

1 About Face

*O why was I born with a different face?
Why was I not born like the rest of my race?*

— William Blake, *Letter to Thomas Butts*

1.1 The Visual Face

Faces play a vital role in our daily lives with their varied repertoire of intricate and often subtle functions. Our faces contain sense organs of which our eyes are the most useful in allowing us to sense the dynamic world around us. But what information do our faces convey to others? One way in which we use our faces to communicate is through the production of audible speech. However, the *visual* face alone conveys a plethora of useful information. Indeed, speech is accompanied by facial motion as a result of which lip-reading is possible. Facial expressions produced by visible deformation of the face provide us with a guide to judging mood, character and intent. We are constantly ‘reading’ one another’s faces and such processes clearly play an important part in social interaction and communication. All these changes in facial appearance due to expression and speech occur on a rather short time-scale. There are of course more enduring visual cues provided by a person’s face and these allow us to estimate such factors as age, gender, ethnic origin and *identity*. It is this last use, *as a visual cue for identification*, that will form the main thread of this book. We have begun by pointing out the other functions of faces because identification is inevitably performed in the presence of all the variations in facial appearance to which these give rise. Whilst your appearance changes all the time, your identity is one thing with which you are stuck!

Faces form a unique class of objects. They share a common structure: *the facial features are always configured in a similar way*. Given their similarity, it is remarkable that we can discriminate between so many different people based upon facial appearance alone. This seems even more remarkable when we consider that we perform recognition in the presence of widely varying viewing conditions.

The perception of faces is highly dynamic, both in space and time and with respect to a given context. In human perception and behaviour it is certainly the case that we use more than just static facial appearances in discriminating between different faces. Visual context constrains our expectation: the object sitting behind a desk is more likely to be human than the object on top of the bookcase! Other sources of information such as body gestures and gait also provide useful cues. It is important to realise that above all, the perception of faces is a spatio-temporal, *dynamic vision* task. Whilst it is true that we can often identify someone from a static photograph, in the real world we usually observe and identify faces in a dynamic setting. Facial appearance alters due to relative motion of the face and the observer. Changes in lighting and the environment also affect appearance. Therefore, face perception involves rather more than the perception of static pictures of faces. In particular, it would seem to require the ability to detect and track faces as they move through cluttered scenes which are themselves often complex and dynamic.

1.2 The Changing Face

The process of identifying a person from facial appearance has to be performed in the presence of many often conflicting factors which alter facial appearance and make the task difficult. It is therefore important to examine the sources of variation in facial appearance more closely. One can consider that there are two types of variation. A face can change its appearance due to either *intrinsic* or *extrinsic* factors. Intrinsic variation takes place independently of any observer and is due purely to the physical nature of the face. Extrinsic variation, on the other hand, arises when the face is *observed* via the interaction of light with the face and the observer. It is the intrinsic variations that one must learn to understand and interpret. This is largely what we mean by the *perception of faces*. Perceptual tasks have to be performed consistently and robustly under all sorts of changes

in external conditions characterised by the extrinsic sources of variation. In general, faces exhibit many degrees of intrinsic variability which are difficult to characterise analytically. Table 1.1. lists some rather obvious perceptual tasks and the corresponding intrinsic sources of variation. These are not independent and the list is by no means exhaustive. If these perceptual tasks are to be performed effectively, intrinsic variability must be analysed and modelled. The visual appearance of a face also varies due to a host of extrinsic factors such as those highlighted in Table 1.2.

Table 1.1 Intrinsic sources of variation in facial appearance.

Source	Possible Tasks
Identity	Classification, known-unknown, verification, full identification
Facial expression	Inference of emotion or intention
Speech	Lip-reading
Sex	Deciding whether male or female
Age	Estimating age

Table 1.2 Extrinsic sources of variation in facial appearance.

