Estimation of Three Dimensional Structure from Passport-style Photographic Images for Enhanced Face Recognition Performance in Humans

Justen Hyde

A thesis submitted for the degree of

Doctor of Philosophy

in the

Department of Electronics

University of York

June 2006

Abstract

Pose and lighting variation are widely considered to be problematic for automatic face recognition, and in most real-world applications human recognition is considered to be more reliable. However, a body of research exists to suggest that this is not the case, and that pose and lighting variation can be equally problematic to humans. The problems posed in human face recognition are presented and the issues of lighting and pose specifically are experimentally explored experimentally, demonstrating a strong relationship between lighting and viewing angle change and increased error rates in an identification task. The effects are shown to be cumulative, with a compound lighting and angle of view change causing performance to fall to near chance. The use of a three dimensional model as a tool to match pose and lighting between images is suggested as a method to improve recognition, and a regularised covariance estimator proposed as a technique for the estimation of shape data from standard photographic images. The training of this estimator is described, and the resulting implemented estimator shown to produce depth images which describe the three dimensional structure of the test images with consistently lower mean square error than the average head model. This demonstrates that the technique provides a better conditional estimator than the global average for facial image plus model data. Subjective testing with the results of this estimator show that when the estimated model is used to match lighting and viewing angle between training and testing cases, error rates decrease and further that error rates are significantly better than the use of the mean head model to perform this correction.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Firstly I would like to thank my family for their encouragement, insight and all their help during my studies, especially my father for sitting through many clumsy expositions of my work as I tried to work out how best to explain, and hence write about, my research. I would like to thank my supervisor John Robinson for his support, enthusiasm and guidance, Dan Parnham for writing the PHP server side script for the distributed regularistion coefficient optimisation software, Enrico Costanza for asking very interesting questions on a daily basis, Tom Heseltine for his assistance with the capture of the York data set, the University of Notre-Dame for the provision of their 3d face data, and the various Department of Psychology academics and students (especially Andy Young and Ruth Harrison) who so kindly gave up some of their time in order to help me to get to grips with an unfamiliar discipline.

Finally I would like to thank my wife Lucy for her love, support and saint-like patience with my many vagaries during the creation of this thesis. I couldn't have done this without her.

Declaration

I declare that all the work presented in this thesis is solely my own except where attributed and cited to another author. Some of the material in the thesis has been previously published by the author. The relevant papers are reproduced with details of original publication in Appendix C.

Introduction

Face recognition when not carried out by a machine is often overlooked in image processing circles. Whilst the issues of varying conditions such as pose, expression and lighting receive a great deal of attention from the perspective of designing automated recognition systems which are reliable, human recognition is neglected.

The implication is either that human recognition is satisfactory as is, or that there is nothing which may be contributed from the image processing point of view. However, the unreliability of human recognition especially when the face to be recognised is unfamiliar to the observer, is well understood by psychologists, even as the actual mechanisms of face recognition remain a area of intense research.

It may not be necessary to fully understand a mechanism to be able to render assistance to it, however. By identifying conditions which degrade human performance, it may be possible to use image processing techniques to artificially remove or reduce the distracting condition – essentially remove noise from the system. This thesis proposes that by the use of a 3d model, viewpoint ("pose") and lighting changes from image to image may be minimised. Further, a suitably accurate model may be produced from a standard passport style photograph by statistical estimation to benefit human recognition, using the relationship between

facial texture and structure to exploit information present in an image which is not being fully utilised by the human visual system.

The starting point for this work conceptually was with the development towards a face-specific 3d coding scheme, as reported in Appendix C. It became evident that the estimation from a texture image of corresponding depth information (initially conceived as the first step in producing an error image for entropy coding) had the potential to become a lossy zero-bit coder – in other words, a structure estimation technique. The value of a robust estimator exceeds the value of such a tightly defined coding scheme, and so the emphasis of the work shifted exclusively to the estimation and evaluation of both the requirements for use by humans and the performance of the estimator.

Chapter one provides background on the psychological aspects, considering the difficulties inherent in face recognition under varying conditions. This chapter frames the problem to be addressed, and provides an introduction to the concepts behind face recognition from a psychological standpoint. The poor reliability of human face recognition (despite day-to-day perception to the contrary) is highlighted, and the type of changes in facial appearance which can cause difficulties in a recognition task discussed. Two of the most commonly varying conditions – lighting and pose – are highlighted, and it is proposed that the use of a three dimensional model to remove these variations may produce improved performance. A subjective test is presented which is designed to gain an understanding of the effect of these changes over the range of conditions that a frontal mask model of a face may be used plausibly, and the results discussed.

Chapter two introduces the 3d estimation problem, and proposes conditional densities as a superior estimation model to regression, with a signal to noise analysis to illustrate the reasons for this. The chapter also discusses the problems caused by the limited set of training data available given the high dimensionality of the data, and suggests a trained regularisation as a solution.

The practical training of this regularised covariance estimator is discussed in chapter three, along with the the preprocessing required to prepare the test data. The quality of the data is also discussed, and further steps required to compensate for the small set size and high dimensionality described. The results of the training process are described, and the optimised system taken forward to chapter four for analysis. As well as mean square error evaluation, subjective tests are reported upon which are designed to evaluate the performance of the estimator with respect to human perception.

Finally, potential for further work is discussed in chapter 5, and concluding comments are made in chapter 6.

1. Human Face Recognition Under Varying Conditions

In this chapter, the mechanics of human face recognition, and some experimental work addressing the understanding of recognition is presented, along with work highlighting some of the shortcomings of human recognition which are not generally considered outside the psychological human face recognition community. The changes possible in a face, in spite of which a human must attempt recognition, are then presented with a discussion of each common type of change, along with techniques to address this change in a bid to compensate and allow more accurate recognition. Finally, an experimental investigation into the effects of lighting and pose change on recognition is detailed, and the results of these experiments are analysed and discussed.

1.1 The mechanisms of human face recognition

The psychology of face recognition in humans is an extremely active field of research, and is still not fully understood. A full review of the field is far beyond the scope of this document, but a summary of some of the more pertinent aspects of the recognition process is beneficial in framing the problems commonly faced by humans attempting face recognition in

common real-world situations.

1.1.1 Distinctiveness and similarity to the mean

As discussed by Bruce and Young in chapter 5 of *Eye of the Beholder*[11] after Valentine and Bruce[69], there is an interesting result when examining the role of "distinctiveness" in the recognition of famous faces, where distinctiveness is a subjective rating on an arbitrary scale. A famous face which is rated as distinctive can be recognised more quickly than a face which is rated as more average in appearance. With unfamiliar faces, a similar improvement in tests of memory is evident when identifying a face in a test set as one present in a training set shown previously. When simply classifying an image as face or non-face, however, faces rated as not distinctive – both unfamiliar and famous - are classified more quickly. Valentine[68] suggests a "face space" as a potential explanation for this.

Any set of data which can be reduced to some parameterisation may be described by a space optimised for the observed parameterisation. "Face space" is simply a term to represent the space described by the parameterisation of face data; a set of coordinates in a face space will refer to a unique face. Furthermore, a set of coordinates in face space can be seen as a perturbation from the mean face. Assuming an approximately gaussian distribution in the space, the majority of faces will be clustered near the mean, with the population density rolling off with distance. Valentine suggests that this distribution may be responsible for the classification versus identification performance reversal reported. A face which is very close to the mean may be readily identified as being face in relatively few dimensions, and with high noise tolerance, since the majority of the information will be identical to the mean with little deviation.

However, distinguishing the face from the dense cluster of related faces local to it in the face space requires much more information. A face which is very distant from the mean may be more difficult to resolve into a valid face (due to its distance from the centre of face space), but once identified is unlikely to be confused with another face due to the sparse local population density.

There are any number of face spaces which may be defined; for example, geometric measurements of the distance between feature points on a set of faces can be used to parameterise the data and thus generate a face space (for illustrative purposes, this could be crudely visualised as a three dimensional space where x is nose length, y is eye separation and z is the distance from chin to nose). Some face spaces will be more accurate and useful than others – the example given above is a very poor description of face space. Appearance-based image processing provides an easy to generate face space: Here faces are defined as points in a high dimensional space, with each dimension being described by a pixel value in an image of a face. This is a representation which is easy to work with as there is no manual re-parameterisation, and the information used to define the parameters is readily accessible (assuming that the images are already digitised for use on a PC) . There are potential problems in this approach: If images are not aligned well then a large proportion of the space is wasted defining the position and orientation of the face in the image rather than any intrinsic properties of the face. Conversely, assuming that the images of faces are reasonably well aligned then all of the information contained within the images is made implicitly available in the space for analysis. The same cannot be said for a less exhaustive parameterisation based on a limited set of measurements. The face spaces arrived at by any of the above

techniques are not absolutely true face spaces (where the space is representative of every possible face) but are approximations to a global face space based upon the training data available and the parameterisation chosen. The more well-chosen the parameterisation, the greater the number of training samples, and the more representative the samples of the whole, then the better the resultant approximation.

These issues are significant in that a technique which estimates the most probable 3d structure for a given two dimensional image will minimise error by tending towards the mean (see chapter 2 for a discussion of the reasons for this) when the training data set is smaller than the dimensionality of the image-based face space, as is almost certain to be the case in any practical situation. The possibility exists, therefore, that an improvement in the mean square error of an "estimated model", where the standard photographic image of a particular face is used to calculate a likely underlying 3d model, over the use of some average model (generated by simply finding the mean 3d model across a sample set) may not be significant in perceptual performance unless there are sufficient training samples to disambiguate individual faces more substantially from the mean.

1.1.2 Familiarity versus unfamiliarity

As discussed by Hancock *et al*[27], face recognition when the subject is recognizing a familiar face has significant differences to unfamiliar face recognition – familiar faces appear to be recognised well even with lighting and viewpoint changes, and with variation in hair, beards, etc., whereas these factors, as discussed below, present significant

problems when a face is unfamiliar. The extent of this difference in performance suggests that familiar faces may not even be recognised using the same mechanisms as unfamiliar faces. Bruce[6] demonstrates the resilience in the familiar case to change compared with the unfamiliar case. Here, faces with a change of pose and expression between training and testing are shown to be recognised equally well as faces with no change (with an approximate 95% success rate) provided that the faces were familiar to the observer. Subjects unfamiliar with the faces, however, dropped from 88% to 54% correct identifications with the same change. Recognition of familiar faces was unaffected, whereas recognition of unfamiliar faces under the same conditions dropped to almost chance, demonstrating the importance of familiarity for robust recognition.

1.1.3 Recognition and shape

Both Patterson and Baddeley[50] and Davies, Ellis & Shepherd[16] demonstrate that a shift in pose between ¾ view and full face does not cause a significant degradation of recognition performance with unfamiliar faces. This suggests that human face recognition involves some mechanism to allow the estimation of differences in viewpoint, either by explicit modelling of apparent 3d structure or some implicit means. Conversely, Bruce *et al*[10] found distinct advantage conferred by the ¾ view, and also a drop in accuracy associated with a shift between full face and ¾ view (or vice versa).

The role of three dimensional structure should not be overstated, however. Bruce *et al*[8] demonstrates that poor recognition performance is achieved when only 3d information without any texture is used, and also

provided more evidence that familiar faces are more easily recognised than unfamiliar. O'Toole *et al*[48] further investigates the contribution of 3d structure and surface information by presenting subjects with images of faces rendered with either a "normalised" structure or "normalised" surface texture. In this context, normalization refers to the use of average data further manipulated to ensure that surface texture features are not misaligned with structural features. Subjects were found to perform less well when either form of normalisation was applied, suggesting that both structure and surface properties are significant for recognition purposes. As an interesting side note, male faces seemed less susceptible to the effects of shape normalisation in terms of recognition, though no definitive reason for this is apparent.

Troje and Kersten[63] suggest that tolerance to pose change is based not upon the construction of some mental three dimensional model, but instead the accumulation of reference images covering increasing ranges of pose. In this study, subjects are presented with a selection of familiar faces (close colleagues) from a variety of angles along with an image of their own face (which they are familiar with in the frontal view) in profile. The speed at which they recognised and correctly named the faces was recorded, and showed while other familiar faces could be recognised equally quickly independent of view, the subject's own face was consistently recognised more slowly in profile than in frontal view. Troje and Kersten conclude that this is because although we are familiar with our own faces in the frontal view (as from our reflection), we rarely see our own face in profile, and so are less primed to recognise it. This suggests at the least that any capability for the construction of a mental 3d model is limited, and cannot abstract from frontal views to profiles. A more extreme

interpretation is that there is no mental three dimensional representation. instead just a collection of representations without any explicit three dimensionality (or without any depth information which can be used mentally to rotate a face for recognition), and that instead a face is recognised based purely on comparison to a large collection of mental images.

1.1.4 Reliability of human face recognition

The accuracy of human face recognition from single photographic images as on passports, driving licenses, etc., is surprisingly low in real-world situations, as demonstrated strikingly by Kemp et al[37]. Here, six subjects in the environment of a supermarket were asked to be alert for fraudulent use of credit cards. A test set of volunteers were issued with four credit cards each; one with a good quality image of their face termed the "unchanged card", one with a image of their face with some minor change to the appearance (such as the addition of spectacles, change of hairstyle, etc) termed the "changed card", one with an image of someone else determined by the experimenters to bear a close resemblance to the bearer – a "matched foil", and an "unmatched foil" featuring an image of someone with an appearance unlike that of the bearer (though race and sex were still matched), again as determined by the experimenters. As can be seen, whilst valid cards were generally well recognised, foils were not well detected. Overall, the average correct decision percentage is relatively low at 67.42% given that completely random decisions should yield 50% correct decisions. An automatic face recognition system with such error rates would not be considered to be operating well enough to be of practical use.

	Unchanged	*Changed*	*Matched Foil*	*Unmatched Foil*	*Average*
% Correct	93.33	86.21	36.36	65.91	67.42

Table 1.1: % Correct decision rates reported by Kemp et al

As a further note, the anecdotal belief that police officers and so on are in some way better at face recognition due to practice and training has little supporting evidence. Burton et al[13] presented both students and police with a set of security video style clips, and then asked them to pick out the faces of the people in the clips from a set of high quality photographs. Whilst subjects familiar with those in the videos performed much better than subjects who were unfamiliar with those in the clips, both students and experienced police officers performed equally badly in the unfamiliar condition.

