cortical anatomy from a large set of subject images, using only a minimal degree of manually labeling, i.e. labeling the OCI. This is the first learning technique to go beyond training labels and automatically learn the appearance of new, unlabeled cortical structure.

While OCI modeling of brain anatomy with respect to a central reference frame such as Talairach may be useful for describing large-scale structures or those near the reference frame origin, this methodology is less useful when modeling small-scale, distant structures such as cortical folding patterns. To understand why, recall that the OCI modeling assumption is that model parts arise from unique underlying anatomical structures. Ambiguous one-to-many or many-to-many geometrically similar feature matches are thus treated as incorrect, resulting in the statistical event $o^{b=0}$ of negative reference frame occurrence. Adopting an OCI in the form of a central reference frame necessarily results in a high number geometrically similar yet ambiguous false feature matches in regions far from the OCI origin as illustrated in Figure 3.17, such as the cortical extremities. This is because the number of such incorrect feature matches permitted by reference frame scale and orientation error thresholds T_θ and T_σ is a function of 1) the distance $dist(o, m_i)$ separating the OCI o and feature m_i and 2) the feature scale σ_i. In the example of 2D imagery, volume renderings of the cortical surface for instance, the probability of false matches can be expressed as:

$$p(m_i^{b=1}|o^{b=0}) \propto \frac{T_\theta T_\sigma dist^2(o, m_i)}{\sigma_i^2}. \tag{3.25}$$

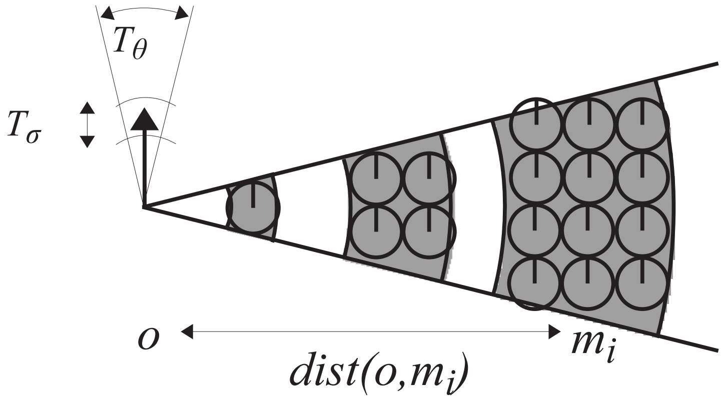


Fig. 3.17 The expected rate of geometrically valid, false correspondences of feature m_i increases with the distance $dist(o, m_i)$ separating a feature m_i and the reference frame o. The three grey arc regions are the areas in which geometrically consistent yet false correspondences could potentially arise for three different values of $dist(o, m_i)$. The number of such false matches is a function of the permitted variability in reference frame scale T_σ, orientation T_θ and $dist(o, m_i)$.

The farther features are located from the reference frame origin, the larger the area in which geometrically similar yet false matches can arise. This is particularly true for features arising from cortical folds, whose characteristic scale is relatively small compared to the distance to Talairach origin located at the anterior commissure. The hypothesis is thus that effective cortical appearance modeling can be achieved by adopting a reference frame defined locally according to stable cortical structures, for example primary sulci such as the central sulcus or the lateral fissure as illustrated in Figure 3.18. In general, defining an OCI reference frame locally/centrally within the specific anatomical region of interest is the best means of ensuring a minimal probability of false correspondence as defined by equation (3.25).

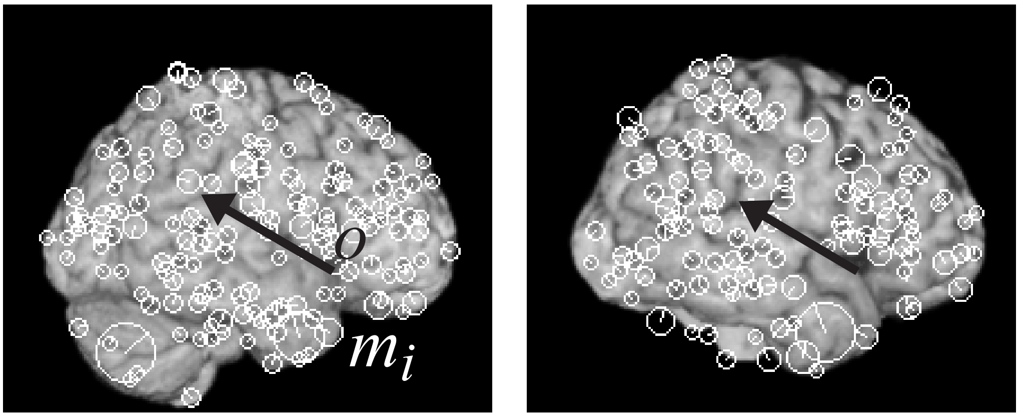


Fig. 3.18 A subset of cortical model parts m_i (white circles) and the reference frame o (black arrow) in lateral volume renderings of MR brain images from two different subjects (left and right images). Model parts derived from scale-invariant features are oriented image regions described geometrically by their location x_i, orientation θ_i and scale σ_i. The OCI reference frame (black arrow) is defined here locally along the lateral fissure, in order to improve modeling in the cortical region. Cortical modeling is more accurate in cortical regions near the OCI origin, as false feature correspondences are more unlikely. The reference frame is specified manually for the purpose of learning but automatically identified during fitting. Note the high degree of cortical folding variability from one subject to the next.

3.3.2 Classifying Subject Traits in Medical Imagery

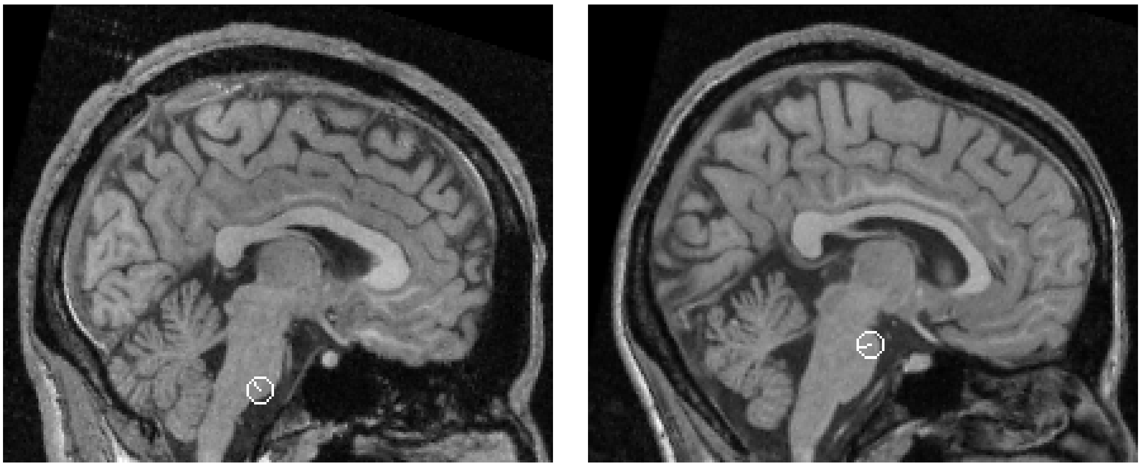
The previous sections described how the OCI model can be used as an anatomical description of a population from a set of medical images, such as the human brain. An important practical use of such an anatomical description is to relate anatomical structure to traits of subjects which define distinct sub-groups of a population. For example, one may ask: given a set of MR images, which regions of the brain are correlated with a trait such as a particular type of pathology? Answering such questions requires first establishing correspondences between images of different subjects, in order to quantify how structures vary with the trait of interest.

Widely used approaches such as morphometry [14] focus on identifying spatial regions within aligned subject images where voxel intensities or deformation fields correlate with traits of interest. The assumption is that one-to-one correspondence has been established between images of all subjects via inter-subject registration prior to analysis. As mentioned, morphometric analysis is typically dependent on the registration technique used, as different techniques generally result in different alignment solutions, particularly in regions where correspondence is ambiguous or difficult to establish. Indeed, in the case where different subject traits manifest themselves in very different image characteristics from one group to another, it would be reasonable to expect that registration may be difficult to achieve.

This thesis proposes identifying image characteristics indicative of subject traits in terms of local image features. Here, this is done by learning a Bayesian trait classifier from stable features of a learned parts-based OCI model of anatomy, as described in Section 3.1.4. The classification of brain images according to traits such as pathology could potentially be used in the context of a computer-aided diagnosis application. More importantly, however, is the ability of technique to link subject traits directly to distinct image patterns from the underlying anatomy of interest, in order to identify anatomical characteristics reflective of the traits. In Figure 3.19 for example, different features can be identified in the brain which are reflective of subject sex.

3.3.3 Inter-subject Registration

A central task in medical imaging is inter-subject registration, which strives to determine a geometrical relationship between two different subjects. An important use of the parts-based OCI model is as a basis for robust, reliable inter-subject registration. Intuitively, fitting the parts-based model to a new image can be seen as describing the new image in terms of a unique combination of learned model parts. Images of different subjects can be compared in terms of the model features that they share, the key notion being that any given pair of subjects shares a unique set of image parts. This information can potentially



A) Masculine feature

B) Feminine feature

Fig. 3.19 Examples of features f_i (white circles) indicative of sex $c : \{c_1 = male, c_2 = female\}$. Feature A) is indicative of male subjects when present, occurring with a log likelihood ratio of $log\ \frac{p(f_i|female)}{p(f_i|male)} = -0.18$. Feature B) is indicative of female subjects when present, occurring with a log likelihood of ratio of $log\ \frac{p(f_i|female)}{p(f_i|male)} = 0.13$.

be used in a number of ways, for example to drive inter-subject registration in regions where images are known to have statistically similar content, or to cluster subject images that share similar image content. As illustrated in Figure 3.20, this can help registration algorithms identify and potentially avoid attempting to determine correspondence in image regions where valid correspondence may not exist, due to inter-subject variability.

3.4 Summary

This chapter presented the theory behind the OCI model of abstract pattern appearance, on which this thesis is based. The probabilistic formulation of the OCI model was described, in addition to algorithms for learning an OCI model from a set of arbitrary training images and fitting the model to new images to detect, localize and classify new pattern instances according to visual traits. Additionally, guidelines were provided for determining an optimal OCI, including an algorithm for automatically determining such an OCI in an iterative, data-driven manner. The overriding focus of the OCI modeling approach is to generalize and unify work done in general object class detection, visual trait identification and classification, and pattern description. Specific application of OCI modeling theory in two

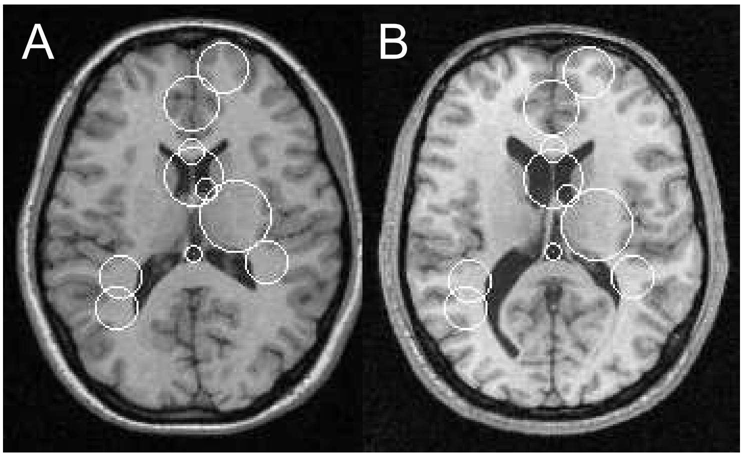


Fig. 3.20 Model-based brain image registration for subjects A and B. Prior to registration, the parts-based model of the brain is first fit to each image to identify features (white circles) that are both 1) reflective of brain imagery and 2) common to the images being registered. In this way, registration can be avoided in regions where valid inter-subject correspondence may not exist due to inter-subject variability, such as the elongated lower left ventricle of subject B.

contexts was described, i.e. modeling 3D object class appearance in the field of computer vision and anatomical modeling in the field of medical imagery.

In computer vision, the OCI model is the first approach capable of modeling appearance of general 3D object classes, such as faces or cars, from arbitrary viewpoints. The model can be learned from a small number of cluttered images acquired from arbitrary viewpoints with minimal manual supervision, and can be used to detect, localize and classify new object class instances in terms of abstract visual traits. The final result is the first model of 3D faces capable of detecting, localizing and classifying 3D faces in terms of sex from arbitrary viewpoints.

In medical image analysis, the OCI model is the first approach to propose describing the anatomy of a population in terms of a collection of generic, spatially localized image patterns, or parts. The anatomy of a complex pattern can be described in terms of a collage of localized parts arising from underlying anatomical structure, where parts are quantified probabilistically in terms of their occurrence frequency, geometry and appearance from a large set of medical images. The final result is the first parts-based description of the human brain in MR imagery, which can be learned automatically from a large set of medical images of different subjects, robustly fit to new subjects in the presence of inter-subject variability and unexpected local perturbation, and used to identify anatomical structure linked to the trait of sex.

The following chapters describe experimentation involving the OCI model. In Chapter 4, the OCI model is demonstrated in the computer vision context as a model of 3D object class appearance. In Chapter 5, the OCI model is used as a description of anatomy in the context of medical image analysis.

Chapter 4

Experimentation in Computer Vision: 3D Object Classes

In this chapter the OCI model is applied in the context of computer vision, where images represent 2D projections of the 3D world. The goal of experimentation is to demonstrate the effectiveness of the viewpoint-invariant OCI model in tasks that require representing the appearance of 3D object classes in 2D projective imagery. The concrete computer vision application investigated in this chapter is a combined system for learning, detecting, localizing and classifying traits 3D object classes in natural imagery acquired from arbitrary viewpoints, such as the faces illustrated in Figure 4.1. The majority of experimentation in this chapter focuses on the class of 3D human faces, due to the ubiquitous nature of face imagery. The OCI model is generally applicable to a wide variety of object classes, however, and experiments involving detection and localization are also performed on the class of 3D motorcycles.

All experimentation in this chapter makes use of scale-invariant features, which are automatically extracted from all images. Although a variety of different features can be used, the scale-invariant feature transform (SIFT) technique [22] is adopted for feature detection and appearance description [1]. The SIFT feature detection method is based on identifying minima and maxima in a difference-of-Gaussian scale-space pyramid, and has been shown to outperform other techniques in terms of detection repeatability [194]. The SIFT appearance representation involves transforming the image content associated with

[1] The SIFT implementation used is publicly available [193].

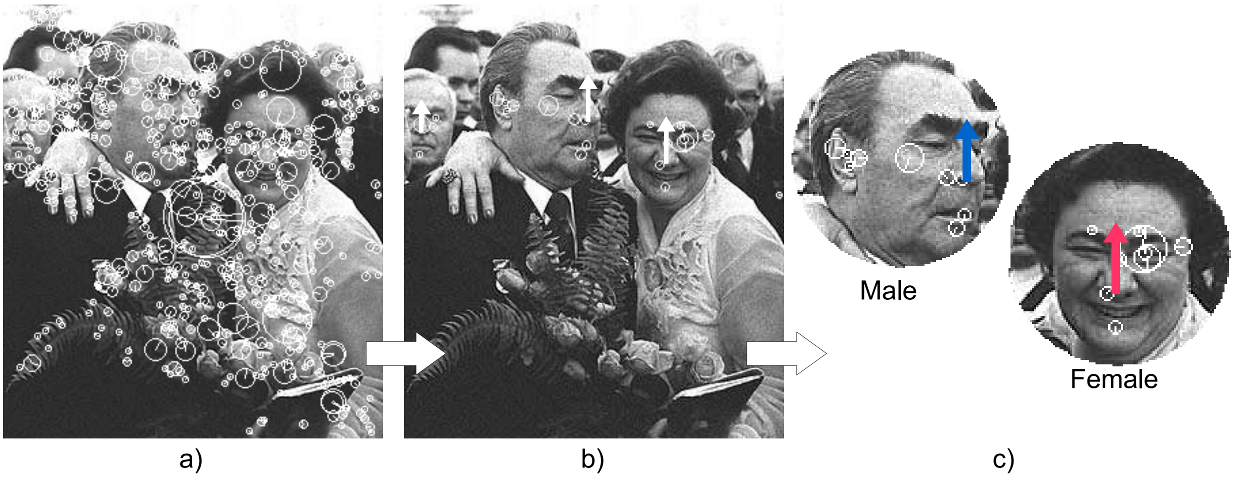


Fig. 4.1 Illustrating the general OCI framework for combined detection, localization and trait classification from arbitrary viewpoints.

features into a histogram of gradient orientations, and has been shown to be superior to other approaches in terms of distinctiveness [81]. Individual histogram bins can be modeled as statistically independent and identically distributed, and similarity can be modeled via the Euclidean distance metric.

The experimentation in this chapter is organized into three main sections. The first, Section 4.1, presents experiments demonstrating OCI model learning, detection and localization in natural, cluttered image face image data captured from arbitrary viewpoints. Section 4.1.1 demonstrates the capacity of the OCI model to learn a viewpoint-invariant representation of 3D face appearance from a set of natural, cluttered images acquired from arbitrary viewpoints, and evaluates the detection/localization characteristic of the OCI model. Section 4.1.2 performs a quantitative comparison of detection performance between the multi-view and viewpoint-invariant appearance representations, showing that the viewpoint-invariant OCI model results in quantitatively improved detection performance over the multi-view model for the case of face imagery. Section 4.1.3 demonstrates the data-driven algorithm described in Section 3.2.3 for deriving an optimal OCI from data, experimental establishing the existence of a stable OCI for the class of face images. The data-driven OCI remains geometrically consistent with the 3D faces in images acquired from arbitrary viewpoints, as predicted by theory, and improves the detection performance of the OCI model.

The second set of experiments, described in Section 4.2, demonstrate the technique for learning and classifying visual traits based on the OCI model of object class appearance presented in Section 3.2.4. The OCI model is used as a framework for learning and classifying the sex of human face images captured from arbitrary viewpoints. This represents the first time this task has been addressed in the literature. Results are based on the standard color FERET database [38] and results establish a first baseline in the literature for sex classification from arbitrary viewpoints. Section 4.2.2 presents a qualitative overview of visual trait learning, demonstrating local facial features that serve as cues of face sex. Section 4.2.3 details an extensive evaluation of sex classification in face images captured from arbitrary viewpoints. A classification error rate of approximately 15% is achieved, and an analysis of error with respect to viewpoint shows that frontal views are classified with lower error than oblique or profile views. Section 4.2.4 evaluates sex classification in the presence of simulated occlusion, thereby demonstrating the capacity of the local OCI model to cope with missing features. Classification performance is shown to degrade gracefully with increased occlusion, and that reasonable classification performance is achieved despite significant occlusion. Section 4.2.5 provides an evaluation of OCI sex classification based on frontal, non-occluded faces in order to compare with results presented in the literature. The OCI classification error rate on frontal faces of 11.2% is higher than other results presented in the literature of $4\% - 10\%$, however other approaches are likely unable to cope with viewpoint change or occlusion.

The final set of experiments in Section 4.3 are designed to demonstrate the generality of the OCI model and to demonstrate the feasibility of a spherical OCI reference frame. Learning, detection and localization trials are performed on the class of 3D motorcycles, based on the standard 2006 PASCAL visual object classes data set [39]. Results show that a viewpoint-invariant OCI model of 3D motorcycles can be learned from cluttered images acquired from arbitrary viewpoints, and used to detect and localize motorcycles with accuracy comparable to other approaches in the literature. A discussion of experimental results follows in Section 4.4.

4.1 Viewpoint-Invariant Learning and Detection of 3D Faces

This section investigates viewpoint-invariant OCI model learning and detection for the class of 3D faces. Due to the ubiquitous nature of face imagery, a large body of literature is devoted to developing high performance face detection systems. This is not focus of this chapter, however. Rather, 3D face modeling is used as an application domain to investigate learning and detection in natural, cluttered imagery acquired from arbitrary viewpoints, multi-view vs. viewpoint-invariant modeling, and the data-driven derivation of optimal OCI reference frames.

4.1.1 Learning and Detecting 3D Faces in Clutter and from Arbitrary Viewpoints

This section examines the feasibility of 1) learning an OCI model from a set of natural, cluttered training images, and 2) using the model to detect new class instances, all in images exhibiting a wide range of appearance variability due to viewpoint change and intra-class variation.

Data: The OCI model is generally applicable to classes of objects from which scale-invariant features can be repeatably extracted. Experimentation focuses on modeling the class of 3D faces due to the abundance of raw data. Image data used consist of 180 examples of faces of different people acquired under of variety of arbitrary imaging conditions and viewpoints, in addition to a set of 43 negative image of scenes not containing faces. Images were obtained from a variety of source including the CMU profile database [37], personal photos and internet image search engines. As illustrated in Figure 4.2, the positive examples contain a high degree clutter, and exhibit a wide variety of appearance variability due to viewpoint change, (sun)glasses, expressions, ethnicity, and other factors.

Model Learning: Prior to learning, OCIs are manually defined and labeled as a line segments from the nose tip to the forehead, as in Figure 3.10. Extreme profile or rear views of the face are labeled by guessing the approximate projection of the OCI, when the front of the face was not visible. Model learning is based on the learning algorithm in Section 3.1.2, based on a total of approximately $16,000$ features. After clustering and discarding redundant features, the model contains approximately $2,800$ features. Model learning is relatively quick, on the order of minutes, and detection on the order of seconds for images of size $\approx 300 \times 200$ pixels, on a Pentium M laptop computer clocked at 1.6 GHz.

Detection trials: Detection trials are performed in a leave-one-out manner. A model is trained using the entire training set except for one face, after which the model is used to detect the face in the remaining face using the OCI model fitting algorithm described in Section 3.1.3. This process is repeated for each image in the data set. In order to

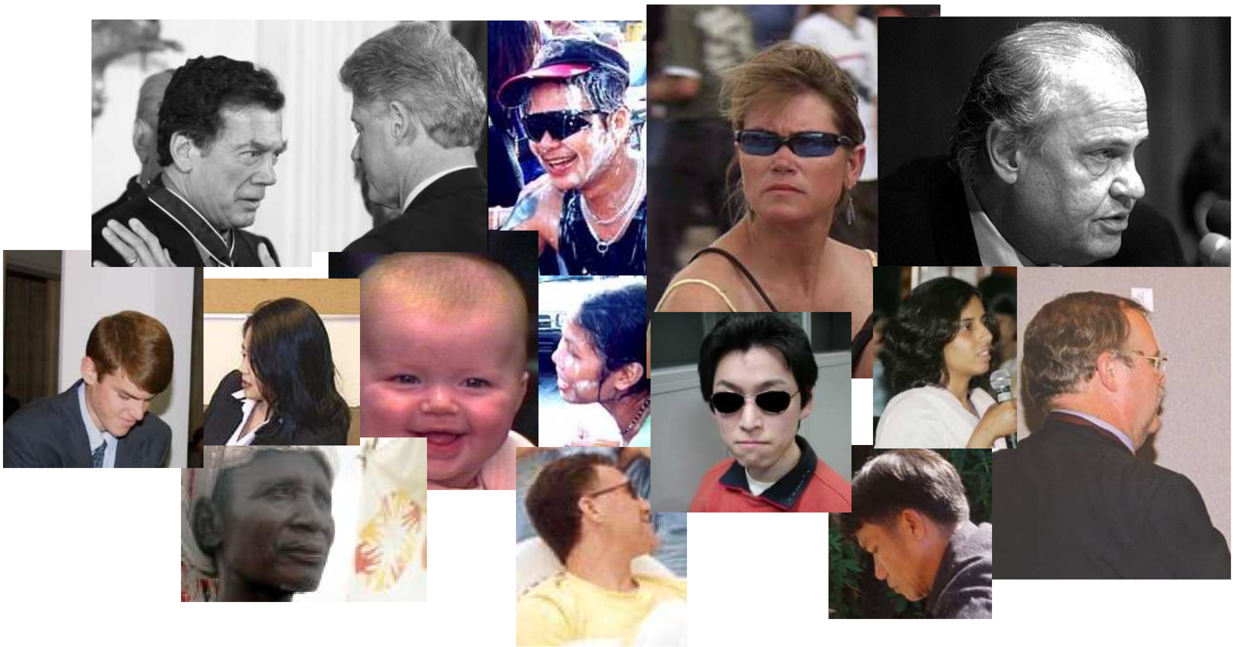


Fig. 4.2 A subset of images used in viewpoint-invariant face learning, detection and localization experimentation. Note the wide range of appearance variability due to viewpoint change, glasses, expressions, etc. Images used were selected from a range of different sources including the CMU profile database [37], personal photos and internet search engines.

evaluate detection and localization performance, a mechanism is required in order to declare OCI hypotheses as correctly or incorrectly localized. Note that this in itself can be a difficult issue to address, as an arbitrary threshold must be established. Measures such as overlap of bounding boxes about the detected object instances can be used [195], however the output of the OCI detection is unique in that it is an oriented, viewpoint-invariant line segment. Here, a detected hypothesis is considered correct if it falls within a scale-invariant threshold $Thres^g$ of the labeled OCI, and unsuccessful otherwise. Here, individual thresholds in $Thres^g$ are set to $T_\sigma = \log(1.5)$ octaves in scale, $T_\theta = 20^o$ in orientation, and $T_x = 0.5/\sigma$ pixels in (x,y) translation, where σ is the scale of the labeled OCI. These values were chosen empirically, larger values permit a greater discrepancy between labeled and detected OCIs. In addition, potential duplicate hypotheses arising from the same face are pruned in a threshold of $Thres^g$ around hypotheses maximizing the Bayes decision ratio in equation (3.12), as illustrated in Figure 4.3. Each detected OCI consists of a unique combination of $10-20$ model features from different training images. This highlights the ability of the OCI model to represent a large range of appearance modes.

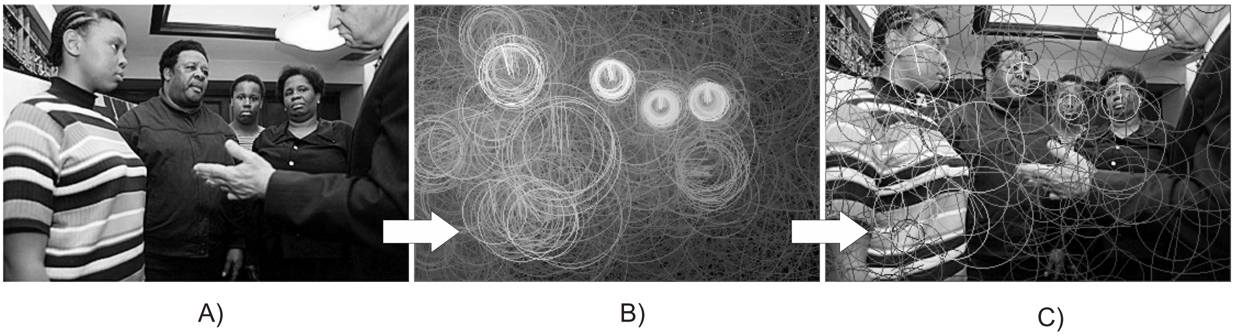


Fig. 4.3 OCI hypothesis generation and pruning based on input image A). In image B), all OCI hypotheses (white circles with radial lines) are shown, where the pixel intensity of each OCI hypothesis shown is proportional to the Bayes decision ratio $\gamma(o, \mathbf{m})$ of the hypothesis. Note how dense clusters of strong OCI hypotheses occur with four of the five faces. Multiple hypotheses in close proximity to a hypothesis with a locally maximal Bayes decision ratio are suppressed or pruned, resulting in a sparse set of hypotheses displayed in image C).

The result of detection trials are illustrated in Figure 4.4 in terms of a true positive vs. false positive detections. Note that this curve is not, by strict definition, a receiver operating characteristic (ROC) curve, as not all face instances are correctly localized. The curve here is based on a total of 180 valid detections and 26,475 false positives, and rises rapidly to a maximum detection rate of approximately 81%. This illustrates that the model is capable of describing approximately 145 of the 180 faces. The maximum detection rate here can be considered conservative, due to the fact that the thresholds $Thres^g$ used to label detections as either correct or false are rather tight. For this reason, several near-correct solutions are labeled as incorrect, e.g. that of Pete Sampras kissing the trophy in Figure 4.4. The detection rate could thus potentially be improved by modifying the means by which OCI hypotheses are declared as correct or incorrect detections. However the goal of this experimentation is not to establish an absolute measure of detection performance, but rather to demonstrate that the OCI model could be learned and used for face detection and localization in difficult, natural imagery acquired from arbitrary viewpoints.

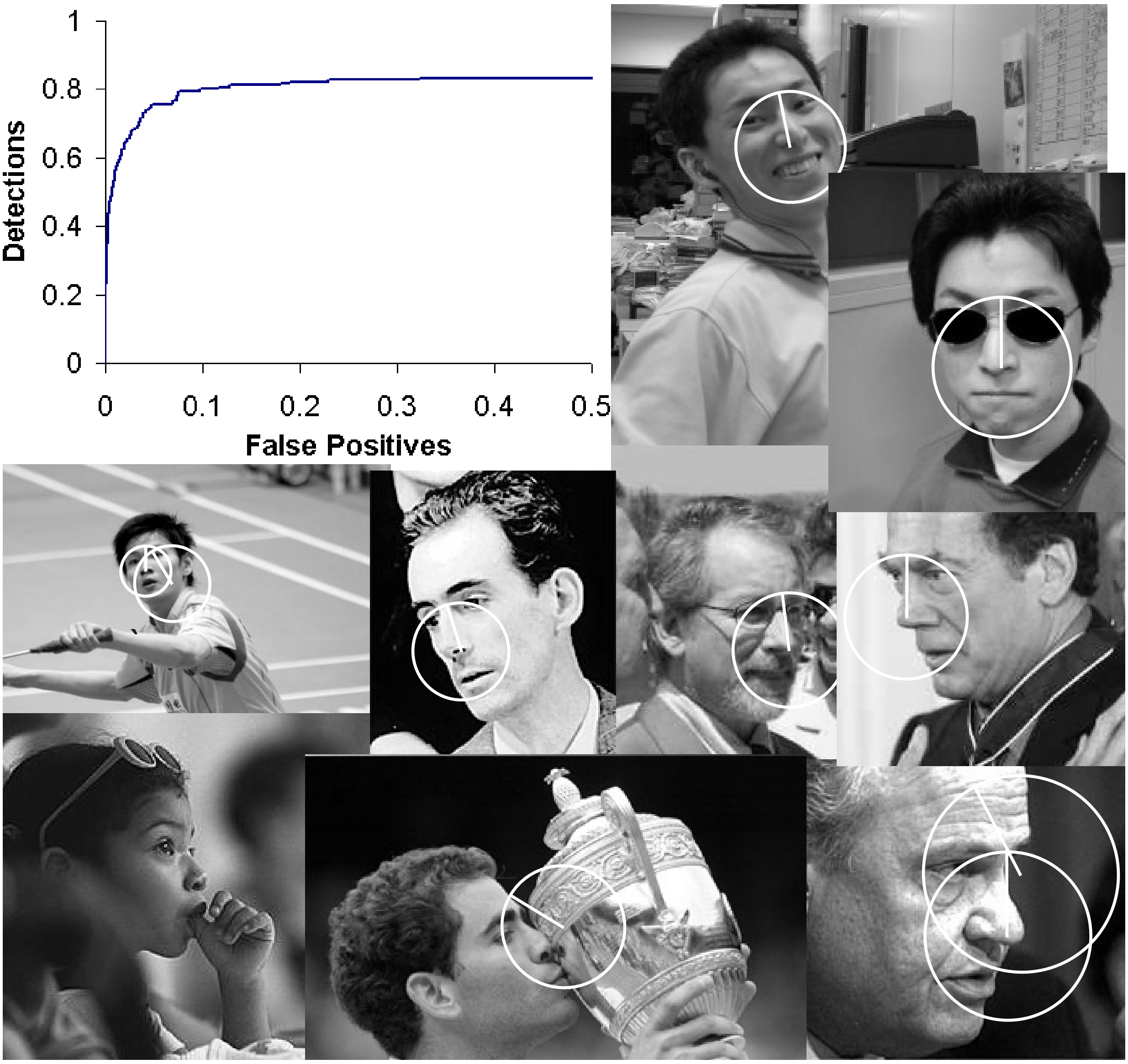


Fig. 4.4 Illustrating the result of viewpoint-invariant OCI face detection on a database of 180 face images, exhibiting a wide range of viewpoint and appearance variability. The curve in the upper left corner reflects the detection vs. false positive characteristic of the OCI model as the detection threshold is varied. The white circles overlaying the images reflect all correct OCI detections, false positives, and missed detections for a subset of images at a positive detection rate of 0.6.