Source	Effects
Viewing geometry	Pose
Illumination	Shading, colour, self-shadowing, specular highlights
Imaging process	Resolution, focus, imaging noise, sampling of irradiant energy distribution, perspective effects
Other objects	Occlusion, shadowing, indirect illumination

Typically, the appearance of a face alters considerably depending on the illumination conditions and in particular due to self-shadowing. The characteristics of the camera (or eye) used to observe the face also affect the resulting image quality. Other objects present in the scene can cause occlusion and cast shadows as well as altering the nature of the incident

light. However, one of the most significant sources of variation is pose change. It is worth emphasising that the pose of a face is determined by the *relative* three-dimensional (3D) position and orientation of the observer. It is, therefore, an extrinsic rather than an intrinsic source of variation because viewing geometry requires the presence of an observer. The main cause of pose change is relative *rigid motion* between the observer and the subject. A face undergoes rigid motion when it changes its position and orientation in 3D space relative to the observer. However, a face can also undergo *non-rigid motion* when its 3D shape changes due for example to speech or facial expression. This results in intrinsic variation of appearance. Whilst these two types of motion usually occur together, it is more convenient to treat them separately. If perceiving faces implies interpreting the intrinsic variations in a manner which is invariant to other changes, such invariances are often best achieved if the extrinsic variations are modelled so that their effects can be negated. For instance, the perception of identity should ideally be pose invariant and as such, the ability to determine the pose of a face can play an important role in perceiving identity. In this context, rigid motion provides strong visual cues for understanding pose. Whilst non-rigid facial motion is also likely to provide useful cues for identification, it clearly plays an even more important role in communication and perception of expressions.

1.3 Computing Faces

Over the last quarter of a century, scientists and engineers have endeavoured to build machines capable of automatic face perception. This effort has been multi-disciplinary and has benefited from areas as varied as computer science, cognitive science, mathematics, physics, psychology and neurobiology. Computer-based face perception is becoming increasingly desirable for many applications including human-machine interfaces, multimedia, surveillance, security, teleconferencing, communication, animation, visually mediated interaction and anthropomorphic environments. Consequently, there has been a strong research effort during recent years in the study and development of computational models, algorithms and computer vision systems for automatic face perception [162, 163, 164, 165, 166].

Broadly speaking, in order to recognise faces or indeed any visual objects, one needs to resolve a *stimulus equivalence* problem. Visual stimuli in the form of images of a particular object or class of objects should have something in common that differentiates them from images of other objects. Such commonalities should exist regardless of most reasonable extrinsic and intrinsic changes. Computationally then, how can we represent and measure characteristics that remain unique to faces under different conditions? In fact, it is rather difficult and computationally unrealistic to consider a single, general solution to the problem of modelling faces undergoing all the intrinsic and extrinsic variations. In this book, we mainly focus on the perception of faces for identification under extrinsic variations such as pose change caused by movement. In other words, we assume that the computations are either approximately invariant to most intrinsic variations such as age and expression, or that intrinsic variation is constrained so as to have little effect on facial appearance. Although there is evidence that the perception of faces involves dedicated neural hardware, face perception has many aspects in common with our perception of moving objects in general. In order to understand the nature of *visual perception*, let us briefly introduce some of the processes involved.

In order to perceive objects, artificial and biological vision systems must solve two general problems: that of *segmentation* (also known as *parsing*), and that of *recognition*. The problem of segmentation involves computation to divide images into regions that correspond to bodies of physical objects in the scene. The problem of recognition is to label such bodies as instances of known objects. The segmentation problem is further addressed by two sub-problems known as the problem of *spatial grouping* (also known as *unit formation*) and the *correspondence problem* [326]. Whilst the process of grouping determines which image elements (pixels) belong to a single physical body, correspondence tries to establish associations over time between image elements that are representations of the same scene entity. Both these tasks are non-trivial to accomplish. This is especially true if the perceived objects are constantly in motion and can be partially occluded due to a change of viewpoint: the problem of the *curse of projection* due to the 3D world being under-constrained in its 2D images. Given highly incomplete information, for example, how does the perceptual process decide if a face seen now is the same face seen in the past?

Dynamic perception of faces is necessarily complex. One can readily expect that visual perception of moving objects such as human faces re-

lies upon a range of computational processes including for instance, the measurement of visual cues such as motion and colour, selective attention, face detection, pose estimation, view alignment, face tracking, modelling of identity and identification. Such *perceptual processes* are closely coupled since the information extracted and the computation involved in each process are intrinsically dependent upon those of the others. However, in order to understand the computations required when perceiving faces, it is convenient and even necessary to decompose such a process into a number of clearly definable tasks. Let us now examine such a decomposition in more detail.