1.2 Impact of change on the recognition problem

As discussed, human recognition of faces is far from fully understood, and many factors can cause degradation of performance, especially where the face in question is unfamiliar to the subject. Generally, the more different a face appears between any two images, the more difficult it will be to correctly identify. There are many natural ways for a facial image to alter – from a change in hairstyle to physical deformation, due to injury or surgery. However, if the area to be recognised from is limited to just the face (i.e., hair and other external features are not considered), and possibility of extreme physical changes is ignored, there are fewer variations likely to occur, namely:

- Change of facial ornamentation (presence or absence of spectacles, sunglasses, etc)
- Presence or absence of makeup
- Change in facial hair
- Change in facial expression
- Change in illumination
- Change in pose
- Ageing

It should be noted that *pose* and *expression* are taken to represent two different variations for the purposes of this discussion. *Expression* includes changes caused by the muscles in the face, and is independent of the view of the face available to the observer, while *pose* is the variation of facial view independent of the physical expression of the face. Facial hair, makeup and spectacles can be introduced to the image relatively easily by an artist, though removing them from an original is considerably more difficult. The automatic introduction of these elements is extremely challenging.

1.2.1 Alterations in *external features*

Any aspects of a facial image outside the area of the face – such as hairstyle and hats – are considered *external* properties of a face. There is nothing that inherently ties any of these to the structure of the face, and hence the identity. They are changed on a regular, if not daily, basis in normal human society. Unfortunately, as discussed above these features are relied upon heavily when unfamiliar face recognition is undertaken. Whilst

there are situations where these features could be usefully applied to a face in order to test identity, in general they should not be relied upon due to this extreme transience.

In addition to these features, there are some non-external features which are also unreliable and have little or no physical relationship to the properties of the face; makeup, facial hair, eye ornamentation and so forth are internal features (in so much as they appear within the face region) but are also transient and at best loosely related to facial structure. They are therefore unreliable cues. For the purposes of this work these transient features are grouped with external features and will not be considered. Addition or alteration of these properties is relatively easy to achieve with the use of alpha keyed images applied over a base face in a two dimensional image, especially if said image is fully forward facing. In three dimensional representations, this is more difficult purely as the overlaying task is more complex.

1.2.2 Ageing

In many applications, ageing is a moot consideration, as the time elapsed between the capture of any two images, or between the capture of an image and the observation of the subject, is insignificant in terms of the effects of ageing; these effects will certainly be trivial compared to those resulting from the other changes listed above. Ageing also involves complex physiological changes. Those most readily brought to mind are thinning and greying of hair, wrinkling of the skin, the change of skin tone and translucency. However, these surface changes are accompanied by a range of structural changes to the face. As discussed in

[11], as humans age the skull shape alters significantly with the majority of change occurring during childhood. Much of this change can be expressed as cardiodal strain when considering the profile of the skull. Cardiodal strain is also evident in apples, kidney beans and many other natural structures where growth is constrained by a single nodal point – such as the stalk of an apple or the brainstem of a human. The effect of the application of this strain to a baby-type skull, which tends somewhat towards the spherical, is to flatten and elongate the shape of the skull towards an almost parallelogramic cross section in the extreme. A result of this change when considering the face viewed from the front is that the ratio of the distance from the forehead to the eyes and the mouth to the bottom of the chin changes appreciably as age varies. Very baby-like faces have extremely large foreheads, and very little mouth-to-chin distances, whereas mature faces have much smaller foreheads, and larger mouth-to-chin distances. As shown by Montepare and McArthur[43], simply moving the facial features as a block up and down in the face has a dramatic impact upon the perceived age of the face manipulated.

As well as these inter-feature changes to the face, there are also many intra-feature structural changes. With age, relative eye size shrinks, and the nose and ears become larger. The combined effects of the changes with the change in position of features have a demonstrable effect upon the perceived age McArthur and Apatow[41], though it is important to note that these changes become less significant to human perception as age increases. By the time adulthood is reached, the rate of change of the face is markedly slower – as the brain ceases the rapid increase in volume characteristic of childhood years, the pressures causing these facial alterations diminish significantly, and so cardiodal strainbecomes far less

significant. The effects of ageing which become more apparent in the soft tissue of the face, wrinkles, surface texture, hair colour; the thinning of lips, etc. become more significant.

There is a large body of work studying the effect of shape changes on perception of faces, some work directed distinctly at change in adults (Burt and Perrett[12]) rather than children, and it follows a large amount of work into the mechanics of artificially ageing a face, and the actual processes of artificial ageing are beyond the scope of this work. It is interesting to note, however, the extent to which structural changes, both to the skull and the soft tissue, are involved with these manipulations. O'Toole *et al* [47, 49] make extensive use of 3d models of heads in order to provide convincing aged faces for perceptual experiments. It is not unreasonable, then, to suggest that while a face image may be aged without a 3d model of the original face, the presence of such a model may assist in convincing ageing.

1.2.3 Variation in facial expression

Variation in facial expression is difficult to represent through a temporally invariant medium such as a still image, the only recourse being animation or multiple still images – which, of course, could be considered very sparse animation. However, this is only the *presentational* aspect of the problem. Equally challenging is the capture of the full range of variation in a human face. The brute force technique – either recording a video, or taking a multitude of still photographs – is in most cases impractical. The length of time taken to gather such data, and the level of cooperation required from any subject, is extensive to say the least. It has

been suggested, initially for video coding applications[61] that an articulated model of a face may produce satisfactory performance. In these schemes, a base image of a face is distorted by mapping it to a 3d model composed of articulated sections, a deformable "skin" surface, and a collection of muscle-analogue actuators. Given that the face is the most muscularly intricate part of the body, the complexity of such modelling should not be underestimated. However, as seen in films such as "Shrek" and "Lord of the Rings", 3d modelling and animation technology is well up to the task, though in these cases the time taken to animate even a short film of a face is still longer than could be desired. Having said all this, it could be argued that facial expression is of secondary importance when considering the identification of a face – as pointed out above, if a face is unfamiliar, use of the internal features for identification tends to be minimal, whereas a familiar face is unlikely to be unrecognised purely because someone is smiling in one image and frowning in another; familiarity could be expected to involve having seen both of these expressions before. Bruce[6] demonstrates the effect that variation in expression combined with variation in pose can have, though does not differentiate between the two variations. Here subjects were shown testing images that had either no change from the training case or a change in both pose (between frontal and ¾ view or vice versa) and expression (smiling to unsmiling or vice versa). For subjects who were unfamiliar with the test faces, correct identification dropped from 88.8% to 54.8% between the unchanged and changed cases. As a preliminary experiment, single changes in either pose or expression were presented to subjects unfamiliar with the faces. Whilst Bruce does not break down the performance into pose change and expression change, the performance drops from 89.6% with no change to 76.0% with one change and 60.5% with two changes. This suggests that neither expression change

nor pose change is negligible.

1.2.4 Pose variation

Possibly the most common change in appearance is change in pose – the angle at which a face is seen. Unless an image is a passport photo, it is unlikely that it shows a face directly from the front. Security camera images, for example, have a tendency to be taken from a high position looking down on the subject as a result of the requirement for the camera to see over crowds and be inaccessible to anyone wishing to disable it. Whilst in a face-to-face meeting it is almost a given that a frontal view will at some point be obtained, when considering photographs and film – especially surveillance footage – a full-faced view is far from guaranteed. As already discussed, resilience to pose changes can be conferred by repeated exposure to the face – though there are limits. If a face has never been seen in profile, then it will not be well recognised in profile. Moses *et al*[44] show recognition to be affected only very slightly with change of pose, as well as change of lighting and expression. This is an unusual result, at odds with much other work. However, subjects participated in a training stage for these experiments, and were not allowed to proceed with the tests until a lower limit of accuracy in the training stage was achieved. There may well be some aspect of the familiar face mechanism responsible for this performance, therefore.

1.2.5 Lighting variation

Figure 1.1: A 3d model with a coloured texture rendered under a range of lighting conditions

As can be seen in figure 1.1, lighting can have a pronounced effect. These images are not only of the same face but of an identical 3d model, with identical surface texture. This suggests that lighting changes can have a significant effect upon appearance. This should not be surprising; in a single image shading is the only cue available to convey information about three dimensional structure, and shading patterns are dependent upon the lighting conditions as well as the shape of the object under scrutiny. For example, the upper and lower images at the right-hand side of figure 1.1 appear to have quite different cheek shapes, though both are in fact renders of exactly the same model from exactly the same position; lighting is the only variation.

Change in pose and lighting are significant as these are almost inevitably different in any real application requiring the verification of identity, either image to image or image to person. While work has been

carried out in order to quantify this to some extent. for the purposes of correcting for this effect a specific case should be considered: Only a frontal mask of the face can be textured using a passport style photograph, and so the limits of plausible pose and lighting variation upon such a model are a concern. By quantifying the relative effect upon recognition of these variations by subjective testing then the level of benefit possible to recognition by pose and lighting matching under extreme circumstances can be estimated. Provided this improvement is not trivial then there is an argument for the use of some 3d model based correction to aid recognition. Davies, Ellis and Shepherd[16] show that even detailed line drawings are inferior to photographic images for recognition, and suggested that the lack of cues from shading to provide both skin texture and depth information as a reason for the poorer performance. Bruce *et al*[8] found that recognition performance based upon 3d structure was dependent upon lighting, though this considered only the extreme case of lighting from above and lighting from below (equivalent to holding a torch beneath the chin). Further work[9] with subjects again recognising faces based on renders of laser scans showed change in lighting condition to be equally detrimental to recognition as a pose change, though again test images were limited to top lighting and extreme bottom lighting conditions. A further conclusion drawn is that the effects of lighting and viewing direction are not independent – the effects of viewpoint appear to be dependent upon lighting condition.

1.3 Subjective testing

It has been demonstrated that subjects will generally be less capable of recognising unfamiliar faces if both lighting and pose change between

the image used to learn a face and the image used to identify a face. If this is the case, a natural extension is to reverse the statement: By matching pose and lighting conditions, recognition accuracy may be increased. For the purposes of this experiment, it is unnecessary to specifically identify or name a face, as this introduces many complications – the relative memorability of names, for example. It is sufficient to recognise a face as being one of a set which have been shown to the subject previously as training images. In order to remove the possibility of expression variation between training and testing images, and indeed any other uncontrolled variation, images can be rendered using identical 3d models. By rotating or re-illuminating a model it is possible to generate a virtual image which has very little difference to a natural image taken under corresponding real lighting and pose conditions. There are, however, difficulties introduced by the use of a 3d model to render test images. As the rendered image is created using a photographic texture image, the lighting conditions of the real photo shoot are implicit in this image. In order to minimise this effect, lighting was rigged to illuminate the faces with a diffuse ambient light, tending the lighting conditions on the face to a "lighting neutral" image, suitable for accurate re-illumination. A further issue implicit in the illumination of the models is the algorithm used to calculate the shading of the model. The technique selected is *ray tracing*, where the light incident upon the object rendered has been traced ray-by-ray from the light source. This provides accurate representations of systems, but is computationally intensive. For this reason, test images must be pre-rendered rather than dynamically generated. Techniques suitable for real-time rendering such as *shadow mapping*, provide inferior approximations to illumination. A further consideration when using the raytracing technique is that the resultant rendered image is dependent upon the intensity and type of any

light sources used, and upon the specific shader models used to render the images. Each object in a rendered scene must be assigned a shader; this defines the way in which the surface of the object reflects or refracts incident light rays. Shaders can be complex, and for truly photo-realistic results, the response of the a shader can be manipulated dynamically over the surface of an object by applying images mapped to different properties of the surface. The mapping which provides the most obvious improvement in the visual quality of the output is a *texture mapping*, whereby the base colour of each part of the objects surface is defined. In this application, the only information available for mapping is the texture – taken from a photograph of the subject being modelled. However, a 3d artist may also add other mappings such as *bump mapping, radiosity mapping* or *alpha-key mapping* to fine-tune the apparent three dimensional texture, refractance, transparency and other optical properties of a surface. For genuinely realistic skin, for example, the most realistic approach is to overlay multiple translucent layers of texture, replicating the way real human skin behaves optically[62,26,70]. In addition, a variety of shading algorithms are available. The most basic raytrace shader generally used is the *phong* shader. This provides a poor approximation to realistic lighting, and is incapable of realistically modelling surfaces that are rough, and hence which diffuse incident light rather than reflecting as a perfect mirror (albeit with attenuation). For all rendering in this work, an *Oren-Nayer* diffuse shader, adjusted to represent relatively roughly textured surfaces, and hence minimise the "shininess" of the skin, has been used along with a *Blinn* specular shader, set to very low reflectance to prevent the appearance of a metallic surface, while retaining some highlights characteristic of human skin. It should be noted that these shaders have been set up subjectively based upon informal experimentation by the author. It is

important to note that this is not the ideal way to shade a human face: facial hair and eyes, for example, should have very different shader properties to skin. However, as these make up relatively small portions of the images, it has been assumed that these errors are negligible. Further, the skill required to set these shaders up ideally is beyond that of the author, and falls firmly into the realm of the talented 3d graphic artist. Additional information would also be required (for example, textures for subcutaneous layers of skin), which could only currently be generated through artistic approximation. It should be noted, however, that whilst the shaders are selected based on ad-hoc experimentation this has limited implications for the results. The shader selected for rendering is common to both training and test images; at no point is a rendered image compared to a natural photograph. Bearing this in mind, provided the shader does not remove information through saturation or produce an image which does not appear to be a naturalistic face to the eye then the choice of shader is to a great extent irrelevant as a common factor.

Another difficulty arises from the use of a single image to texture the entire model. This image is correct only for the render of the model where the pose of the model is identical to that of the texture image. As the image has no z-axis information, a pose change of 90 degrees will expose smearing in the z-axis, where one pixel of texture information is stretched over an area apparently larger in the new view than in the original view. This can be mitigated by the use of high-resolution texture images to some degree, but more effectively by limiting the maximum change of viewing angle. This must be limited regardless of the texturing consideration, as only a limited amount of the head is captured in a face model – specifically the mask of the front of the face, with the sides and top of the head

generally absent in the model. This leads to unrealistic sharp edges around the face. These disconcerting edges can be masked somewhat by the use of relatively low-level lighting for the rendering process.