4.1.2 Multi-view vs. Viewpoint-Invariant Modeling

Experimentation in this section aims to investigate the question as to whether it is preferable to model in terms of a set of distinct views (i.e. the multi-view representation) or in terms of a single model over the entire viewpoint range (i.e. the viewpoint-invariant representation). The hypothesis presented in Section 3.2.2 was that linking features directly to the 3D geometry of the object class via the viewpoint-invariant OCI model, instead of to specific views, would improve detection. This hypothesis is tested through experimentation in the context of 3D face learning, detection and localization.

Training data: Model learning is based on the standard, publicly available FERET face image database [38], consisting of images of 994 unique subjects of various ethnicity, age, sex, acquired from various viewpoints, illumination conditions, with/without glasses, etc. As the FERET database contained images acquired from carefully controlled viewpoints, it provides a good basis for multi-view modeling and therefore comparison of the multi-view and viewpoint-invariant representations. For the purpose of training, 497 different subjects (half of all FERET subjects) are randomly selected. For each subject, a viewpoint from the range of -90 to 90 degrees (i.e. left profile to right profile) provided by the FERET database is randomly selected. This results in a total of 497 images, which are processed in grey scale at a resolution of 256x384 pixels. Each image results in 150-300 SIFT features. Training OCI geometries are manually labeled from the base of the nose to the forehead.

Testing data: Detection performance is evaluated on a subset of images from the CMU profile database [37], containing a total of 95 faces. The database is a challenging detection test set containing face images under arbitrary viewpoints amid a high degree of background clutter. A subset of testing faces which are of higher resolution are selected, as low resolution faces (i.e. < 40 pixels) produced few SIFT features and are thus difficult to detect. Ground truth OCI locations are labeled as line segments from the base of the nose to the forehead, as in training images. A list of the CMU images used in testing can be found in Appendix A.

Learning: For the purpose of multi-view learning, the range of viewpoint is quantized into 3 distinct views: frontal, oblique and profile, as illustrated in Figure 4.5. The multi-view representation is generated by learning 3 distinct single-view OCI models from the images in each of the 3 view ranges. The viewpoint-invariant OCI representation is generated by learning a model from all 497 images. Model learning is performed by the process described in Section 3.1.2.

Detection: Detection is performed by fitting the respective models to features extracted in the test images, as described in Section 3.1.3. To suppress multiple detection

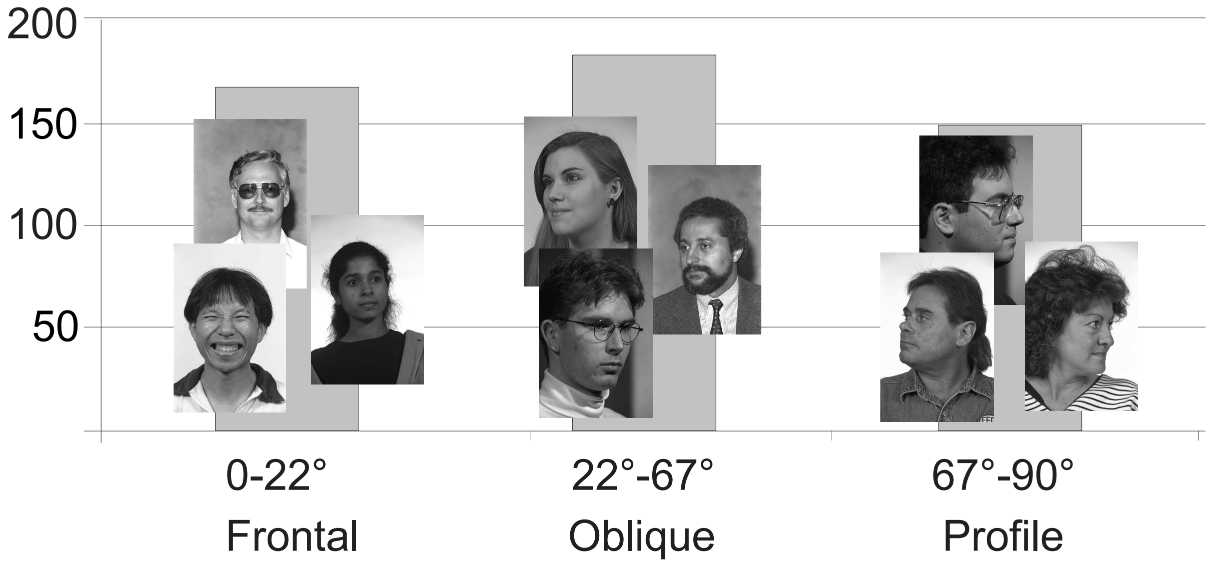


Fig. 4.5 A histogram of the three viewpoint ranges used in multi-view modeling: frontal, oblique and profile. The 497 test images are approximately equally distributed across all views. Note that as the model exploits the mirror symmetry of the face, a 180 degree range of viewpoint can be covered by the three views.

hypotheses arising from the same face, all hypotheses in a proximity of a hypothesis with a locally maximal Bayes decision ratio are removed, where proximity is defined by the geometrical OCI agreement threshold $Thres^g$ as using in model learning. Note that the detection output is considered for all three view models of the multi-view representation, while the viewpoint-invariant model produces only a single detection output.

Figure 4.6 illustrates the precision-recall curves for viewpoint-invariant and multi-view detection. In general, the higher the precision-recall curve, the better the performance of detection. Two different precision-recall curves can be quantitatively compared by the average precision (AP) measure [114], defined by the arithmetic mean of the precision evaluated at 11 recall increments from 0.0 to 1.0. Here, the viewpoint-invariant representation achieves an AP of 0.26, outperforming the multi-view representation which achieves an AP of 0.24. Note that the viewpoint-invariant representation performs particularly well in the region of high detection precision, where detection is most reliable.

As hypothesized, the superior detection performance of the viewpoint-invariant representation is due to the higher false positive rate of the multi-view approach. This is

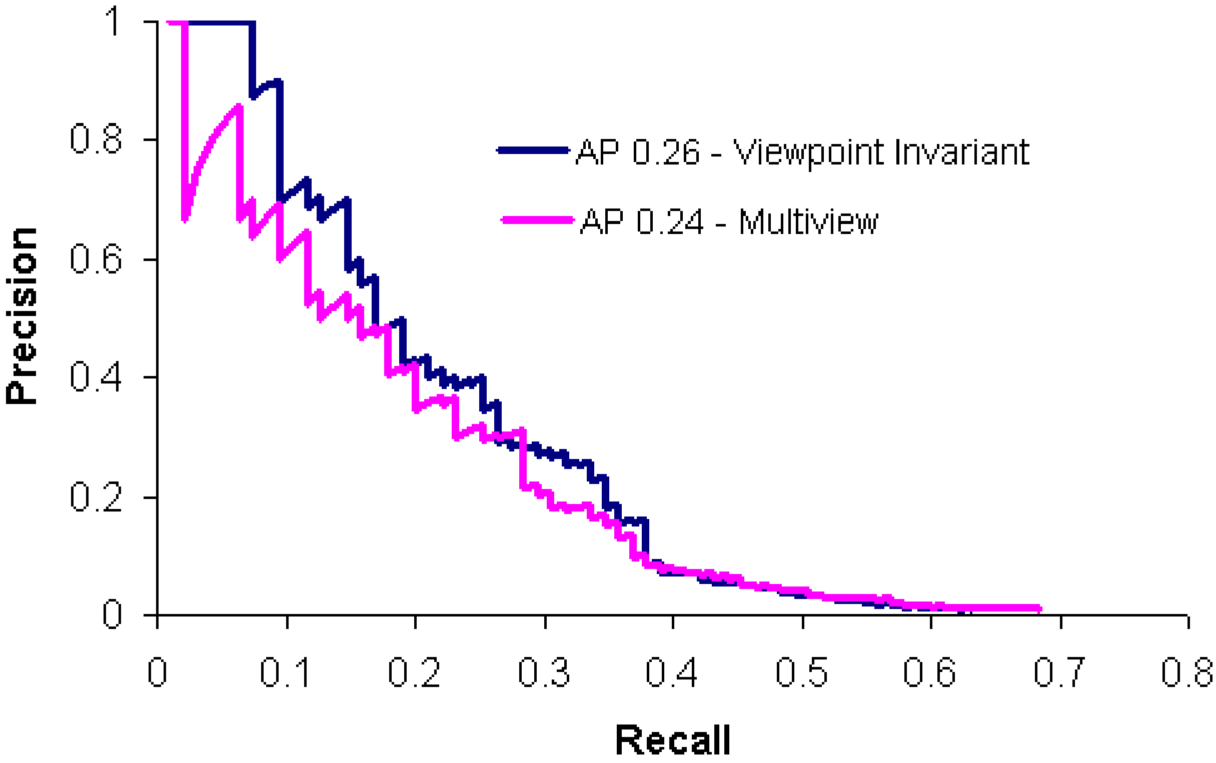
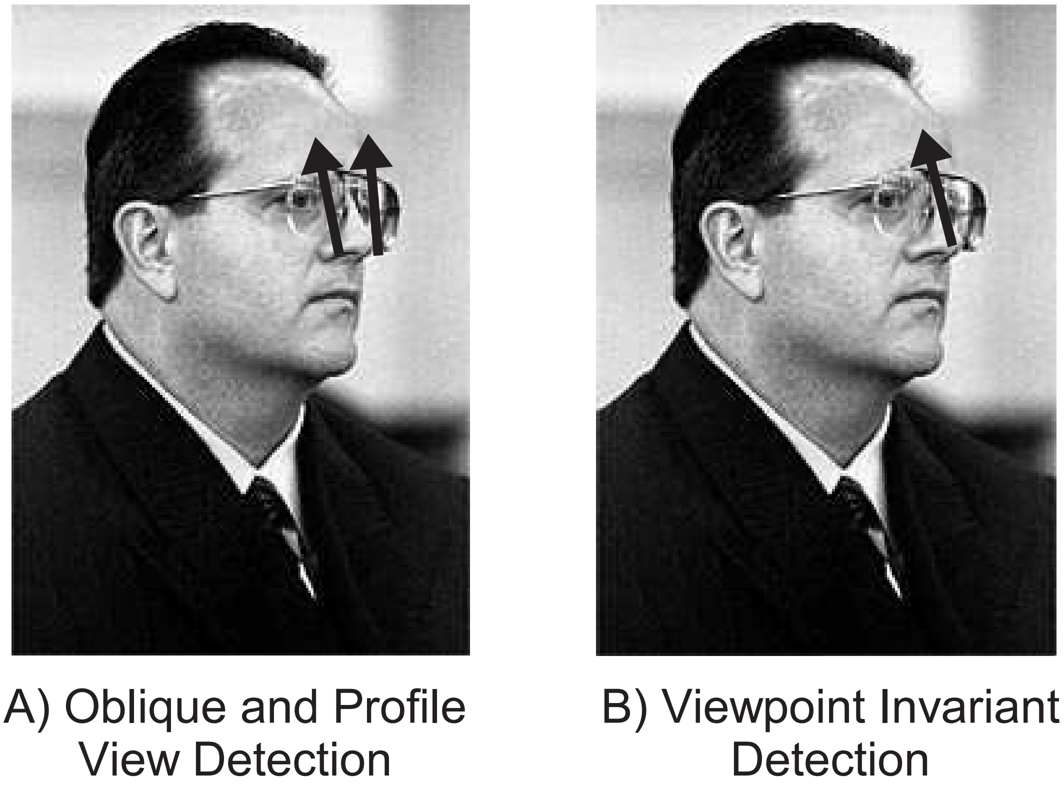


Fig. 4.6 Precision-recall curves for viewpoint-invariant and multi-view detection. Viewpoint-invariant detection generally outperforms multi-view detection, with respective average precision (AP) values of 0.26 and 0.24. The AP measure is defined by the arithmetic mean of the precision evaluated at recall increments of 0.1 [114]. Note here that the discrepancy between the two curves is more pronounced for low values of recall.

particularly true when the viewpoint of a testing image falls in between the two training views of the multi-view representation, as illustrated in Figure 4.7. In this situation, the same image features are fit to different view models, producing distinctly different geometrical interpretations with respect to the geometry of the object class instance. In contrast, the viewpoint-invariant model tends to produce a single strong hypothesis as to the object class within the image. This is because features are not linked to fixed views, but rather to a viewpoint-invariant property of the object class itself, the OCI.

Although the tendency of the multi-view model to produce multiple hypotheses for the same object class instance can be addressed to an extent [4], it is generally difficult to know whether two neighboring hypotheses arise from two distinct object class instances or



A) Oblique and Profile View Detection

B) Viewpoint Invariant Detection

Fig. 4.7 The viewpoint of the face shown above falls somewhere between the oblique and profile views of the multi-view model. As a result, the oblique and profile views of the multi-view representation produce two strong, distinct interpretations as to the geometry of the object class instance, as shown in A). The viewpoint-invariant representation produces a single strong interpretation, as shown in B), as image features are related directly to the geometry of the object class instance and not to fixed views. Note that the frontal view of the multi-view representation did not produce a significant detection hypothesis for this image.

simply different interpretations of the same object class instance from different views, as illustrated in Figure 4.8. Furthermore, this difficulty is much less prevalent when starting from a viewpoint-invariant representation, where each feature effectively serves as its own best view.

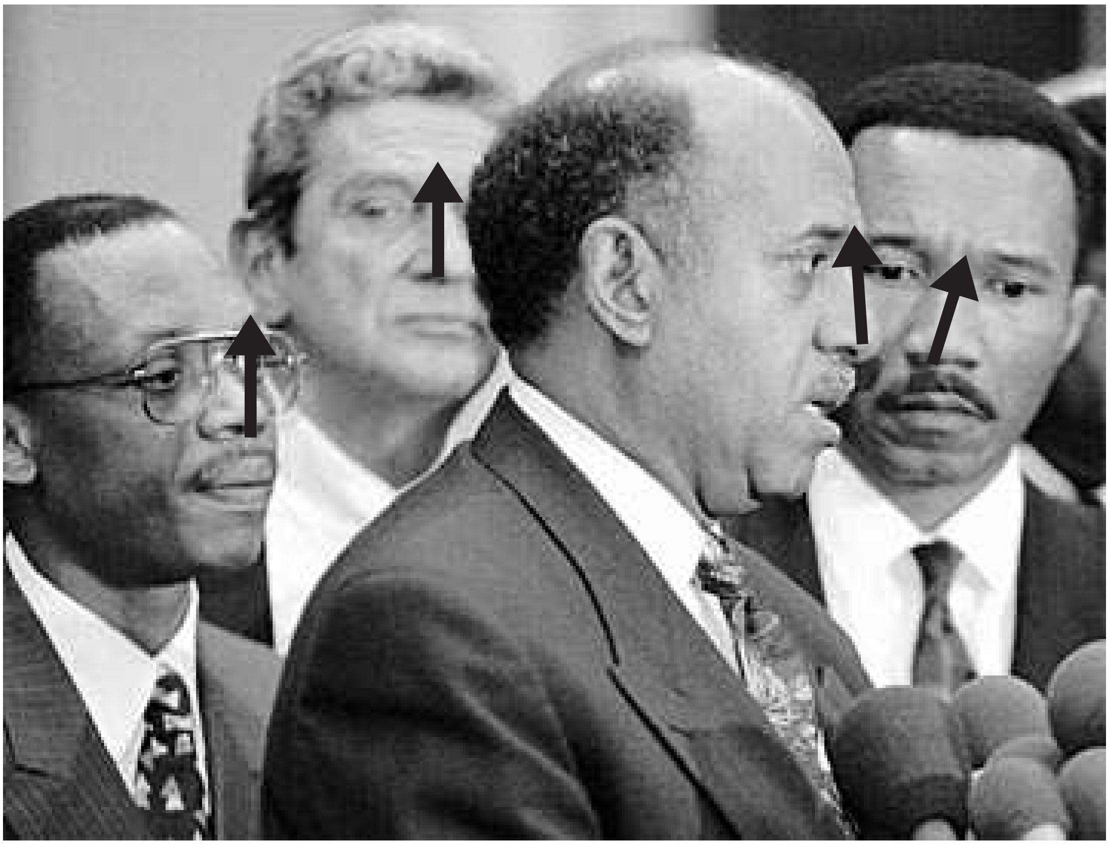


Fig. 4.8 The two overlapping faces to the right illustrate a situation where it is difficult to determine whether multiple detection hypotheses are the result of different interpretations of the same object class instance or two distinct object class instances.

4.1.3 Determining an Optimal Viewpoint-invariant OCI

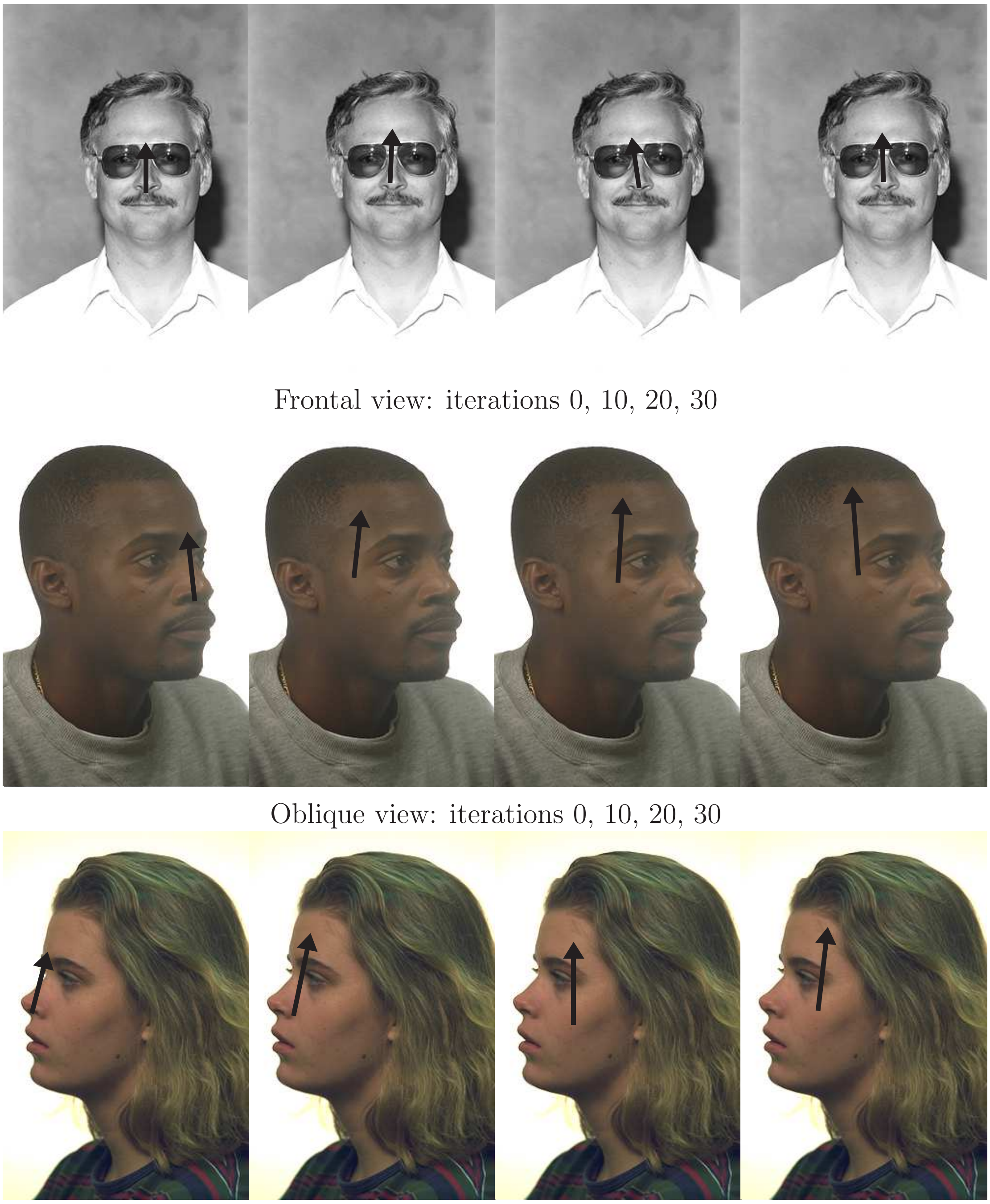
Experimentation in this section is intended to address the question as to what is an optimal OCI, and more specifically, can an optimal invariant reference frame be derived for an object class in an iterative, data-driven manner? This section tests the iterative algorithm proposed in Section 3.2.3, which attempts to learn an optimal OCI by alternately learning a set of model features $\mathbf{m}$ from labeled training OCIs $\{o_j\}$, then re-estimating OCI labels by fitting the learned model back to the training images.

The experimentation in this section follows from that in Section 4.1.2, involving precisely the same experimental setup. The FERET face image OCIs are initially labeled as line segments from the base of the nose to the forehead, and the iterative process is repeated until OCIs appear to converge in the training images, which requires approximately 30 iterations. For frontal face training images, the re-learned OCI labels become slightly larger in scale but do not change significantly in orientation or location. In all oblique and profile images, however, the OCI labels retreat noticeably from the nose back to the cheeks, as illustrated in Figure 4.9. These new labels, determined by an iterative, data-driven process, are consistent with the projection of a single 3D line segment central to the 3D human head onto the image plane. As such, they minimize the distance between image features and the projected OCI location over a 180 degree range of viewpoint.

In order to evaluate the new optimal OCI definition, detection trials are re-run. The difficulty, however, is that the ground truth OCI geometries in the testing images must be relabeled according to the new OCI definition. As the new optimal OCI definition has been automatically determined, the human labeler him/herself must first learn it by visualizing examples of the new OCI in training images, such as those in Figure 4.9 iteration 30, before producing ground truth labelings in the testing images. The detection results are shown in Figure 4.10, where the optimal OCI results in improved detection performance over the initial manual OCI in terms of average precision for both multi-view and viewpoint-invariant representations.

4.2 Detecting, Localizing and Classifying Visual Traits of 3D Faces

This section investigates the use of the OCI model for the combined task of detection, localization and visual trait classification. Although the classification approach generalizes to a variety of object classes and visual traits, experimentation here involves the trait of sex in face images, due to the large body of work addressing this task and the availability of public data sets. Although a large body of research has examined sex classification of faces from frontal face images, this experimentation is the first to address learning and classifying human faces in terms of sex from arbitrary viewpoints. It is also the first to embed viewpoint-invariant detection, localization and classification in a single framework.



Frontal view: iterations 0, 10, 20, 30

Oblique view: iterations 0, 10, 20, 30

Profile view: iterations 0, 10, 20, 30

Fig. 4.9 The result of iteratively learning and re-estimating OCI labels in training images, for 0, 10, 20 and 30 iterations. In iteration 0, all OCIs are manually initialized as line segments from the base of the nose to the forehead. Little change occurs for OCIs after 30 iterations in frontal views, which are already approximately central to image features arising from the face. In quarter and profile views, OCI locations recede to the cheeks, minimizing the average distance to image features characteristics of these views (e.g. ears, cheeks, eyes). Note that the OCIs in all views remain consistent with 3D geometry of the object class, corresponding to the 2D projections of the same 3D line segment located within the center of the 3D head.

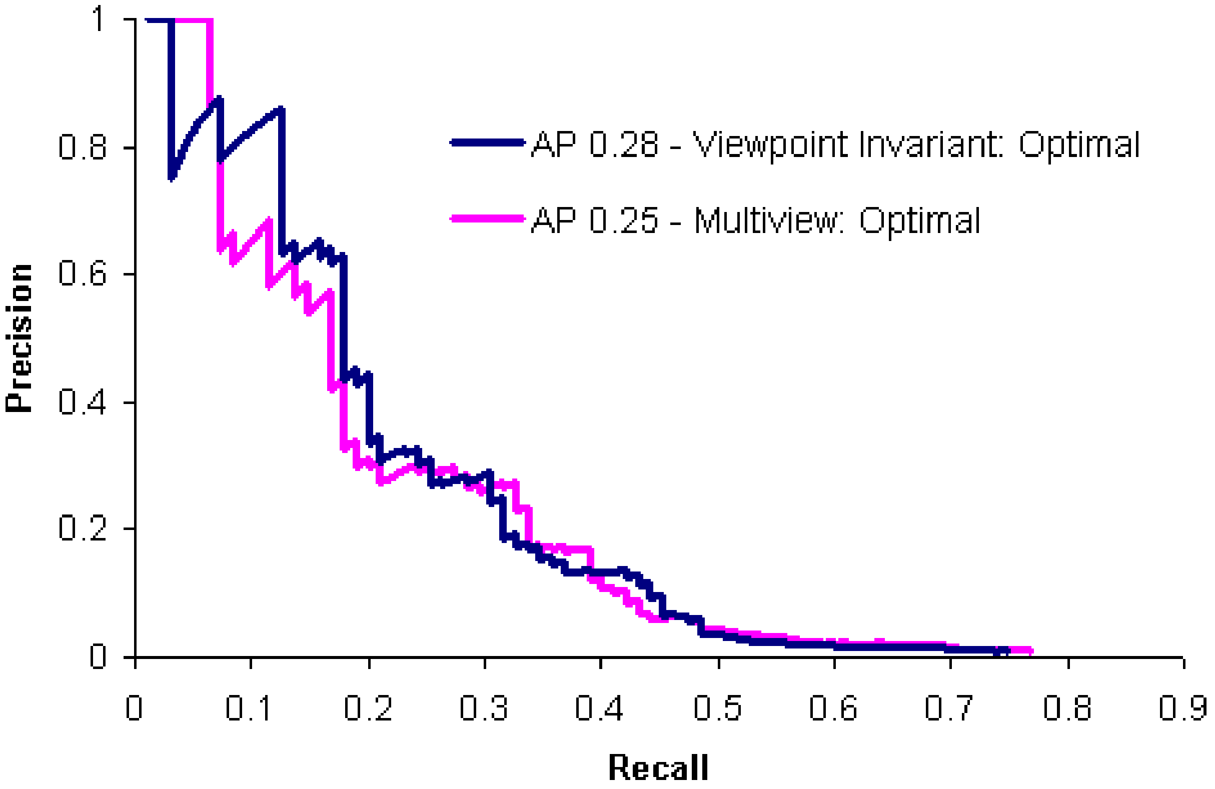


Fig. 4.10 Precision-recall curves using the optimal OCI for both viewpoint-invariant and multi-view detection. Using the optimal OCI definition improves detection for both viewpoint-invariant and multi-view representations from the initial definition, which show average precision (AP) improvements of 0.26 → 0.28 for viewpoint-invariant detection and 0.24 → 0.25 for multi-view detection.

4.2.1 Experimental Details

Data: An ideal testing scenario for sex classification would be a public database of face images acquired from arbitrary viewpoints in difficult natural scenarios, with accurate sex labels. Unfortunately, such as database does not currently exist. Challenging data sets for benchmarking face detection performance exist, but are lacking when it comes to classifying visual traits. The CMU database [37], for example, contains no ground truth sex labels, and manually determining labels can be ambiguous. As most images were gleaned from similar sources, diversity of faces in traits such as ethnicity, age and sex is limited. For instance, out of ≈ 600 faces in the database, 1 in 6 appear to be female, making it difficult to evaluate classification for both genders. Furthermore, several faces are duplicated within the database, e.g. prominent politicians and personalities. Trials of combined detection, localization and classification were performed using images from the CMU profile database [37] as illustrated in Figure 4.11. Qualitatively reasonable results were obtained, particularly for larger faces from which a good number of SIFT features can be extracted.

Experimentation in this section thus focuses on evaluating the performance of detection, localization and sex classification based on the standard, publicly available color FERET face image database [38] for both training and testing. Although the FERET database does not necessarily represent a challenging scenario for detection and localization, it is a good basis for evaluating and comparing trait classification, as FERET faces are diversely sampled and labeled in terms of viewpoint, sex, age and ethnicity. A database of 994 images is created, one for each FERET subject, where each subject image is chosen at random from a 180 degree viewpoint range (i.e. from left to right profile images). In this way, no subjects are duplicated in either testing or training data, in order to evaluate the generality of the approach. Figure 4.12 illustrates the viewpoint distributions for males and females in the data set. The male:female ratio in the database is approximately 3:2 (591:403). Images are converted to grey scale and processed at a resolution of 256x384 pixels.

Model Learning: A viewpoint-invariant OCI model of 3D faces is learned from training data using the learning procedure outlined described in Section 3.1.2. OCIs are labeled as line segments from the base of the nose to the forehead in all training images.

Classifier Training from Modeled Features: Once the OCI face model has been learned, model feature occurrences identified in the training set along with FERET sex labels are used to estimate likelihood ratios of the Bayesian trait classifier as described in Section 3.1.4. In estimating likelihood ratios via equation (3.16), an empirically determined Dirichlet regularization parameter of $d_{i,j} = 2$ is used, which maximizes training set classification performance. Larger values of $d_{i,j}$ lead to overly regularized likelihood ratios and suboptimal classification, and smaller values lead to noisy likelihood ratios and classification.

Detecting and Localizing in Testing Images: Once the appearance model and classifier have been learned, fully automatic 3D face detection and localization proceed on testing images, in order to identify the OCI instance in each of the testing images maximizing the Bayesian decision ratio in as described in Section 3.1.3.

Classifying Traits in New Images: Once an OCI instance is detected in a new

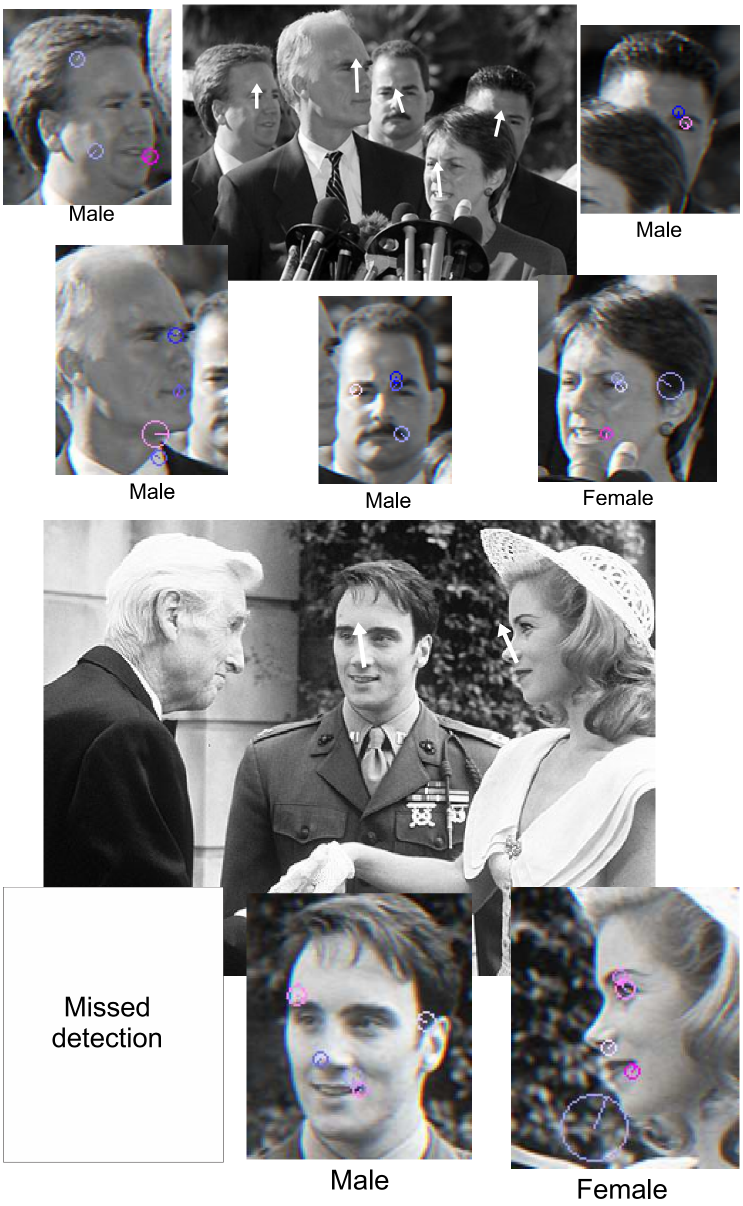


Fig. 4.11 Examples of face detection, localization and sex classification in cluttered images of the CMU dataset [37]. White arrows indicate detected and localized OCI locations. Colored circles overlaying thumbnail face images indicate the image features operative in localizing and classifying the particular face. Features indicative of male characteristics are blue, and those indicative of female characteristics are pink, where the color saturation is proportional to the magnitude of the log likelihood ratio. Of the 8 faces correctly localized here, 7 are correctly classified in terms of sex.