Perceptual Grouping and Focus of Attention

In any visual scene there is always a large amount of information to process. A necessary computational task is *perceptual grouping* which *focuses attention* on areas in the field of view where faces are likely to be present. This task is essentially one of determining small attentional windows within the visual field where further computation should be directed. Pre-attentive visual cues such as motion and colour are useful for focusing attention. Notice that focus of attention need not involve determining whether faces are present or where exactly faces are located in the scene.

Detection

The function of determining the presence of a face and locating it within the scene is *face detection*. This task requires a model to discriminate faces from all other visual objects or patterns. It does not require any of a face's intrinsic variability to be interpreted. In particular, it does not involve identification. Face detection is also known as *basic-level* or *entry-level* recognition. It can also be considered to involve the tasks of face image segmentation and face alignment.

Tracking

In a dynamic and cluttered visual scene, the location and appearance of a face can change continually. The task of following a face through a visual scene requires *tracking* which in essence involves establishing *temporal correspondence*. Attentional windows and detected face image regions need to be constantly updated, maintaining an appropriate degree of correspon-

dence over time so that points of reference within these windows and regions are consistent.

Identification

Beyond entry-level recognition, the task of *identification* requires a function to discriminate between different faces. In the study of object recognition in general, such a task is often regarded as being *sub-ordinate level* or *within-category* recognition where all faces together constitute a category. It might seem that the problem of identification has just been defined. In fact, the exact nature of the identification task can vary quite significantly. Consider, for instance, a database consisting of a set, \mathcal{Y} , of M known faces. Several different identification tasks can be envisaged. In fact, at least four tasks can be defined as follows.

- (1) **Classification:** The task is to identify a face under the assumption that it is a member of \mathcal{Y} .
- (2) **Known-Unknown:** The task is to decide whether or not a face is a member of \mathcal{Y} .
- (3) **Verification:** The identity of a face is supplied by some other non-visual means and must be confirmed using face images. This is equivalent to the known-unknown task with $M = 1$.
- (4) **Full identification:** The task is to determine whether a face is a member of \mathcal{Y} and if so to determine its identity.

It seems clear that any computational treatment of the *dynamic perception of faces* will involve functions that must at least perform the above tasks well. As a result of decomposing face perception in this way we must address another issue in perceptual processing: that of integration.

Perceptual Integration

Despite that the perception of faces, in common with that of other dynamic visual objects, can be conveniently modelled as an assembly of sub-tasks performed independently by a set of functions, such functions can only be effective and even computationally viable if they are closely coupled. This requires the task of *perceptual integration* or *perceptual control*. Closely coupled information processing implies that the performance of any individual process is highly correlated to and dependent upon the effectiveness of the others. In contrast, conventional approaches to vision have often modelled

the computations as independent sequential processes. This has been motivated by the need for simplicity and tractability. However, overwhelming neurobiological evidence suggests that perception is only effective if it is performed by closely coupled, co-operative processes with feedbacks [143, 144].

Visual Learning and Adaptation

Humans are born with a certain innate knowledge of faces and subsequently learn to distinguish between different faces. How much knowledge of faces must be *hard-wired* into the process of face perception and how can this process then *bootstrap* using this knowledge so as to learn to recognise many different faces? Since visual information is always subject to noise and occlusion due to the generally ill-posed nature of inverse projection, model learning is often more difficult than expected. In fact, learned models need to be updated and tuned to specific recognition and tracking tasks. The ability to adapt previously learned models reflects another aspect of perceptual integration in which learned or hard-wired models are improved during the process of performing a perceptual task. The perception of moving objects and faces can also benefit greatly if recognition not only tracks and matches visual appearance with known models, but also interprets patterns of behaviour. This has been shown to be important in the perception of moving objects in general [244, 246].

Other functions are also deemed necessary although they may not be so intuitively apparent. For example, an effective system should exhibit the following functionality:

Learning from examples

A face must be known before it can subsequently be identified. It will therefore be necessary for a system to acquire identity models from observations. Such a *learning from examples* approach may be necessary not only for identification, but also for the other functions described above.