Figure 1.2: A 3d model made from a frontal view of a face, in the lower pictures the model has been rotated slightly. In the left hand, unshaded images, the lines where the model creation has failed as the face surface angles away from the camera are evident. Using strong directional illumination to shade the face, as shown in the right hand-column, minimises the perceptual effect of these edges at the cost of generating a slightly "theatrical" image.

Figure 1.3: The left-hand image is the source of the texture for the renders in figure 1.2. The right shows the same subject from a more extreme angle, revealing much data missing from the rendered images and the frontal natural image.

As the models are recorded as a collection of points, and rendered by mapping triangles of texture image onto triangles formed by neighbouring points, distinct sharp edges are apparent in rendered images, especially under varying lighting conditions. It could be argued that these represent "fair" noise, as these artefacts are common to all rendered images. However, to ensure the images are as naturalistic as possible, the mesh can be converted to a collection of smooth curves, using the surface subdivision tool in 3D Studio Max, the package used to render the images. This provides a marked improvement in the appearance of the faces.

Finally, as the same model is used to generate both testing and training views, a like-to-like comparison, with variation in neither lighting nor pose, is unfair as both images would be identical. While these results could be gathered, they would be of limited use: it is unclear how much the mechanism by which they would be recognised is face recognition, and how much would be rather the identification of an exact replica of an arbitrary image.

1.3.1 Data selection

The possibility exists that the subjects may perform recognition not based upon the faces shown to them, but on spurious artefacts or genuine tells in the images. For example, if only one individual in the image set is captured with eyes closed and a broad smile, it is possible that this individual will be recognised based upon these features rather than In order to provide the most natural images to the subjects, the data used to generate the images has been manually selected from the York data set, using only front facing, neutral images with open eyes. Only the forty most complete models have been used in order to prevent the shape of missing areas of the face from being used as an identity cue. Due to these considerations, and the limited size of the data set, individuals were not selected on age, race or sex basis. As a result of this, the test data is almost exclusively white caucasian, is biased towards male individuals, and does not represent a controlled spread of ages (there are no minors, for example). As discussed in 1.3.4, a distinctiveness survey was conducted in order to remove data from any images which were subjectively found to be unusual compared to the other images in the set. See chapter 3 for more detail on the York data set. The York data set was created with a fixed number of poses, allowing data to easily be selected which shares orientation and expression characteristics (front facing, eyes open, neutral expression). From the set of front facing images, the least noisy (based upon visual inspection) was selected for each individual.

1.3.2 Posing and rendering

Scenes were arranged in 3d Studio Max as shown in figure 1.4. with the nose tip of each face aligned in x, y and z axes manually. The rotation of the face was not altered from the initial value in any axis for any of the faces, as it was assumed that all faces were captured in approximately the same orientation relative to the 3d capture device, and were selected as such as described above. These two sets of schematic images show the full scene set-up used to generate the four test stimuli required from each face. The lighting elements can be seen as the two cylinders (which each show the beam of a flat emitter at the end of the cylinder furthest from the face model), one to the lower left and one to the upper right of the face. Note that a low intensity point source emitter is positioned directly in front of the face. This is not a directional source, and due to its positioning does not introduce shadows to the image. It is present simply to soften the current shadows and make the darkest areas of shadow less extreme. Lighting set-up is identical in both schematics; the only change is the position of the camera used to render the image. The rendered images are produced by enabling each of the two directional lights in turn, then changing the camera position and repeating the process. This produces as set of images as shown in 1.5.

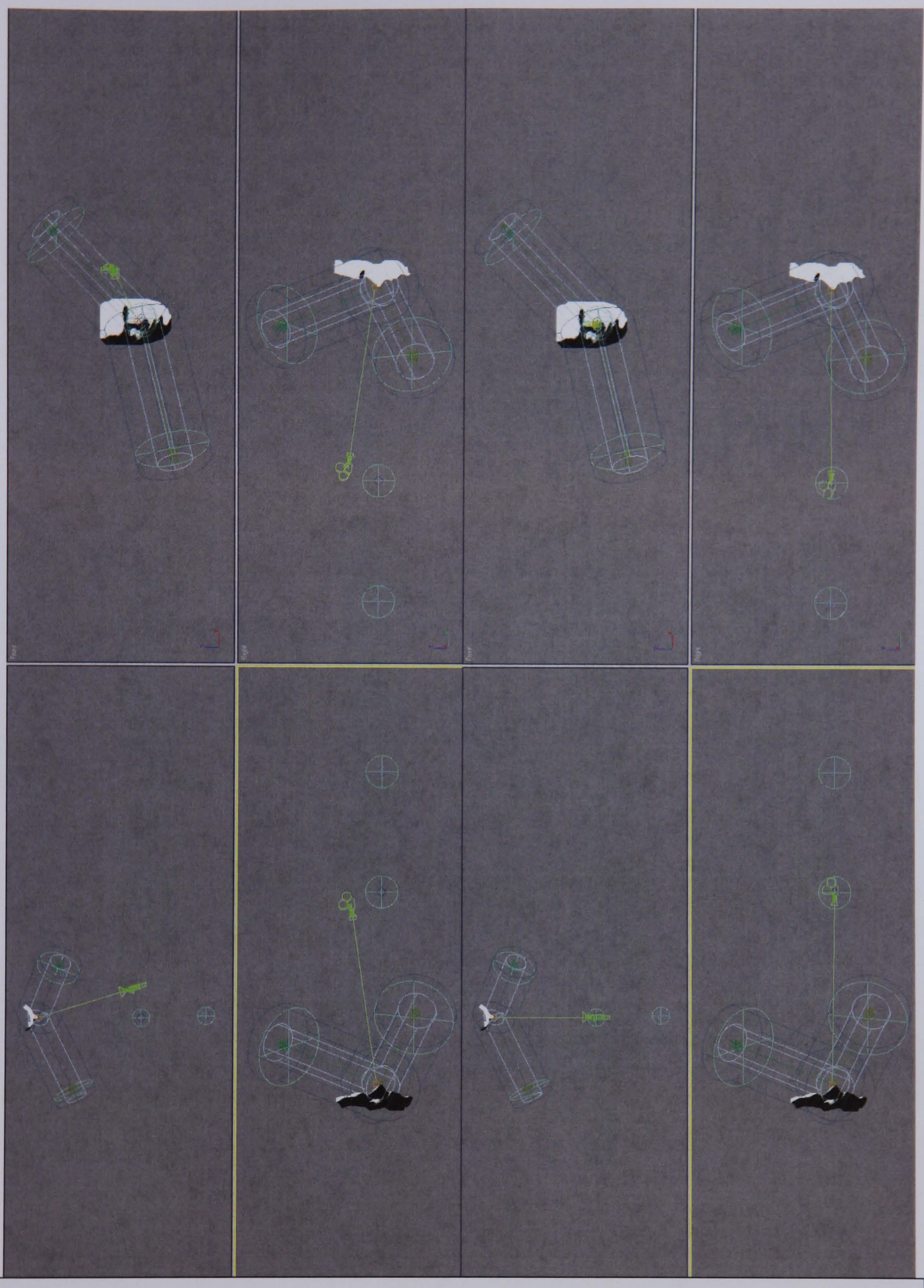


Figure 1.4: The render set-up used to generate the images. The cylinders represent directional lighting. Lights were enabled and disabled in order to generate the illumination conditions, and the camera moved between the positions shown to generate the pose conditions

Figure 1.5: The four rendered images of a face used in the test set.

From left to right in figure 1.5, the conditions are:

- Lower left light source enabled, upper right light source disabled, frontal camera
- Lower left light source disabled, upper right light source enabled, frontal camera
- Lower left light source enabled, upper right light source disabled, non-frontal camera
- Lower left light source disabled, upper right light source enabled, non-frontal camera

By moving the camera rather than rotating the model, the light sources provide the same shading on the face for the equivalent frontal and non-frontal situations; it is only the angle through which the shaded face is observed that varies. In total, images of 40 faces were used, each being rendered four times under the conditions described As a frontal mask of the face is to be used there is clearly a greater range of plausible lighting conditions than pose conditions, and so it is to be expected that the lighting variation may have a larger effect upon recognition than pose change.

1.3.3 Procedure

This experiment was designed to quantify the effect upon the reliability of recognition of varying the lighting conditions, pose or both lighting and pose between training and testing view. Note that some change is always present. The reason for this is to remove the situation where the training and testing images are identical (which may involve different memory processes). Whilst different models of the same face could be used to circumvent this problem, there would be differences beyond lighting and pose between training and testing, introducing extra degrees of freedom in the test data which could confound analysis of the results.

The experiment is a simple recognition test, where subjects are exposed to a watch list of faces in the training stage, and are then presented with a set of test images, including both images of the faces in the training set and distractor images of faces not present at the training stage. The subjects are asked, for each image in turn, to indicate whether the face was present in the training set or not.

In order to meet the above conditions, test subjects were separated into two groups. Group one introduced a change in pose consistently between the appearance of each individual when re-presented to a subject at the test stage whereas group two consistently introduced a change in lighting when re-presented. For both groups, half of the individuals presented to each subjects had both lighting and pose changes between stages; for the other half only lighting or pose (depending upon the group) varied. Since all the images were rendered as described above, all pose and lighting changes were identical. Each group contained four test scripts,

used to specify which images to be used for a given test. The scripts were written such that in each groups every face was included in the training set for two of the scripts, with different conditions in each of these instances. Within each script both training and test sets contained equal numbers of each render type shown in figure 1.5 in order to ensure that no render condition was presented disproportionately in the test. Within the four scripts for each test group all faces were presented as both training face and distractor an equal number of times to reduce the effect of the individual distinctiveness of each face. Each test subject used only one script, and each script was used an equal number of times. Subjects carried out the experiment only once, and different subjects were used in each group, in order to prevent familiarisation with the test data. Whilst the scripts determined which training and testing images were used for a given subject, the order in which both the testing and training images were presented was randomised at run-time by the test software.

The result of this is that all faces were presented in all conditions, as both stimulus and distractor, an equal number of times in an order randomised for each subject. Therefore, any bias based on image order or the individual distinctiveness of a face should be greatly diminished.

The test was carried out in a darkened room using a CRT monitor with brightness and contrast optimised for the test images in order the images in such a way that dark areas of the face were visible and light areas were not saturated. The control of the room lighting was in order to remove the effects of relative ambient light levels (for example, in a very bright room the darker regions of the facial image may no longer be

distinguishable from background black). Subjects were instructed that the test took place in two steps; firstly they were to be presented with a series of images of faces, each face to be shown for five seconds, preceded by three seconds each of a mid-grey blank image and a mid-grey blank image superimposed with a fixation crosshair centred upon the point at which the centre of the face image will appear. These timings were chosen as long enough to provide the subject with an opportunity to inspect the face but not so long that the attention of the subject would wander. Several similar experimental set-ups have been recorded, though the time allowed for training varies considerably; for example Davies et al[16] allow 12 seconds, Bruce[6] allowing 5 seconds and Patterson and Baddeley[50] 28 seconds. In all of the above, however, when recognition latency is measured a figure of between one and three seconds is recorded. As the images are shown sequentially at training before any testing is performed, a short time was deemed more practical. All images were centred horizontally and vertically in the screen. The software ran in full-screen mode, and any part of the screen not used to present the test data was coloured black to match the background of the rendered images to remove any other stimulus from appearing on the screen, and to give the most convincing view of the face. Subjects were asked specifically to remember the faces they were to be shown, and were made aware that they would be asked to identify the faces they had been presented with in the second part of the test.

After viewing these images, subjects were then told that they were to be presented with another set of images of faces and were asked to press the space bar if they thought a face had been present in the first set of images. They were asked to press any other key if they thought it was a new face. Images were again shown in turn with mid-grey then a fixation

point preceding each image. However, in this stage the face images themselves would persist until the subjects made a decision. Subjects were instructed that accuracy was more important than speed, and that they should take as much time as they needed over each image to come to a decision that they were happy with. This was to remove any idea that the subject was under time pressure, and to allow them the time they felt necessary to make a decision rather than hitting keys without considering the image due to a misconception that speed of response was important.

Selection of subjects was not controlled, and was based primarily upon availability. All subjects were adult, though no personal information at all was recorded for reasons of confidentiality, with results referencing only test ID numbers. No control for ethnic grouping was applied, though as the subjects were drawn from the same demographic as the individuals represented in the data set, it could be argued that the test set likely to be a close match to the ethnicity of the subjects. Recognition of faces across racial divides has been demonstrated repeatedly to provide poorer performance than recognition within the same race; i.e., a white caucasian will recognise other white caucasian faces more accurately than those of any other race[3,7,39]. This effect may be present in the results reported here, though for the reasons outlined it is anticipated that any such effect will be negligible. Results were recorded by the test software and automatically tabulated thus for each subject:

Test number 5	Same Lighting	Different Lighting	Total
Correct Identification	6	5	11
False Positive	*N/A*	*N/A*	7
True Negative	*N/A*	*N/A*	13
False Negative	4	5	9

Table 1.2: An example result from the test run, in this case for subject number 5. Note the N/A entries: these are significant as they prevent certain analyses of the data, but cannot be avoided due to the changed / not changed nature of the variables under test.

Note that the true negative and false positive responses must by definition not have a same / different breakdown; as this test is concerned only with change in appearance from training to test, any response to a distractor must be undefined in these terms. This is significant, as a signal detection test (discussed in the Results section below) becomes impossible to perform. Test number is used to differentiate between subjects (the example shows the results for subject 5), but no personal details are attached to the number. The test number is used to select the script for the experiment, allowing the images used to be identified outside the test.

In addition, a sample of six subjects selected at random were asked upon completing the test to participate in a distinctiveness survey. The reason for this was to detect any unusually distinctive faces which may be recognised well regardless the conditions they are presented under to allow the response to these faces to be removed from the results. This survey involved presenting the subject with a printed page upon which one image of all of the faces in the test is present, all under identical lighting and pose

conditions. Subjects were asked to rate each face according to how distinctive they considered it, from very non-distinctive (1) to highly distinctive (5).

1.3.4 Results

The average results for each test group are shown in tables 1.3 and 1.4.