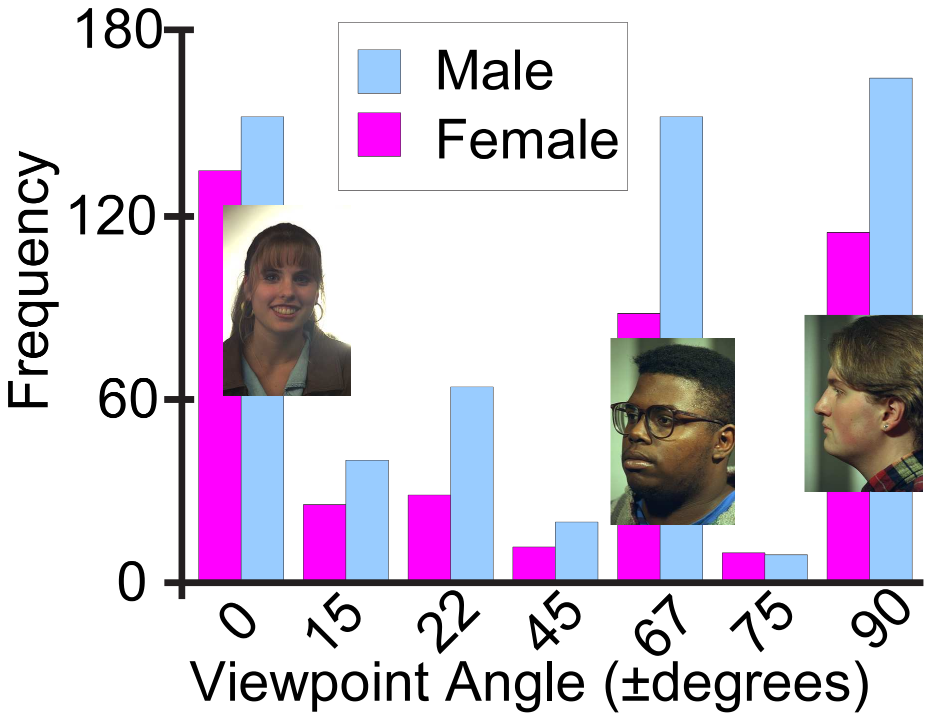


Fig. 4.12 Histograms illustrating the viewpoint distributions for males and females for the 994 unique FERET subject randomly selected for experimentation. The distributions for males and females approximate their expected distributions given the 591:403 male:female ratio in the FERET data set.

image, model features associated with the instance are then used to determine sex using the Bayesian classifier in equation (3.13). As faces are either male or female, determining face sex is a two-class problem and thus $\psi(male) = \psi(female)^{-1}$. A single threshold ψ^* on $\psi(c)$ can be used such that faces are classified as either male if $log\ \psi(male) > \psi^*$ or as female if $log\ \psi(male) < \psi^*$. As likelihoods and prior trait ratios are corrected for training set bias in the training process, all classification results are based on a threshold of $log\ \psi^* = 0$.

4.2.2 Identifying Visual Cues of Sex

Humans are generally capable of determining visual traits such as the sex of a face image with reasonable certainty. What is more difficult is to identify the visual cues that are operative in making the determination. Most faces contain a variety of cues that could be construed as either male or female, and it is their ensemble which determines the final decision. The local feature-based approach provides insight in terms of what local image cues are most important in determining visual traits, insight which is not possible from other representations based on global features or templates. By sorting features according to their sex likelihood ratios, the image regions most telling regarding the trait of sex can be visualized as in Figure 4.13. Note that visual cues linked to sex are generally found in all viewpoints. Ear features are often indicative of males, as they are less visible due to generally longer female hair. Several features around the mouth are indicative of males, indicative of beards or facial stubble. Females are distinguished by features arising from hairlines, eyes (possible from makeup) and lips. In contrast, certain model features arising from nostrils or cheeks, although very common in the class of face images, were generally less informative regarding sex. Note that although the male:female ratio in training data was 6:4, approximately twice as many sex-related features were identified for males as for females, suggesting a greater number of visual cues characteristic of the male sex.

Figure 4.14 illustrates instances of three different model features identified in different faces. Note that certain features, such as those in Figure 4.14 a) and b), persist over a range viewpoint variation. Other features, such as the nasal feature in Figure 4.14 c), are specific to fixed views. Figure 4.15 illustrates how individual faces generally contain features indicative both genders. As such, faces and facial regions can be quantified in terms of degrees of masculinity and femininity.

4.2.3 Classifying Sex from Arbitrary Viewpoints

In order to evaluate sex classification from arbitrary viewpoints, 15 different trials of training, localization and classification are performed. 5 training set sizes of 100, 200, 300, 400 and 500 face images are used, and for each size, 3 training sets are randomly selected from the 994 images. In this way, both cross validation and training efficacy can be investigated. Figure 4.16 illustrates the classification error as a function of training set size. As expected, classification error decreases with an increase in the size of the training data set, from 27% error based on 100 training faces to 15.9% error based on 500 training faces. This is primarily due to the emergence of more rare, yet sex-informative features.

Examples of correctly localized and classified faces, including localized OCIs and detected features, are illustrated in Figure 4.17. Note that the OCIs are correctly localized

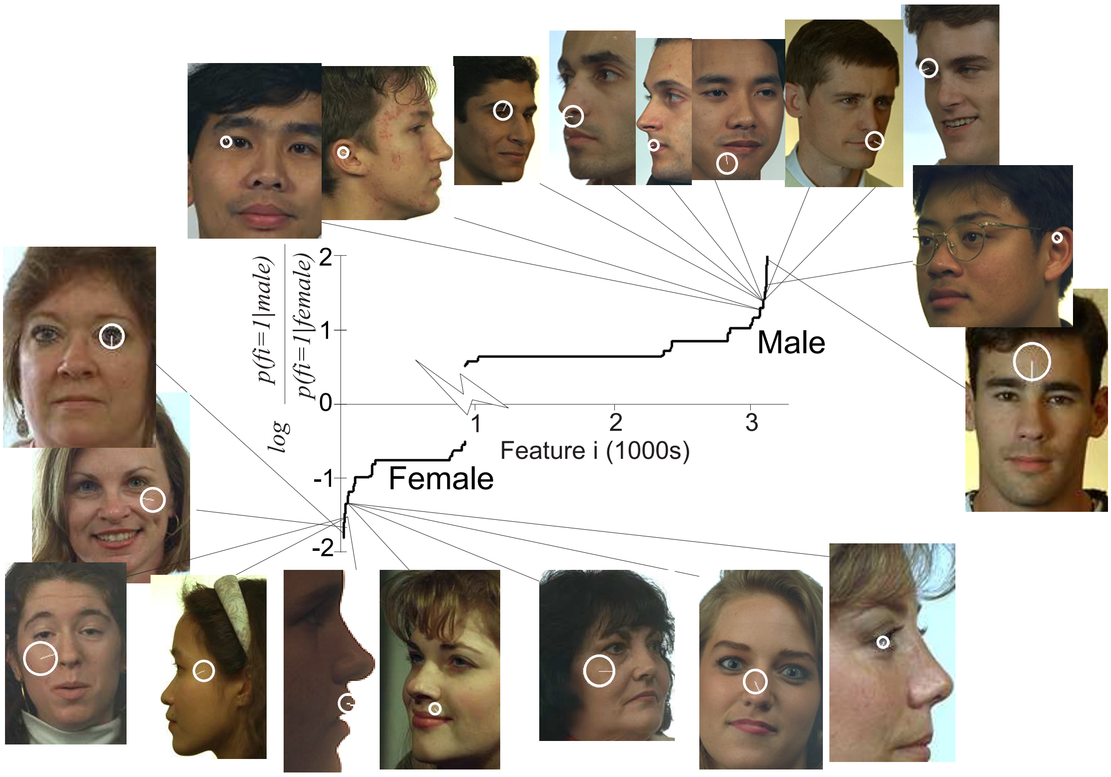
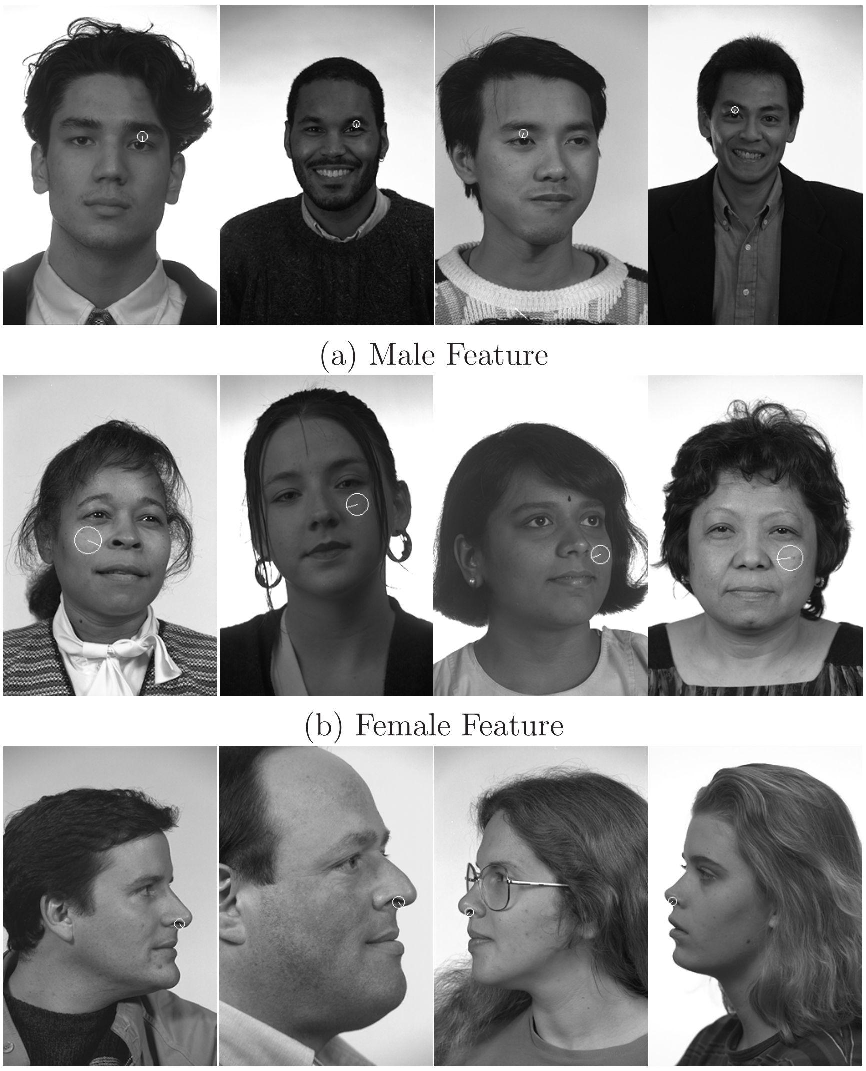


Fig. 4.13 Visual cues indicative of face sex, in the form of scale-invariant features. Features are sorted in increasing order of their log likelihood ratio $ln\frac{p(f_i=1|male)}{p(f_i=1|female)}$. Of approximately 15,000 features in a viewpoint-invariant face model learned from 500 randomly selected FERET images, approximately 3000 features bear information regarding sex (i.e. $|ln\frac{p(f_i=1|male)}{p(f_i=1|female)}| > 0.5$). Features to the left and below occur more frequently in female subjects and those to the right and above more frequently in male subjects. Face images shown illustrate instances of sex-informative features (white circles) with absolute log likelihood ratios ranging from 1.3 to 2.0. Although the male:female ratio in the training data was 6:4, approximately twice as many sex-reflective features are associated with males.

near their expected locations along the nose. Note also that faces generate contain features indicative of both male and female sex.

Examples of misclassified images can be see in Figure 4.18. Note that the faces in Figure 4.18 B) and C) appear relatively ambiguous with respect to sex. For trials based on 500 training images, approximately 3.6% of error cases are due to poor model localization, where the discrepancy between the localized and labeled OCIs was greater than a threshold in scale, orientation and location of $log(1.5)$, 20 degrees and OCI scale/2 pixels, respectively.



(a) Male Feature

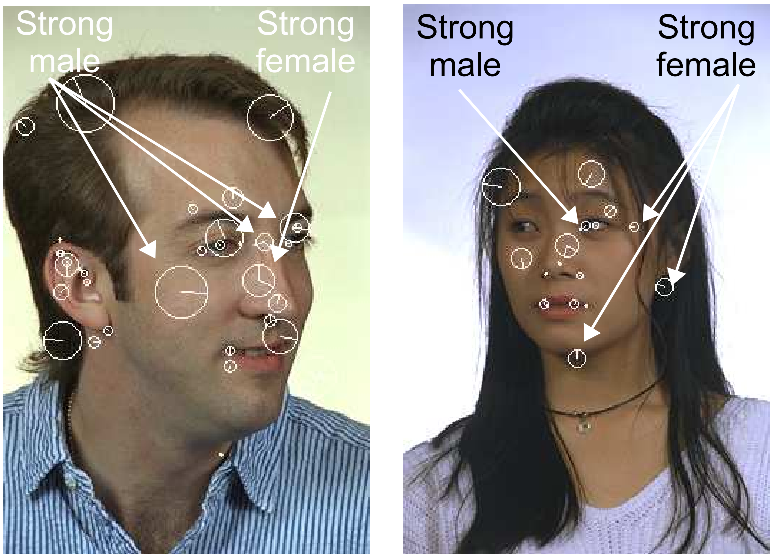
(b) Female Feature

(c) Neutral Feature

Fig. 4.14 Different instances of the same local feature bearing sex information (white circles). Note that the training data set has a male:female ratio of approximately 3:2. The eyelid feature shown in (a) occurs in 15 males and 1 female, and is indicative of male faces. The cheek feature shown in (b) occurs in 0 males and 8 females, and is indicative of female faces. The nasal feature shown in (c) occurs in 16 males and 10 females, and bears no significant information regarding sex.

Images with extreme background/foreground contrast is a primary cause of poor model localization, for example Figure 4.18 A), as few recognizable SIFT features can be extracted.

Figure 4.19 illustrates classification error as a function of *log* $\psi(c)$ for training set



(a) (b)

Fig. 4.15 Classification of the visual trait of sex from local features (white circles). A given face instance consists of a set of local features, a subset of which are reflective of either sex, and it is their ensemble which determines the final decision. To illustrate, a feature is described as strongly male or female if its likelihood ratio of co-occurring with the indicated sex in training images is greater than 2:1. Of the 63 model features detected in image (a), 15 are strongly male and 1 is strongly female, suggesting a male face. Of the 31 features detected in image (b), 7 are strongly female and 1 is strongly male, suggesting a female face. Many features, although very common in the class of face images, are uninformative regarding sex.

sizes of 100, 300 and 500 face images. In all cases, the minimum classification error is achieved near $log\ \psi^* = 0$ as predicted. The width of the trough surrounding classification error minima is related to the number of sex-informative features operative in classification. Note how the width increases with an increase in the training set size. The reduction in the minimum classification error is small going from 300 to 500 training images, suggesting that a further increase in the number training images may not significantly improve classification performance. Classification error asymptotes to the left and to the right of the graph at rates of 0.6 and 0.4, the respective error rates of a classifier guessing either female or male.

The correct classification rates for male and female faces vs. the classification threshold $log\ \psi(c)$ trials based on 500 training images is illustrated in Figure 4.20. Note that the threshold of $log\ \psi^* = 0$ results in a slightly better classification rate for male faces than

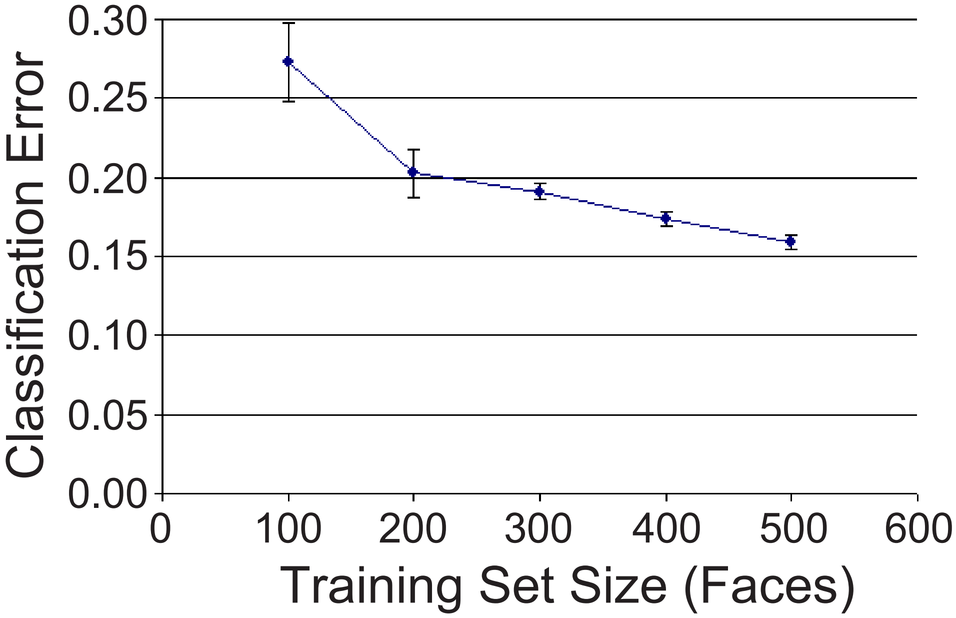


Fig. 4.16 Classification error as a function of the number of faces used in training. For each training set size, 3 different sets of training data are randomly selected, and classification is performed on the remaining face images not used in training. The points and error bars indicate the mean and the standard deviations of the classification error for the indicated training set size. Note that the both the classification error mean and standard deviation diminish with an increase in training set size. The minimum classification error achieved is 15.9% for a training set size of 500 images. Classification results shown reflect a threshold of $log\ \psi^* = 0$

for female faces. This is consistent with results in the literature reporting better classification on male faces [8], possibly due to the fact that in general, a greater number of sex-informative features occur in male faces. The point at which the classification error rate is equal for both male and female face images occurs at $log\ \psi(c) = -1.1 \pm 0.1$, slightly lower than the threshold of $log\ \psi^* = 0$. In practice, threshold $log\ \psi^*$ can be lowered or raised to bias classification error in favor of either males or females, while maintaining near minimal combined error.

Figure 4.21 illustrates the distribution of sex classification over three ranges of viewpoint from frontal to profile views. Note that the error rate for profile views is almost double that of frontal views, suggesting that profile views are less informative regarding sex.

A question one might ask is whether the Bayesian classifier proposed for visual traits

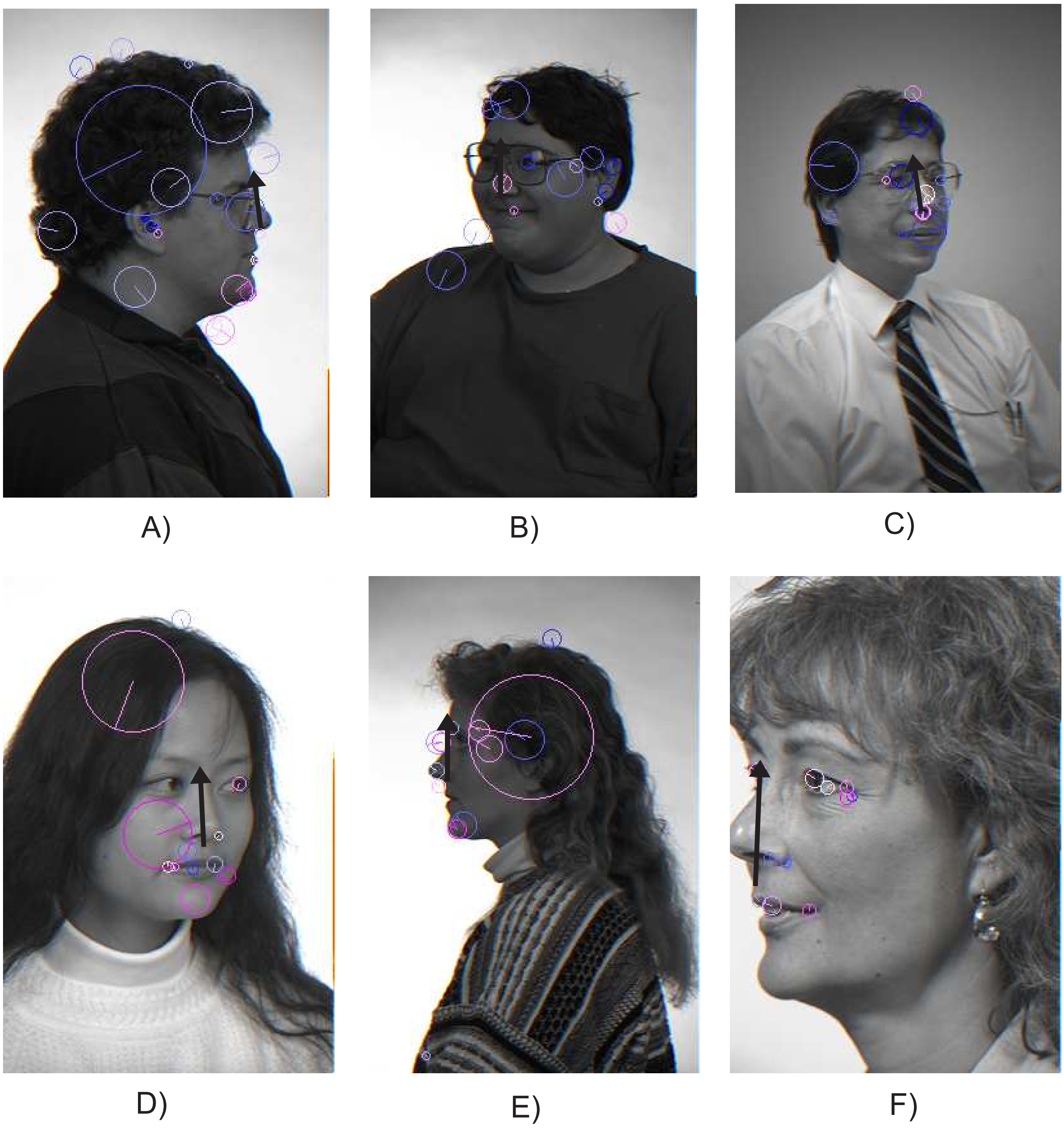


Fig. 4.17 Several examples of correctly classified faces in trials involving 500 training images (best viewed in color). The localized OCIs are indicated by black arrows along the nose. Features shown (circles) are those automatically detected and used for classification. Blue features indicate male characteristics and pink features indicate female characteristics, where the color saturation is proportional to the magnitude of the log likelihood ratio. Images (a)-(c) are correctly classified as male, and Images (d)-(f) are correctly classified as female.

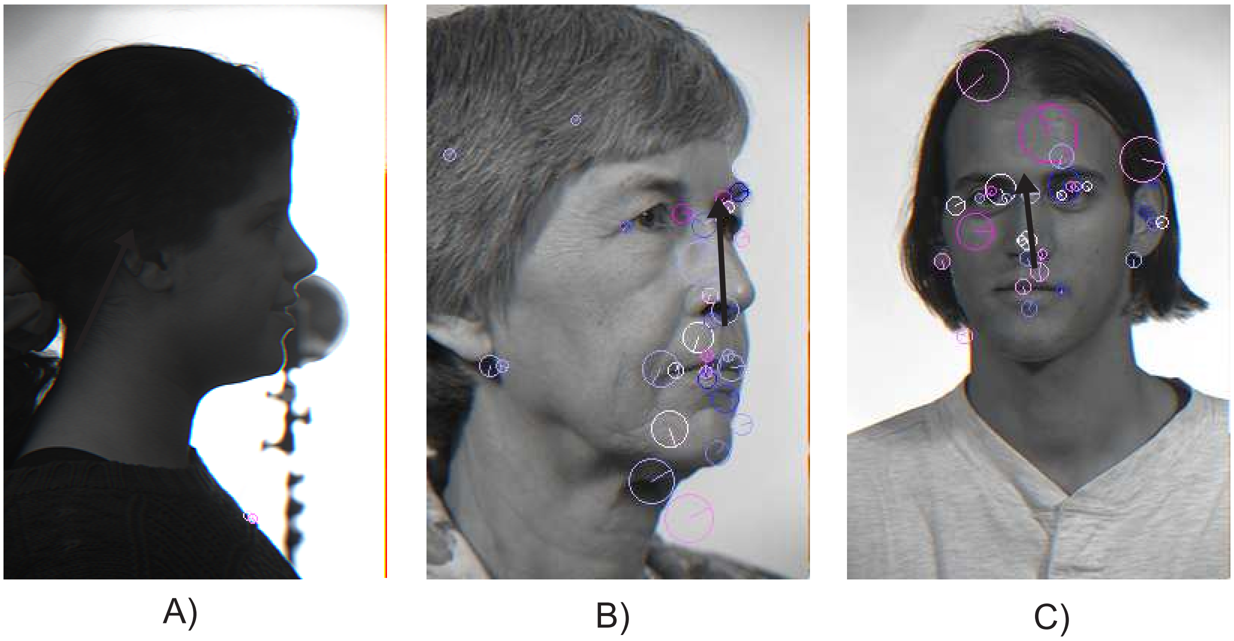


Fig. 4.18 Several examples of sex misclassification in trials involving 500 training images (best viewed in color). The localized OCIs are indicated by black arrows along the nose. Features shown (circles) are those automatically detected and used for classification. classification. Blue features indicate male characteristics and pink features indicate of female characteristics, where the color saturation is proportional to the magnitude of the log likelihood ratio. Image (a) is misclassified as male due to model localization failure. Here, strong backlighting results in poor image contrast in the face, and thus few scale-invariant features. Image (b) is a female face misclassified as male, due to an excess of strong male characteristics. Image (c) is a male face misclassified as female, due to an excess of female features.

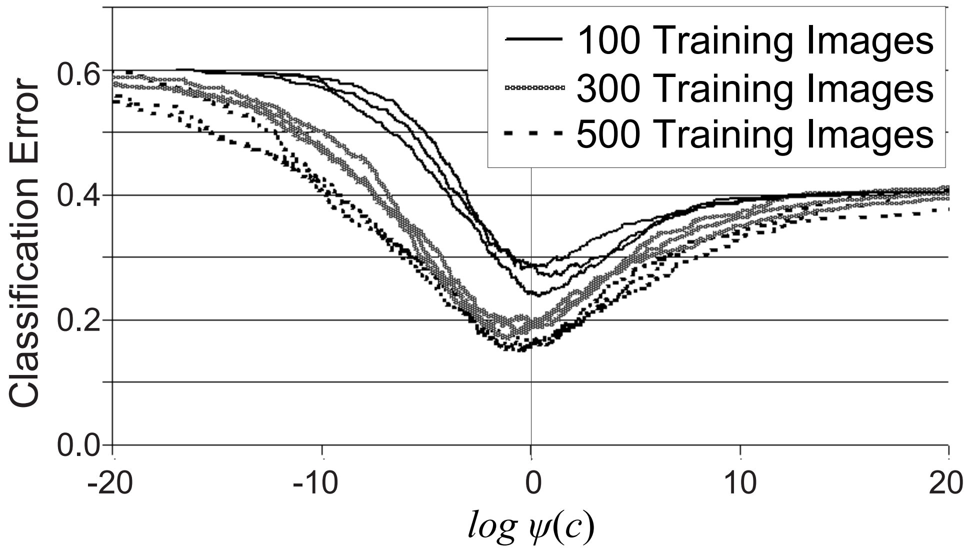


Fig. 4.19 Classification error as a function of the classification threshold $log\ \psi(c)$ for trials based on randomly selected training sets of 100, 300 and 500 face images. The three curves shown for each training set size correspond to the three randomly selected data sets. Note how the reduction in classification error going from 300 to 500 training face is marginal compared to going from 100 to 300 training faces, indicating that the minimum possible error is likely achieved somewhere near 500 training faces. The minimum classification error in all cases is achieved near $log\ \psi(c) = 0$. The classification error asymptotes to the right and to the left of the graph at rates of approximately 0.4 and 0.6, the approximate ratios of female and male images in the data.

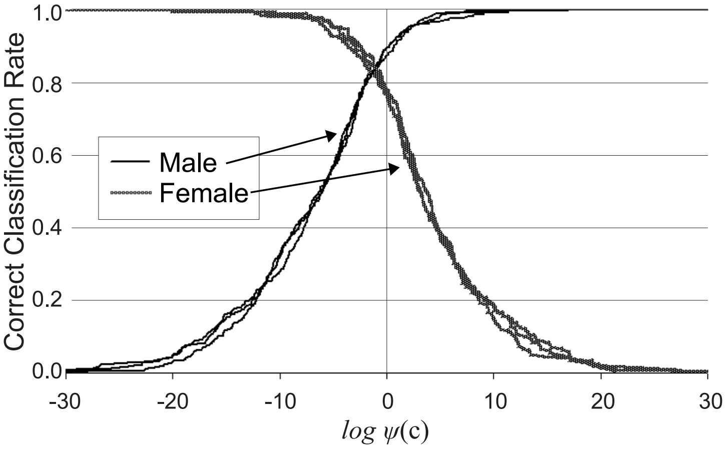


Fig. 4.20 Correct classification rates for males and females for randomly selected training sets of 500 face images vs. classification threshold $log\ \psi(c)$. Note that thresholding classification at $log\ \psi(c) = 0$ results in a slight bias in favor of correctly classifying male faces.

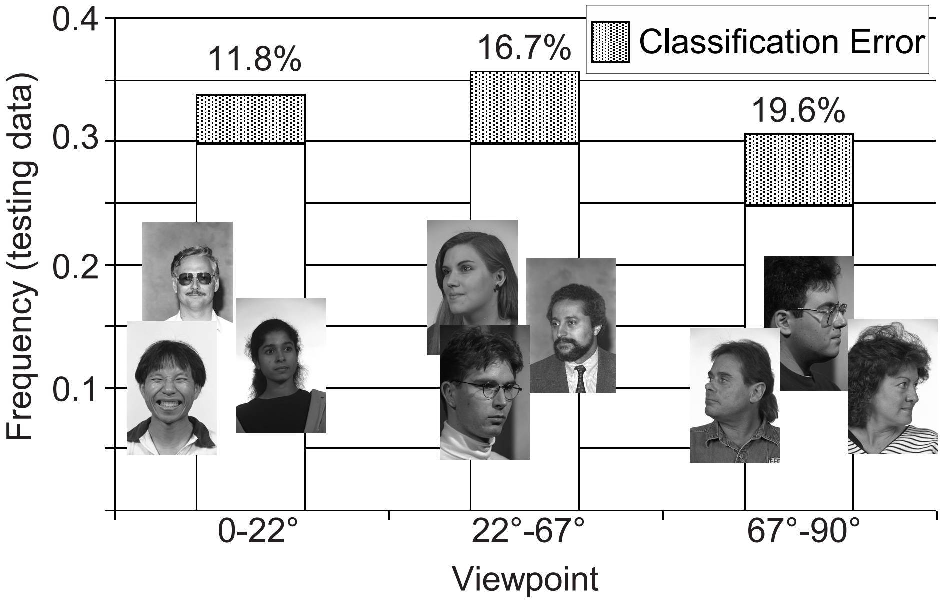


Fig. 4.21 Classification error as a function of viewpoint based on randomly selected training sets of 500 face images. Face viewpoint is separated into three ranges from frontal to profile. The images shown illustrate examples of faces in the indicated viewpoint range. The height of the columns is proportional to the frequency of the corresponding viewpoint in the testing data. The shaded regions illustrate the relative proportion of misclassified data. The percentages above each column are the classification error rates for the associated viewpoint range.

in this thesis is optimal. Treating classification as a black box task, it is possible that other classification techniques could result in superior performance. Here, this question is investigated by performing sex classification using the popular SVM classifier [196], based on precisely the same feature data used in the three classification trials involving sets of 500 training and 494 testing images.

Briefly, SVM classification is based on identifying hyperplanes which maximally separate feature data arising from different classes. The input to the SVM classifier is equivalent to that of the Bayesian classifier, i.e. a set of positive model feature occurrences for each face image to be classified. The SVM implementation used is the open source LIBSVM package [197]. The SVM is defined by a kernel function, the radial basis function (RBF) kernel [8] and the linear kernel are popular, general choices. The best SVM classification results obtained here involve the RBF kernel, and these are reported. The RBF kernel is defined by two free parameters, C and γ, where C is related to the classification error margin and γ to the width of the RBF kernel. A search over these parameters is performed in order to determine the parameter combination which results in the best SVM sex classification performance.

The results are shown in Table 4.1, demonstrating that Bayesian classification generally outperforms that of the SVM in terms of combined error. Note that for the results shown here, the classification error of female faces is lower for the SVM than for the Bayesian classifier. This does not reflect superior classification, however, as SVM parameters are tuned for maximum combined error.

Classifier	Combined error	Male error	Female Error
SVM	$18.1 \pm 1.5\%$	$15.7 \pm 2.8\%$	$21.3 \pm 6.2\%$
Bayesian	$15.9 \pm 0.4\%$	$11.0 \pm 1.2\%$	$23.0 \pm 1.2\%$

Table 4.1 Error rates for sex classification over viewpoint for SVM and Bayesian classification. The average classification error rate and the standard deviation are shown. Bayesian classification outperforms SVM classification in terms of combined classification error.