Understanding Pose

Among all the extrinsic and intrinsic sources of variation in facial appearance, pose changes are particularly problematic for the tasks of face detec-

tion, tracking and identification. One way of coping with pose change is to make pose explicit by estimating it. Although it may seem unnecessary to estimate pose explicitly in order to perform recognition in a manner that is invariant to pose, it will be useful computationally to consider *pose understanding* as a required task.

Real-time Computation

We must also consider performance issues related to computational complexity and accuracy. Many aspects of dynamic perception require computation to be performed on-line under certain time constraints and even in real-time. Real-time performance is often not a luxury but rather a necessity for visual interpretation of moving faces in dynamic scenes because correct perception needs to be performed within the spatio-temporal context. It is worth noting here that real-time does not imply that processing must be performed at full video frame-rate as long as computation proceeds correctly. In addition to constraints on computational speed, there will also be constraints of accuracy imposed on a system. Systems should be robust with graceful degradation of performance rather than complete failure.

Computer vision systems for automatic face recognition are now beginning to be deployed outside the laboratory in applications such as access control [190]. These systems typically assume a single face imaged at high resolution in frontal or near-frontal view. The environments in which they operate are highly constrained compared to the scenarios in which humans perform recognition. In fact, face recognition is a task which the human vision system seems to perform almost effortlessly in our everyday lives where the conditions are far more varied and challenging than those often artificially imposed for automated systems. Beyond the practical applications, another reason for building artificial face recognition systems is the belief that many of the computational and cognitive issues raised may provide insights into how object recognition in general is performed by the brain.

1.4 Biological Perspectives

This book is largely concerned with a computational approach to building artificial systems capable of perceiving and recognising dynamic faces. It is important to point out that the approach taken is not necessarily biologically plausible and we make no claims about furthering understanding of

face recognition and perception in any biological visual systems. However, it will be interesting and possibly even enlightening to draw some parallels with biological systems. We will therefore pause towards the end of each chapter to reflect upon evidence from psychophysical and neurobiological studies. The reader should note that such studies are often somewhat inconclusive by their very nature.

In general, psychophysical and behavioural studies investigate our ability to recognise faces under various experimentally controlled conditions. The conclusions which can be drawn are often rather limited. In particular, most studies have focused on examining static *picture recognition* rather than real *face recognition* [38]. Equally intriguing and interesting is evidence from neurobiological studies that there exist cells in the cortex dedicated to the interpretation and recognition of faces. Visual neurons in the superior temporal sulcus (STS) area of the temporal cortex of monkeys (and sheep) have been found which respond with good specificity to different aspects of facial information [263]. Some face-selective cells respond differently to different people's faces, although not usually only to one person's face [16]. Others are sensitive to eye gaze and pose. The responses of such cells could provide information upon which behavioural responses to different people's faces are based.

1.5 The Approach

Provided that perceptual integration can be properly addressed, dynamic perception of faces can be decomposed into a set of visual tasks which in turn are performed by appropriate computational functions. The aim of this book is to examine how this decomposition can be achieved effectively and to suggest appropriate functions. In recognition of the fact that such a decomposition is perhaps rather artificial, we also examine the means for perceptual integration and control. Furthermore, since computation is mediated by representation, it is necessary to address the role of representation and to determine how best to model faces so as to facilitate perception and recognition. It should be realised that the adoption of a particular representational scheme will constrain the choice of approach taken in determining the computational framework for building a system for face recognition. Representation schemes are discussed in Chapter 2.

A further aim of this book is to illustrate the means for engineering an effective artificial system capable of *learning to model* and subsequently recognising dynamic faces from example images. Chapter 3 introduces the reader to statistical machine learning and to a number of learning techniques used later in the book.

Part II describes methods for performing the tasks of perceptual grouping, focusing of attention, face detection and face tracking. These are performed in the presence of large pose changes and Chapter 6 is devoted to coping with this particular form of extrinsic variation.

In Part III, the task of identification is explored. Initially, this task is eased by constraining the pose to be approximately frontal and upright. Chapter 8 deals with identification under large pose changes and Chapter 9 considers the role of temporal information in identification.

Part IV addresses the problem of perceptual integration and explores potential roles for computing faces within the context of dynamic vision in broader terms. It describes ways in which tasks can be closely coupled for the perception of faces. The appendices include a treatment of databases for development, evaluation and application as well as a look at the current state of commercial applications. Finally, we give detailed descriptions of many of the algorithms used.