Average response Group 1	***Different pose, same lighting***	***Different pose, different lighting***	***Total***
True positive	7.45	5.7	13.15
False positive	*N/A*	*N/A*	6.4
True negative	*N/A*	*N/A*	13.6
False negative	2.55	4.3	6.85

Table 1.3: Average results for group 1 (pose always varies)

Average response Group 2	***Different lighting, same pose***	***Different lighting, different pose***	***Total***
True positive	6.3	5.25	11.55
False positive	*N/A*	*N/A*	6.45
True negative	*N/A*	*N/A*	13.15
False negative	3.7	4.75	8.45

Table 1.4: Average results for group 2 (lighting always varies)

Typically the results of experiments such as this are analysed by application of signal detection theory to calculate the discriminability index *d'*. This is a useful measure of how well a stimulus can be discriminated from noise by a decision making mechanism – such as a human – as it requires no model of the mode of operation of the decision maker, and

requires no modelling of the signal upon which the decision is based. The only assumption required is that the distribution of decisions is normal about some mean. The calculation of d' is impossible for this experiment, however, as it requires an estimate of both a correct identification and a false acceptance for a given condition. However, this experiment is designed to investigate the effect of a change in lighting, pose or both from the training to testing stage. The probability of a correct decision for each of these conditions can be estimated by dividing the number of correct decisions for each case by the number of actual training faces re-shown for each condition. A probability of false acceptance for each condition is by definition inestimable, as a falsely accepted image has no training case to be different or identical to.

A less robust measure of the relative ease of identification of faces under each condition is the raw percentage of correctly identified faces under each condition. As shown in table 1.5, varying lighting appears to have a substantially greater impact upon recognition than varying pose, though it should again be noted that the change in lighting possible before compromising the plausibility of the image when working with a frontal mask is far greater than the change in pose. A compound change in condition of both lighting and pose results in an even greater drop in performance, with subjects recognising faces less than 5% better than chance.

	Same Lighting, Different Pose	***Same Pose, Different Lighting***	***Different Pose, Different Lighting***
% Correct Identifications	74.50%	63.00%	54.75%

Table 1.5: % Correct identifications for each of the three test conditions

The results of the distinctiveness survey (shown in figure 1.6) clearly indicate two faces which were much more distinctive than the others in the set. Inspection of the images, reproduced here, show the reasons for this score; The other faces are a mix of caucasians of both sexes whereas face 138 is clearly of different ethnicity. Face 146 is very unusual in both shape and age, looking relatively dissimilar to the other faces.

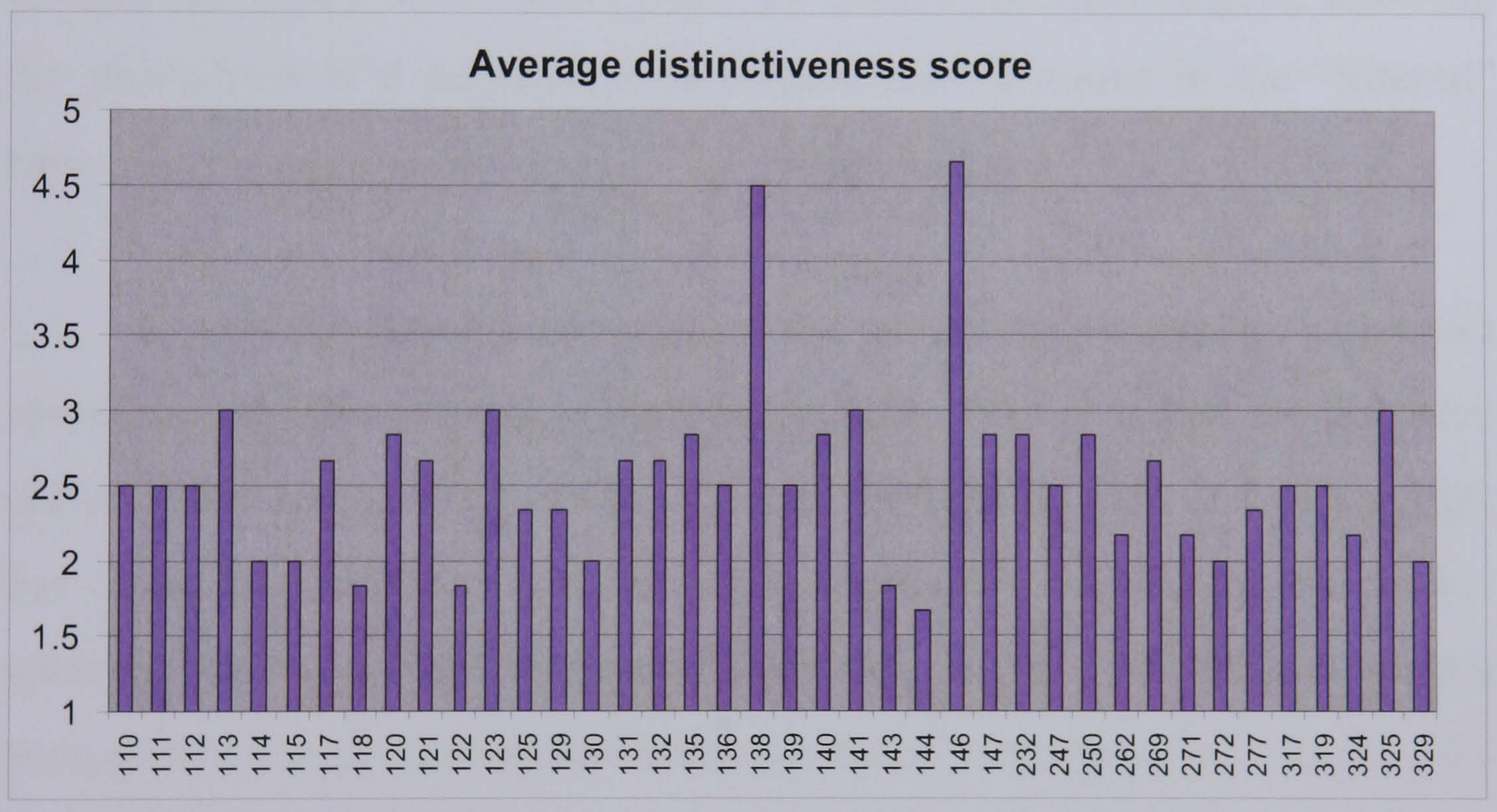


Figure 1.6: Distinctiveness survey results. Note the two unusually high results.

Figure 1.7: Face 138 (left) and 146 (right) – the two samples with unusually high distinctiveness according to the survey results

As these faces are so distinctive, it is possible that they can be accurately recognised despite changes in condition, and so could be skewing the outcome of the experiment. Responses based upon these images can easily be removed from the results for each subject, allowing the production of a percentage of correct identifications in the "filtered" case.

A problem which is inherent in the use of the percentage of correct identifications as a measure of accuracy rather than d' is that the apparent performance of a subject can be affected strongly by their individual bias. For example, if a subject answers positively to all of the test stimuli, by this measure they would have performed perfectly as false acceptances are not accounted for at all. To provide some measure of compensation for this, the results from each individual subject can be normalised to remove their bias. As there are an equal number of distractors and true stimuli in the test set, an unbiased subject could be expected to respond positively 50% of the time. If a subject were to respond positively in only 20% of cases, these positive results could be considered more significant, as the subject has a greater tendency towards answering negatively. Dividing each positive response category by the sum of all positive responses and multiplying by

half the total number of responses normalises the performance of the subject such that individual bias is no longer present in the data though the detection probability is still estimable. As can be seen in figure 1.8, both of the above processes have only a slight effect upon the mean correct decision rate.

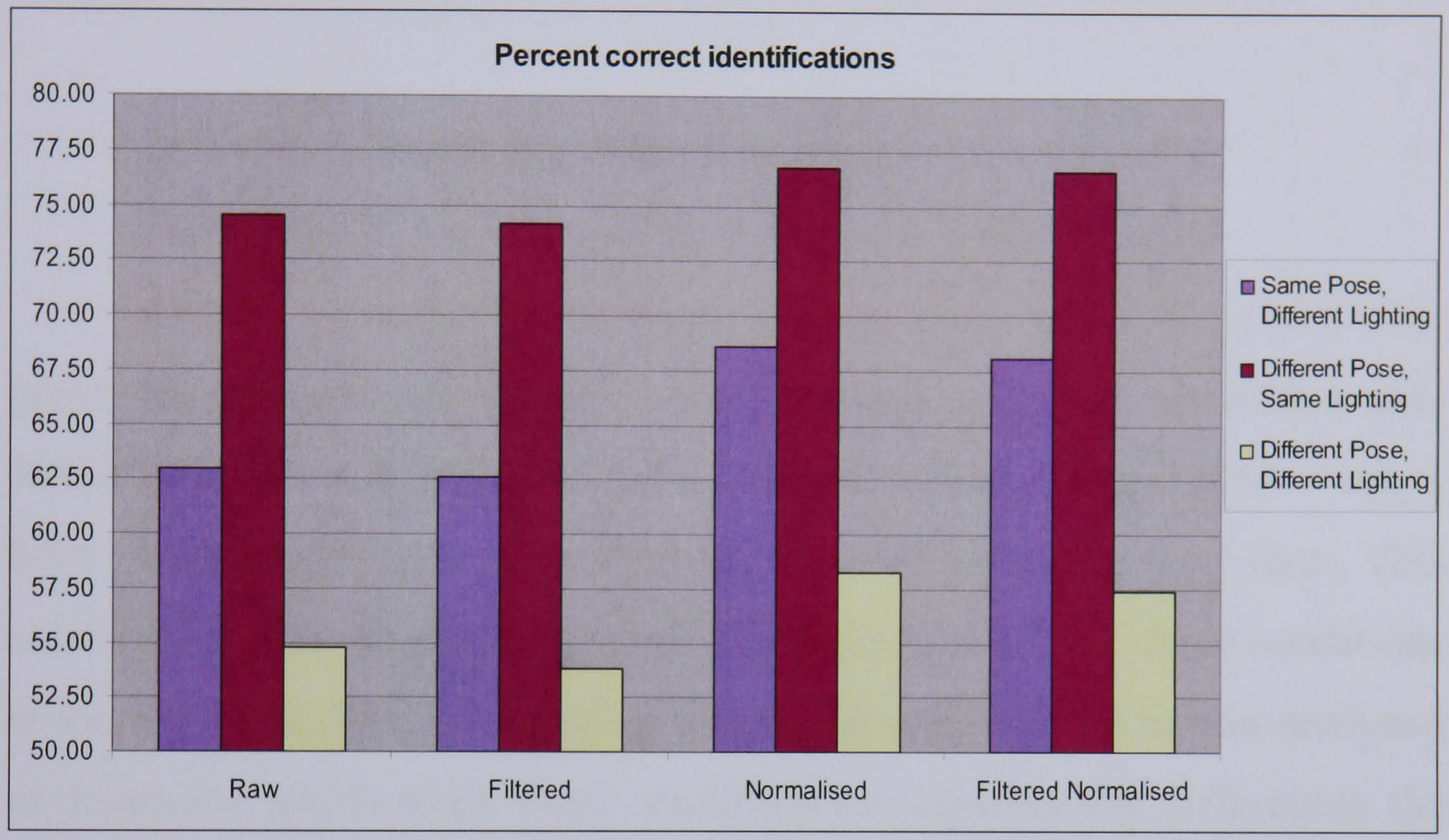


Figure 1.8: % Correct identifications before and after both normalisation and the filtering out of results from unusually distinctive stimuli

1.3.5 Analysis

With a compound change in both pose and lighting, the raw results show a mean identification rate which is only slightly above the 50% level which could be expected if decisions were made randomly. As expected, a change in only either pose *or* lighting rather than a change in both results in significantly better performance. Varying the lighting direction has a more pronounced effect than variation of pose, though as discussed this is to be expected due to the greater shift possible in lighting

when using a frontal mask style model to render the images. Comparing lighting to pose directly is difficult, as a like for like comparison is problematic to define. Given the application in question, however, these results are interesting in that there is greater scope for illumination to be varied to match an observed condition than pose, and this correction alone can yield approximately a 20% increase in successful identification.

1.3.5.1 Analysis of the raw data

In order to determine if the conditions under which the faces were presented produced a statistically significant effect upon recognition, a one-factor ANOVA test with three conditions (lighting change, pose change, both change) has been performed on the raw percent correct data. This resulted in a statistically significant difference between the three conditions at the 5% level. Post-hoc analysis allows the results to be further analysed to determine which of the three conditions are significantly different to the each other. A Tukeys pairwise comparison allows the probability that the mean score is the same to be calculated for each pair of means. At the 5% level, there is a significant difference between the lighting varies and pose varies cases and also between the pose varies and both vary cases. Therefore, the results from lighting-changed and pose-changed vary from each other, as do the pose-changed and both-changed. There is, however, less variation between the lighting-changed and both-changed conditions. This suggests that the effect of pose change is largely subsumed in the both-vary case by the effect of lighting changes. This lends credence to the idea that over the ranges available with a depth-mapped "mask" model lighting variation can be more significant than pose variation – though whether this is true in the case where the magnitude of variation in lighting

is equal to the magnitude of variation in pose is unknown. This is not entirely relevant, however, as the limits of pose variation are defined by the rotation possible with an incomplete "mask" model, not by some arbitrary concept of "same"-ness with lighting variation. As the two are fundamentally different and cannot be directly compared it could be argued that the a reasonable way of defining "equality" is to use the limits of plausible rendering with the models available; rendering beyond these extremes is impossible, and artificially constraining the system within the limits pointless.

1.3.5.2 Analysis of normalisation and discriminability filtering

As described above, the raw data may be normalised to remove the individual bias of each subject and may also have the results from faces rated as unusually distinctive removed. Do these procedures have any impact upon the data, or was the impact of both bias and the more unusual faces negligible?

Discriminability filtering yields the small anticipated drop in performance across all conditions, whereas normalisation increases the success rate of the subjects. Neither significantly alters the relative performance across conditions. The increased success rate after applying normalisation suggests that subjects tend to be pessimistic when carrying out such tasks, and when uncertain will tend to guess negatively. However, comparing the raw, filtered, normalised and normalised filtered responses within each set using a one-way ANOVA suggests there is no significant difference between them, as summarised in table 1.6. Individual subject

bias and the distinctiveness of individual unusual faces within the set can be assumed have a negligible effect upon the results. The analysis of the raw data is, therefore, valid and the compensation for individual bias and discriminability is not necessary.