Interestingly, while Bayesian classification outperforms SVMs for determining sex from face images, the opposite has been found for classifying images according to distinctly different scene categories such as faces, buildings and cars [127]. One hypothesis for this difference is as follows: SVMs generally exploit dependencies or co-occurrences between image features when determining hyperplanes separating classes. Such inter-feature dependencies are likely more prominent in distinctly different scene categories than in face images of different sexes, where the majority of features occur frequently and relatively independently in both sexes. A Bayesian classifier based on the assumption of conditionally independent features avoids making the assumption of feature inter-dependencies, and therefore results in improved performance.

4.2.4 Classifying Sex from Occluded Faces

Classifying object class instances according to visual traits in arbitrary scenes is complicated by occlusion, as features useful for classification may not be visible. In the case of face images, occlusion can arise from a number of factors, such as sunglasses, hats, hairstyles, scarves, crowds. The effect of occlusion has not been previously investigated in the context of face classification, as most previous work has assumed that facial features required for classification are visible and precisely localized. Local feature-based classification is capable of coping with a significant degree of occlusion, as only a subset of features are required.

Occlusion is tested by artificially obscuring each FERET testing image with a black occluding circle, and then performing classification trials using the three classifiers trained on 500 images described in the previous section. The black circle was placed in the center of the images, thereby obscuring a variety of different facial regions in different images, as faces are approximately (but not precisely) centered in the FERET dataset. The degree of occlusion was varied by incrementing the radius of the occluding circle from 0 to 80 pixels. The occluding circle border was blended smoothly into the face images using a small Gaussian kernel (with a standard deviation of 2 pixels), to simulate a more natural occluding contour. Examples of occluding circles in various face images are shown in Figure 4.22.

Figure 4.23 illustrates classification error as a function of occluding circle radius. Note that classification performance degrades gracefully with an increase in the occluding circle radius, as sex-informative features can still be extracted and used for classification in non-occluded regions of the face. Even at for relatively large occluding circles of radius 40 pixels, classification error is approximately 25%. At an occluding radius of 80 pixels, classification error reaches approximately 0.4, the rate of female faces in the data set. Note that the standard deviation of classification error does not generally change significantly with the degree of occlusion, indicating that error variability is generally determined by the amount

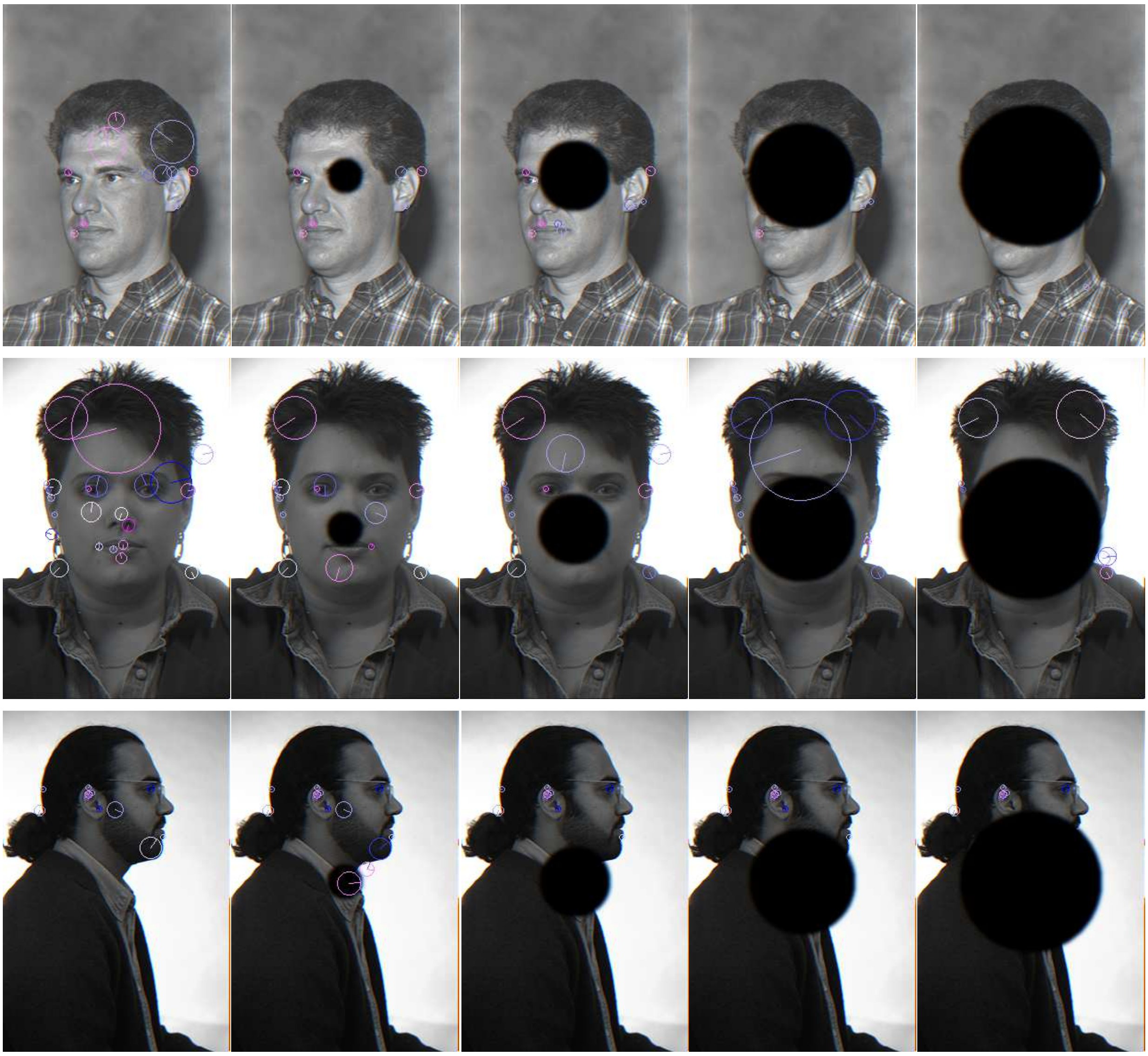


Fig. 4.22 Examples of images used in occluded face classification (best viewed in color). Occlusion is simulated by a black occluding circle centered in the image, with radii of 0, 20, 40, 60 and 80 pixels from left to right. Features shown are those automatically detected and used for classification. Features indicative of male characteristics are blue, and those indicative of female characteristics are pink, where the color saturation is proportional to the magnitude of the log likelihood ratio.

of training data and not the number of features identified in the image. One exceptionally low classification error for an occluding radius of 40 pixels resulted in higher standard deviation in classification error.

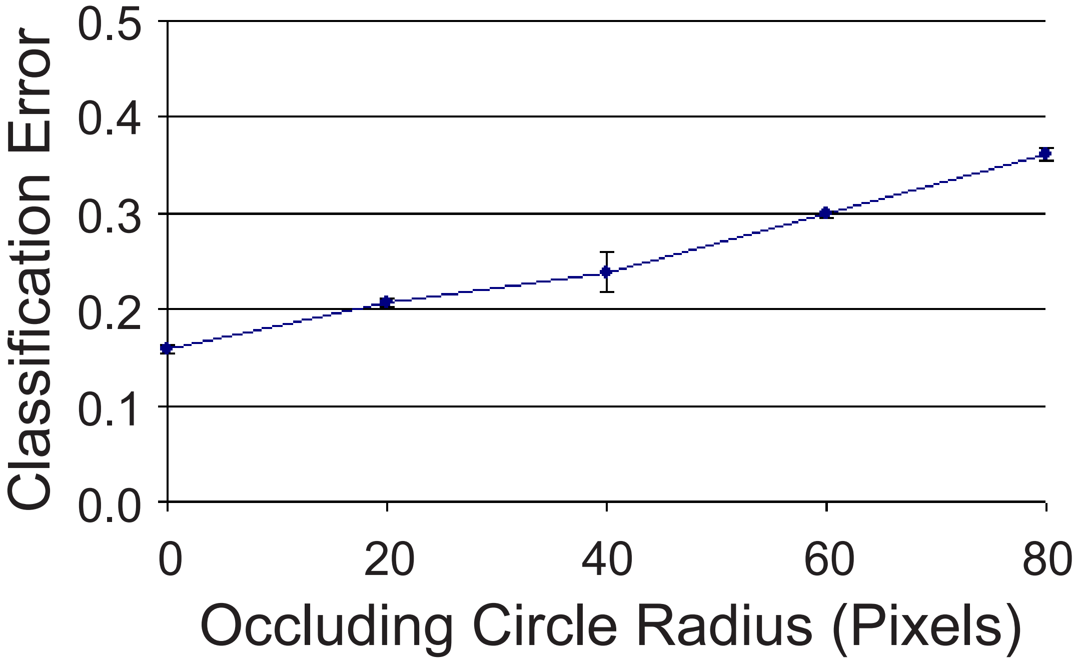


Fig. 4.23 Classification error as a function of the degree of occlusion in the image. For each occluding circle radius, 3 different classification trials are performed using the three classifiers in the previous section trained on 500 images, and tested on occluded training images not used in training. The points and error bars indicate the mean and the standard deviations of the classification error for the indicated radius. Classification error starts at 15.9% with no occlusion and rises to approximately 36.2% for an occlusion radius of 80 pixels.

Figure 4.24 illustrates classification error as a function of the classification threshold *log* ψ^* for occluding circles radius values of 0, 40 and 80 pixels, for classifiers based on 500 training faces. It can be seen that minimum classification error increases gracefully as the occluding circle increases in radius. Again, the width of the trough surrounding classification error minima reflects the number of sex-informative features operative in classification. The shrinking width associated with increased occlusion indicates that fewer and fewer features are available for classification. Additionally, as the number of face features decreases, the performance of face instance localization degrades, and at worst latches onto random noise when no face features remain. In such circumstances, classification error at $log\ \psi^* = 0$ approaches a rate of 0.4, the error rate of a classifier which simply guesses that each face is male.

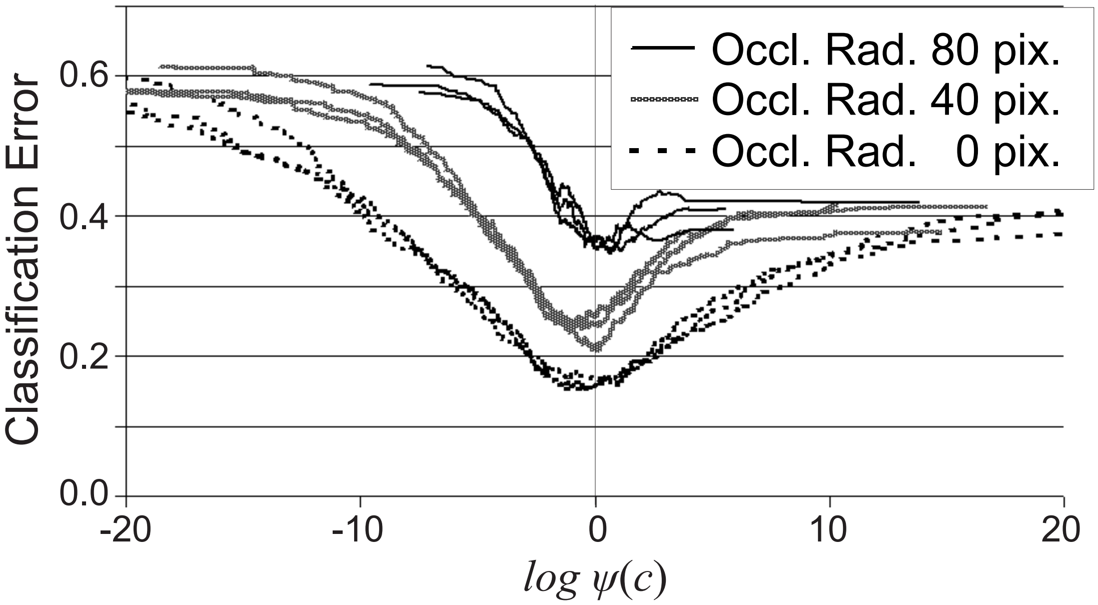


Fig. 4.24 Classification error as a function of classification threshold $log\ \psi(c)$ for 3 trials based on 500 randomly selected training faces, for occluding circle radii of 0, 40 and 80 pixels. Once again, minimum classification error occurs near a threshold value of $log\ \psi^* = 0$. Classification performance degrades gracefully as the degree of occlusion increases. The magnitude of $log\ \psi(c)$ decreases as occlusion increases, as fewer and fewer features are observable.

4.2.5 Classifying Sex from Frontal Views

In order to compare the classification results to those present in the literature, training and testing was also performed on a restricted set of frontal faces. Using the 925 standard FERET frontal images labeled "*_fa.*", a model was trained from a randomly selected subset of 500 images. Table 4.2 lists error rates reported in the literature for the combined task of detection, localization and classification. For the more difficult combined task, the OCI model-based Bayesian sex classifier in Table 4.2 (a) achieves an error rate of 11.2% for training based on 500 randomly selected FERET subjects. This is lower than the rate 21% achieved by Shakhnarovich et al. [162] in Table 4.2 (b) which makes use of a boosted decision tree of Haar wavelet features for both detection and classification. However, experimentation for this approach is based on proprietary training and testing databases however, and an exact comparison is not possible.

Error rates for classification only (i.e. faces pre-aligned and/or cropped prior to classification) represent an ideal-case baseline and are included for completeness in Table 4.2(c). While the OCI model-based classification error rate for frontal faces is higher than error rates reported in the literature for these approaches, there are important additional factors that must be considered. Approaches based on global image features and non-occluded, frontal faces would likely incur significantly higher error rates in the presence of partial occlusions due to scarves, sunglasses, or changes in viewpoint. As shown in [5], even minor errors in face localization possibly to such factors can significantly increase classification error.

Task	Method	Error Rate
Combined detection, localization and classification	(a) OCI model	11.2%
	(b) Shakhnarovich et al. [162]	21%
Classification only	(c) Various [5, 6, 7, 8, 9]	4%-10%

Table 4.2 Published error rates for combined detection, localization and classification for frontal faces. The Bayesian classifier trained on 500 faces in (a) achieves an error rate of 11.2%, lower than error rate of 21% for the approach of Shakhnarovich et al. [162] in (b) trained on approximately 3000 faces. The result for (b) is based on a proprietary database, however, so a precise comparison cannot be made. Results for classification only reported in the literature (c), i.e. frontal, upright, pre-aligned and cropped faces, represent an ideal-case baseline and are included for completeness.

Note that for the local OCI model-based approach, the error rate for frontal face classification is significantly lower than for classification from arbitrary viewpoints, reflecting the difficulty of coping with viewpoint. Note, however, that the error rate for 500 frontal images is similar to the error rate of 11.8% obtained for the frontal viewpoint range in Figure 4.21.

4.3 Detection and Localization of 3D Motorcycles

This section investigates detection and localization based on the object class of 3D motorcycles, in order to demonstrate the generality of the viewpoint-invariant OCI model to 3D object classes other than faces. Additionally, experimentation makes use of a spherical OCI which extends viewpoint invariance to an entire view sphere around the object class of interest.

Data: In order to compare with results in the literature, experimentation is performed using the training and testing motorcycle images from the PASCAL 2006 [39] data set. The training data set consists of 235 training images containing 275 positive motorcycle examples. The testing set consists of 2,686 images containing 274 examples of motorcycles.

Learning: In the previous sections the assumption is made that 3D faces are typically viewed from a coronal plane about the face, and thus an OCI in the form of a line segment is sufficient for modeling. However, in the PASCAL data set motorcycles are generally viewed from a wider range of viewpoints, including frontal, lateral and overhead views. As mentioned in Section 3.2, OCI modeling can be extended to an entire view sphere around the object class by adopting an OCI in the form of a 3D sphere. In this scenario, the OCI consists of a centroid and radius whose projection remains consistent with the location and size of the object class within the image plane, as shown in Figure 4.25. Manually labeling requires roughly estimating the center and radius of the spherical OCIs in images of motorcycles acquired from arbitrary viewpoints. Note that labeling need not be precise, as the OCI model learning procedure is robust enough to cope with a degree of labeling imprecision. Model learning is performed according to the learning algorithm described in Section 3.1.2. Note that evaluating geometrical consistency between spherical OCIs for the purpose of model learning and detection is equivalent to that of linear OCIs, except that the orientation component T^{θ} is unused.

Detection: Detection is performed using the model fitting technique described in Section 3.1.3. Note here that the criterion used to declare an OCI model detection/localization as correct is the not 50% bounding box overlap proposed in [39], but rather the geometrical consistency threshold $Thres^{g}$ between detected and ground truth OCIs. However, given the threshold used here ($T^{x} = 0.5/\sigma$, $T^{\sigma} = log(1.5)$), results should be comparable.

Several examples of motorcycles correctly detected and localized are illustrated in Figure 4.26. Note the high degree of clutter and variation in viewpoint. Several examples of

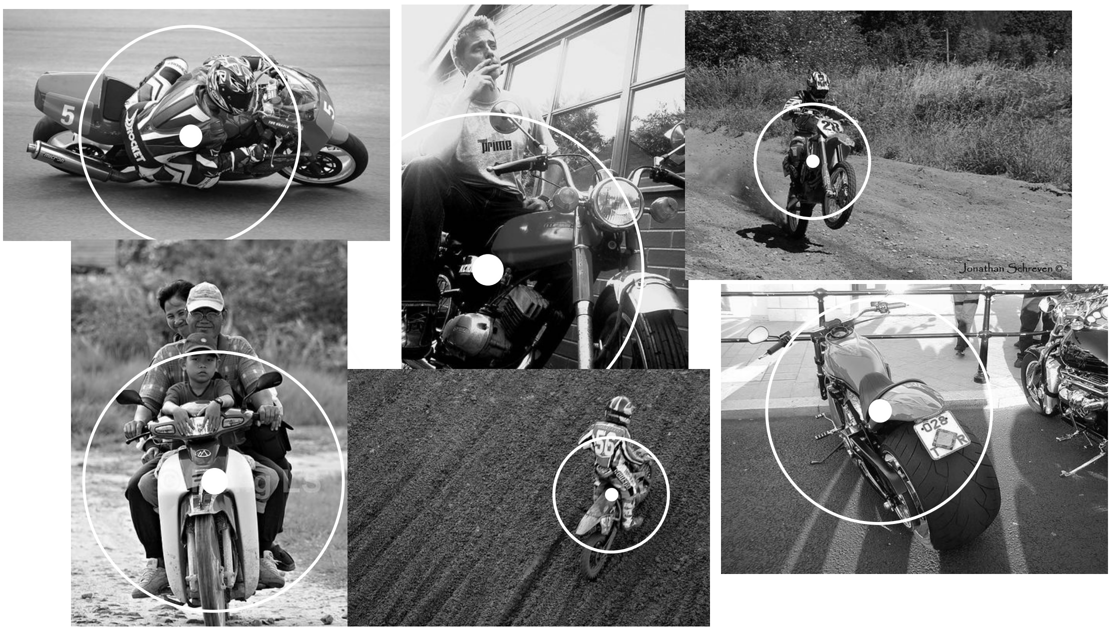


Fig. 4.25 Examples of motorcycle training images from the PASCAL 2006 dataset [39] and spherical OCI labels (circles). The OCI here is a 3D sphere centered on the motorcycle, and serves as a geometrical reference frame for modeling motorcycle appearance over an entire view sphere. Note the variability in the appearance, shape and viewpoint of motorcycles.

incorrectly detected or localized motorcycles are illustrated in Figure 4.27. False detections can arise from background clutter (lower left image) as well as motorcycle-like objects such as the bicycle (upper left image). Missed detections typically occur as the result of occlusion (upper right image) or low resolution images leading to few SIFT features (lower right image).

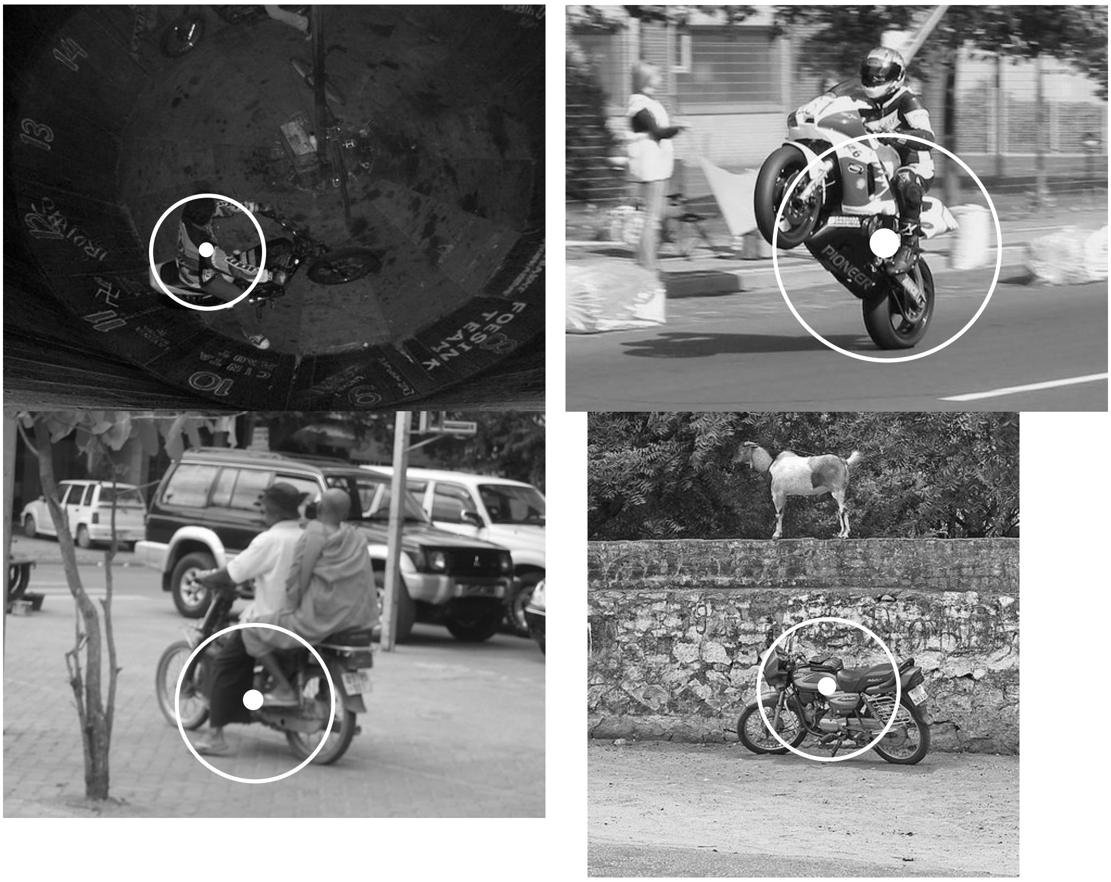


Fig. 4.26 Examples of correct motorcycle detection and localization. Note the degree of clutter and variability of viewpoint.

Detection results for the PASCAL 2006 dataset [39] have been reported for five other methods. The method of Dalal et al. [86] uses a sliding window detection approach, based on histograms of oriented gradient features specially constructed for detection, in combination with SVMs. Laptev et al. [198] also use sliding windows with histograms of oriented gradient features, with AdaBoost classification, and combine side and frontal views into a single detector. Viitaniemi et al. [199] use a tree-structured self-organizing map, based

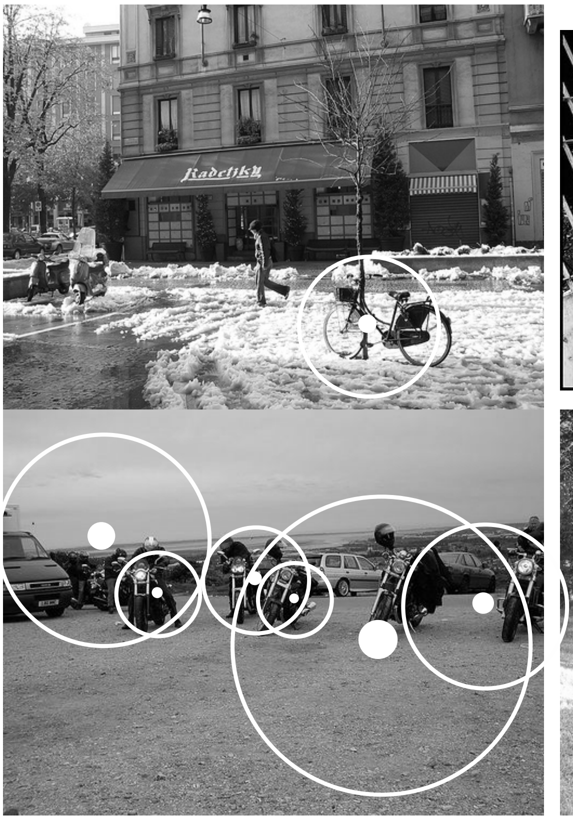


Fig. 4.27 Examples of difficult motorcycle detection and localization. The upper left image illustrates one false positive detection as a bicycle is mistaken for a motorcycle, and two missed detections as the motorcycles in the background are undetected. The upper right image illustrates a missed detection, as the majority of the motorcycle is occluded. The lower left image illustrates a cluttered scene with several spurious false positive solutions along with two correct detections. The lower right image illustrates one correct detection and one missed detection, as the small size of the motorcycle in the image results in fewer SIFT features.

on many different types of content descriptors used in the MPEG7 video content-based retrieval literature [200]. Shotton et al. [201] use a collection of 10 boosted classifiers, based on texture features. The approach of Fritz et al. [202] is based on the implicit shape model [122] using Hessian-Laplace scale-invariant feature detection with shape context descriptors [81]. Other results for motorcycle detection have been reported on same set of testing data, but are not directly comparable as they are based on different sets of training images [4, 155].

The precision-recall characteristics of detection methods trained and tested on the PASCAL motorcycle dataset, including that of the OCI model, are illustrated in Figure 4.3, and Table 4.3 lists average precision values. The OCI model results in an average precision of 0.159, which is lower than four other methods. It is important to note, however, that the average precision does not necessarily reflect the effectiveness of a given modeling technique per se. The choice of image features used by the model can have a significant impact on detection performance [123], as can algorithm-independent feature selection techniques such as boosting [128] or bootstrapping [131]. These techniques are not extensively investigated here for the purpose of OCI model-based motorcycle detection, and it is conceivable that the OCI model precision-recall characteristic could be further improved. For example, note that in Figure 4.3, the OCI model recalls approximately 80% of all true positives, while other methods recall only approximately 60%. The geometrical constraints used in localizing object instances and in pruning multiple detection hypotheses could be modified in order to improve precision-recall. For example, detection could be constrained to consider only upright motorcycles, a technique which is implicitly applied in many approaches. As most motorcycles in the PASCAL database are upright, the true positive rate would not change significantly, however many false positives would discarded, those corresponding to upside down motorcycles for instance. Applying such a constraint would reduce the generality of the OCI model, however, which is generally capable of detecting motorcycles at arbitrary image orientations. These aspects are avenues left for future investigation.

A final note is that while sliding window approaches result in the best average precision here, they have the drawback that they are not invariant to in-plane image rotations. Given that the majority of motorcycles are upright in the PASCAL 2006 images, this may not heavily impact detection performance, but it does reduce detection generality. An explicit search can be made by performing detection over a range of orientations, however this incurs a significant computational cost and will necessarily result in a higher false positive

Method	Average Precision
Dalal et al. [86]	0.390
Laptev et al. [198]	0.318
Viitaniemi et al. [199]	0.265
Shotton et al. [201]	0.175
OCI Model	0.159
Fritz et al. [202]	0.153

Table 4.3 Average precision values for detection and localization of motorcycles from the PASCAL 2006 Visual Object Classes Challenge [39].

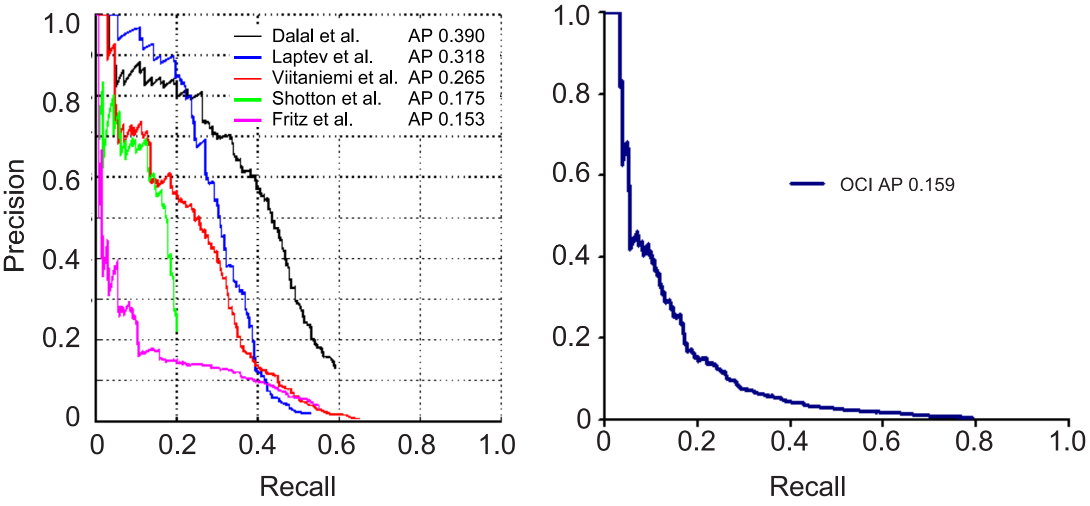


Fig. 4.28 The precision-recall characteristics of motorcycle detection and localization. The graph to the left illustrates precision-recall characteristics of the five other methods reporting detection results on the PASCAL 2006 motorcycles dataset [39]. The graph to the right illustrates the precision-recall characteristic of the OCI model.

rate and therefore lower average precision. Models that account for in-plane orientation such as the OCI model and the implicit shape model result lower average precision here, but do not have these drawbacks.

4.4 Discussion

This chapter described experimentation relating to OCI modeling of 3D object classes appearance in 2D projective imagery, based on the classes of human faces and motorcycles.

Experimentation was divided into three main sections, the first to demonstrate the feasibility of learning and detecting faces in natural, cluttered images acquired from arbitrary viewpoints, the second to evaluate the entire system in the context of detection, localization and sex classification of faces, and the third to compare detection and localization of motorcycles with other modeling approaches in the literature.

Section 4.1.1 outlined experiments demonstrating the feasibility of learning a viewpoint-invariant model from cluttered images acquired from arbitrary viewpoints. The detection performance can be considered good given the degree of variability of the faces and the relatively low number of images used in training (179 faces). As mentioned, face detection models aiming for optimal detection performance typically require tens of thousands of face examples taken in controlled settings, i.e. cropped, no clutter and organized according to viewpoint. It is unlikely that such models can be learned from natural training data with little or no manual supervision.

Section 4.1.2 described experiments demonstrating that the viewpoint-invariant representation for relating image features to a single viewpoint-invariant reference frame outperforms the multi-view representation, in terms of the average precision measure of detection performance. As predicted, the multi-view representation produces a higher number of strong, false detection solutions. It is possible that strategies could be employed to reduce these false positives in a multi-view setting, as is commonly done. However, the primary difficulty with the multi-view model would still be present, i.e. the same image feature being related to multiple, different views.

Section 4.1.3 details experiments demonstrating that an optimal OCI definition can be learned from data in a data-driven manner. Algorithm convergence was achieved experimentally after approximately 30 iterations of OCI label re-learning. An analytical proof guaranteeing convergence is left for future work. Detection performance improves using the optimal re-estimated OCIs, both viewpoint-invariant and multi-view representations. The optimal reference frame has a geometrical interpretation which is consistent with respect 3D faces, that of a 3D line central to the head, thereby minimizing the error in predicting the reference frame geometry from image features, as predicted by theory. These results demonstrate that stable image interpretation consistent with 3D geometry can be automatically from 2D images, without explicit knowledge of viewpoint. This opens the door to fully unsupervised learning of 3D appearance models. The key to doing this would be in establishing initial reference frame labels sufficiently accurate to ensure convergence.

Section 4.2 described a detailed analysis of combined detection, localization and sex classification of 3D faces represent the first results of their kind in the literature. Individual image features are shown to correspond to visual cues of face sex from a wide range of viewpoints. The OCI model obtains a minimum sex classification rate of 15.9% from arbitrary viewpoints. Classification error is generally lower in frontal views than from side views. The Bayesian classification approach proposed in this thesis is shown to be superior to SVM classification, another popular technique which could potentially be used. Sex classification was also shown to be feasible under a significant range of simulated occlusion. Sex classification for the restricted case of non-occluded frontal faces was performed in order to compare with results in the literature. The OCI model obtains a frontal error rate of 11.2%, which is higher than rates of 4%-10% reported by other approaches on the same data. However, there is strong reason to believe that these approaches would not be viable either occluded faces or faces viewed from arbitrary viewpoints.