	Lighting varies	***Pose varies***	***Both vary***
P(Same)	0.75	0.88	0.67

Table 1.6: Result of one-way ANOVAs on each set of results, showing the probability that the results for each type share the same mean within the type across raw, normalised and filtered results

1.4 Conclusion

In this chapter an overview is given of the challenges presented to the human facial recognition which are present in many real world recognition tasks. The low accuracy of such recognition when a face is not well known to the subject is discussed, and a review of the prior work carried out in the identification of problematic cases for recognition, presented. Of particular interest is the body of work identifying pose and lighting changes as a source of error in the recognition process. In light of this work, it is suggested that if change of pose and lighting condition has a demonstrable effect upon the accuracy of recognition then matching pose and lighting will provide a benefit to the recognition process. In order to demonstrate this, the use of 3d models, based on frontal-mask data and hence with greater possible variation in lighting, is suggested. Rendering techniques have been discussed, and the rendering method used to produce test images described. By the experimental process described, it has been shown that both pose and lighting have a significant effect upon

recognition, and a compound change in both pose and lighting causes recognition rates to drop to almost chance. However, matching pose provides approximately a 7.5% absolute increase in accuracy, and matching lighting provides an approximate 20% absolute increase. The reason for the greater increase due to light matching is theorised to be due to the greater variation in lighting condition possible when operating with this type of model compared to the degree of pose variation. It can be concluded that an accurate 3d model used to match pose and lighting condition between training and testing views has a significant beneficial effect.

If an accurate model can be estimated from a single photographic image, therefore, a technique for enhancing the performance of recognition in many applications is possible.

2. Estimation of Three Dimensional Structure by Statistical Methods

This chapter begins with a review of shape-from-shading techniques, and forms an argument for an alternative approach to the usual image irradiance technique in the case of human faces. It introduces the techniques generally employed when analysing facial images to reduce the dimensionality of the data, and describes the reasons commonly cited for their use. The representation of a facial image as a point in an n-dimensional space is described, and general statistical techniques for the estimation of a known value from an unknown value for a given distribution described using low dimensional examples, leading to the introduction of principal components analysis (a technique often used in appearance based face image processing). An alternative approach, often applied outside the image processing field, of conditional distribution estimation of unknown values is introduced with a novel signal-to-noise analysis demonstrating the value of this approach. Finally, the issue of sparse sample data is addressed, with a trained regularisation technique being proposed as the solution.

2.1 Shape-from-shading

Shape-from-shading is an area of active research which has grown initially from photogrammetric work in the 1950s-60s which attempted to use shading as a shape cue, specifically for the recovery of topographical information from lunar photographic images (Van Diggelen[17] and later Rindfleisch[56], for example) and hence made assumptions based upon the specific application (for example, based upon specific properties of the lunar maria in the latter work). The term "shape-from-shading" was first coined by Horn[28], who also provided the general form of the problem which has been the foundation for much, if not all, later work in the area. Shape-from-shading is based upon the premise that for an image of a given object, under known (or well estimated) lighting conditions, the topology of the surface of the object can be inferred based upon the the projection of light through the system from the source to the image; effectively a related problem to that of 3d image rendering (in image rendering, a known light source and a known object are used, and the light propagated through the system to form the unknown image; shape-from-shading techniques address essentially the same system but with a known light source and a known image and an unknown model). Whilst strictly the light source may be unknown for an arbitrary image, it is generally assumed in shape-from-shading work that the light source is known – or, indeed, in some cases assumptions about the light source are fundamentally built into the technique[52, 53].

Horn[28] suggested that the shape-from-shading problem could be posed as a non-linear first order partial differential equation derived from the Image Irradiance Equation (equation 1)

$$R(n(x,y))=I(x,y)$$

Equation 1

Where I is the image intensity (i.e., pixel value) at x, y of the input image, R is the reflectance function which relates the orientation of the normal at point x, y to intensity and n *is* the unit normal to the surface at x, y, z (z, the range, being implicit), defined as:

$$n(x,y)=\frac{1}{\sqrt{1+p(x,y)^2+q(x,y)^2}}(-p(x,y),-q(x,y),1)$$

Equation 2

Where p and q are the gradients in z at (x, y) along the x and y dimensions respectively such that the z-axis gradient at point x, y is defined as $\nabla u(x,y)=(p(x,y),q(x,y))$. Assuming the surface in question is Lambertian with unit albedo, then the reflectance function may be rendered

$$R(n(x,y))=\omega\cdot n(x,y)$$

Equation 3

where ω is a unit vector which describes the direction of a light source at infinity. Equation 1 can, therefore, be rewritten as

$$I(x,y)\sqrt{1\text{abs}\nabla u(x,y)^2}+(\omega_1,\omega_2)\cdot\nabla u(x,y)-\omega_3=0$$

Equation 4

i.e., a first order differential equation, the solution of which will yield the shape of the object. However, there are several issues with this, which

subsequent work has striven to address, not least the fact that there are multiple unknowns in the equation. Furthermore, as noted, the equation assumes a lambertian surface of the object and also requires a known lighting condition – in practise, in any uncontrolled lighting situation, the latter is not true, and the former is not the case for human skin[70, 26]. Indeed, the broader, unstated assumption that the reflectance of the object is constant is clearly not the case for the human face, with eyes and facial hair providing the most obvious examples of surface reflectance which differs from the skin of the face.

A further problem is the ambiguity of some shading patterns, as shown in figure 2.1. This bas-relief ambiguity, discussed by Bellhumeur *et al.*[4], is present even in the human visual system, and is resolved in this case by the application of further assumptions – that lighting tends to be from above, for example, and that faces are convex structures. This latter assumption is exploited effectively by the "hollow face" illusion, whereby a hollow mask of a face, viewed from the rear, appears still to be facing the observer. This illusion persists until the mask is rotated to the point that the profile of the mask begins to be visible, thus dispelling the illusion. Castelan and Hancock[14] tackle this problem for face data specifically, using local shape based methods to enforce convexity where appropriate. Of interest are the example images where the unconstrained shape-from-shading system incorrectly creates a concave nose; without an abstract understanding of what a face is, there is not the absurd decision that it appears to a human with an understanding of what constitutes a face.

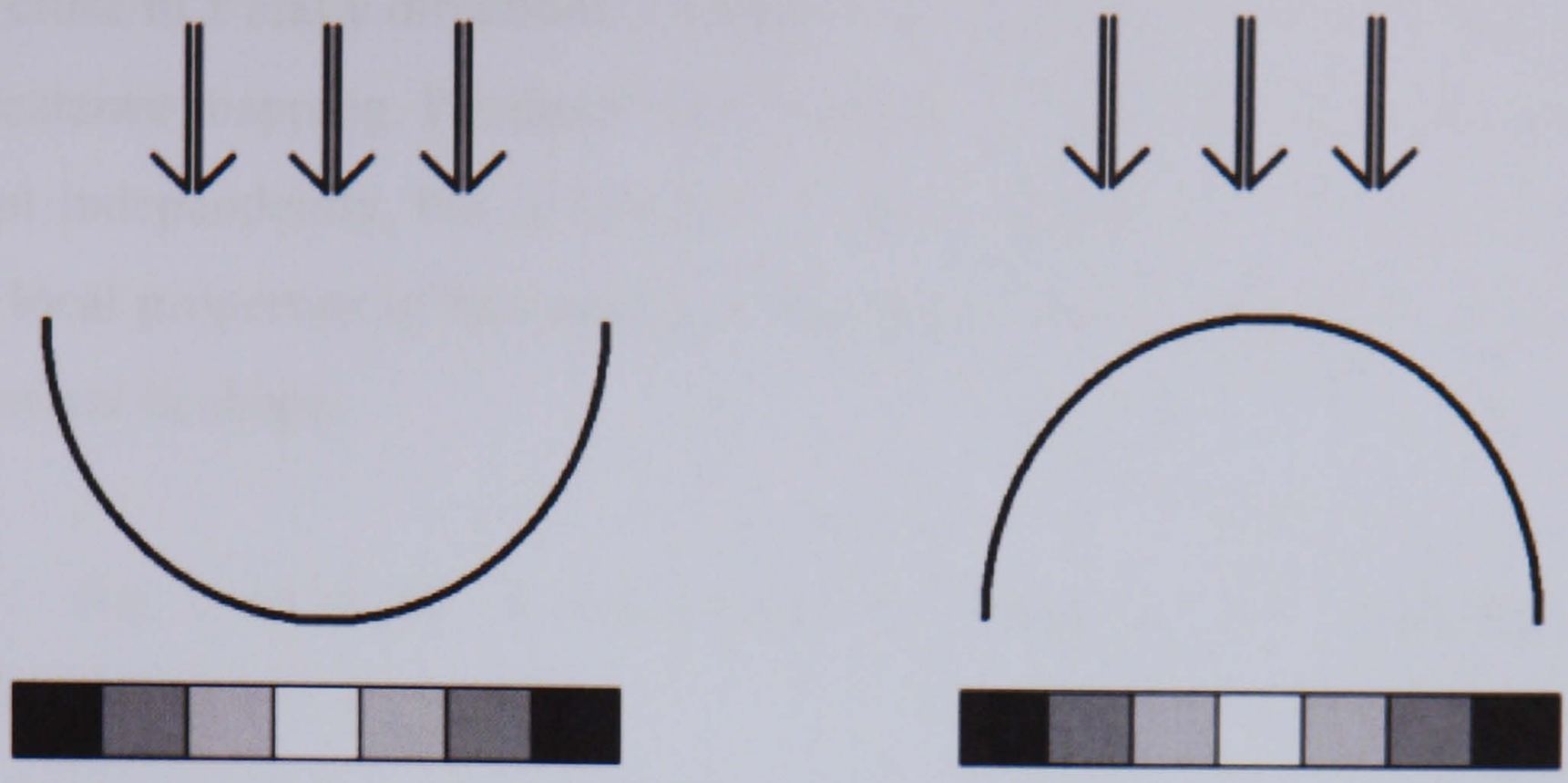


Figure 2.1 : The bas-relief ambiguity. With identical illumination (denoted by the arrows) the two shapes above produce identical intensity patterns.

Recent surveys of shape-from-shading techniques[71, 19] provide an extensive assessment of the state of the art, and separate current techniques into broad categories. Minimisation or Optimisation approaches attempt to address the issue of the extra unknowns in the system by applying additional constraints to generate an error term which may then be minimised. Both Ikeuchi and Horn[33] and Brooks and Horn[5] apply brightness and smoothness constraints – the former requiring that the resulting shape produces the same intensity value as the input image, and the latter ensuring that the surface of the result is smooth. An issue is implicit in this approach in so far as the smoothness constraint is not based upon any intrinsic property of the system, and so may be poorly weighted in relation to the brightness constraint leading to over- or under-smoothed results. Other approaches[23,29] apply integrability rather than smoothness as a constraint, requiring that the result recover surfaces for which $Z_{xy}=Z_{yx}$, or that the intensity gradients of the reconstruction and the original image

are close in x and y directions[72]. Linear approches use a linearisation of the reflectance mapping. Pentland[51] for example, calculates the normal of each point independently, but is forced to make a substantial assumption about the local properties of the surface – that at any point the surface is locally spherical in shape.

The conclusion of the survey by Zhang *et al*[71]. includes the following, telling, list:

1. *All the SFS algorithms produce generally poor results when given synthetic data*
2. *Results are even worse on real images, and*
3. *Results on synthetic data are not generally predictive of results on real data*

The remaining form of solution to the shape-from-shading problem was suggested initially by Horn[28], and at in the quoted survey did not met with markedly superior results to the other techniques. Solution of this form are generally termed propagation techniques[40, 58] since they start from a singular point – a point at which the intensity is at a maximum, and is surrounded by points of lower intensity which can therefore be assumed to be a point at which the surface normal is pointing directly towards the camera, given that the light source is coaxial with the camera. The surface then be calculated by propagating the shape information from this known point. Prados and Faugeras[54] demonstrate a substantial step forward using a propagation technique, by re-posing the basic modelling assumption. By positing a perspective, rather than orthogonal projection and a light source at the optical centre – a situation which closely models high powered flash

photography. They demonstrate the technique on a variety of real facial images, with relatively successful, though far from natural-looking, results.

2.1.2 Face-specific shape-from-shading techniques

Given that the solutions proposed to the traditional shape-from-shading problem it is clearly of benefit, when considering only facial images, to attempt to apply some intrinsic property of faces as one of the constraints for a minimization method. Atick *et al*[2] describe such an approach inspired by the theory that human interpretation of three dimensional structure may be prototype based, limiting the search space by the use of PCA to produce a low-dimensional parameterisation (using only the first 200 components) of the facial model on the basis that the vast majority of the components are insignificant and so irrelevant to the search process. Nandy & Ben-Arie[45] propose approaching the problem using recognition as a first step; dividing the face up into parts, each parameterised into a low dimensional representation using PCA independently, and performing recognition using trained neural networks to infer the shape of each area. The parts are then recombined to form a whole. Vetter and Blanz[66] suggest the use of linear combinations of training faces in 2d to match a novel face (again using an iterative error reduction process), using image correspondencies to relate the training samples to each other and to their 3d models, hence connecting the novel face to a matched 3d model. Many alternative techniques[42, 59] rely upon multiple images or video, and so address a fundamentally different situation to this work.

Given the approximations that are inherent in the general shape-from-

shading approach, this thesis proposes that the explicit modelling of the light source – face – image system could be improved upon by the use of an implicitly trained system based entirely upon the statistically inferred relationship between image and structure, based upon the fact that faces form a definite sub-set of images. By building such a model, issues such as the reflectance properties of skin, lighting situations with more than one source (practically guaranteed in any real image taken without a flash), and consistent resolution of hill-valley ambiguities may all be implicitly built into the system, providing the training data is representative of the conditions under which reconstruction images are captured.

2.2 High dimensional versus low dimensional representation of faces

The techniques described so far attempt to characterise the face as some kind of special or unusual data such as by adding local convexity constraints or dividing the face into feature regions, by explicitly modelling some of the physical properties of the light–face–camera system using a variety of assumptions about the reflectance properties of the face and the nature of the illumination. In addition to these alternatives, the reduction of the dimensionality of the space may be achieved by automated methods such as PCA. There is an unstated assumption underlying those techniques; namely that explicit high-level modelling, using some conceptual basis, provide a distinct benefit when attempting to describe or infer the relationship between a two dimensional photographic image of a face and the three dimensional structure of the face shown in the image. Assuming that some relationship between the two exists, as inferred by Hyde and Robinson[30], then the more correct this modelling of the relationship, the

more accurately a 3d model may be inferred from a photographic image. This assumption that a conceptually derived model is of benefit is neither a foolish nor irrational, though the logical extension of this argument is not routinely considered. The demonstrable success of these techniques shows that such modelling has some merit, but the question must be asked: What of any statistical relationship in the data which is not represented in a given proposed model? Should sufficient information be lost in these implicit, potentially non-intuitive relationships, the use of a structured model may hinder performance.

The complexity of the system suggests that this issue is significant. A useful way of quantifying this complexity is to calculate the unoptimised dimensionality of the system. Assuming that model of a face is to be represented as a greyscale depth map of n by m dimensions and is paired with a photographic texture image of the same n by m dimensions, also reduced to greyscale rather than full colour, then the model dimensionality is n by m by two, with a range of 0 to 255 in each dimension. This limitation is admittedly arbitrarily imposed on the depth image, as the depth values here are quantised and clipped in order to fit into the same range as an image even though there is no physical requirement to do so. It could be argued that the 0-255 limitation on the range of pixel values is equally arbitrary, however. The reason for general use of this value range for greyscale images is twofold: firstly, it is computationally convenient (as 0-255 can be represented in eight bits), and secondly due to the ubiquity of this format it is not possible to easily display images with a greater range than this on standard equipment. This does not mean that the extra information provided in a higher bit-depth or unquantised image would be irrelevant or useless. Some standard image formats do allow for more

quantisation levels though these are not commonly used for the above reasons. However, adhering to convention and limiting the data in this way has benefits – unquantised data, or data quantised to many more levels, would be much more impractical to manipulate mathematically, for example, and even with quantised, clipped data the space described by the arbitrary face model-image pair is phenomenally large. Even a very small image of 50 by 50 pixels, for example, leads to a space of 5,000 dimensions. In any practical application this dimensionality will be much larger – the United Kingdom Passport Service[67] for example specifies a minimum image size of 2126 by 1654 pixels, leading to an approximately 7 million dimensional space.

Prior art has arguably concentrated upon techniques for the reduction of this dimensionality – either implicitly or explicitly – for three reasons:-

- Most of the useful information should be concentrated in relatively few dimensions, provided that the dimensions used to represent the data are chosen intelligently. Eigenspace-based analysis techniques such as principal components analysis (PCA) attempt to provide optimally ordered, optimally powerful dimensions, and make up a large body of work in this field[2, 64, 35]. Given that the significant dimensions are relatively few, and dimensions are ordered by significance, a viable model of the system can safely ignore the majority "insignificant" dimensions.

- Dealing with n-dimensional hyperspaces is mathematically slightly more complex, but conceptually difficult to the point

that mental visualisation of the hyperspace generally fails to be constructive, and may also be actively detrimental to the understanding of operation in said space. Operations and geometries in very high dimensional space are often counter-intuitive. The most natural way for humans to deal with extremely high-dimensional data is to abstract, building logical models of the system to represent either symbolically or by encapsulating aspects of the system in discrete sub-models. This is a poor argument for statistical dimensional reduction, but can be a result of the use of high-level modelling of faces due to the limitations of the designer.

- Computational complexity of systems with high dimensionality can be a significant issue, causing the processing time for the operations used by a given analysis to become impractical. If it can be convincingly argued for a particular case that the reduction of dimensions does not significantly degrade the performance of the model whilst the analysis time can be reduced sufficiently to be useful then dimensional reduction becomes beneficial.

The former of these points will be addressed later, from section 2.4.4. Regarding the second, this is indeed a sensible way to model some systems. However, it is arguable that an abstract modelling of the system is perfectly achievable without any reduction of dimensionality provided that the abstraction does not rely upon any modelling based explicitly upon conceptual analysis of the system: a truly abstract statistical model can allow an implicit model to be developed whilst employing far fewer a

priori assumptions. For example, many of the shape-from-shading techniques discussed assume Lambertian reflectance which has been shown to be not the case for human skin[26, 70]. This has the potential to force resulting reconstructions away from the actual most probable data, and must be considered a problem.

In order to avoid this, a generic technique should be applied which has the fewest presuppositions possible – ideally none based upon the properties of "face" data. Rather, these properties should be built into a model by some training process based upon brute force statistical sample-based training. In this case, the expected relationships – if statistically valid – should be present in the resulting trained model along with less obvious implicit relationships. The model may not of itself provide any insight into the relationship, as all relationships will most probably remain implicit even in the trained model, but they will be incorporated and useful as a black box.

2.3 Re-representing the image/model pair as a point in n dimensional space

It is worth taking a moment to consider what is meant by the "n dimensional" space alluded to above, where n is defined by the number of pixels in the image, and to properly understand what this space describes. Consider a photographic image of a face. Rather than a picture, it could be interpreted as an array of numbers. This is exactly how a computer stores an image in memory, with black being represented as 0,0,0 and white being represented as 255, 255, 255 in most common image formats. As shown in figure 2.2, an image can easily be represented as a one-dimensional array.

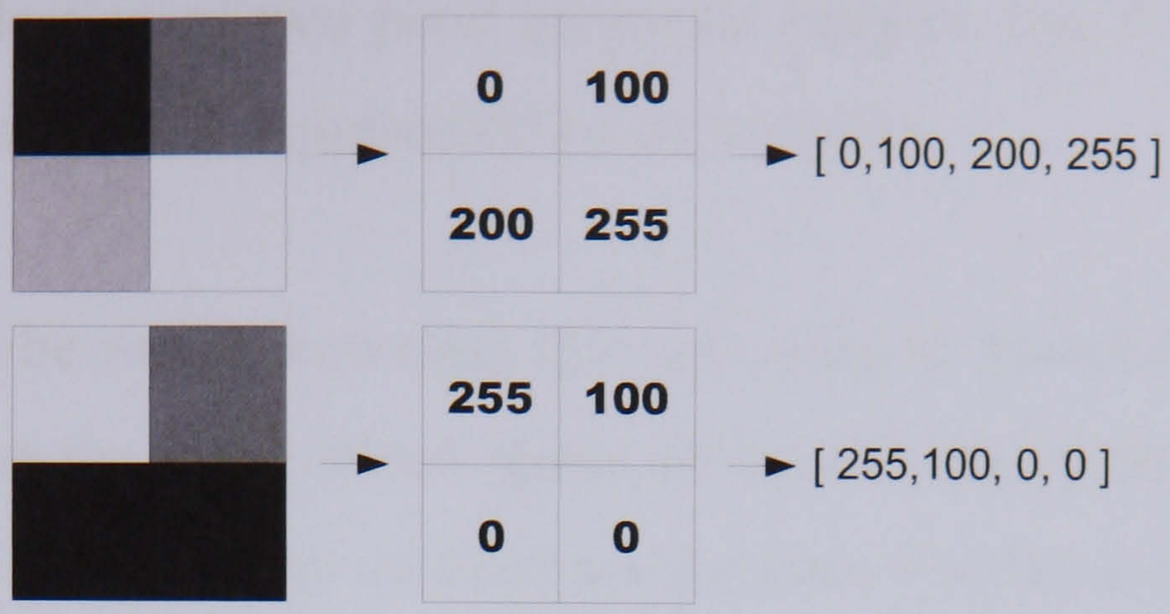


Figure 2.2: Representing images as one dimensional arrays. The images on the left are two pixels wide by two pixels high. They can be represented either as images, or as two dimensional arrays of values or as a one dimensional vector.

Using this notation any image can, provided the image is of the expected number of dimensions, be defined by an array of numbers. Novel images may be generated by varying each element in the array. An alternative way of looking at these strings of numbers, rather than as pixel values, is to see them as coordinates by which it is possible to navigate through all possible images – so by adjusting elements in the array it is possible to move through image space, arriving at the point in image space which is occupied by the desired image. Any image is defined by a point in an image space which has a distinct dimension for each pixel, and conversely any point in image space defines a unique image. If an image is represented in this way, it becomes clear that there is no fundamental difference between the image data and any data set which is defined over n dimensions. Taking it to the simple extreme, the volume and mass of a

collection of rocks could be represented in a similar space to that used to represent a collection of two pixel greyscale images, and would be a two dimensional space easily represented by an x y plot.

It should be noted, however, that this n by m dimensional space is much larger than the space which spans all front-facing facial images. The space is limited by the range of valid pixel values – in the case of standard 8 bit images, from 0 to 255 in each dimension. However, this space spans all possible images – all possible images of *faces*, assuming any level of correlation between face images, must be encompassed by a relatively tiny volume of this space. Assuming a gaussian distribution of faces in image space, all possible faces will be described by a *d* dimensional hyperellipsoid centred upon the mean image, where *d* is less than or equal to the dimensionality of the image. Given this assumption of gaussian distribution, and this generic data style of representation, a range of standard statistical tools are opened up for application to the problem of reconstruction: rather than building a three dimensional model based explicitly on shape-from-shading or similar techniques, the most probable data for any given image can be calculated on a purely statistical basis.

2.4 Generic statistical modelling

In order to better understand the process of calculating the expected values of a model from an image, it is useful to take the example of a much lower dimensional problem. Consider the situation whereby the height of sons is to be estimated from the heights of their fathers. To build a statistical model describing this relationship, a training set must be created by sampling a large random selection of fathers and sons from the

population. If it is assumed that the heights of both fathers and sons conform to gaussian distributions, and also that the heights of sons are related to the heights of their fathers then a distribution similar to that shown in figure 2.3 would be expected.

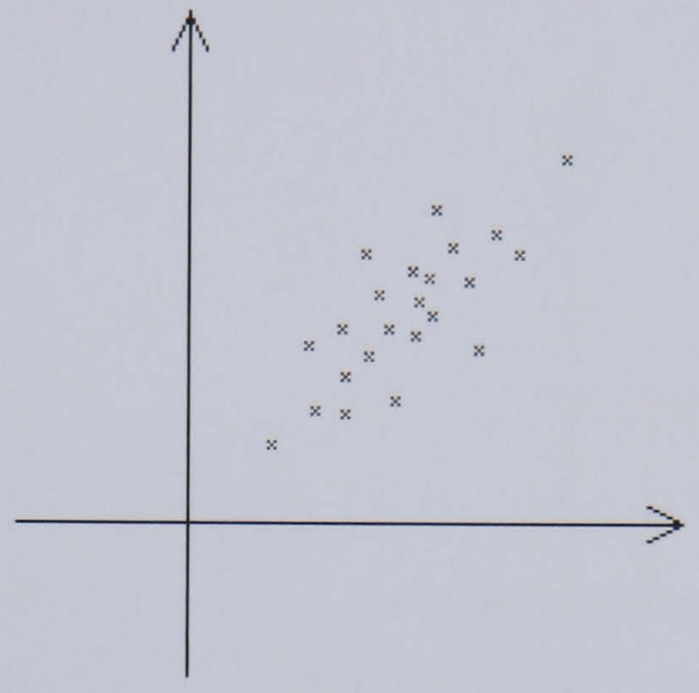


Figure 2.3: Sketch of a scattergraph representing two correlated variables, each having a gaussian distribution about some mean. For the purposes of the illustrative example, let the X axis represent the heights of fathers and the Y axis represent the heights of sons

2.4.1 Estimation based on the global mean

The most naïve estimate of the height of a son given the height of a father would be to find the most probable height of all sons irrespective of the heights of the fathers – i.e., the mean height of the sons. An estimate of this value is plotted in figure 2.4a It should be clear both from the sketch and from the process of estimation that this implicitly assumes that there is no correlation between the two axes, and as shown in figure 2.4b the predicted heights of the sons of very short or very tall fathers using this technique fall well outside the area within which all of the

samples lie. Taking a cross section of probability along the dashed projection line will yield a peak at a distance from the calculated value. Whilst the most probable height of a son *for a father of unknown height* has been accurately estimated, the combination of this son with a father of given height may be extremely improbable.

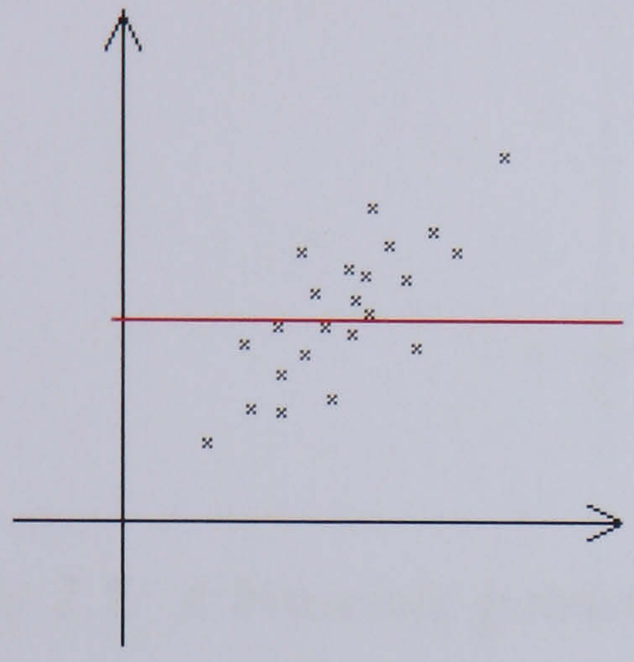


Figure 2.4a: The mean Y-value (Height of sons) has been calculated, and indicated by the horizontal line. Using this technique, any given X value (height of father) is projected to the mean y value to calculate the expected corresponding value.

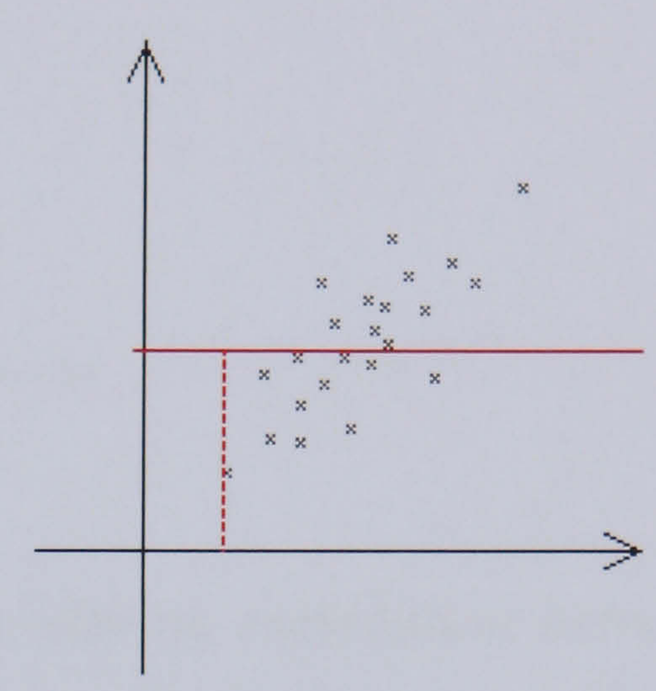


Figure 2.4b: As demonstrated by the dashed line of projection, this technique can engender results which are at considerable variance with the sample data.