As the FERET database used in classification contains reasonably uncluttered faces, the performance of the detection and localization aspects of the combined task could not be fully appreciated. There is a need for a new public database combining faces in difficult, cluttered scenes with accurate trait labels. Qualitative results showed that the system was capable of detecting, localizing and classifying sex on examples from the CMU profile database, see Figure 4.1. The database was not selected in order to evaluate aspects such as visual trait classification, however. No ground truth labels are available, and the diversity of traits such as sex, ethnicity and age is limited (e.g. faces are largely male).

Section 4.3 described experimentation involving learning, detection and localization of motorcycles, based on the benchmark PASCAL 2006 training and testing data set [39]. While OCI face modeling utilized an linear OCI offering invariance to viewpoint change in a plane, motorcycle modeling employed a spherical OCI which offers invariance to viewpoint change over an entire view sphere. Results demonstrate that an OCI model of motorcycles can be learned from a difficult natural image set, and used to identify motorcycles in new imagery with precision-recall performance comparable to other approaches trained and tested on the same data.

There are a variety of future directions that could be taken for OCI modeling in computer vision. The OCI defines a principled means of defining correspondence between images of different 3D object class instances acquired from different viewpoints. Although the OCI model in this thesis proposed modeling appearance based on local scale-invariant features, the OCI could thus be used as the basis for viewpoint-invariant modeling based on other image features, such as Haar wavelets [129] or global principle components [28]. While the OCI model is shown to be effective at detecting and localizing faces in this section, detection performance could potentially be improved in a number of ways. For instance, algorithm-independent machine learning techniques such as boosting [128], bagging [130] or bootstrapping [131] could be applied, and combinations of different invariant features types could be used [123]. Such techniques could also potentially be used to improve the performance of OCI model-based trait classification, by potentially identifying more trait-informative features or feature combinations. Trait classification could be extended to modeling continuous trait variables such as age in terms of continuous-valued likelihoods.

4.5 Summary

This chapter presented experimentation applying OCI modeling theory to the field of computer vision, as the first integrated system for detecting, localizing and classifying visual traits of 3D object classes from arbitrary viewpoints. Experimentation showed that OCI model models of faces and motorcycles could be learned from sets of natural, cluttered images taken from arbitrary viewpoints and identified in new images. The viewpoint-invariant OCI representation was shown to outperform a multi-view representation in terms of detection performance. An iterative algorithm was shown to converge to an optimal OCI definition which improved detection performance. Experimental results were presented using the OCI model to detect, localized and classify faces in terms of sex from arbitrary viewpoints and in the presence of occlusion, the first time the visual trait of sex has been classified from arbitrary viewpoints.

Chapter 5

Experimentation in Medical Imaging: MR Brain Anatomy

This chapter presents experimentation applying the OCI model in the context of medical image analysis, as described in Section 3.3. The goal of experimentation is to examine how the OCI model can be used to describe the anatomy of the human brain, and used to answer anatomical questions such as: How does a subject relate to its population? Which anatomical structures are common in a population, which are rare and which are reflective of subject traits such as pathology, age or sex?

Experimentation in this chapter is based on T1-weighted MR brain images from the International Consortium for Brain Mapping 152 (ICBM) data set [40, 203], consisting of 152 volumes of normal subjects, 88 male and 66 female, aged 24.6 ± 4.8 years, of (x,y,z) resolution (181x217x181) voxels [1]. This image set is provided courtesy of the Montreal Neurological Institute. All images used are processed in 2D, as slices through the brain or as volume renderings of the cortical surface. As in the previous section, all images are first processed by automatic scale-invariant feature extraction using the scale-invariant feature transform (SIFT) technique [22], based on a publicly available implementation [193]. Although experimentation is based on 2D imagery, the OCI model is independent of image dimension and can be applied to 3D brain volumes or 4D time series data, this is discussed later.

The first set of experiments in Section 5.1 is designed to demonstrate the primary

[1] Informed consent was obtained by all human subjects used in the study, and the protocol was approved by the ethics board of the Montreal Neurological Institute and Hospital

medical imaging contribution of this thesis, the use of the OCI model as a parts-based anatomical description of anatomy as outlined in Section 3.3. There are three main components to this set of experiments. Section 5.1.1 details the resulting learned model of brain anatomy, showing the types of anatomical structures identified by the OCI learning algorithm and their frequencies within the population. Section 5.1.2 provides a quantitative evaluation of OCI model fitting to new images not used in model learning, showing that model features can be identified and localized in new images with accuracy similar to 3 human raters. Section 5.1.3 quantitatively compares the stability of OCI model to that of the global active appearance model (AAM) of Cootes and Taylor [29], a well-established approach in the literature. It is shown that in the OCI model fitting can be fit more robustly and with greater accuracy to both normal subject images and subject images containing artificial intensity perturbations reminiscent of pathology.

The rest of the experimentation in this chapter demonstrates the remaining medical imaging contributions of this thesis. Section 5.2 examines the use of the OCI model in determining anatomical structure characteristic of subject traits described in Section 3.3.2. It is shown that brain structures indicative of sex can be identified in a set of training images, and that new brain images not used in training can classified according to sex with an accuracy of 80%. Section 5.3 demonstrates the use of the OCI model in describing the highly variable cortical surface, using the lateral fissure as a local reference frame as described in Section 3.3.1. Validation of 10 frequently occurring model parts performed by an expert neuroanatomist indicates that in 77% the model parts correctly indicate the same underlying cortical structures. Section 5.4 demonstrates how the OCI model can be used as a basis for robust inter-subject registration. A discussion of results follows in Section 5.5.

5.1 Parts-based OCI Modeling of MR Brain Imagery

This section presents experimentation adapting the general OCI model to describing human brain anatomy in terms of a set of distinct, local parts. The goals of experimentation here are to investigate the result model learning, to evaluate the accuracy of OCI model-to-subject fitting, and to evaluate and compare the stability of OCI model-to-subject fitting to that of the AAM [29].

5.1.1 Learning an Anatomical Model

Data: Experimentation is based on 2D sagittal slices of T1-weighted MR brain images from the ICBM 152 data set [203]. Slices are taken at y=4 in Montreal Neurological Institute (MNI) stereotactic space coordinates [204], slightly off the sagittal midplane. This particular slice is chosen as it contains a large number of important, recognizable brain structures, including cortical regions, the corpus callosum and basal ganglial structures, and is often used in the study of the human brain.

Learning: Learning involves automatically identifying a set of informative model parts $\mathbf{m}$ and estimating the parameters of their appearance, geometry and occurrence frequency distributions, based on a set of subject images. Prior to learning, each training image is processed by automatically labeling an OCI reference frame o^g in the form of the standard Talairach AC-PC line, as described in Section 3.3. Note that other definitions could be used depending on the image context, the sole constraint being that the reference frame represent a stable scale-invariant structure shared by all subjects being modeled. Labeling can be done manually by defining a single line segment corresponding to o^g in each subject image, or in an approximate manner via linear registration of MR volumes into the same stereotactic space. The latter approach is adopted here, using MR volumes preregistered into the Montreal Neurological Institute (MNI) stereotactic space [204], and thus no manual intervention is necessary.

The result of learning is a set of spatially localized parts, including their occurrence frequency, geometrical variability and appearance variability. This set serves a natural and intuitive representation for describing anatomical variability. Figure 5.1 illustrates a graph of the likelihood of model part occurrence in the population, i.e. $p(m_i^{b=1}|o^{b=1})$, sorted in order of descending likelihood. Note that a small number of model parts occur in a relatively high percentage of subjects, indicating structure that is stable and repeatably identifiable in many members of the population. Conversely, a relatively large number of model parts are related to subject-specific characteristics or noise, occurring in only a small number of subjects. While such rare parts are of limited use in describing the anatomy of the population, they may be of interest in linking subtle anatomical features common to small groups of subjects sharing similar anatomical traits.

Note that in general, no model parts are detected in all 102 subject brains. This is not to say that common anatomical structures such as the pons or corpus callosum are not

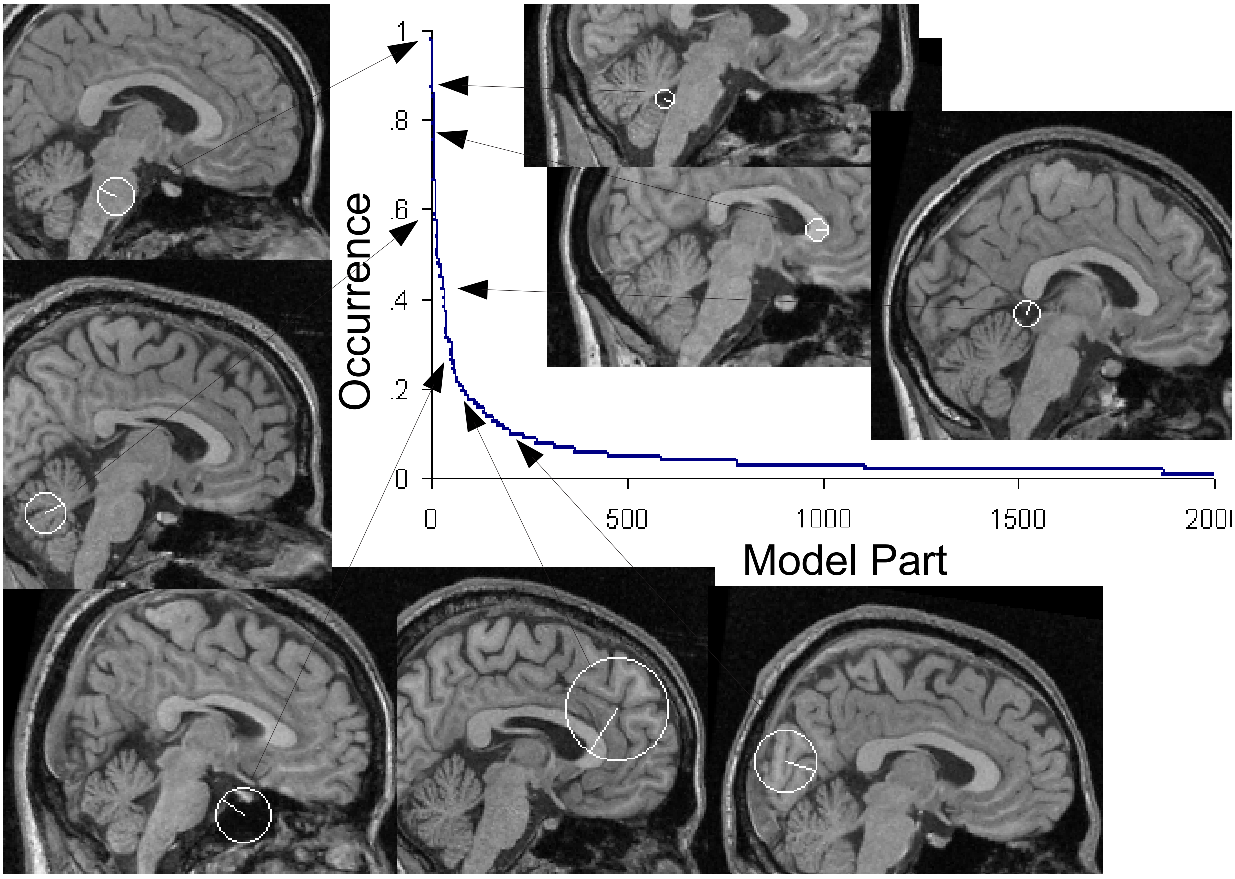
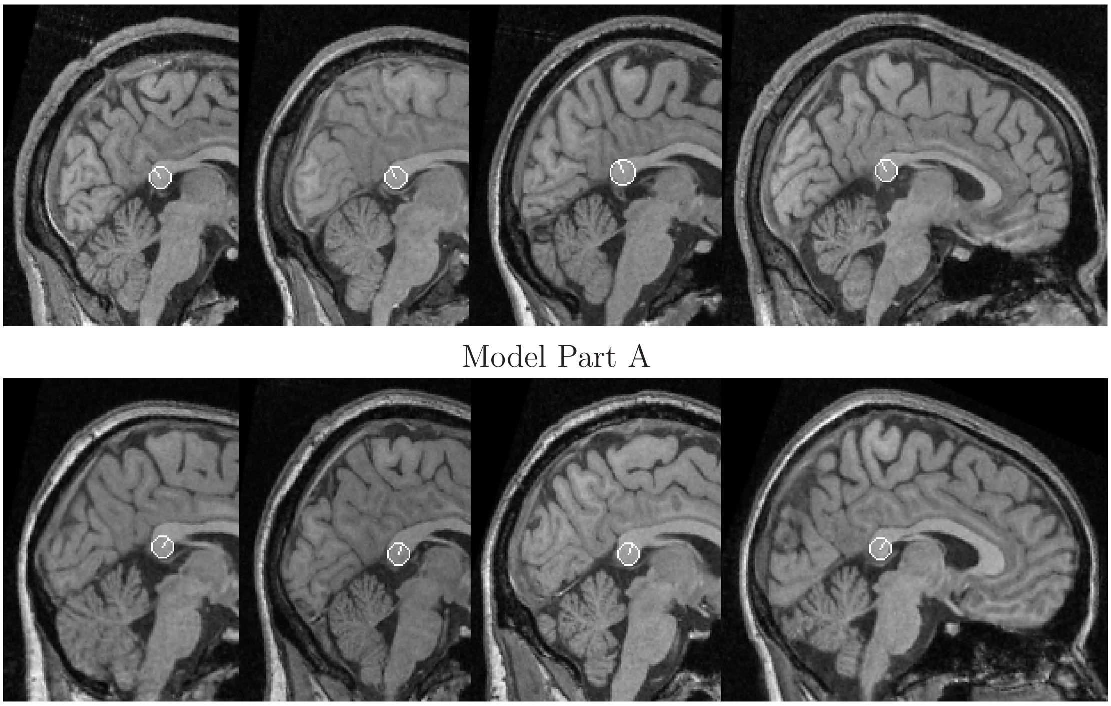


Fig. 5.1 A graph of learned model parts sorted by descending occurrence frequency $\pi_i^3 = p(m_i^{b=1}|o^{b=1})$. The images illustrate parts which occur at the indicated frequency within the population. Note that part occurrence drops off sharply, indicating that a relatively small number of model parts are common to all brains, whereas a large number are specific to a small number of brains.

present in all brains, but that scale-invariant features arising images of these structures tend to cluster into several distinct modes of appearance and geometry. As a result, they are identified as distinct model parts. In Figure 5.2 for example, the same section of the corpus callosum in brain images results in two model parts with distinct modes of orientation. This is a characteristic of the learning approach, which seeks to identify a set of maximally distinctive image patterns that arise from regularity in underlying brain anatomy.



Model Part A

Model Part B

Fig. 5.2 Illustrating instances of two different model parts, labeled A and B, arising from the same anatomical structure. In general, the same underlying anatomical structure can give rise to multiple model parts in the learning process, due to the interaction between the feature detector used and the image characteristics of the anatomical structure. Here, the corpus callosum splenium results in features (white circles) with two significant orientation modes, which are grouped into two distinct model parts by the learning process. The occurrence frequencies of model parts A and B are $\pi^3_A = 0.49$ and $\pi^3_B = 0.55$, respectively.

Figure 5.3 illustrates examples of model parts identified in different subjects in the presence of significant inter-subject variability. Model parts can be potentially useful in a number of ways. The following sections demonstrate that model parts serve as a basis for robust, stable model-to-subject registration, and that a subset of common parts can be used to quantify model fitting accuracy. Model part statistics can be used in order to interpret the result of model fitting in a meaningful, quantitative manner. For example, the distinctiveness in equation (3.7) represents the certainty with which a model part can be identified in a new image, and the geometrical likelihood in Section 3.1.1 represents the variability that can be expected in localizing a model part in space, orientation and scale. For the purpose of anatomical study, model parts could be grouped and assigned semantic labels according to their underlying anatomical structures and tissues, after which point model-to-subject registration could be used to propagate these labels in an automatic, robust manner to hundreds of new subjects. It is possible that distributions of part appearance and geometry could serve as useful indicators of subjects traits, for example abnormality, pathology or sex. For example the geometry of features lying on the corpus callosum could potentially serve as robust indicators of schizophrenia as in [15]. Learning brain structure characteristic of sex is investigated later in Section 5.2. Part geometry and occurrence variables could potentially be used improve morphometric analysis [14] by improving inter-subject alignment and indicating where inter-subject alignment may be valid or invalid. This is demonstrated later in Section 5.4.

5.1.2 Model-to-subject Registration

The goal of model-to-subject registration is to determine how a new subject relates to its population. The OCI model does this by identifying the features in the subject image that are representative of the population, as determined through learning. This section describes model-to-subject registration trials where the model learned from 102 subjects described in Section 5.1.1 is automatically fit to the remaining 50 new test subjects excluded from learning, via the process described in Section 3.1.3.

No gold standard exists for evaluating the ground truth accuracy of inter-subject registration [171]. Automatic part registration accuracy is thus measured with respect to manual part registration established by three different human raters as in [171]. Since there are many model parts, none of which are generally identified in all subjects, accuracy evaluation is based on a set of 25 test parts that occur frequently and throughout the brain during model learning. Fitting trials identify a subset of these 25 parts in each test subject, which serves as the basis for fitting evaluation. The number of test parts identified

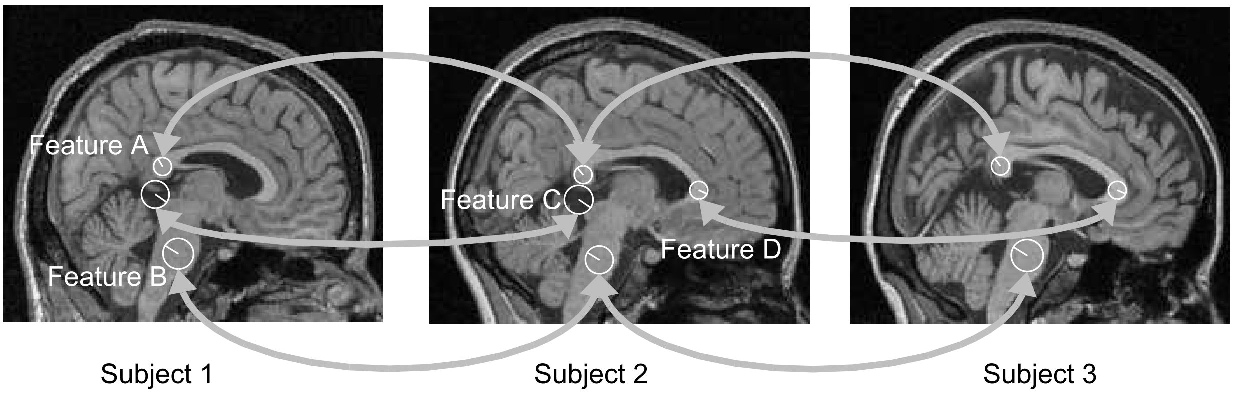


Fig. 5.3 Robust inter-subject correspondences identified despite a high degree of inter-subject variability based on scale-invariant features. There are generally many feature correspondences between any given pair of subjects. Here, a small subset of features occurring across a set of three subjects are shown for illustration purposes. Feature A located in the pons and feature B from the corpus callosum splenium are identified in all subjects 1, 2 and 3. Note that due to inter-subject variability and the characteristics of the feature detector used, not all features are identified in all subjects. Feature C reflective of the ambient cistern is only identified in subjects 1 and 2, due to the large ambient cistern in subject 3. Feature D arising from the corpus callosum genu is only identified in subjects 2 and 3, due to the more rounded genu shape in subject 1.

per subject image is 10 on average and ranges from 4 to 16. The number of instances of each test part identified over all 50 trials is 20 on average and ranges from 4 to 47. In total, 516 part instances are considered. Figure 5.5 illustrates a set of four test subject images which together contain at least one instance of each test model part.

Since model parts are defined via a fully automatic learning procedure, and may not necessarily match obvious anatomical structures, human raters themselves must first be taught the appearances and geometries of the parts before attempting to localize them in new images. To do this, an interactive application was developed in order to show the raters images of model part instances identified in different subjects during the model learning process as illustrated in Figure 5.4. Specifically, ten images such as those in Figure 5.2 of a single model part are shown in a looping video sequence. The rater is asked to watch the videos, and then determine the model part locations in all test subject images within which the part is identified during model fitting. Note that model parts contain a much richer description than simple spatial location, as they include orientation and scale information in addition to a measure of distinctiveness. These aspects could also be established and verified by human raters, however to do so is difficult and labor-intensive and evaluation is thus restricted to part location in this study. The measure of fitting quality used is the target registration error (TRE) [205] of model parts, calculated between image locations identified by the model and human raters (model-to-rater) and between raters (inter-rater). The TRE generally measures the discrepancy between different locations of a 'target' identified by different methods in an image, and is defined as the Euclidean distance between image points.

The TRE for each test model part averaged over all 50 test images is illustrated in Figure 5.5. Overall, the inter-rater and model-to-rater TREs are similar, indicating that individual model parts can be automatically fit with similar accuracy to human raters on a part-by-part basis. Localization is more precise for certain parts than for others, for both inter-rater and model-to-rater cases. This is primarily due to part scale, as large-scale parts such as those arising from cerebral lobes are intrinsically more difficult to localize with precision than small-scale features. For part W, the inter-rater TRE is somewhat lower than the model-to-rater TRE, indicating some disagreement between human raters and automatic model fitting for this particular model part. Subsequent investigation revealed that model part W had a relatively high intrinsic geometrical variability, as reflected in term $p(m_i^g|o^b, o^g)$, whereas human raters tended to agree as to where they felt the part should

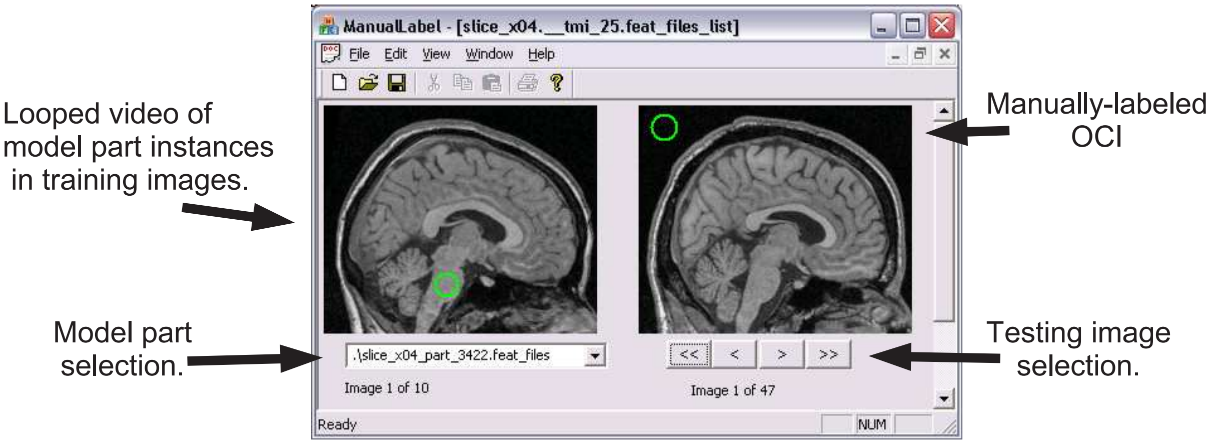

ManualLabel - [slice_x04.__tmi_25.feat_files_list]
File Edit View Window Help
Looped video of model part instances in training images.
Manually-labeled OCI
Model part selection.
.\slice_x04_part_3422.feat_files
Image 1 of 10
<< < > >>
Image 1 of 47
Testing image selection.
Ready
NUM

Fig. 5.4 Illustrating an interactive application developed to facilitate ground truth of model parts automatically identified in new images. A looped video of model part instances identified in training images is shown on the left, allowing the human rater to learn the appearance characteristics of the particular model part. Testing images are shown on the right, where the human rater labels the part location in order to establish manual ground truth.

occur. Note that model-to-rater agreement could be forced by tightening the geometrical consistency constraint in model learning as described in Section 3.1.2. In the current framework, however, the geometrical uncertainty has already been quantified by the density $p(m_i^g|o^b, o^g)$ and accounted for in model fitting.

It is also of interest to know the error with which the model can be registered to individual test images. The TRE for each of the 50 test images averaged over identified test model parts is illustrated in Figure 5.6. Here, agreement between inter-rater and model-to-rater TRE indicates that automatic model fitting is similar in accuracy to human raters on a per-image basis. The average inter-rater and model-to-rater TREs over all images are similar, 2.3 and 2.4 voxels, respectively. The two images shown below the graph in Figure 5.6 illustrate two test subjects for which the TRE is somewhat higher than average for both inter-rater and model-to-rater cases, indicating increased fitting difficulty for both human and machine. On the left, this is due to the fact that the feature detector has fused model parts X and Y into a single region with two dominant orientations. On the right, model part W has been mismatched to a similar-looking yet incorrect feature in the absence of a feature detected at the correct location just beneath the cerebellum. As previously mentioned, the high geometrical uncertainty associated with part W is quantified in term

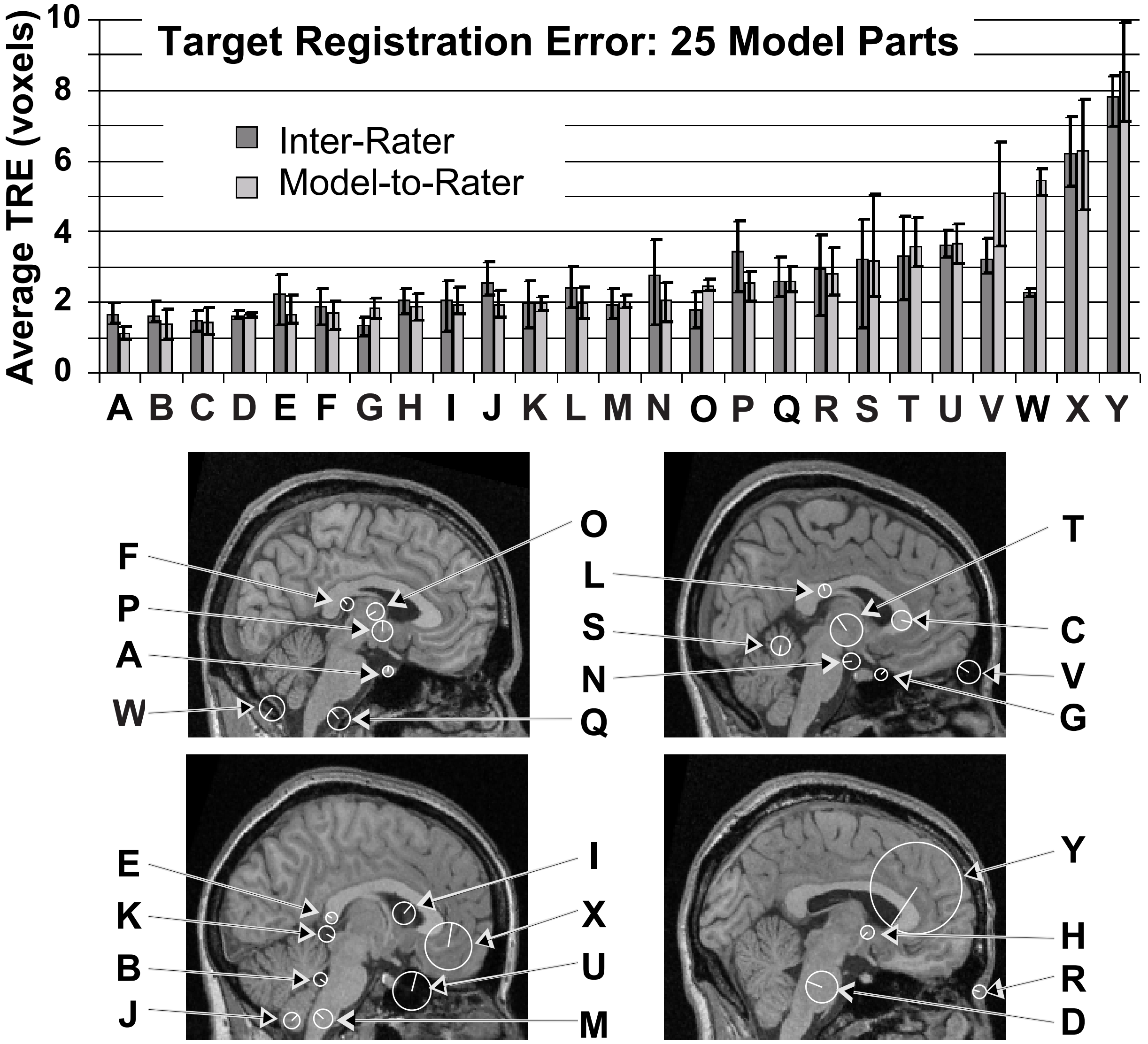


Fig. 5.5 The inter-rater and model-to-rater TRE for 25 test model parts, averaged over 50 test subject images and sorted in order of ascending model-to-rater TRE. The height of the bars represents the mean TRE, and error bars indicate the minimum and maximum TRE. The 4 images below illustrate instances of the indicated test part in images. Note that agreement between inter-rater and model-to-rater TRE indicates that individual model parts can be automatically localized with similar precision to human raters, validating model fitting on a part-by-part basis. Note also that the TRE varies from one part to the next, primarily due to the scale of the part in question, where the larger the part scale, the greater error associated with its localization.

$p(m_i^g|o^b, o^g)$ and its influence on fitting accuracy is already discounted by the model. The TRE measure in Figure 5.6 does not reflect this uncertainty, however, as the errors of all identified model parts are weighted equally.

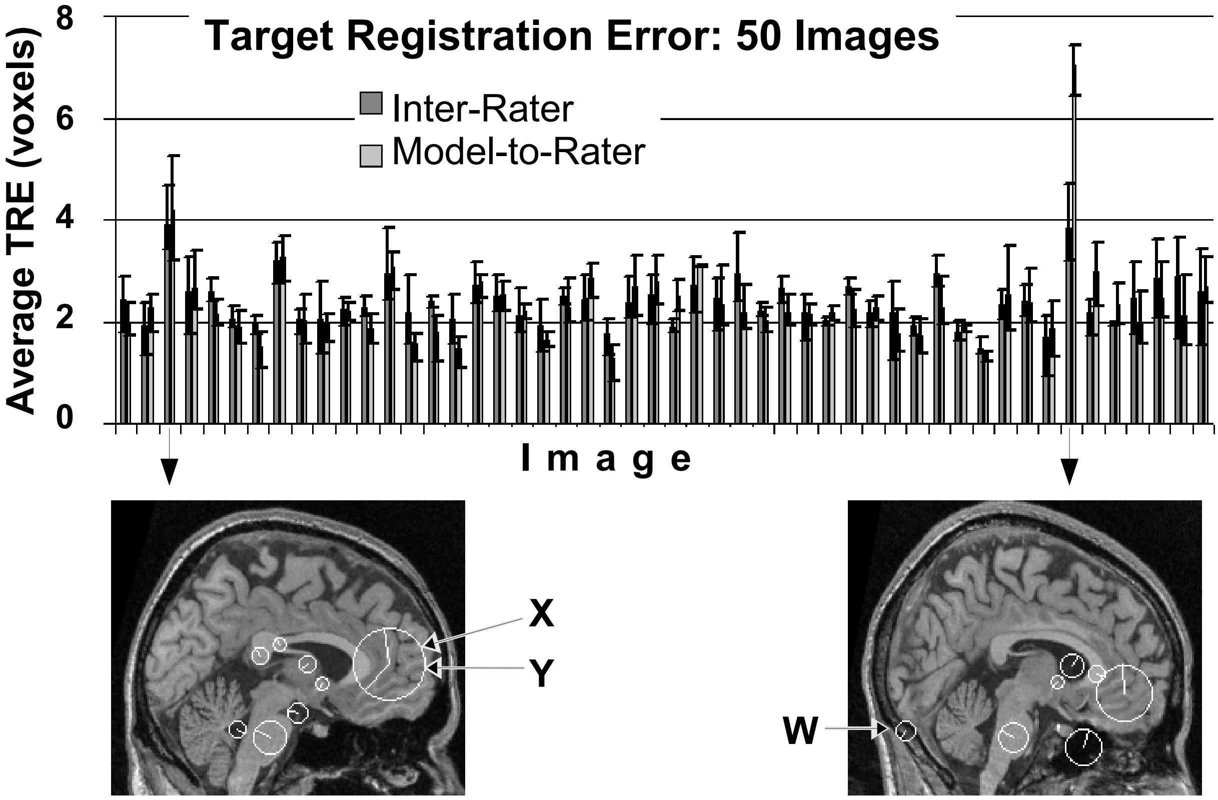


Fig. 5.6 The inter-rater and model-to-rater TRE for each test subject image averaged over identified test model parts. The height of the bars represents the mean TRE, and error bars indicate the minimum and maximum TRE. Note that general agreement between inter-rater and model-to-rater TRE indicates that the model can be automatically fit with similar precision to human raters. The images below illustrate two test subjects for which the TRE is noticeably higher than average, for both inter-rater and model-to-rater cases. On the left, this is due to the fact that the feature detector has fused parts X and Y into a single region with two dominant orientations. On the right, the part W has been mismatched to a similar-looking yet incorrect feature.