The only cases where this global mean y-axis value will be guaranteed to be representative of any given x-axis value are where the long axis of the distribution lies exactly perpendicularly to the x-axis (and hence the expected value for any probe value is coincident with the global mean) or where there is no correlation between the axes, leading to a

distribution similar to that shown in figure 2.5. This could be seen as a plot of the heights of men and their neighbours sons, for example, where there is no relationship between the data but both data sets have a gaussian distribution about some mean.

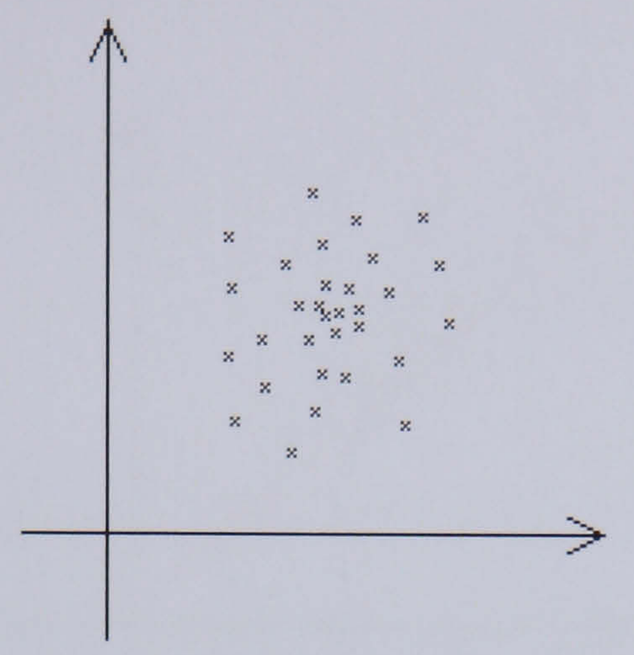


Figure 2.5: A bivariate gaussian distribution with no correlation between the two variables. In this case, the expected y value given known x is the global mean, regardless of the value of x.

2.4.2 Estimation by regression

A more sophisticated approach may be taken based upon the feature whereby as the level of correlation between the two gaussian distributions increases, the plot of the distribution becomes more elongated along a major axis. By regressing the distribution to find this axis, it may be used to provide a projection line, with the heights of sons being equal to *f*(Height of father) where *f* describes the line of the major axis. As can be seen, this technique allows the estimation of a value which lies much close to the body of the samples even for outlying x-axis values.

It should briefly be noted that nonlinear regression techniques also

exist which plot higher than first order regression lines through a space. These are not of interest in this case, however, as the initial assumption that the data conforms to some multivariate gaussian precludes the use of higher-order regression; a gaussian distribution by definition is best described by linear regression.

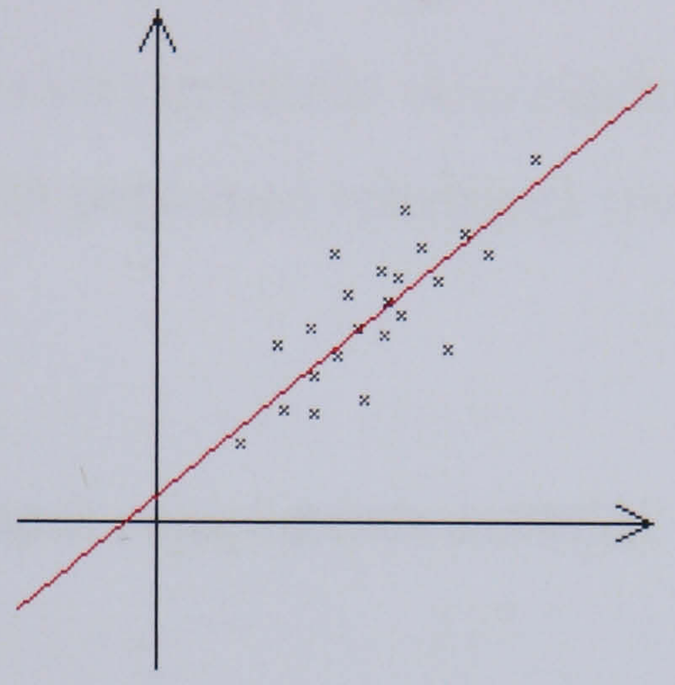


Figure 2.6: A linear regression line can be plotted through the sample space. A well fitted line will describe the major axis of said space. The expected y-value of a given x-value may be read off by projecting from the x axis to the regression line. The regression line does not necessarily intersect the Y axis at the origin.

This is a common solution to the task of expected value estimation. An implicit assumption upon which this calculation is based is, however, often not considered. The technique is based upon the reduction of the dimensionality of the space describing the distribution to a line, thus making possible a one-to-one mapping from known variable to unknown variable. This implicitly assumes that the maximum amount of information is contained along this line, and that any variation in any dimension perpendicular to this line is insignificant and can be ignored. This is effectively a low dimensional application of principal components analysis

(PCA) in that the principal of two possible components has been identified, and the minor components have been discarded. Before taking issue with the veracity of this technique, it is useful to consider PCA as the higher-dimensional analogue of linear regression and equivalent to this technique when dealing with high dimensional sample spaces.

It should be noted that regression is an extremely old, and well-studied phenomenon, with published references stretching back as far as 1886[25].

2.4.3 Principal components analysis

Whilst there are alternative regression techniques available, especially in low dimensional spaces, PCA has many benefits – it is a well defined technique which operates effectively in any number of dimensions. Qualitatively, PCA transforms a sample distribution in *n* measured dimensions into a sample distribution in n optimally powerful, mutually perpendicular dimensions. These new dimensions are also ranked in order of significance, and for a given space can provide not only the optimally significant vectors (eigenvectors) of the space, but also the relative significance of each vector – the eigenvalues. The most significant component – the major axis of the elliptical gaussian distribution in the earlier example – has the largest eigenvalue. It is up to the user to decide the point at which the vectors calculated cease to be significant, based upon the relative eigenvalues of each component. At this point, the user must make a choice between accuracy (which requires the use of as many dimensions as possible) and the computational simplicity provided by using fewer dimensions.

A limitation of PCA is that the technique relies upon either richly sampled data (typically requiring many more samples than there are possible degrees of freedom in the original population) or some technique of reliably estimating the covariance matrix of the population from sparse data. At the theoretical absolute lower limit, it is impossible for PCA to calculate more than s-1 eigenvectors, where s is the number of samples in the space (though such sparse sampling would lead to low certainty of accurate vector generation. The reason for this limitation is that to describe a one dimensional line, two points are required. For a two-dimensional plane, three points, and so on. Techniques exist to circumvent this limitation, such as the linear algebra manipulations suggested by Turk and Pentland[64]. These techniques cannot invent data, and merely serve to add uncorrelated noise which will gather in dimensions of lowest significance and allow the algebra required for the calculation of the transform to be soluble in the Turk and Pentland rearrangement. In this case the vectors beyond the s-1 principal vectors are guaranteed to be junk vectors. The base assumption is that whilst the samples are insufficient to model the entire space, they do contain representative enough information to allow accurate calculation of all significant vectors, with only dimensions with insignificant variation in the population being lost – though the insignificance of these dimensions is by no means guaranteed. Given the extremely high dimensionality of texture plus model face space, and relative scarcity of samples available, this issue is not negligible.

2.4.4 Limitation of low-dimensional modelling of a multivariate gaussian distribution

Whilst linear regression is a widely used technique, the assumption that the sample distribution may be accurately represented in a lower dimensional space (a line in the simple two dimensional example) is demonstrably dangerous. Returning to the two dimensional scatter plot for ease of conceptualisation, the scattered samples can be seen as describing an elliptical distribution. The indicated ellipse is not a strict bound of the distribution (as a gaussian distribution extends infinitely, though tends towards infinitely small probability), but a locus of equiprobability. Similar loci may be plotted within and without the initial ellipse, forming a set of concentric ellipses centred upon the mean of the distribution. Ellipses closer to the mean describe loci of greater probability than loci further from the mean, and hence the loci may be considered as contour lines describing a probability gradient which peaks at the mean.

Consider the situation described in figure 2.7, where a known x value is used as a probe into a bidimensional gaussian distribution, with the distribution centred upon the origin (this can be achieved by deducting the mean value from all samples should the distribution not be centred upon the origin). Assuming the minor axis of the distribution to be negligible and hence projecting up to the major axis of the space results in an expected Y value of Y_m, and together these define a point on the outer equiprobable locus shown. However, it is apparent that the line of the x projection passes inside this locus, and cuts through the body of the ellipse of which this is the perimeter. However, some point upon the x projection will coincide with a contour of equiprobability which no other point on the x projection

will lie inside. This minimal contour is the smaller ellipse in figure 2.7 and has a higher significance than the larger contour. The optimal expected value of Y, therefore is Y_{opt}, as this is the point of maximum probability on the projection line where x is equal to the known probe value.

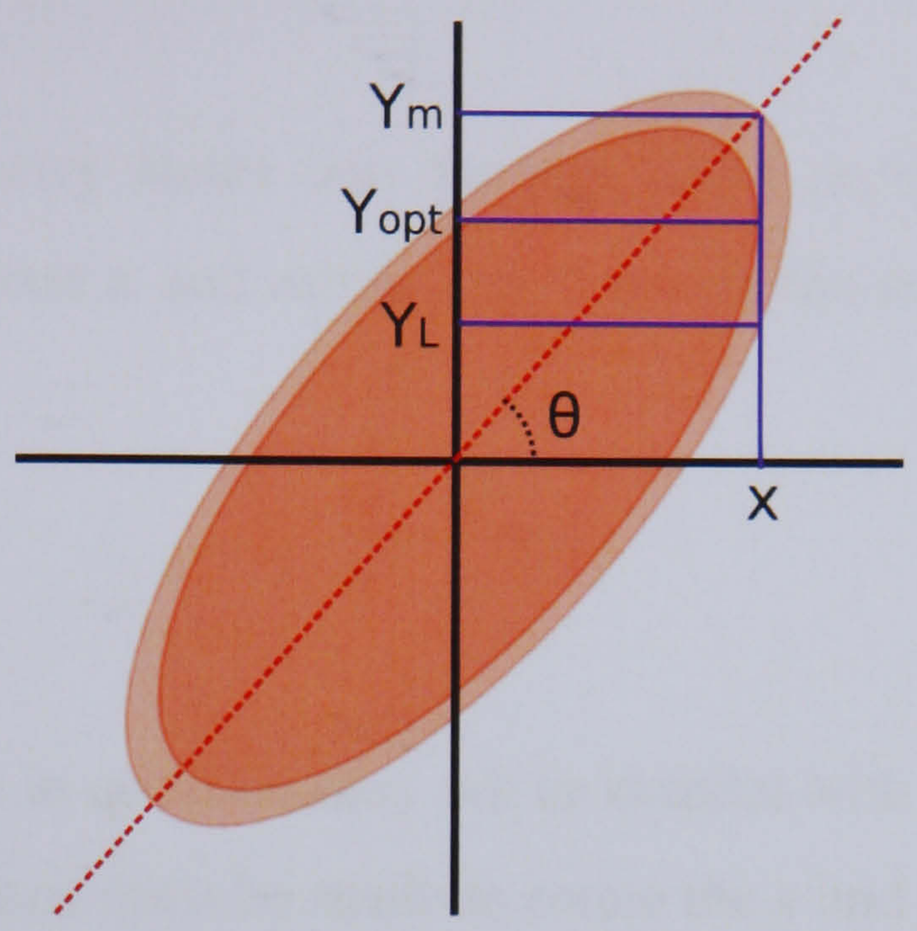


Figure 2.7: Estimation of values from a multivariate gaussian distribution. Note that Y_{opt} is the point of maximum likelihood for X=x, not Y_m the point which is chosen when the minor axis of the elliptical distribution is assumed to be negligible. The ellipses describe contours of equiprobability, with probability higher for ellipses nearer the mean (this assumes that the mean is at the origin)

The optimal estimated value for probe value x (y_{opt}) may be considered to be a noise free signal. The value predicted by projection to the major axis of the elliptical contour described by the training sample may be considered to be the noise free signal plus some noise value. Therefore the signal to noise ratio (SNR) may be calculated in terms of the orientation of the major axis of the contour relative to the axis from which the probe is projected and the eccentricity of the ellipse.