5.1.3 Parts-based OCI Modeling vs. Global Modeling

This section compares parts-based OCI modeling to the global modeling approach common in the literature. The OCI model describes a set of brain images as a collection of spatially localized, conditionally independent parts. In contrast, global models such as the AAM [29] typically assume one-to-one correspondence between all subject images, and represent the population terms of global modes of covariance about a mean, in which spatially distinct regions are coupled in a linear relationship and are thus statistically dependent. Forcing such a global model to fit in locations where correspondence is invalid can adversely affect the entire fit, including locations where valid correspondence exists. The hypothesis is that OCI model fitting is therefore more robust to unexpected inter-subject variability on a local scale than the AAM, as the OCI model specifically accounts for such variability and avoids forcing the assumption of one-to-one correspondence in locations where it is invalid. To test this hypothesis, AAM and OCI model fitting are compared, where an AAM [2] and an OCI model are trained on same set of 102 subjects, and fit to the same independent test set of 50 subjects. In the remainder of this section, the OCI model is referred to as the parts-based model (PBM), in order to emphasize the focus on modeling appearance in terms of spatial local features or parts.

The AAM and the PBM differ significantly in both training and fitting. AAM training requires manually determining a set of point correspondences in all 102 training images, after which a linear Gaussian model over image intensity and point location is estimated. Establishing manual point correspondence is tedious, requires a human to decide which and how many points to use, and is subject to inter-rater variability. PBM learning is fully automatic and requires no manual intervention, as features are determined by an automatic detector. AAM fitting is an iterative process, starting from an initial guess and occasionally falling into suboptimal local maxima when outside of a particular 'capture range'. During experimentation, multiple restarts are performed in order to obtain the best possible AAM fitting solutions. In contrast, the PBM fitting produces a robust, globally optimal solution, even in the presence of image translation, rotation and scale change.

Directly comparing AAM and PBM fitting to new subjects is difficult for several reasons. First, the fitting solution output of the two approaches is fundamentally different: the AAM produces a smooth mapping from model to subject, whereas the PBM identifies the

[2] The AAM implementation used is publicly available [206].

presence and local geometries of a discrete set of modeled parts. Second, there is no gold standard for evaluating the ground truth accuracy of inter-subject registration [12]. Third, little guidance exists in selecting a set of manual AAM landmarks leading to an optimal model. The two models are thus compared in terms of their stability in the presence of an artificial perturbation. To this end, two different sets of fitting trials are performed: the first is based on 50 normal test subjects. The second is based the same subjects with the addition of a localized, artificial intensity perturbation. Model fitting stability can then be evaluated in terms of the per-image TRE averaged over all model locations identified before and after perturbation. For the PBM, locations are based on the locations of the identified test parts defined in the previous section, and for the AAM, locations are based on the landmarks defining the model.

The first step in comparing the two models is to establish a baseline in terms of AAM model fitting accuracy. This is done by constructing an AAM based on the set of 102 training subject images, which is then fit automatically to the remaining 50 test subjects. The AAM is defined by six manually chosen landmark points as illustrated in Figure 5.7. The results of automatic AAM fitting trials are compared to manually-labeled solutions established by a single human rater in terms of the TRE, and illustrated in Figure 5.7. Note that in 2 of 50 trials, the TRE is exceptionally high, indicating that the AAM has converged to suboptimal, incorrect solutions. An average TRE less than 10 voxels is determined as the definition of a successful model fitting trial, and all subjects for which the TRE of model fitting is greater than this threshold are excluded. Note that as illustrated in Figure 5.6, all PBM fitting trials are successful by this definition. The average per-image TRE for successful fitting trials before perturbation is 2.3 voxels for the PBM and 3.8 voxels for the AAM fitting trials.

Having established that both the AAM and the PBM can be successfully fit a set of 48 subjects, where a successful fit is defined by an average TRE of less than 10 voxels, the stability of fitting in the presence of perturbation can be evaluated. Given that accurate, stable solutions are obtained in fitting to the 48 normal test subjects, how do these solutions vary when a localized artificial perturbation is introduced in each of the test subjects? The perturbation considered consists of a single black circle inserted at random locations in each test subject image, thereby modifying the image appearance locally in a manner reminiscent of a tumor or a resection. The circle is of radius 16 voxels, occupying approximately 2 percent of the slice of size 217x181 voxels. The image intensity at the circle center is

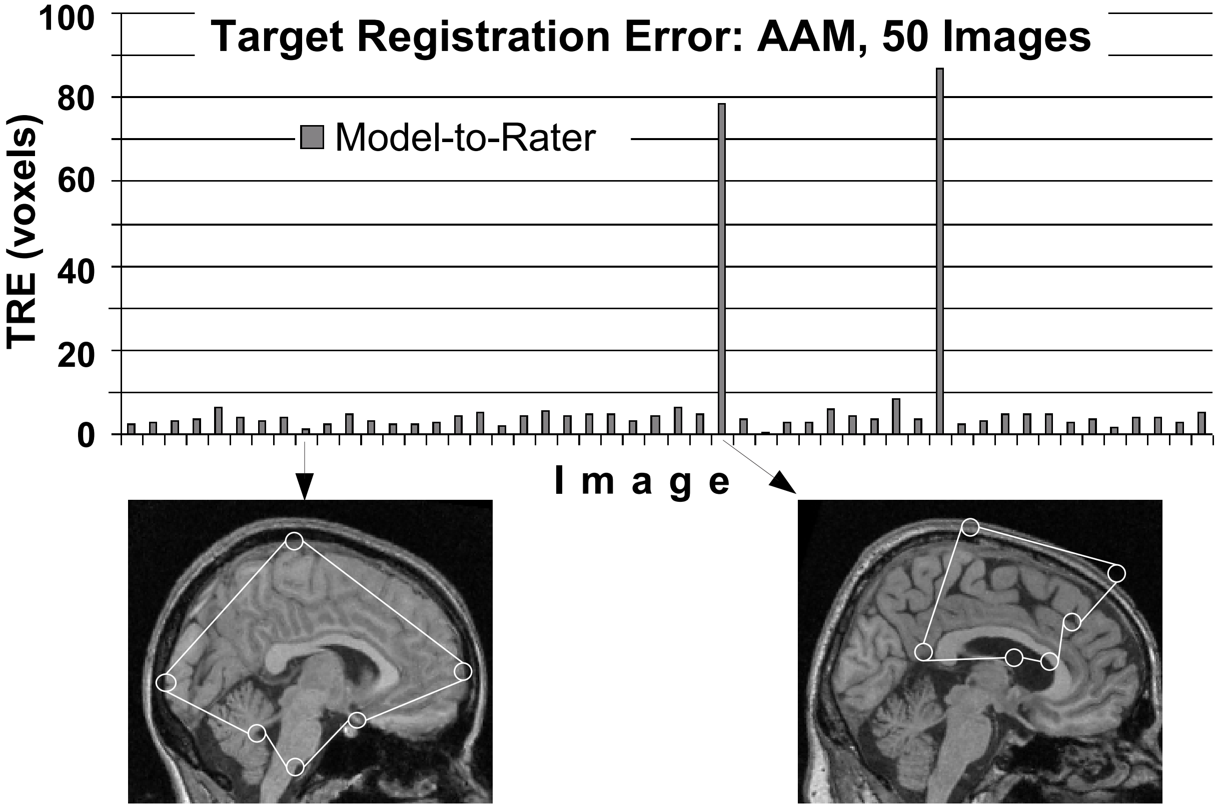


Fig. 5.7 The model-to-rater TRE of AAM fitting for 50 test subject images, averaged over all landmark locations identified in each image. As illustrated in the lower left image, the AAM created for the purpose of experimentation is defined by 6 manually selected points, 3 lying on distinct subcortical structures and 3 on cortical extremities. The modeled image content lies within the convex hull of the 6 points, and includes both cortical and sub-cortical structure. The TRE is calculated from the 6 landmark point locations determined both via automatic AAM fitting and manual identification by a single human rater, and the height of the bars represents the average TRE of all 6 points. Note that in 2 of the 50 subjects, the model incorrectly converges to suboptimal incorrect solutions with extraordinarily high TRE. The lower left and right images illustrate examples of successful and unsuccessful AAM fitting trials, respectively, where a successful fitting trial is deemed to be one in which the average TRE is less than 10 voxels. The unsuccessful trials are due to the inability of the AAM to cope with normal, unexpected inter-subject variability, such as the large, diagonally-oriented ventricles of the subject in the lower right image.

0, with a border blended smoothly into the original image in order to simulate a more natural image boundary. Intuitively, since the perturbation has only affected the image intensity in a local image region, the fitting solution should not change except perhaps in a neighborhood surrounding the perturbed area. The measure of model fitting stability adopted is the per-image TRE between fitting solutions obtained before and after the perturbation, which is referred to here as the original-to-perturbed TRE.

The original-to-perturbed TRE for both the AAM and the PBM is illustrated in Figure 5.8. Note that a large number of AAM fitting trials fail to converge to similar solutions before and after perturbation, as evidenced by exceptionally high original-to-perturbed TRE. As hypothesized, all AAM fitting solutions undergo a global change after perturbation, extending to image regions obviously unaffected by the perturbation, as shown in Figure 5.9. This contrasts sharply with the PBM fitting solutions. While the *minimum* original-to-perturbed TRE for the AAM is 2.9 voxels, the *maximum* for the PBM is 0.5 voxels and is barely visible on the graph in Figure 5.8, indicating that PBM fitting is stable in the presence of unexpected local variation. As seen in the example in Figure 5.9, PBM fitting solutions are virtually identical before and after perturbation, with the exception of fewer matches in a local neighborhood around the perturbation. This is because the perturbation is recognized as new and unmodeled image structure by the PBM, and can be safely ignored. The size of the neighborhood affected by the perturbation is defined both by the scale of the perturbation and the extent of the filter used in feature detection, which in the case of SIFT features is the width of a truncated Gaussian filter.

5.2 Determining Subject Traits from MR Brain Imagery

The previous section demonstrated how the OCI model can be used to describe brain anatomy in terms of a collection of localized image parts. A useful extension of the OCI model is to link these parts to subject traits of interest, such as age or pathology, in order to understand how anatomy co-varies with such traits. The experimentation in this section investigates the question of how the human brain varies with the trait of sex. While this question has a relatively limited clinical application, it can be used to benchmark classification performance [21] and to demonstrate the strength of the data-driven approach to automatically identifying specific image features associated with arbitrary traits of interest. One of the most obvious indicators of sex in the brain is overall volume, as male brains

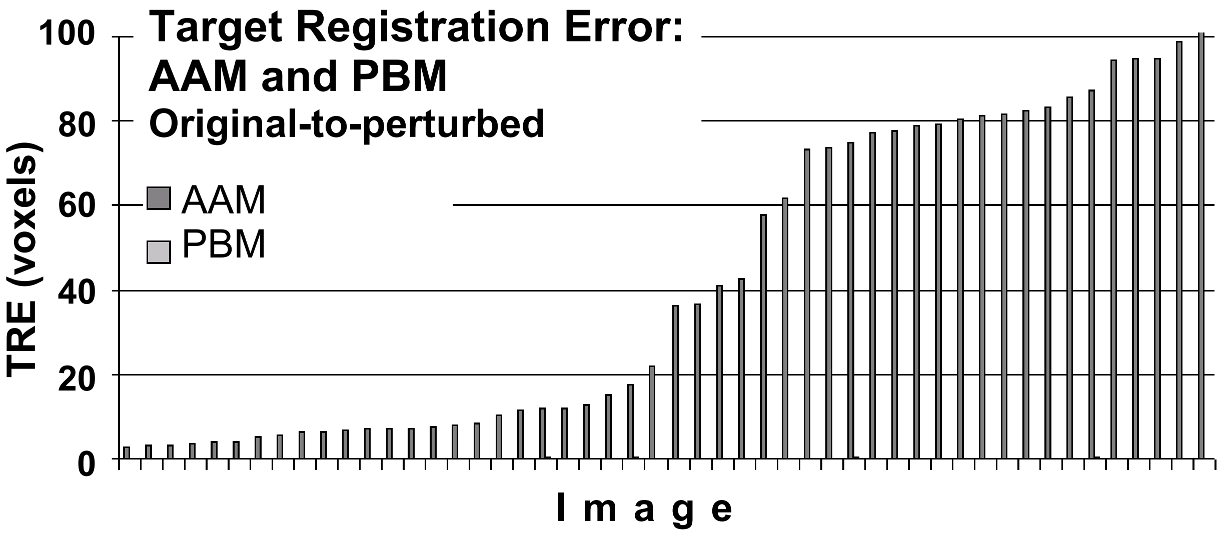
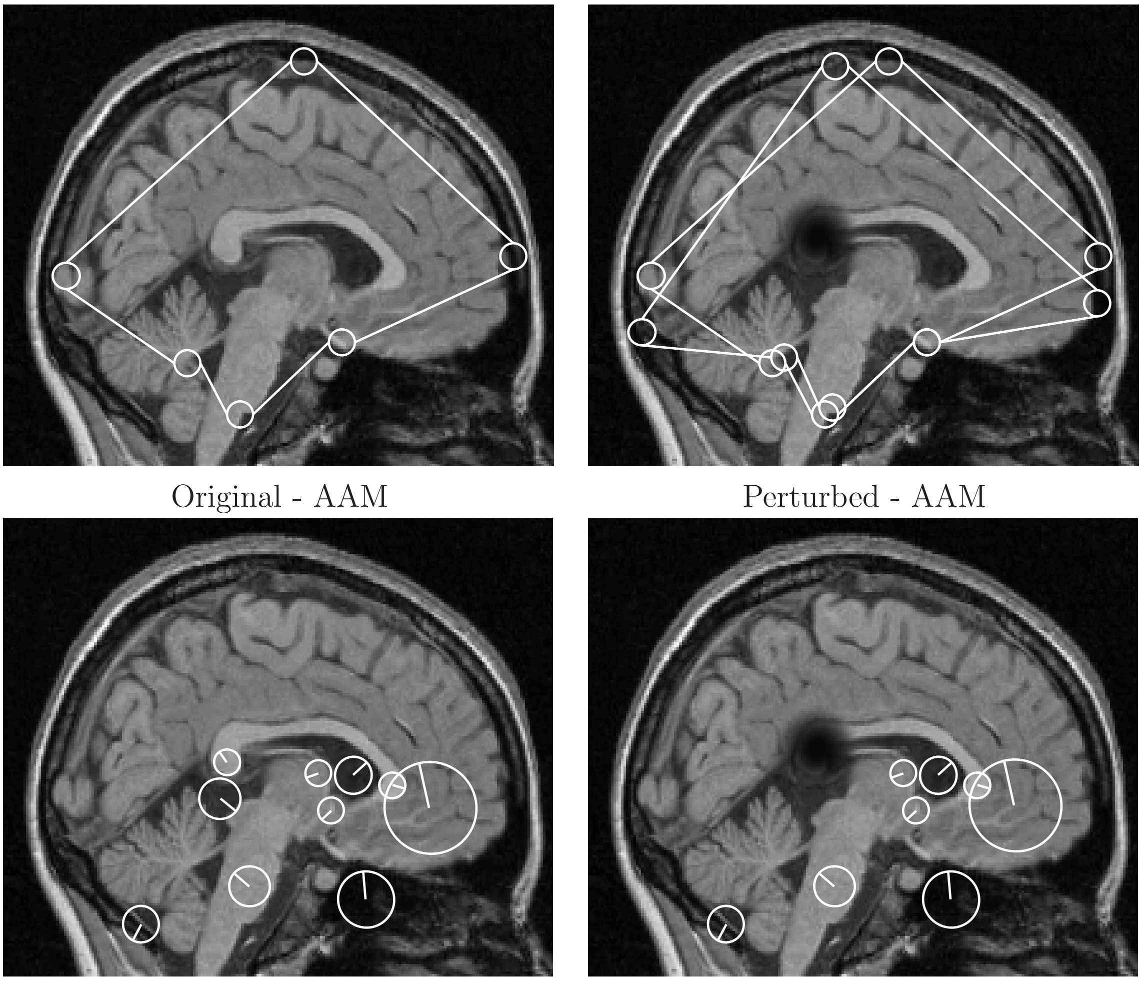


Fig. 5.8 The original-to-perturbed TRE for AAM and PBM fitting, averaged over AAM landmarks and identified PBM test parts in each of 50 test images, and sorted according to increasing AAM TRE. Note that AAM fitting is generally unstable, as a local image perturbation generally induces a global perturbation in the fitting solution, as evidenced by non-zero AAM TRE. In many cases, AAM solutions become completely invalid, resulting in exceptionally high TRE. While the original-to-perturbed TRE for the AAM ranges from [2.9-110] voxels, the PBM TRE ranges from [0.01-0.5] voxels and is barely visible on the graph.



Original - AAM

Perturbed - AAM

Original - PBM

Perturbed - PBM

Fig. 5.9 Illustrating the stability of AAM and PBM fitting in the presence of a local perturbation. The upper left image illustrates the original AAM fit to a new subject. The upper right image illustrates the original and the perturbed AAM fitting solutions to the same subject, after the introduction of a local, artificial perturbation in image content akin to the effect of a tumor or a resection. Note that the local perturbation gives rise to a global change in the AAM solution, extending to brain structure unaffected by the change. The two lower images illustrate PBM fitting for the same two images. Note that the PBM fit remains virtually unaffected except in a neighborhood of the perturbation, where two model test parts disappear due to the appearance of the novel, unmodeled image structure. The original-to-perturbed TRE for this test subject is 8.4 voxels for the AAM and 0.03 voxels for the PBM.

are generally larger than female brains, but this is of limited interest since it could merely reflect the general size difference between males and females. As a result, researchers have attempted to establish differences in scaled image space, with brains normalized via linear registration including scale [172] in order to better understand structural differences between the sexes. A similar methodology is adopted here, learning the features of brain anatomy relating to the trait of sex. Experimentation is based on the OCI model used in the previous section, learned from mid-sagittal slices of 100 subjects from the ICBM 152 data set. The experimental details are as follows.

Learning: Learning proceeds as a two-step process. First, a set of statistically significant features is learned from a subset of 102 subjects, consisting of 59 males and 43 females, as described in the previous section. Likelihood ratios of the Bayesian classifier are estimated for all feature instances in training data along with their corresponding sex labels, as described in Section 3.1.4. An empirically-determined Dirichlet regularization parameter of $d_{i,j} = 20$ samples is used. Note that this parameter is larger here than in the case of face image classification in Section 4.2.1, due to the fact that the same image features occur much more frequently in brain images than in face images.

Figure 5.10 illustrates examples features sorted according their likelihood ratios. While the majority of features are uninformative regarding sex, 179 out of 1549 features bear some degree of information regarding sex (i.e. $|log_{10}(\frac{p(m_i=1|female)}{p(m_i=1)|male)})| > 0.1$), meaning they are more likely to occur in either males or females. To highlight a few examples, feature A located in the corpus callosum splenium represents a particular pattern that is more likely to occur in male than in female subjects, supporting other research findings of sex-related differences in splenium shape [207]. No strong sex-related differences are found in feature B, nor in feature C located in the corpus callosum genu, supporting the absence of sex-related differences in the genu in scaled image space reported in [172]. Feature D represents a pattern indicative of female subjects.

Classification: Once the trait likelihood ratios have been calculated, they can be used to classify new subject images in order to assess the generalizability of subject trait learning. Using the Bayesian classifier described in Section 3.1.4, a correct sex classification rate of 80% is achieved on the 50 ICBM subject images not used in model learning. Note that this classification result is based on single sagittal slices comprising a fraction of the total MR brain volume data. Other techniques based on cortical volumes have shown slightly higher classification rates of 85% based on brain structures in cortical volumes [21].

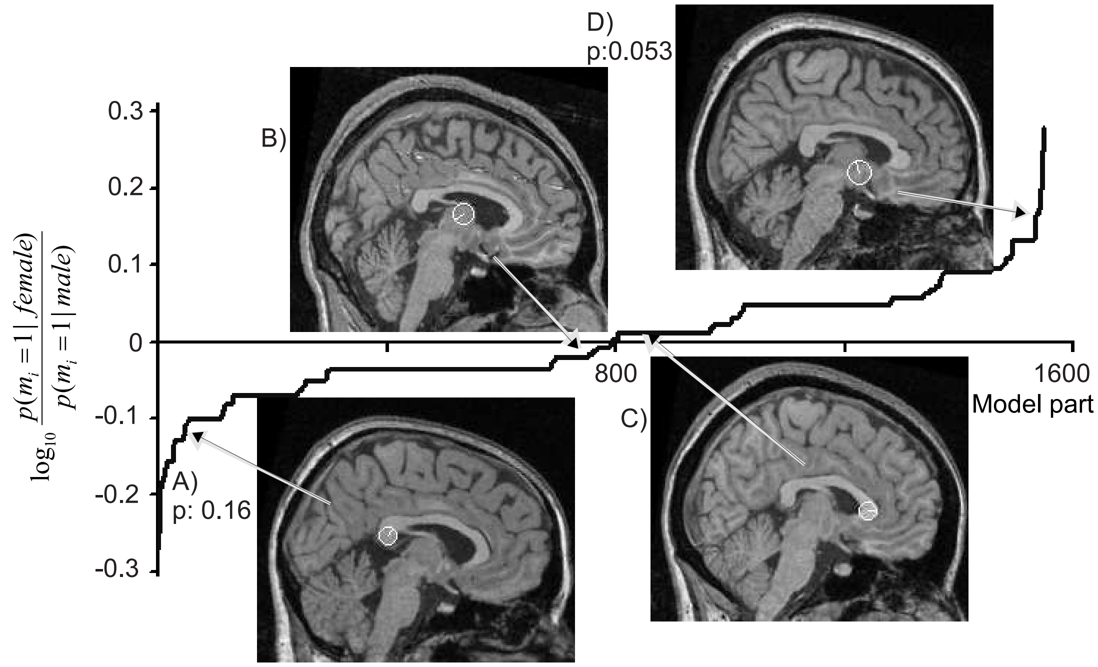
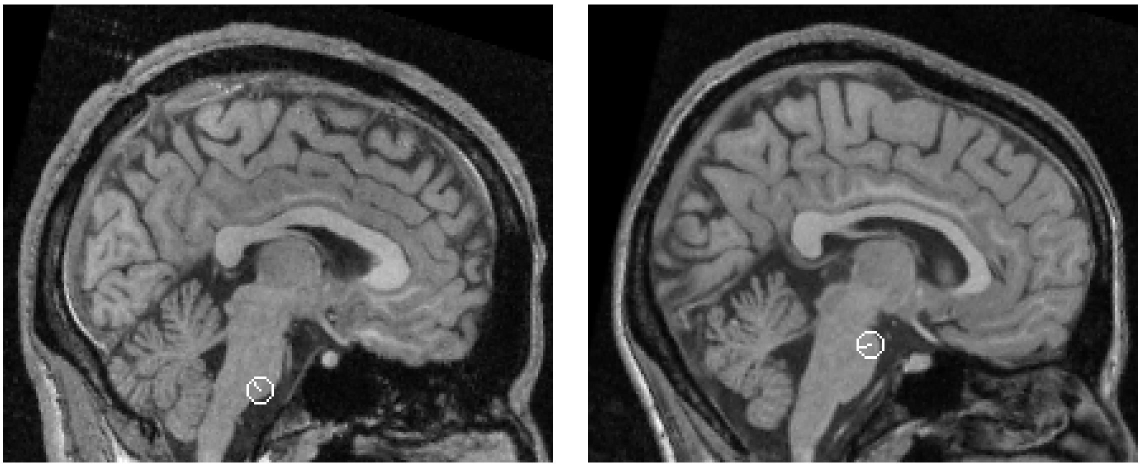


Fig. 5.10 Features sorted according to their informativeness regarding sex, based on the event of feature occurrence $m_i = 1$. 179 out of 1549 model features bear information regarding sex (i.e. $|log_{10}(\frac{p(m_i^{b=1}|female)}{p(m_i^{b=1}|male)})| > 0.1$). Features with significantly non-zero log likelihood ratios indicate image structure potentially related to the trait of interest. Those lying to the left of the graph occur more frequently in male subjects and those to the right more frequently in female subjects. Feature A) is indicative of male subjects and feature D is indicative of female subjects, with statistical significance indicated by the associated p-values. Features B) and D) occur frequently in brain slices but are uninformative regarding subject sex.

It is reasonable to expect that applying the approach to full 3D volumes would yield improved classification. A similar technique could be applied to a set of normal and diseased brains to identify anatomical features associated with the disease in an exploratory, data-driven fashion. This could have important ramifications for the future development of computer-aided diagnosis tools. Note that although a variety of black-box methods could be used to potentially classify brain imagery, e.g. perceptron networks [175], the method presented here using scale-invariant features can be used to identify the specific features and anatomical structures operative in classification, in addition to their occurrence statistics in the training population.

A strength of the approach presented here is that feature likelihood ratios could potentially be used to discover new structural morphologies related to subject traits. Figure 5.11 illustrates several model parts identified that suggest possible sex-related variations in the shape of the pons. Other such cases were identified in other regions throughout the brain and skull. As the link between these model features and the trait of sex has been established in a completely data-driven manner, further research is required in order to validate and understand the of neuroanatomical implications these findings.



A) Masculine, p: 0.037

B) Feminine, p: 0.18

Fig. 5.11 Examples of features indicating potential sex-related variations in the pons. Feature A) is indicative male subjects when present, occurring with a female:male log likelihood ratio of -0.18. Feature B) is indicative of female subjects when present, occurring with a female:male log likelihood of ratio of 0.13. The statistical significance of each feature by the associated p-values.

5.3 Learning a Parts-based Description of Cortical Appearance

The previous sections showed how the OCI model can be used to describe the anatomy of the brain, primarily in the relatively stable subcortical regions. Describing the anatomy of the cerebral cortex, in contrast, represents a special case, highly challenging modeling task. The cortex is of special interest to the neuroanatomical community, as cortical folding patterns are closely tied to functional regions of the brain. Unlike the relatively stable subcortical anatomy, however, the anatomy of the cortex is highly variable and difficult to analyze, even for human experts.

The goal of experimentation in the section is to identify similar, stable cortical structures shared by different subjects despite the exceptional high degree of inter-subject variability in the cortex. Experimentation involves learning an OCI model of cortical appearance using OCI reference frames localized in the cortical surface as described in Section 3.3.1.

Data: Experimentation in the section is based lateral volume renderings of the cortical surface of MR images in the ICBM 152 database. Prior to rendering, the skull is stripped from all subjects using a public implementation of the brain extraction tool (BET) [208]. Renderings of the cortical surface are then generated using the public MRIcro software program [209]. Scale-invariant features are automatically extracted in all images using the SIFT technique [22].

Learning: Learning was based on 98 different subjects. Left hemispheres are mirrored to obtain 196 aligned images of resolution 217x181. Each volume rendering results in approximately 500-600 scale invariant features. The lateral fissure is adopted as the local cortical OCI reference frame, more precisely the posterior horizontal ramus [186] originating from the triangular inferior frontal gyrus and running along the lateral fissure as illustrated in Figure 5.12. The lateral fissure is an obvious feature that can be labeled in an efficient, approximate manner with minimal expert knowledge, as model learning can tolerate a degree of labeling imprecision. Using this local reference frame, modeling error is expected to be minimized in cortical regions near the reference frame origin, such as Broca's area. In humans, Broca's area is responsible for language processing, speech production and speech comprehension [210]. In macaque monkeys, the corresponding region is responsible for control of orofacial actions [211]. In general, other such local reference frames could be similarly defined according to stable local structures present in all or most subjects.

Figure 5.13 illustrates the set of model parts automatically identified during the learning

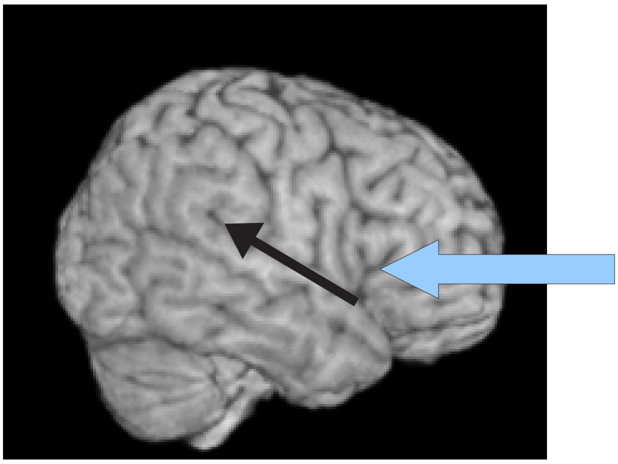


Fig. 5.12 Illustrating the lateral fissure cortical reference frame (black arrow). The lateral fissure can be used as a local reference frame to learn a description of the surrounding cortex, for instance Broca's area (indicated by the solid blue arrow).

process. Although several model parts correspond to cortical lobes, the majority arise from gyral bulges/corners or sulcal intersections. Once learned, model parts can be used to describe the variation in anatomy of the cortex over a population. The strength of the parts-based description is that the same anatomical region of the brain can be explained in terms of distinctly different appearance modes, as missing features or multiple distinct morphologies common in the cortex are explicitly modeled. Not all model parts identified during learning are necessarily anatomically significant. As mentioned, the potential for false feature correspondences increases as a function of feature size and distance to the reference frame origin.

To exhaustively verify that all occurrences of each model part in fact represent valid occurrences of the same anatomical structure in different subjects is a very difficult task, as an expert must learn to label thousands of different features in different images. Validation is thus performed on a limited set of 10 model parts, each associated with 10 occurrences in different subjects. Parts selected occur in more than 25% of subjects and are located within a radius of 55 pixels from the reference frame origin. Spurious features in the fringes of the image are disregarded. For each of the 10 model parts, a neuroanatomical expert was presented with 10 different occurrences identified in different images by model learning. For each part, the expert first decided which underlying cortical structure was associated with the majority of the 10 part instances. Afterwards, each individual occurrence was

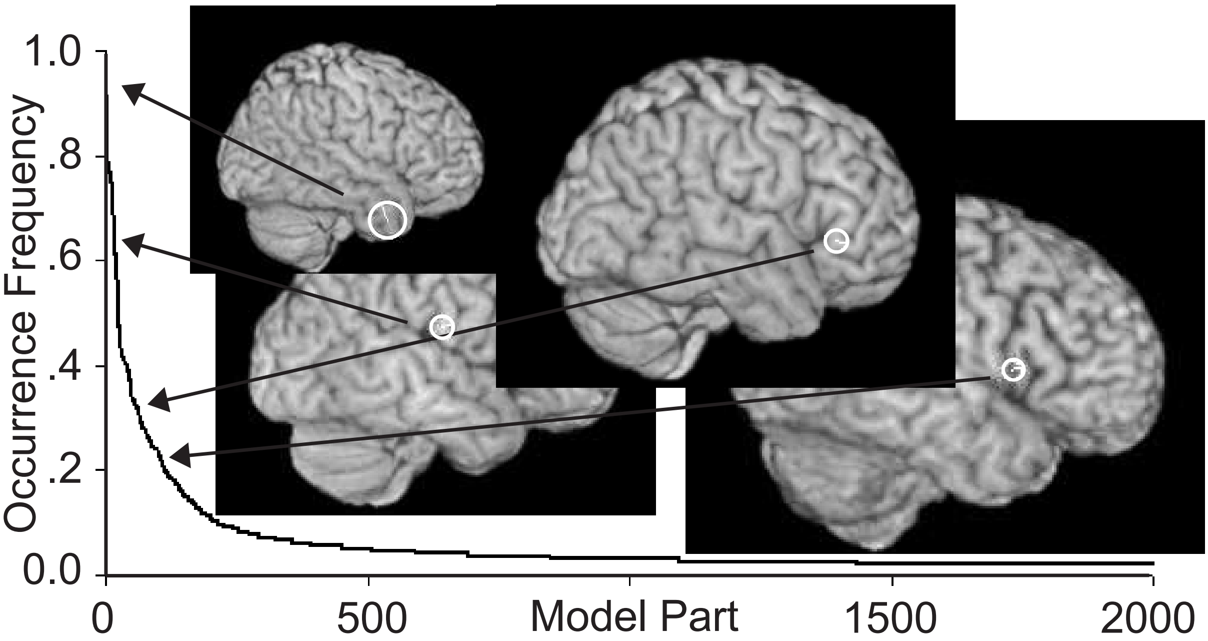


Fig. 5.13 Illustrating learned cortical model parts sorted in descending order of their occurrence frequency in training subjects. Note that a relatively small number of parts occur in many subjects, e.g. parts occurring in stable structures such as cortical lobes. A large number of parts represent cortical image patterns that are specific to small groups of subjects.

labeled as either correct (i.e. properly identified structure), incorrect or unsure. Figure 5.14 illustrates the results of validation: out of 100 different model part occurrences labeled A-J, 77 were labeled as correct, 7 as unsure and 16 as incorrect. Note how model inaccuracy is associated with human rater uncertainty: parts A-E have low numbers of incorrect labels and no unsure labels, while parts I and J have high numbers of both incorrect and unsure labels.

Figure 5.15 illustrates occurrences of model parts B, D and H in 3 different subjects. Note the extreme difficulty in correctly identifying corresponding cortical features, even for a human expert. Here, parts B and D represent different appearance variations of a similar section of the postcentral gyrus, illustrating the ability of the model to learn multiple appearance modes.

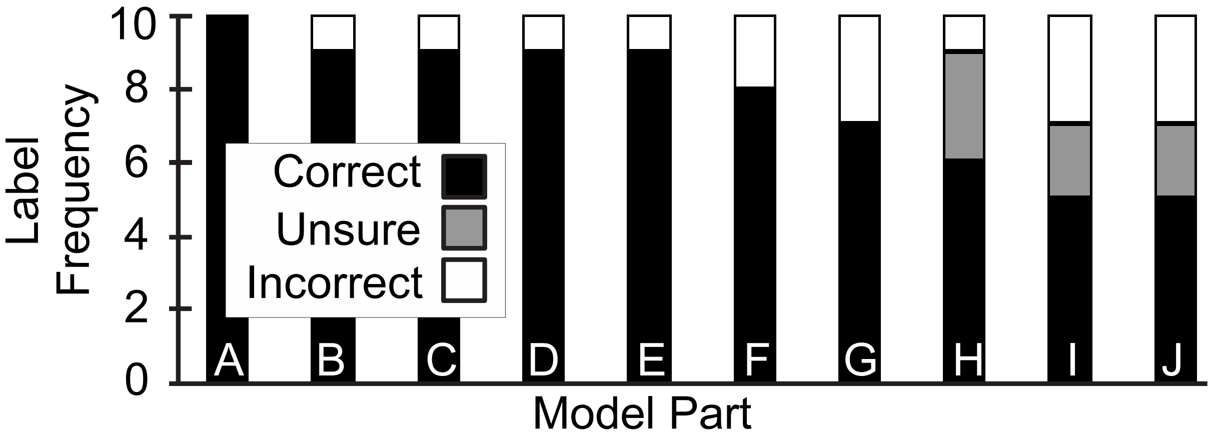


Fig. 5.14 A histogram of expert labeling results for 10 different cortical model parts A-J, based on 10 occurrences of each part in different images. Of 100 occurrences, 77 are labeled as correct, 7 as unsure and 16 as incorrect by a neuroanatomical expert. Incorrect occurrences typically arise in similar yet incorrect nearby structure. Note how model inaccuracy is associated with human rater uncertainty. Images of part occurrences used in validation can be seen in Figure 5.15.

5.4 Model-based Inter-subject Registration

Inter-subject registration involves determining a mapping between images of different subjects, and is generally complicated by the fact that one-to-one correspondence may not

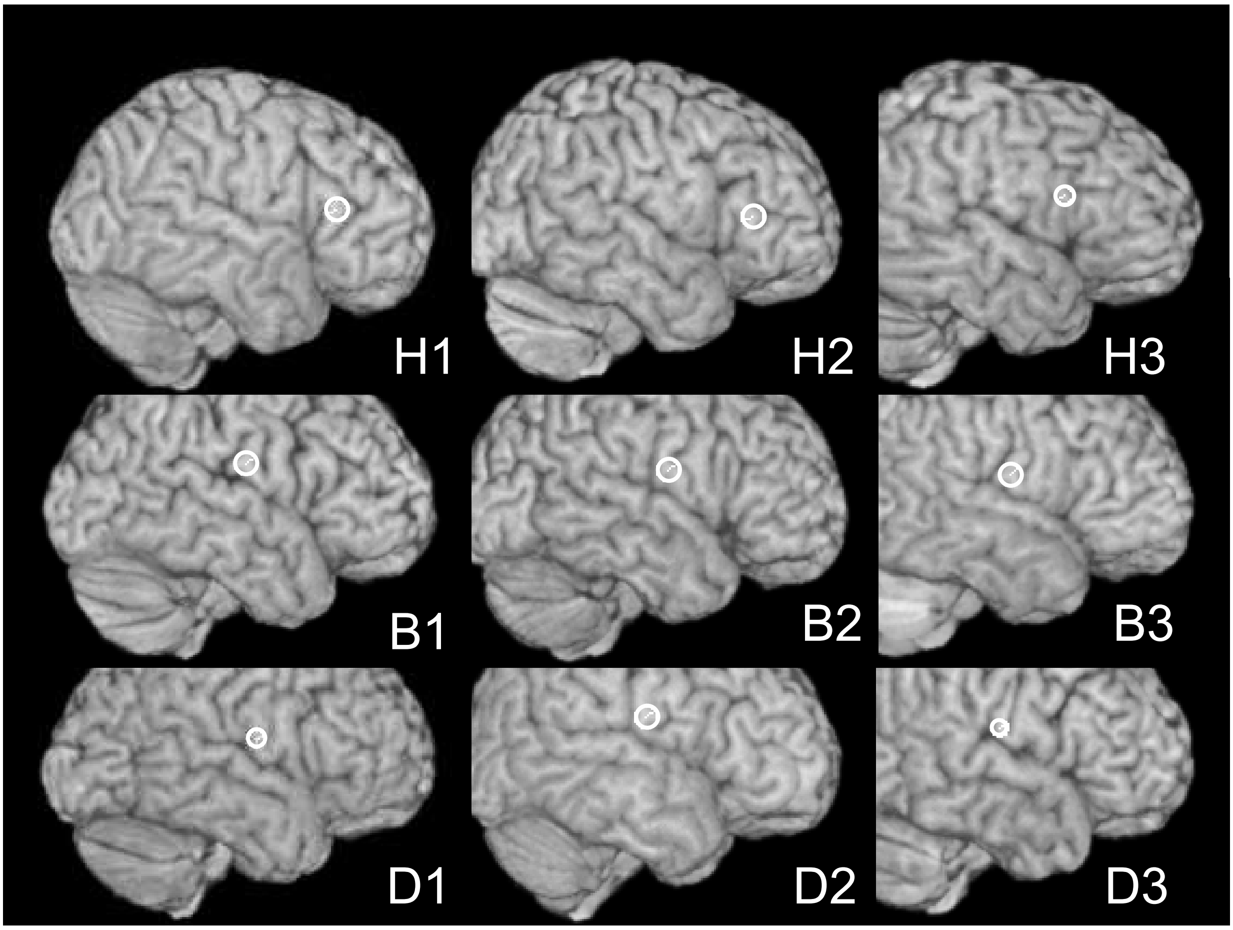


Fig. 5.15 Three occurrences of cortical model parts H, B and D in different subjects. Part H occurs anterior to the ascending sulcus in the pars triangularis of Broca's area, in 24% of subjects. Note that the image H3 was labeled an incorrect instance. Parts B and D represent two distinct appearance variations of a similar section of the postcentral gyrus, and occur in 58% and 62% of subjects, respectively. Note that model parts are not mutually exclusive and can independently co-occur in the same subject.

generally exist. A good strategy for inter-subject registration is to base registration on image features which have been identified as distinctive or characteristic of the population in question, based on a learned model of appearance. Such features can be identified in a single subject or model template and then registered to the next, similar to the strategy proposed by [182]. In the case where one-to-one correspondence does not exist, however, it is more reasonable to identify features common to both subjects [31], thereby avoiding registration of anatomical structure that may not be present in both subjects.

This section provides a brief demonstration of how the OCI model can be used to improve registration between images of different subjects, identifying regions where valid correspondence is likely to exist. An illustrative comparison of registration trials based on these two feature selection strategies is performed. The hypothesis is that registration based on model features shared by two subjects will be more successful at identifying meaningful inter-subject correspondences features of a single subject.

Data: Data consists of 20 coronal slices of ICBM 152 volumes, defined by the plane z=14 in MNI stereotactic space coordinates [204]. Scale-invariant features are automatically extracted using the SIFT technique [22].

Learning: An OCI model of coronal slices is learned from 102 subject images using the learning process described in Section 3.1.2. The OCI reference frame is defined as the projection of the Talairach AC-PC line in the coronal plane.

Fitting: The OCI model is fit to images of new subjects to be registered, using the fitting process described in Section 3.1.3, thereby identifying the same model features in different subjects.

Registration: Registration is performed by determining the displacements of features from an image of subject A to and image of subject B, based on a sum-of-squared-differences similarity measure regularized by an elastic prior between features. For each pair of subjects, registration trials based on the following two feature selection strategies are performed:

1. **Selecting model features of one subject:** The OCI model is fit to subject A, and the 10 most frequently occurring model features identified are registered from subject A to subject B.

2. **Selecting model features common to both subjects:** The OCI model is fit to both subjects A and B, and the 10 most frequently occurring model features identified

and shared by both subjects are registered from subject A to subject B.

As illustrated in Figure 5.16, where subjects to be registered exhibit similar anatomical structure, most subjects share similar features and both feature selection strategies result in reasonable registration. In the presence of significant inter-subject variability however, selecting model features common to both subjects allows registration to avoid regions in which a valid solution may not exist, such as the enlarged ventricular region of subjects A and B in Figure 5.16.

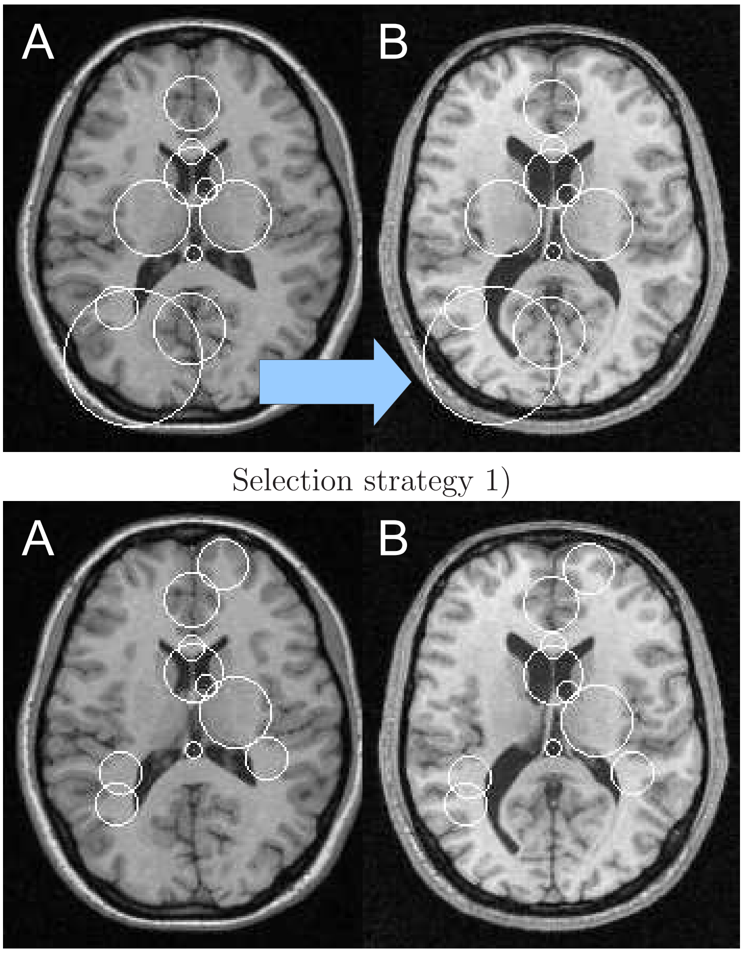
5.5 Discussion

This chapter presented experimentation adapting the OCI model as a statistical parts-based description of the brain in MR imagery. The parts-based OCI model represents images of a population as a collection of spatially localized image regions, or model parts, each of which consists of an appearance, a geometry and an occurrence frequency. The model specifically addresses the case where one-to-one correspondence between subjects does not exist due to anatomical variability, as model parts are not required to appear in all images.

Section 5.1 investigated OCI model learning on a subset of 102 subjects, establishing that a set of stable model parts could be identified in a large set of different subjects. OCI model fitting trails were performed, demonstrating that automatic model fitting could localize image features in new images with accuracy comparable to human raters, both on a part-by-part and an image-by-image basis. The stability of OCI model fitting was shown to be quantitatively superior to that of the global active appearance model in the case of unexpected artificial perturbation, illustrating the advantage of modeling appearance in terms of local, independently observable parts.

Section 5.2 presented experimentation in relating anatomical brain structure to the subject trait of sex. Image features relating to the trait of sex were identified from a model trained on 2D sagittal MR brain images of 102 different subjects, and relationships between sex and features in the corpus callosum support findings in literature. Additionally, sex classification on 50 brains not used in learning achieves a correct classification rate of 80% from sagittal slices comprising a fraction of the total brain volume.

Section 5.3 details experimentation focusing on parts-based modeling of the highly variable cortical surface, where the OCI model is adapted to the cortex using the lateral fissure



Selection strategy 1)

Selection strategy 2)

Fig. 5.16 Illustrating model-based feature selection for registering image A to image B. In each image pair, 10 features are selected and registered from subject A to subject B. In selection strategy 1), the most likely model features of subject A are registered to subject B. In selection strategy 2), the most likely features common to both subjects A and B are registered from subject A to subject B. Notice that a valid registration solution may not exist in the lower left region, due to an enlarged ventricle of subject B. Basing registration on features common to both subjects, this ambiguous region can be avoided, resulting in more meaningful registration when one-to-one correspondence between both subjects may not exist.

as a local reference frame. Experimentation shows that a large set of anatomically meaningful model parts can be automatically learned from new, unlabeled cortical regions in set of 196 lateral volume renderings. Expert human rater validation of 10 different model parts, each occurring in 10 in different subject images, revealed that 77% of automatically determined part correspondences represented valid correspondence of underlying cortical structures. Many additional cortical correspondences were identified but were not included in this thesis due to the time required for expert validation.

Section 5.4 provided a brief demonstration as to how the OCI model could be used to guide inter-subject registration in the case where the assumption of one-to-one correspondence may not be valid, by robustly identifying image regions that are both representative of the population and shared by the subjects being registered.

In summary, the parts-based model represents several important advancements with respect to statistical appearance typically applied to quantifying anatomical brain variability. These are as follows:

1. The model can be constructed via a fully automatic machine learning procedure capable of dealing with a large image set containing significant inter-subject variation.

2. The model can be robustly fit to new subjects to obtain a globally optimal solution, in the presence of significant inter-subject variability in addition to global image translation, rotation and scale changes.

3. Model fitting is stable, in the sense that a localized image deformation results in a localized change in the fitting solution.

4. All subjects of a population can be modeled simultaneously without making *a-priori* classifications as to which subjects are 'normal'.

5. The spatially localized model parts identified by the model offer an intuitive means of describing and communicating anatomical variability within a population.

6. Distinct anatomical parts can be linked directly to subject traits of interest such as pathology, age or sex, and used to understand and discover their links to anatomy.

7. The model is general and applicable to describing the anatomy of a population in a wide variety of different contexts.

The technique presented could potentially be extended in a number of ways. Development of robust invariant feature detectors in 3D will lead to cortical modeling in full MR volumes. Scale-invariant features have been derived from a variety of image characteristics such as edges, phase and entropy, all of which could potentially be used together in a single OCI model to enrich the descriptive power of the model. Additionally, features with a higher degree of invariance, such as affine invariants, have been developed and may be more appropriate for describing elongated features such as cortical sulci.

There are a number future directions to scale-invariant feature-based learning of subject traits. The approach is described using discrete random variables of traits, but likelihood ratio / MAP estimation strategy is equally applicable to continuous variables such as age. Although this thesis applies the approach to automatically determining the brain features indicative of sex, the framework could be applied to addressing a wide variety of general questions, for example determining characteristics of brain variation due to pathologies. Ongoing investigations are being performed to determine how cortical patterns vary with traits such sex, age and pathology, and differ between brain hemispheres. Other work is aimed at identifying anatomical differences between healthy control subjects and subjects afflicted by Parkinson's disease (PD). This could potentially result in the discovery of reliable biomarkers that can be used to understand the disease and its progression, a major challenge in PD research.

The experimentation described in this chapter involved 2D brain imagery, which greatly facilitated the evaluation of OCI modeling as existing 2D implementations of invariant feature detection and active appearance modeling could be used. The OCI modeling theory is independent of image dimension, however, and can be applied to modeling images of higher dimensionality such as 3D brain volumes or 4D temporal data based on invariant feature implementations in higher dimensions. This would be an important step in describing the true 3D nature brain structure and how it changes over time, to track disease progression for instance. While extracting the scales and locations of invariant features in higher dimensions is straightforward using N-dimensional scale-space pyramids, the more difficult question is dealing with orientation parameters. As mentioned, the number of orientation parameters required to characterize features and the OCI model in N-dimensional space is $N(N-1)/2$. Identifying these orientations by detecting peaks in image derivative histograms becomes difficult, as multiple histogram peaks are required. A potential solution might be to simply use the principle derivative orientations for individual features. It may be that explicitly modeling all orientations may be unnecessary.

5.6 Summary

This chapter presented experimentation applying OCI modeling theory to the analysis of medical imagery, as a parts-based model of anatomical appearance. The model can be used to learn the appearance of brain images from a large set of subjects with no manual supervision, quantifying appearance probabilistically in terms of the occurrence frequency, geometry and appearance of a set of natural, local patterns. The strength of the parts-based model is the ability to address the case where one-to-one correspondence does not occur between all subjects of a population, a major difficulty in many medical image analysis contexts. Experimentation was presented showing how the parts-based model can be learned from a large set of MR brain images, robustly fit to new subjects in situations where global models fail and used to identify anatomical structure related the trait of subject sex. Furthermore, experimentation showed that model features can be linked to subject traits, and used as a basis for identifying anatomical characteristics reflective of traits, here the trait of sex. Learning a parts-based description for the special case of the highly variable cerebral cortex was possible by basing modeling on an appropriate locally defined OCI reference frame, here the lateral sulcus. Finally, a demonstration was presented showing how a learned anatomical model can be used as the basis for inter-subject registration.

Chapter 6

Conclusion and Future Directions

The proliferation of imaging devices in recent times, in addition to high capacity and high bandwidth storage and communication media, has led to an explosion of image data. The need to organize this data in a meaningful way has spurred intensive efforts to develop techniques for automatically learning, detecting, localizing and classify appearance patterns in arbitrary imagery. Designing a general computational model to meet this need is a challenging task, as the model must be capable of effectively describing a wide range of pattern variability in appearance and geometry, lend itself of computationally efficient algorithms required to process large amounts of image data, and generalize to describing a wide variety of different types of image patterns. Computational approaches taken often vary according to the goals of the specific context, for example in computer vision or medical image analysis.

This thesis presents a new, general approach to modeling appearance patterns arising from classes of visually similar objects. The primary theoretical contribution is a new probabilistic model of pattern appearance, which represents pattern of interest in terms of an object class invariant (OCI), a high level geometrical structure that is 1) uniquely defined for each pattern instance and 2) invariant to nuisance parameters of the imaging process. The OCI is not observed directly from the image, but inferred from generic scale-invariant features via a probabilistic model, which can be learned from image data. As nuisance parameters are not modeled explicitly but that dealt with via model invariance, the OCI model remains computationally efficient. Furthermore, the OCI model can be used to effectively describe image patterns in a wide variety of different imaging contexts, due to its general nature.

The major contribution of this thesis is the generalization and unification of modeling techniques in the distinct yet related research fields of computer vision and medical image analysis. Computer vision deals primarily with 2D imagery representing projections of the 3D world. A central focus is effectively describing the appearance of classes of similar objects in natural, arbitrary imagery, despite intra-class appearance variability and appearance changes due to changes in viewpoints. The OCI model accomplishes this by describing the appearance of object classes in a manner invariant to changes in viewpoint. A central focus of medical image analysis is to quantitatively describe the variability of anatomical structure of a population. A major challenge is to cope with inter-subject variability, where the mapping of anatomical structure from one subject to another may ambiguous or non-existent. The OCI model accomplishes this by describing subject anatomy in terms of a collage of localized features or 'parts'. This parts-based description explicitly models the situation where one-to-one correspondence between subjects does not exist, as individual parts do not occur in all subjects but rather with a probability in a population. The following sections summarize the significance of the OCI modeling technique and future research directions in the respective contexts of computer vision and medical imaging.

6.1 Computer Vision

The primary contribution of this thesis to computer vision is the first integrated system for learning, detecting, localizing and classifying visual traits of 3D object classes in natural imagery captured from arbitrary viewpoints. Experimentation demonstrates this system in the context detecting, localizing faces and motorcycles, and classifying human faces in terms of sex. This work represents the first viewpoint-invariant model of 3D object class appearance based on local image features. Unlike other approaches which model the variable of viewpoint explicitly via multi-view techniques, the OCI model treats viewpoint as a nuisance parameter. This simplifies model learning from arbitrary imagery, as no explicit knowledge of viewpoint is required. Multi-view approaches require viewpoint information, either via manual labeling [3] or by capturing different images around the same object class instances [155, 156, 4], neither of which may be available in arbitrary, natural image databases such as the Internet. Experimentation in this thesis shows that the viewpoint-invariant OCI model results in quantitatively superior detection performance in comparison with an equivalent multi-view model, in the context of face detection.

The simplified learning afforded by the OCI model leads to the contribution of visual trait learning and classification from arbitrary viewpoints. Visual traits are abstract characteristics of object classes that can be determined from image data, and used to describe or subclassify objects. While other approaches to determining visual traits have investigated single, non-occluded views, the OCI model is the first approach to address trait classification from arbitrary viewpoints and in the presence of occlusion. Experimental results establish the first baseline in the literature for face sex classification from arbitrary viewpoints. Furthermore, experimentation establishes that reasonable sex classification results can be achieved despite a significant degree simulated occlusion.

The majority of experimentation in this thesis is based on a supervised learning algorithm for OCI models, where a human decides which images are to be grouped into an object class and how they related geometrically in terms of the OCI. Unsupervised learning of OCI models in a completely data-driven fashion is the next logical step, where patterns are naturally clustered into OCI models with no manual supervision. To this end, an algorithm is developed in order estimate and optimal OCI in a data-driven manner, by iteratively learning model parts based on initial OCI labels, then re-estimating the OCI geometry based on learned model parts. Experimentation shows that this algorithm drives the OCI definition to a stable structure centered with respect to image features arising from the underlying object class, thereby minimizing OCI localization error as predicted by theory. Furthermore, the OCI geometry in the 2D image plane remains geometrically consistent with that of the projected 3D object class in images captured from in arbitrary viewpoints. In the case of faces, for instance, the data-driven OCI is consistent with a 3D line segment located centrally within the head. This represents empirical evidence for the existence of a viewpoint-invariant OCI in the case of human faces. Experimentation shows that the OCI modeling based on the data-driven OCI results in quantitatively improved detection performance in comparison with modeling based on a manually-defined OCI.

The work in this thesis makes use of an OCI in the form of a 3D line segment for face images and a 3D sphere for the class of motorcycles. The linear OCI provides invariance to viewpoint change in a plane surrounding the object and the spherical OCI provides invariance to viewpoint changes in a view sphere surround the object. Other geometrical definitions could be used, for instance a collection of perpendicular linear OCIs could potentially be used. An issue is whether or not all object classes exhibit meaningful viewpoint-invariant reference frames. It would seem so for rigid object classes such as cars or 3D scenes. A collection of reference frames would likely be required to efficiently encode deformable or articulated object classes, similar to the geon approach [26]. Highly deformable object classes such as towels would pose a difficulty, modeling of such classes may be more dependent on texture and context than on geometrical description.

An important issue that arises in appearance modeling is whether to model sources of variability explicitly or via model invariance. The OCI modeling approach describes appearance in a manner invariant to viewpoint, as opposed to explicitly modeling variation due to viewpoint changes. This approach is feasible for the tasks of detection, localization and classification because explicit knowledge of object viewpoint is unnecessary. While these tasks can be performed in a viewpoint-invariant manner, other tasks require or make use of viewpoint information. For instance, specific object recognition in the human visual system is thought to be viewpoint-dependent whereas object categorization is thought to viewpoint-invariant [212]. Further research is required to investigate the link between viewpoint modeling and viewpoint-invariance in these situations.

The approach in this thesis is based on modeling classes of similar objects in terms of distinctive image features. Many classes of objects may not share distinctive features, for example cups are defined by simple geometrical features such as bounding contours. New types of invariant image features, particularly contour-based features, could be potentially important in extending OCI modeling to such objects. The work in this thesis focuses on modeling the appearance of 3D object classes in 2D projective imagery. Future work will see the OCI model applied in different data dimensionalities. The OCI model could potentially be applied for speech representation and recognition, to learn robust models for words from noisy 1D sound samples. Likewise, it could potentially be used to model space-time patterns based on visual events in 3D space-time video sequences, where work has begin addressing space-time feature detection [183].

6.2 Medical Imagery

The primary contribution of this thesis to medical image analysis is the first parts-based model of anatomy, that can be learned from a set of images of a population, robustly fit to new subjects and used to identify anatomical image characteristics reflective of subject traits such as sex or pathology. The system is applied in the context of modeling the anatomy of the human brain in slices of MR imagery. The OCI model is the first approach in the literature to model brain anatomy in terms of a collage of local, conditionally independent image features or parts, which do not occur in all subjects but with probability in a population. In contrast, the majority of other approaches in the literature are based on the assumption of global one-to-one correspondence between different subjects. This assumption is difficult to justify, however, particularly considering factors such as multi-morphic anatomical structures, pathology, surgical resection and general inter-subject variability. By avoiding the assumption of one-to-one correspondence, the OCI model can avoid confounding the analysis of different underlying anatomical tissues where correspondence is ambiguous or non-existent, and is capable of describing situations where the same anatomical region exhibits multiple modes of morphology, in the cortex for instance [185]. Experimentation demonstrates that OCI model fitting is significantly more robust and stable than fitting of the global active appearance model of Cootes and Taylor [29], particularly in the presence of unexpected local perturbation.

The OCI model is also applied to the case of learning the anatomy of the cerebral cortex. The anatomy of the cortex is of special interest due to importance of cortical folding patterns in delineating functional regions of the brain, but is exceeding difficult to describe due to the high degree of variability in cortical folding from one subject to the next. While other techniques attempt to reproduce manual labelings of a human expert, the OCI model is the first technique proposed for learning new, unlabeled cortical structure.

Experimentation demonstrates how the OCI model could potentially be used for model-based registration of images of new subjects, by robustly identifying OCI model features shared by the subjects to be registered. Such features represent regions where statistically significant correspondence exists, and can be used to bias registration in order to avoid fitting where correspondence may not exist. Finally, experimentation shows that the OCI model can be used to determine anatomical characteristics of the brain which indicative of subject sex, in the form of learned model parts.

The parts-based description of medical image anatomy provides an interpretation of the underlying data which is fundamentally different from that offered by global modeling approaches commonly used to analyze medical imagery. Individual subjects are described as a collage of distinct model parts, instead of as smooth global mappings between a model template and a subject. This interpretation necessarily implies discontinuities between adjacent or overlapping features, an aspect which does not arise in models based on smooth mappings. It is precisely such discontinuities, however, which make it possible to account for occlusion or multiple modes of appearance, phenomena which are by definition discontinuous and difficult to describe using smooth, global mappings.

There are many future directions open to parts-based modeling of brain anatomy. Subjects could be grouped according to their model parts, in order to identify sub-populations sharing similar anatomical characteristics. This could be done via clustering techniques. For instance, groups of features could be clustered in order to identify distinct morphological modes in regions of the cortex, as is commonly done by manual raters [185]. Different types of invariant features incorporated into the model could prove useful for modeling different aspects of brain anatomy. For instance, affine-invariant features may prove more effective at modeling elongated structures in the cortex such as sulci or giri.

Due to the generality of the OCI model, it should prove useful in a wide variety of other medical imaging domains for the study of anatomical structure within a population, an important area of computational anatomy. The experimentation presented in this thesis is based on 2D imagery, as the primary goals are to validate the parts-based model and to contrast parts-based and global modeling, which is greatly facilitated using publicly available 2D implementations of invariant feature detectors and active appearance models. An important extension of the OCI model in medical imagery is extending modeling to 3D volumetric and 4D temporal images. This will require the development of robust feature detectors in these image dimensionalities, preliminary work has shown feature locations and scales can be identified using an image pyramid-based peak detection methodology similar to 2D implementations. Remaining work will involve effectively coping with additional parameters of feature orientation and automatically establishing inter-subject correspondences.

Appendix A

Subset of CMU Face Images

This appendix outlines the subset of images from the CMU profile database [37] used to test OCI model detection and localization in Sections 4.1.2 and 4.1.3. These are as follows:

afghanistan.1.tchv altschul.1.tchv annan.2.tchv assemblyman.1.tchv bailout.1.tchv bbosmith.1.tchv billy.1.tchv bosnia-ngos.1.tchv brezhnev.1.tchv brief.2.tchv briefs.1.tchv brother.1.tchv bulls-celtics.2.tchv bulls-pacers.2.tchv bush.1.tchv chile.3.tchv chiquita.1.tchv clapton.1.tchv cockell.1.tchv congress.1.tchv cosby-trial.2.tchv cruise.tchv cuny.1.tchv currie.2.tchv dole.1.tchv einstein.1.tchv elderly.2.tchv elderly.3.tchv energy.1.tchv english-writer-4.3.tchv filmmakers.3.tchv floyd.1.tchv garment.2.tchv germ-sale.1.tchv gloup.1.tchv gold.1.tchv gould.1.tchv gymnastics.2.tchv hearings.1.tchv hford.tchv homeless.1.tchv homeless.2.tchv hongkong.1.tchv huac.tchv hurricane_dennis_1.tchv insomnia.1.tchv japan.2.tchv

Appendix B

Inter-feature Geometrical Transform

The linear transform $t_{ij} : m_i^g \rightarrow m_j^g$ mapping the scale-invariant geometry of feature i to that of feature j (or the geometry feature i to the geometry of an OCI o^g) in an image I is defined as follows:

$$\begin{bmatrix} log(\sigma_j) \\ \theta_j \\ x_j \\ y_j \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \log d\sigma_{ij} \\ 0 & 1 & 0 & 0 & d\theta_{ij} \\ 0 & 0 & 1 & 0 & r_{ij}\sin(\phi_{ij}) \\ 0 & 0 & 0 & 1 & r_{ij}\cos(\phi_{ij}) \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} log(\sigma_i) \\ \theta_i \\ x_i \\ y_i \\ 1 \end{bmatrix} \tag{B.1}$$

with parameters $d\sigma_{ij}$, $d\theta_{ij}$, r_{ij} and ϕ_{ij}. This can be seen in Figure B.1.

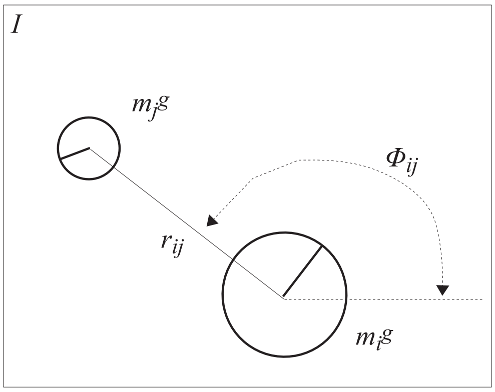


Fig. B.1 Inter-feature geometrical transform from feature m_i to m_j.

When geometry $m^g_{i'}$ of feature i is identified in a new image I', it can be used to predict the feature geometry $m^g_{j'}$ in I', as illustrated in Figure B.2.

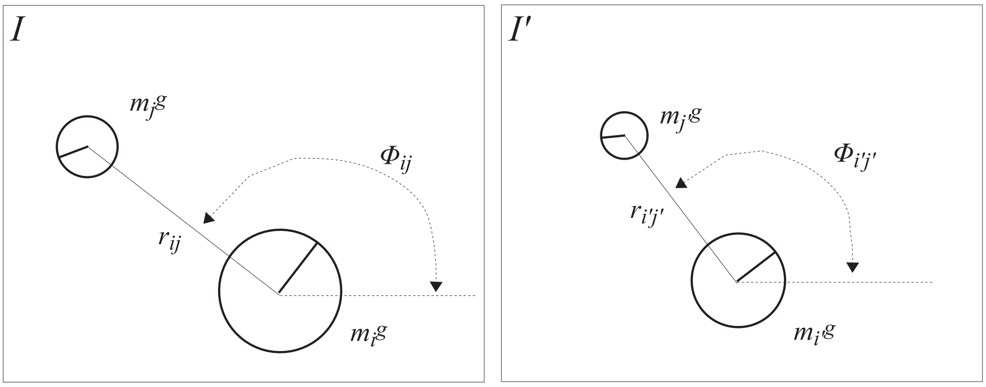


Fig. B.2 Predicting $m^g_{j'}$ in image I', based on $m^g_{i'}$ in image I', and $m^g_{j'}$ and $m^g_{j'}$ in image I.

This can be done by defining a transform $t_{i'j'} : m^g_{i'} \rightarrow m^g_{j'}$ based on transformation t_{ij} in equation (B.1) as follows. Let $m^g_{i'}$ be the geometry feature i in image I'. The parameters of transform $t_{ii'} : m^g_i \rightarrow m^g_{i'}$ are first estimated:

$$\begin{bmatrix} log(\sigma_{i'}) \\ \theta_{i'} \\ x_{i'} \\ y_{i'} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \log d\sigma_{ii'} \\ 0 & 1 & 0 & 0 & d\theta_{ii'} \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} log(\sigma_i) \\ \theta_i \\ x_i \\ y_i \\ 1 \end{bmatrix} \tag{B.2}$$

Note that $r_{ii'} = 0$, and $t_{ii'}$ is therefore defined by two parameters $d\sigma_{ii'}$ and $d\theta_{ii'}$. Now, the relative scale and orientation of features i and j do not change from I to I', and thus:

$$d\sigma_{i'j'} = d\sigma_{ij}, \tag{B.3}$$

$$d\theta_{i'j'} = d\theta_{ij}. \tag{B.4}$$

The distance and angle separating feature i and j have undergone the scale and orientation changes from I to I'. Thus,

$$r_{i'j'} = d\sigma_{ii'} r_{ij}, \tag{B.5}$$

$$\phi_{i'j'} = d\theta_{ii'} + \phi_{ij}. \tag{B.6}$$

The transform $t_{i'j'} : m^g_{i'} \rightarrow m^g_{j'}$ is thus:

$$\begin{bmatrix} log(\sigma_{j'}) \\ \theta_{j'} \\ x_{j'} \\ y_{j'} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \log d\sigma_{ij} \\ 0 & 1 & 0 & 0 & d\theta_{ij} \\ 0 & 0 & 1 & 0 & d\sigma_{ii'} r_{ij} \sin(d\theta_{ii'} + \phi_{ij}) \\ 0 & 0 & 0 & 1 & d\sigma_{ii'} r_{ij} \cos(d\theta_{ii'} + \phi_{ij}) \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} log(\sigma_{i'}) \\ \theta_{i'} \\ x_{i'} \\ y_{i'} \\ 1 \end{bmatrix} \tag{B.7}$$

References

[1] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of models for visual recognition," *IJCV*, vol. 71, no. 3, pp. 273–303, 2006.

[2] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their localization in images," in *ICCV*, pp. 370–377, 2005.

[3] A. Kushal, C. Schmid, and J. Ponce, "Flexible object models for category-level 3D object recognition," in *CVPR*, 2007.

[4] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool, "Towards multi-view object class detection," in *CVPR*, 2006.

[5] S. Baluja and H. A. Rowley, "Boosting sex identification performance," *IJCV*, vol. 71, no. 1, pp. 111–119, 2007.

[6] Z. Yang, M. Li, and H. Ai, "An experimental study on automatic face gender classification," in *ICPR*, pp. 1099–1102, 2006.

[7] A. Jain, J. Huang, and S. Fang, "Gender identification using frontal facial images," in *ICME*, 2005.

[8] B. Moghaddam and M. Yang, "Learning gender with support faces," *IEEE TPAMI*, vol. 24, no. 5, pp. 707–711, 2002.

[9] S. Gutta, H. Wechsler, and P. Phillips, "Gender and ethnic classification of human faces using hybrid classifiers," in *Int. Conf. on Automatic Face and Gesture Recognition*, pp. 194–199, 1998.

[10] D. L. Collins and A. C. Evans, "ANIMAL: Validation and applications of nonlinear registration-based segmentation," *IJPRAI*, vol. 11, no. 8, pp. 1271–1294, 1997.

[11] J. P. Thirion, "Image matching as a diffusion process: an analogy with Maxwells demons," *MIA*, vol. 2, no. 3, pp. 242–260, 1998.

[12] D. Reuckert, *Nonrigid Registration: Concepts, Algorithms and Applications*, ch. 13, pp. 281–301. CRC Press, 2003.

[13] J. Mangin, D. Riviere, A. Cachia, E. Duchesnay, Y. Cointepas, D. Papadopoulos-Orfanos, D. L. Collins, A. C. Evans, and J. Regis, "Object-based morphometry of the cerebral cortex," *IEEE TMI*, vol. 23, no. 8, pp. 968–983, 2004.

[14] J. Ashburner and K. J. Frison, "Voxel-based morphometry-the methods," *NeuroImage*, vol. 11, no. 23, pp. 805–821, 2000.

[15] A. W. Toga, P. M. Thompson, M. S. Mega, K. L. Narr, and R. E. Blanton, "Probabilistic approaches for atlasing normal and disease-specific brain variability," *Anat Embryol*, vol. 204, pp. 267–282, 2001.

[16] P. M. Thompson, J. N. Giedd, R. P. Woods, D. MacDonald, A. C. Evans, and A. W. Toga, "Growth patterns in the developing human brain detected using continuum-mechanical tensor mapping.," *Nature*, vol. 404, no. 6774, pp. 190–19, 2000.

[17] U. Grenander and M. I. Miller, "Computational anatomy: An emerging discipline," *Quarterly of Applied Mathematics*, vol. LVI, no. 4, pp. 617–693, 1998.

[18] F. L. Bookstein, "voxel-based morphometry should not be used with imperfectly registered images," *NeuroImage*, vol. 14, pp. 1454–1462, 2001.

[19] J. Ashburner and K. J. Friston, "Comments and controversies: Why voxel-based morphometry should be used," *NeuroImage*, vol. 14, pp. 1238–1243, 2001.

[20] C. Davatzikos, "Comments and controversies: Why voxel-based morphometric analysis should be used with great caution when characterizing group differences," *NeuroImage*, vol. 23, pp. 17–20, 2004.

[21] E. Duchesnay, A. Cachia, A. Roche, D. Riviere, Y. Cointepas, D. Papadopoulos-Orfanos, M. Zilbovicius, J. Martinot, J. Regis, and J. Mangin, "Classification based on cortical folding patterns," *IEEE TMI*, vol. 26, no. 4, pp. 553–565, 2007.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[23] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, 2004.

[24] A. Yuille, D. Cohen, and P. Hallinan, "Feature extraction from faces using deformable templates," in *CVPR*, pp. 104–109, 1989.

[25] K. Yow and R. Cipolla, “Detection of human faces under scale, orientation, and viewpoint variations,” in *Int. Conf. on Automatic Face and Gesture Recognition*, 1996.

[26] I. Beiderman, “Recognition-by-components: A theory of human image understanding,” *Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.

[27] M. Turk and A. P. Pentland, “Eigenfaces for recognition,” *CogNeuro*, vol. 3, no. 1, pp. 71–96, 1991.

[28] S. K. Nayar, S. A. Nene, and H. Murase, “Subspace methods for robot vision,” *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 750–758, 1996.

[29] T. Cootes, G. Edwards, and C. Taylor, “Active appearance models,” *IEEE PAMI*, vol. 23, pp. 681–684, June 2001.

[30] S. Ullman, M. Vidal-Naquet, and E. Sali, “Visual features of intermediate complexity and their use in classification,” *Nature Neuroscience*, no. 5, pp. 682–687, 2002.

[31] M. Toews and T. Arbel, “Detection over viewpoint via the object class invariant,” in *ICPR*, vol. 1, pp. 765–768, 2006.

[32] M. Toews and T. Arbel, “Detecting, localizing and classifying visual traits from arbitrary viewpoints using probabilistic local feature modeling,” in *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, vol. 4778, pp. 154–167, LNCS, 2007.

[33] M. Toews and T. Arbel, “Detecting and localizing 3D object classes using viewpoint invariant reference frames,” in *ICCV Workshop on 3D Representation for Recognition*, 2007.

[34] M. Toews, D. L. Collins, and T. Arbel, “A statistical parts-based appearance model of inter-subject variability,” in *MICCAI*, vol. I, pp. 232–240, 2006.

[35] M. Toews and T. Arbel, “A statistical parts-based appearance model of anatomical variability,” *IEEE TMI - Special Issue on Computational Neuroanatomy*, vol. 26, no. 4, 2007.

[36] M. Toews and T. Arbel, *Parts-based Appearance Modeling of Medical Imagery*. CRC Press, To Appear: 2008.

[37] CMU Face Group, “Frontal and profile face databases.” http://vasc.ri.cmu.edu/idb/html/face/.

[38] “Color FERET face database.” www.itl.nist.gov/iad/humanid/colorferet.

[39] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool, "The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results." www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

[40] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer, "A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM)," *Philos Trans R Soc Lond B Biol Sci*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[41] J. Koenderink and A. van Doorn, "Receptive field families," *Biological Cybernetics*, vol. 63, pp. 291–297, 1990.

[42] R. N. Bracewell, *The Fourier Transform and its Applications.* McGraw-Hill, 3 ed., 2000.

[43] B. Schiele and J. L. Crowley, "Recognition without correspondence using multidimensional receptive field histograms," *IJCV*, vol. 36, no. 1, pp. 31–50, 2000.

[44] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene analysis," *IEEE Transactions on Computers*, vol. C-20, pp. 562–569, 1971.

[45] J. Canny, "A computational approach to edge detection," *IEEE TPAMI*, vol. 8, no. 6, pp. 679–698, 1986.

[46] L. Kitchen and A. Rosenfeld, "Gray level corner detection," *PRL*, vol. 1, pp. 95–102, December 1982.

[47] H. P. Moravec, "Visual mapping by a robot rover," in *Proc. of the 6th International Joint Conference on Artificial Intelligence*, pp. 598–600, 1979.

[48] C. Tomasi and J. Shi, "Good features to track," in *CVPR*, pp. 593–600, 1994.

[49] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, 1988.

[50] Y. Lamdan, J. T. S. Schwartz, and H. J. Wolfson, "Affine invariant model-based object recognition," *IEEE Trans. on Robotics and Automation*, vol. 6, no. 5, pp. 578–588, 1990.

[51] D. W. Thompson and J. L. Mundy, "Three dimensional model matching from an unconstrained viewpoint," in *International Conference on Robotics and Automation*, pp. 208–220, 1987.

[52] D. Forsyth, J. L. Mundy, and A. Zisserman, "Invariant descriptors for 3-d object recognition and pose," *IEEE TPAMI*, vol. 13, no. 10, pp. 971–991, 1991.

[53] T. Lindeberg, "Feature detection with automatic scale selection," *IJCV*, vol. 30, no. 2, pp. 79–116, 1998.

[54] L. Bretzner and T. Lindeberg, "Feature tracking with automatic selection of spatial scales," *CVIU*, vol. 71, pp. 385–392, September 1998.

[55] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE TPAMI*, vol. 19, pp. 530–535, May 1997.

[56] G. Carneiro and A. Jepson, "Multi-scale phase-based local features," in *CVPR*, vol. 1, pp. 736–743, 2003.

[57] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *ICCV*, pp. I: 525–531, 2001.

[58] T. Kadir and M. Brady, "Saliency, scale and image description," *IJCV*, vol. 45, no. 2, pp. 83–105, 2001.

[59] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," *CVIU*, vol. 94, no. 1-3, pp. 3–27, 2004.

[60] T. Tuytelaars and L. Van Gool, "Matching widely separated views based on affinely invariant neighbourhoods," *IJCV*, vol. 59, no. 1, pp. 61–85, 2004.

[61] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *ECCV*, pp. 128–142, 2002.

[62] J. Matas, S. Obdrzalek, and O. Chum, "Local affine frames for wide-baseline stereo," in *ICPR*, vol. 4, pp. 363–366, 2002.

[63] A. Baumberg, "Reliable feature matching across widely separated views," in *CVPR*, pp. 774–781, 2000.

[64] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *BMVC*, 2002.

[65] D. Marr and E. Hildreth, "Theory of edge detection," *The Royal Society Proceedings-B*, vol. 207, pp. 187–217, 1980.

[66] A. Opelt, A. Pinz, and A. Zisserman, "A boundary-fragment-model for object detection," in *ECCV*, 2006.

[67] J. Burns, R. Weiss, and E. Riseman, “View variation of point-set and line-segment features,” *IEEE TPAMI*, vol. 15, no. 1, pp. 51–68, 1993.

[68] A. Vedaldi and S. Soatto, “Features for recognition: Viewpoint invariance for non-planar scenes,” in *ICCV*, vol. 2, pp. 1474–1481, October 2005.

[69] F. Bookstein, “Principle warps: thin-plate splines and the decomposition of deformations,” *IEEE TPAMI*, vol. 11, no. 6, pp. 567–585, 1989.

[70] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis.* John Wiley & Sons, 1998.

[71] B. t. H. Romeny, *Front-End Vision and Multi-Scale Image Analysis.* Kluwer Academic Publisher, 2003.

[72] D. J. Field, “What is the goal of sensory coding?,” *Neural Computation*, vol. 6, no. 4, pp. 559–601, 1994.

[73] L. Sirovich and M. Kirby, “Low-dimensional procedure for the characterization of human faces,” *Journal of the Optical Society of America*, vol. 4, no. 3, pp. 519–524, 1987.

[74] D. Willshaw, O. Buneman, and H. Longuet-Higgins, “Non-holographic associative memory,” *Nature*, vol. 222, pp. 960–962, 1969.

[75] B. A. Olshausen and D. J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?,” *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[76] B. Olshausen and D. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, pp. 607–609, 1996.

[77] P. Common, “Independent component analysis, a new concept?,” *Signal Processing, special issue on higher-order statistics*, vol. 36, no. 3, pp. 287 – 314, 1994.

[78] A. J. Bell and T. J. Sejnowski, “The independent components of natural scenes are edge filters,” *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[79] D. Lowe, “Object recognition from local scale-invariant features,” in *ICCV*, pp. 1150–1157, 1999.

[80] K. Mikolajczk and C. Schmid, “A performance evaluation of local descriptors,” in *CVPR*, vol. 2, pp. 257–263, 2003.

[81] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE TPAMI*, vol. 27, pp. 1615–1630, October 2005.

[82] P. Moreels and P. Perona, “Evaluation of features detectors and descriptors based on 3D objects,” *IJCV*, vol. 73, no. 3, pp. 263–284, 2007.

[83] K. Yan and R. Sukthankar, “PCA-SIFT: a more distinctive representation for local image descriptors,” in *CVPR*, vol. 2, pp. 506–513, 2004.

[84] K. Mikolajczyk, B. Leibe, and B. Schiele, “Local features for object class recognition,” in *ICCV*, 2005.

[85] G. Hua, M. Brown, and S. Winder, “Discriminant embedding for local image descriptors,” in *ICCV*, 2007.

[86] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*, 2005.

[87] S. Winder and M. Brown, “Learning local image descriptors,” in *CVPR*, 2007.

[88] A. Roche, G. Malandain, N. Ayache, and S. Prima, “Toward a better comprehension of similarity measures used in medical image registration,” in *MICCAI*, pp. 555–566, 1999.

[89] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* Wiley, 2nd ed., 2001.

[90] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *IJCV*, vol. 40, no. 2, pp. 99–121, 2004.

[91] P. Viola and W. Wells, III, “Alignment by maximization of mutual information,” *IJCV*, vol. 24, pp. 137–154, September 1997.

[92] A. Collignon, *Multi-modality medical image registration by maximization of mutual information.* PhD thesis, Catholic University of Leuven, Leuven, Belgum, 1998.

[93] J. Pluim, J. Maintz, and M. Viergever, “Mutual-information-based registration of medical images: a survey,” *MedImg*, vol. 22, pp. 986–1004, August 2003.

[94] A. Roche, G. Malandain, X. Pennec, and N. Ayache, “The correlation ratio as a new similarity measure for multimodal image registration,” in *MICCAI*, pp. 1115–1124, 1998.

[95] M. Toews, D. L. Collins, and T. Arbel, “Maximum a posteriori local histogram estimation for image registration,” in *MICCAI*, pp. 163–170, 2005.

[96] P. Rogelj, S. Kovacic, and J. C. Gee, “Point similarity measures for non-rigid registration of multi-modal data.,” *CVIU*, vol. 92, pp. 112–140, October 2003.

[97] F. Jurie and E. Nowak, "Learning visual similarity measures for comparing never seen objects," in *CVPR*, 2007.

[98] R. Hartley and A. Zisserman, *Multiple View Geometry.* Cambridge, New York: Cambridge University Press, 2nd ed., 2003.

[99] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision.* New Jersey: Prentice-Hall, Inc, 1998.

[100] H. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, 1981.

[101] F. Riggi, M. Toews, and T. Arbel, "Fundamental matrix estimation via tip - transfer of invariant parameters," in *ICPR*, vol. 2, pp. 21–24, 2006.

[102] M. Toews and T. Arbel, "Entropy-of-likelihood feature point selection for image correspondence," in *ICCV*, pp. 1041–1047, 2003.

[103] A. Pope and D. G. Lowe, "Probabilistic models of appearance for object recognition," *IJCV*, vol. 40, no. 2, pp. 149–167, 2000.

[104] W. E. L. Grimson and D. P. Huttenlocher, "On the sensitivity of geometric hashing," in *ICCV*, pp. 334–338, 1990.

[105] D. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[106] W. E. L. Grimson and D. P. Huttenlocher, "On the sensitivity of the hough transform for object recognition," *IEEE TPAMI*, vol. 12, pp. 255–274, March 1990.

[107] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[108] P. Moreels, M. Maire, and P. Perona, "Recognition by probabilistic hypothesis construction," in *ECCV*, pp. 55–68, 2004.

[109] D. Lowe, "Local feature view clustering for 3D object recognition," in *CVPR*, pp. 682–688, 2001.

[110] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints," in *CVPR*, pp. 272–277, 2003.

[111] J. Pearl, *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers, Inc., 1988.

[112] G. Edwards, T. Cootes, and C. Taylor, "Face recognition using active appearance models," in *ECCV*, pp. 581–595, 1998.

[113] S. M. Pizer, P. T. Fletcher, S. Joshi, A. Thall, J. Z. Chen, Y. Fridman, D. S. Fritsch, A. G. Gash, J. M. Glotzer, M. Jiroutek, C. Lu, K. E. Muller, G. Tracton, P. Yushkevich, and E. L. Chaney, "Deformable M-Reps for 3D medical image segmentation," *IJCV*, vol. 55, pp. 85–106, Nov.-Dec. 2003.

[114] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool, "The pascal visual object classes challenge 2006 (voc2006) results," tech. rep., September 2006.

[115] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *ECCV*, pp. I: 18–32, 2000.

[116] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *ICCV*, 2005.

[117] D. Crandal, P. Felzenswalb, and D. Huttenlocher, "Spatial priors for part-based recognition using statistical models," in *CVPR*, vol. 1, pp. 10–17, 2005.

[118] G. Carneiro and D. G. Lowe, "Sparse flexible models of local features," in *ECCV*, vol. III, pp. 29–43, 2006.

[119] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *CVPR*, pp. 264–271, 2003.

[120] E. Bart, E. Byvatov, and S. Ullman, "View-invariant recognition using corresponding object fragments," in *ECCV*, pp. 152–165, 2004.

[121] S. Helmer and D. G. Lowe, "Object recognition with many local features," in *Workshop on Generative Model Based Vision*, 2004.

[122] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proceedings of the Workshop on Statistical Learning in Computer Vision*, 2004.

[123] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE TPAMI*, vol. 28, no. 3, 2006.

[124] G. Dorko and C. Schmid, "Object class recognition using discriminative local features," tech. rep., INRIA - Rhone-Alpes, February 2005.

[125] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *ICCV*, 2005.

[126] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *ICCV*, vol. II, pp. 90–96, 2003.

[127] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual categorization with bags of keypoints," in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[128] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[129] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, vol. 1, pp. 511–518, 2001.

[130] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, 1996.

[131] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE TPAMI*, vol. 20, no. 1, pp. 39–51, 1998.

[132] H. Schneiderman and T. Kanade, "Object detection using the statistics of parts," *IJCV*, vol. 56, no. 3, pp. 155–177, 2004.

[133] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Mathematische Annalen*, vol. 69, pp. 331–371, 1910.

[134] M. Jones and P. Viola, "Fast multi-view face detection," in *Technical Report TR2003-96*, pp. 1–10, MERL, 2003.

[135] P. Wang and J. Quiang, "Learning descriminant features for multi-view face and eye detection," in *CVPR*, 2005.

[136] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE TPAMI*, vol. 20, no. 1, pp. 23–38, 1998.

[137] K. Mikolajczyk, B. Leibe, and B. Schiele, "Multiple object class detection with a generative model," in *CVPR*, 2006.

[138] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," vol. 28, no. 4, pp. 594–611, 2006.

[139] B. Evgeniy and S. Ullman, "Cross-generalization: learning novel classes from a single example by feature replacement," in *CVPR*.

[140] T. Cootes, K. Walker, and C. Taylor, "View-based active appearance models," in *Int. Conf. on Face and Gesture Recognition*, pp. 227–232, 2000.

[141] G. Gill and M. Levine, "A single classifier for view-invariant multiple object class recognition," in *BMVC*, 2006.

[142] M. Weber, W. Einhauser, M. Welling, and P. Perona, "Viewpoint-invariant learning and detection of human heads," in *Int. Conf. on Automatic Face and Gesture Recognition*, pp. 7–20, 2000.

[143] I. Beiderman and P. C. Gerhardstein, "Recognizing depth-rotated objects: Evidence and conditions for 3D viewpoint invariance," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, pp. 1162–1182, 1993.

[144] M. J. Tarr and H. H. Bulthoff, *Object Recognition in Man, Monkey and Machine.* MIT/Elsevier, 1998.

[145] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *CVPR*, pp. 762–769, 2004.

[146] T. Arbel and F. P. Ferrie, "Entropy-based gaze planning," *Image and Vision Computing*, vol. 19, no. 11, pp. 779–786, 2001.

[147] D. Weinshall and M. Werman, "On view likelihood and stability," *IEEE TPAMI*, vol. 19, no. 2, pp. 97–108, 1997.

[148] J. Koenderink, "The internal representation of solid shape with respect to vision," *Biological Cybernetids*, vol. 32, no. 4, pp. 211–216, 1979.

[149] S. J. Dickinson, A. P. Pentland, and A. Rosenfeld, "Qualitative 3-D shape reconstruction using distributed aspect graph matching," in *ICCV*, (Osaka,Japan), pp. 257–262, IEEE Computer Society Press, Dec. 1990.

[150] J. Burns, R. Weiss, and E. Riseman, "The non-existence of general-case view-invariants," in *Geometric Invariance in Computer Vision*, pp. 120–131, 1992.

[151] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *ECCV*, pp. 69–81, 2004.

[152] T. Binford, "Visual perception by computer," in *IEEE Conference on Systems, Man and Cybernetics*, 1971.

[153] J. Ponce, D. Chelberg, and W. Mann, "Invariant properties of straight homogeneous generalized cylinders and their contours," *IEEE TPAMI*, vol. 11, no. 9, pp. 951–966, 1989.

[154] G. Dorko and C. Schmid, "Selection of scale-invariant parts for object class recognition," in *ICCV*, pp. 634–640, 2003.

[155] S. Savarese and L. Fei-Fei, "3D generic object categorization, localization and pose estimation," in *ICCV*, 2007.

[156] P. Yan, S. M. Khan, and M. Shah, "3D model based object class detection in an arbitrary view," in *ICCV*, 2007.

[157] N. Ahuja and S. Todorovic, "Learning the taxonomy and models of categories present in arbitrary images," in *ICCV*, pp. 1–8, 2007.

[158] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *CVPR*, 2006.

[159] H.-C. Kim, D. Kim, Z. Ghahramani, and S. Y. Bang, "Appearance-based gender classification with gaussian processes," *PRL*, vol. 27, pp. 618–626, 2006.

[160] C. BenAbdelkader and P. Griffin, "A local region-based approach to gender classification from face images," in *CVPR*, 2005.

[161] A. Lapedriza, D. Masip, and J. Vitria, "Are external face features useful for automatic face classification?," in *CVPR*, 2005.

[162] G. Shakhnarovich, P. A. Viola, and B. Moghaddam, "A unified learning framework for real time face detection and classification," in *Int. Conf. on Automatic Face and Gesture Recognition*, 2002.

[163] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. Collins, A. Evans, G. Malandain, N. Ayache, G. Christensen, and H. Johnson, "Retrospective evaluation of intersubject brain registration," *IEEE TMI*, vol. 22, pp. 1120–1130, September 2003.

[164] R. Bajcsy and S. Kovacic, "Multiresolution elastic matching," *Computer Vision, Graphics and Image Processing*, vol. 46, pp. 1–21, 1989.

[165] M. Bro-Nielsen and C. Gramkow, "Fast fluid registration of medical images," in *Visualization in Biomedical Computing*, pp. 267–276, 1996.

[166] J. Gee, D. Haynor, M. Reikvich, and R. Bajcsy, "Finite element approach to warping of brain images," in *SPIE Medical Imaging 1994: Image Processing*, vol. 2167, pp. 18–27, 1994.

[167] F. Bookstein, "Thin-plate splines and the atlas problem for biomedical images," in *IPMI*, pp. 326–342, 1991.

[168] C. J. Twining, T. Cootes, S. Marsland, V. Petrovic, R. Schestowitz, and C. J. Taylor, "A unified information-theoretic approach to groupwise non-rigid registration and model building," in *IPMI*, 2005.

[169] S. Joshi, B. David, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," *NeuroImage*, vol. LVI, no. 23, pp. 151–160, 2004.

[170] R. Beichel, H. Bischof, F. Leberl, and M. Sonka, "Robust active appearance models and their application to medical image analysis," *IEEE TMI*, vol. 24, pp. 1151–1169, September 2005.

[171] D. Rueckert, A. F. Frangi, and J. A. Schnabel, "Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration," *IEEE TMI*, vol. 22, no. 8, pp. 1014–1025, 2003.

[172] E. Luders, K. L. Narr, Z. Eran, P. M. Thompson, and A. W. Toga, "Gender effects on callosal thickness in scaled and unscaled space," *Brain Imaging*, vol. 17, no. 11, pp. 1103–1106, 2006.

[173] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. M. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimentional shape transformations and machine learning methods," *NeuroImage*, vol. 21, pp. 46–57, 2004.

[174] P. Cachier, J.-F. Mangin, X. Pennec, D. Riviere, D. Papadopoulos-Orfancs, J. Regis, and N. Ayache, "Multisubject non-rigid registration of brain MRI using intensity and geometric features," in *MICCAI*, pp. 734–742, 2001.

[175] D. Riviere, J.-F. Mangin, D. Papadopoulos-Orfanos, J.-M. Martinez, V. Frouin, and J. Regis, "Automatic recognition of cortical sulci of the human brain using a congregation of neural networks," *MIA*, vol. 6, pp. 77–92, 2002.

[176] L. M. Lui, Y. Wang, T. F. Chan, and P. M. Thompson, "Automatic landmark tracking applied to optimize brain conformal mapping," in *ISBI*, 2006.

[177] K. Rohr, "On 3D differential operators for detecting point landmarks," *Image and Vision Computing*, vol. 15, no. 3, pp. 219–233, 1997.

[178] W. R. Fright and A. D. Linney, "Registration of 3-d head surfaces," *IEEE TMI*, vol. 12, no. 3, pp. 515–520, 1993.

[179] K. Rohr, H. S. Stiehl, R. Sprengel, T. M. Buzug, J. Weese, and M. H. Kuhn, "Landmark-based elastic registration using approximating thin-plate splines," *IEEE TMI*, vol. 20, no. 6, pp. 526–534, 2001.

[180] Y. Shi, F. Qi, Z. Xue, K. Ito, H. Matsuo, and D. Shen, "Segmenting lung fields in serial chest radiographs using both population and patient-specific shape statistics," in *MICCAI*, 2006.

[181] D. Burschka, M. Li, R. Taylor, G. D. Hager, and M. Ishii, "Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery.," *MIA*, vol. 9, no. 5, pp. 413–439, 2005.

[182] W. Guorong, F. Qi, and D. Shen, "Learning best features for deformable registration of MR brains," in *MICCAI*, 2005.

[183] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV*, pp. 372–387, 2003.

[184] W. Cheung and G. Hamarneh, "N-sift: N-dimensional scale invariant feature transform for matching medical images," in *ISBI*, 2007.

[185] M. Ono, S. Kubik, and C. D. Abernathy, *Atlas of the Cerebral Sulci.* New York: Thieme Medical, 1990.

[186] T. P. Naidich, A. G. Valavanis, and S. Kubik, "Anatomic relationships along the low-middle convexity: Part I-Normal specimens and magnetic resonance imaging," *Neurosurgery*, vol. 36, no. 3, pp. 517–532, 1995.

[187] P. Hellier and C. Barillot, "Coupling dense and landmark-based approaches for non rigid registration," *IEEE TMI*, vol. 22, no. 2, pp. 217–227, 2003.

[188] S. Zheng, Z. Tu, A. L. Yuille, A. L. Reiss, R. A. Dutton, A. D. Lee, A. M. Galaburda, P. M. Thompson, I. Dinov, and A. W. Toga, "A learning based algorithm for automated extraction of the cortical sulci," in *MICCAI*, pp. 241–248, 2006.

[189] J. Talairach and P. Tournoux, *Co-planar Stereotactic Atlas of the Human Brain: 3-Dimensional Proportional System: an Approach to Cerebral Imaging.* Stuttgart: Georg Thieme Verlag, 1988.

[190] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE TPAMI*, vol. 24, no. 5, 2002.

[191] M. I. Jordan, *An Introduction to Probabilistic Graphical Models.* In preparation, 2003.

[192] E. Jaynes, "Prior probabilities," *IEEE Transactions on systems, science, and cybernetics*, vol. SSC-4, no. 3, pp. 227–241, 1968.

[193] D. Lowe, "SIFT keypoint detector." http://www.cs.ubc.ca/ lowe/keypoints/.

[194] C. Schmid, R. Mohr, and C. Bauckhage, “Evaluation of interest point detectors,” *IJCV*, vol. 37, pp. 151–172, June 2000.

[195] S. Agarwal, A. Awan, and D. Roth, “Learning to detect objects in images via a sparse, part-based representation,” *IEEE TPAMI*, vol. 26, no. 11, pp. 1475–1490, 2004.

[196] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[197] C.-C. Chang and C.-J. Lin, “LIBSVM – a library for support vector machines.” www.csie.ntu.edu.tw/ cjlin/libsvm.

[198] I. Laptev, “Improvements of object detection using boosted histograms,” in *BMVC*, 2006.

[199] V. Viitaniemi and J. Laaksonen, “Techniques for still image scene classification and object detection,” in *International Conference on Artificial Neural Networks*, 2006.

[200] J. Laaksonen, M. Koskela, and E. Oja, “Picsom self-organizing image retrieval with mpeg-7 content descriptors,” *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 841–853, 2002.

[201] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost: Joint appearance, shape and context modelling for multi-class object recognition and segmentation,” in *ECCV*, pp. I:1–15, 2006.

[202] M. Fritz, B. Leibe, B. Caputo, and B. Schiele, “Integrating representative and discriminant models for object category detection,” in *ICCV*, 2005.

[203] D. L. Collins, N. Kabani, and A. Evans, “Automatic volume estimation of gross cerebral structures,” in *4th International Conference of Functional Mapping of the Human Brain* (A. Evans, ed.), 1998.

[204] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, “Automatic 3D inter-subject registration of MR volumetric data in standardized talairach space,” *Journal of Computer Assisted Tomography*, vol. 18, no. 2, pp. 192–205, 1994.

[205] J. V. Hajnal, D. L. Hill, and D. J. Hawkes, *Medical Image Registration.* CRC Press, 2003.

[206] T. Cootes, “AM Tools.” http://www.isbe.man.ac.uk/ bim/software.

[207] C. DeLacoste-Utamsing and R. Holloway, “Sexual dimorphism in the human corpus callosum,” *Science*, vol. 216, no. 4553, pp. 1431–1432, 1982.

[208] S. Smith, “Fast robust automated brain extraction,” *Human Brain Mapping*, vol. 17, pp. 143–155, 2002.

[209] C. Rorden, “MRIcro.” www.sph.sc.edu/comd/rorden/mricro.html.

[210] W. Penfield and L. Roberts, *Speech and Brain Mechanisms.* Princeton Univ Press, 1959.

[211] M. Petrides, G. Cadoret, and S. Mackey, “Orofacial somatomotor responses in the macaque monkey homologue of Broca’s area,” *Nature*, vol. 435, pp. 1235–1238, 2005.

[212] S. Edelman, “Class similarity and viewpoint invariance in the recognition of 3D objects,” *Biological Cybernetics*, vol. 72, no. 3, pp. 207–220, 1995.