Using trigonometry it is possible to define y_m in terms of x and θ thus (note that from this point the ellipse is assumed to be centred upon the origin):

$$y_m = x\tan(\theta) \qquad \textit{Equation 5}$$

Ellipse geometry states that for any point on the perimeter of an ellipse with major axis a and minor axis b where the major axis lies along the x axis that

$$\frac{\hat{y}^2}{b^2}+\frac{\hat{x}^2}{a^2}=1 \qquad \textit{Equation 6}$$

As the ellipse in question may not be coaxial with the x axis, the following substitutions must be made to rotate the x and y values of a given point to conform to the prerequisite of the above equation 6:

$$\hat{x} = x\cos(\theta)+y\sin(\theta) \qquad \textit{Equation 7}$$
$$\hat{y} = y\cos(\theta)-x\sin(\theta) \qquad \textit{Equation 8}$$

It should also be noted that y_{opt} can be defined as being halfway between y_L, the lower point at which the projection line intersects the contour, and y_m (as the ellipse which touches the x projection at this point is similar to the larger ellipse of which the probe is projected up to the major axis - perimeter intersection, shares the major axis vector and is also centred upon the origin). Therefore:

$$y_{opt} = \frac{y_m + y_L}{2} \qquad \textit{Equation 9}$$

And finally, SNR can be defined as

$$SNR = 20\log\left(\frac{y_{opt}}{y_m - y_{opt}}\right) \qquad \textit{Equation 10}$$

Substituting equation 9 into equation 10:

$$SNR = 20\log\left(\frac{\frac{y_m + y_L}{2}}{y_m - \frac{y_m + y_L}{2}} = \frac{y_m + y_L}{y_m - y_L}\right) \qquad \textit{Equation 11}$$

By substituting equations 7 and 8 into equation 6, the following quadratic is produced which describes the two points at which the x projection line intersects the perimeter of the larger ellipse:

$$\hat{y}^2 a^2 + \hat{x}^2 b^2 - a^2 b^2 = 0 \qquad \textit{Equation 12}$$

Therefore:

$$a^2(y\cos(\theta) - x\sin(\theta))^2 + b^2(x\cos(\theta) + y\sin(\theta))^2 - a^2 b^2 = 0$$

Equation 13

Rearranging to a quadratic in y:

$$\begin{gathered} y^2(a^2\cos^2(\theta) + b^2\sin^2(\theta)) \\ + y(2x\sin(\theta)\cos(\theta)(b^2 - a^2)) + x^2(a^2\sin^2(\theta) + b^2\cos^2(\theta)) \\ - a^2 b^2 = 0 \end{gathered}$$

Equation 14

The two roots of this equation may be calculated using the general quadratic solution, where A is the y^2 coefficient, B is the y coefficient and C the y-independent coefficient. In this case, the higher of the two roots is known (as one of the roots is y_m and one is y_L), so:

$$y_m = \frac{-B + \sqrt{B^2 - 4\mathrm{AC}}}{2A} = x\tan(\theta) \quad \text{and} \quad y_L = \frac{-B - \sqrt{B^2 - 4AC}}{2A} \qquad \textit{Equation 15} \qquad \textit{Equation 16}$$

Substituting these into equation 11:

$$SNR = 20\log\left(\frac{-B + \sqrt{B^2 - 4AC} - B - \sqrt{B^2 - 4AC}}{-B + \sqrt{B^2 - 4AC} + B + \sqrt{B^2 - 4AC}} = \frac{-B}{\sqrt{B^2 - 4AC}}\right)$$

Equation 17

However, by rearranging the equation determining the higher root of the quadratic it is possible to state that:

$$\sqrt{B^2-4AC}=2Ax\tan(\theta)+B \qquad \textit{Equation 18}$$

And hence:

$$SNR=20\log\left(\frac{-B}{2Ax\tan(\theta)+B}=\frac{-1}{2\frac{A}{B}x\tan(\theta)+1}\right) \qquad \textit{Equation 19}$$

where:

$$A=a^2\cos^2(\theta)+b^2\sin^2(\theta) \qquad \textit{Equation 20}$$
$$B=2x\sin(\theta)\cos(\theta)(b^2-a^2) \qquad \textit{Equation 21}$$

Simplifying the non-unity term in the denominator:

$$2\frac{A}{B}x\tan(\theta)=\frac{a^2\cos^2(\theta)+b^2\sin^2(\theta)}{\sin(\theta)\cos(\theta)(b^2-a^2)}\tan(\theta)=\frac{a^2+b^2\tan^2(\theta)}{b^2-a^2}$$

Equation 22

Substituting back in to equation 19:

$$SNR=20\log\left(\frac{-1}{\frac{a^2+b^2\tan^2(\theta)}{b^2-a^2}+1}\right) \qquad \textit{Equation 23}$$

The eccentricity of an ellipse is defined as:

$$e=\sqrt{1-\frac{b^2}{a^2}} \qquad \textit{Equation 24}$$

Equation 23 can therefore be rearranged to replace all a and b terms with e terms thus:

$$SNR = 20\log\left(\frac{-1}{\frac{1+\frac{b^2}{a^2}\tan^2(\theta)}{\frac{b^2}{a^2}-1}+1}\right) = 20\log\left(\frac{-1}{\frac{1+(1-e^2)\tan^2(\theta)}{-e^2}+1}\right)$$

$$SNR = 20\log\left(\frac{e^2}{(1-e^2)(1+\tan^2(\theta))}\right) \qquad \textit{Equation 25}$$

This function can be charted to give an impression of the effect upon the quality of major-axis projection as an estimator as the eccentricity and rotation of the gaussian sample distribution vary.

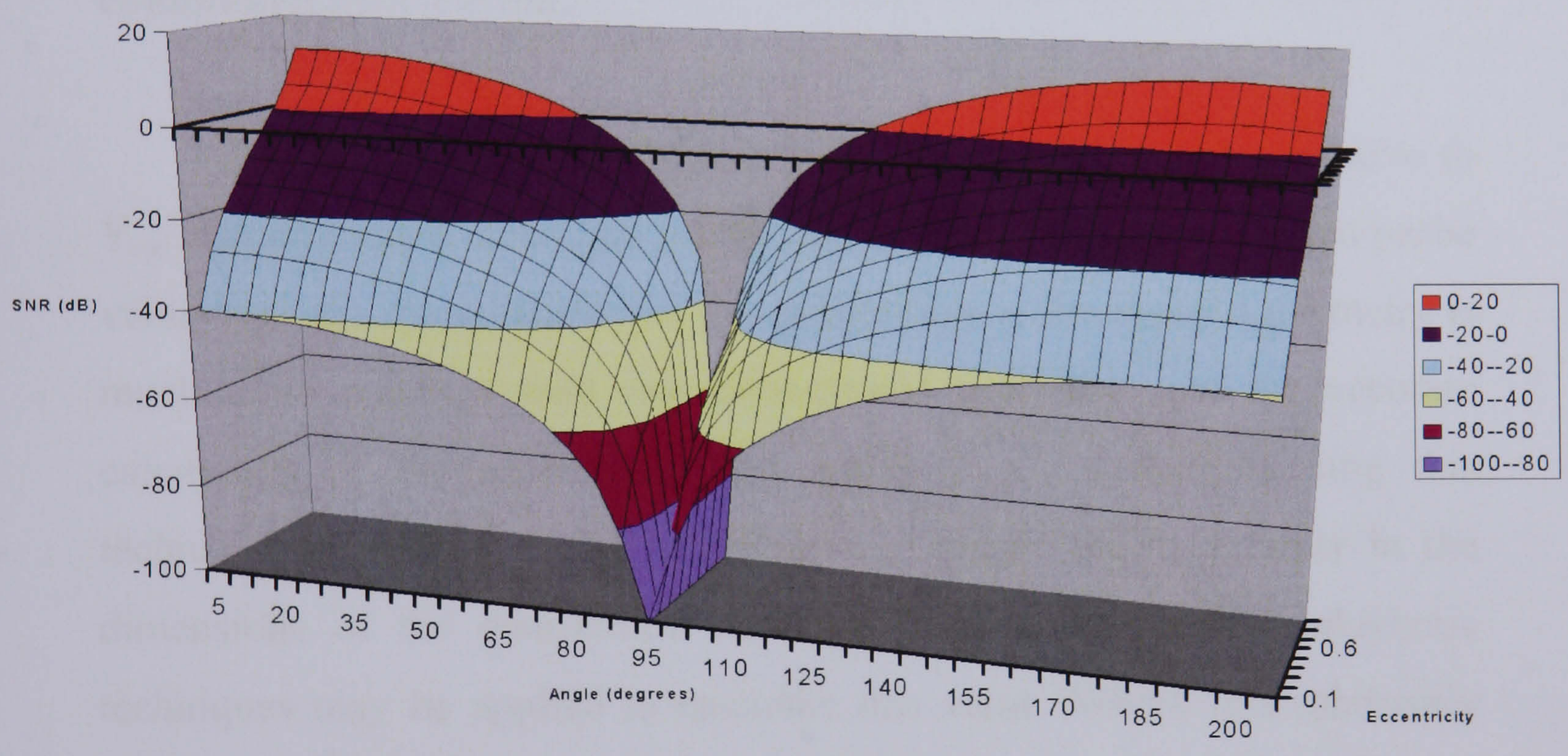


Figure 2.8: Signal to noise interpretation of the estimation (assuming the maximally probable point where $X=x_{probe}$ is signal, and deviation from this is noise) as the angle of the major axis and the eccentricity of the distribution vary. SNR tends asymptotically to infinity as eccentricity tends to 1, negative infinity as eccentricity tends to 0. SNR also tends to negative infinity as the angle of the major axis tends to 90 degrees though the measure is meaningless at exactly 90 degrees.

As can be seen, signal to noise ratio varies between positive and negative infinity, but for the majority of cases the performance of this estimator is very poor. It can be concluded from this that the major-axis projection estimator is severely compromised by dimensions of unknown eccentricity and rotation, and in the presence of such unmodelled "negligible" dimensions should not be used. Instead, some estimation of the properties of these dimensions must be used. It should also be noted that where the major axis lies exactly along the x axis, this result is meaningless (as both signal and noise are zero).

2.4.5 Estimation of optimal values from a known probe by conditional distribution

The above geometric argument for the inferiority of Y_m relative to Y_{opt} also provides a techinque for the calculation of Y_{opt} from a given probe value from the characteristics of the distribution. *n* dimensional geometry is much more complex than two dimensional geometry, and the accurate calculation of the eccentricity and angle of a distribution using this technique relies upon sufficient samples to reduce the uncertainty in the dimensions of the distribution to zero. However, alternative algebraic techniques may be applied to calculate this value. Whilst less inherently easy to visualise than the geometric solution as an argument for the merits of the technique, this can provide a more practical analysis method for the general n dimensional problem. The following argument is based upon that reported by Hyde and Robinson in [31].

Consider the random vector $\boldsymbol{X}$ distributed as an n dimensional multivariate normal with expected value vector μ and covariance matrix Σ,

ie $N_n\ (\mu, \Sigma)$, and a sample from the distribution $\boldsymbol{P}$ in which some of the coefficients are known and some are not. For this sample the unknown values are to be estimated. This can be considered a probe. and is equivalent to the known x value for which a corresponding y value was required in the geometric example. A binary "mask" vector $\boldsymbol{M}$ may be used to indicate the known values in $\boldsymbol{P}$ for which a permutation matrix $\boldsymbol{R}$ can be defined which will reorder $\boldsymbol{M}$ in to a column of q zero values followed by p non-zero values (where $p+q = n$). This permutation matrix, when applied to the probe, moves all unknown values to the top q positions of the vector.

$$\boldsymbol{P}_{perm} = \boldsymbol{R}\,\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_1 \\ \cdots \\ \boldsymbol{P}_2 \end{bmatrix} \qquad \textit{Equation 26}$$

Where $\boldsymbol{P}_1$ is a column vector of q undefined values and $\boldsymbol{P}_2$ is a column vector of p known values. Similarly,

$$\boldsymbol{X}_{perm} = \boldsymbol{R}\,\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \cdots \\ \boldsymbol{X}_2 \end{bmatrix} \qquad \textit{Equation 27}$$

The parameters of the distribution must also be reordered to match $\boldsymbol{RP}$ and hence become:

$$\boldsymbol{\mu}_{perm} = \boldsymbol{R}\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \cdots \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_{perm} = \boldsymbol{R}\,\boldsymbol{\Sigma}\,\boldsymbol{R} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} \vdots \boldsymbol{\Sigma}_{12} \\ \cdots \vdots \cdots \\ \boldsymbol{\Sigma}_{21} \vdots \boldsymbol{\Sigma}_{22} \end{bmatrix} \qquad \textit{Equation 28} \quad \textit{Equation 29}$$

As in[36] a matrix $\boldsymbol{A}$ can be defined with submatrices thus:

$$\boldsymbol{A} = \begin{bmatrix} \underset{q\times q}{\boldsymbol{I}} & \vdots & \underset{q\times p}{-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}} \\ \cdots\cdots\cdots & \cdots & \cdots\cdots\cdots \\ \underset{p\times q}{0} & \vdots & \underset{p\times p}{\boldsymbol{I}} \end{bmatrix} \qquad \textit{Equation 30}$$

i.e. the upper left and lower right submatrices are identity matrices and the lower left submatrix is composed entirely of zero values.

The random vector:

$$A(\Sigma_{perm}-\mu_{perm})=A\begin{bmatrix}X_1-\mu_1\\ \cdots\cdots\\ X_2-\mu_2\end{bmatrix}=\begin{bmatrix}X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ X_2-\mu_2\end{bmatrix}$$

Equation 31

is a linear transformation of the normally distributed random vector X_{perm} and therefore is itself normally distributed with mean:

$$E\left[A(X_{perm}-\mu_{perm})\right]=A\,E\left[(X_{perm}-\mu_{perm})\right]=0$$

Equation 32

and covariance matrix :

$$A\Sigma_{perm}A^T$$

Equation 33

As covariance matrices are symmetric, the identities $\Sigma_{22}^T=\Sigma_{22}$ and $\Sigma_{12}^T=\Sigma_{21}$ may be used to calculate equation 34:

$$A\Sigma_{perm}A^T=\begin{bmatrix}I & \vdots & -\Sigma_{12}\Sigma_{22}^{-1}\\ \cdots & \cdots & \cdots\\ 0 & \vdots & I\end{bmatrix}\begin{bmatrix}\Sigma_{11} & \vdots & \Sigma_{12}\\ \cdots & \cdots & \cdots\\ \Sigma_{21} & \vdots & \Sigma_{22}\end{bmatrix}\begin{bmatrix}I & \vdots & 0\\ \cdots & \cdots & \cdots\\ (-\Sigma_{12}\Sigma_{22}^{-1})^T & \vdots & I\end{bmatrix}=\begin{bmatrix}\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \vdots & 0\\ \cdots & \cdots & \cdots\\ 0 & \vdots & \Sigma_{22}\end{bmatrix}$$

Equation 34

This is the motivation for the choice of A: the covariance matrix has upper right and lower left corner submatrices which are both zero matrices. This is significant as it means that and $X_2-\mu_2$ have zero covariance, and are therefore independent. This means that the quantity $X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$ may be considered as a distinct $q\times q$ multivariate normal distribution. Vector X_2 may take the probe values P_2, and by substitution into equation 32 it may be stated that:

$$E[X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(P_2-\mu_2)]=0$$

Equation 35

As $\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{P}_2-\boldsymbol{\mu}_2)$ is known, the mean of $\boldsymbol{X}_1$ (i.e., the expected value of the unknown data corresponding to known data $\boldsymbol{P}_2$) is given by:

$$E[\boldsymbol{X}_1]=\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{P}_2-\boldsymbol{\mu}_2) \qquad \textit{Equation 36}$$

Hence it is possible to calculate the expected values corresponding to a probe directly for an arbitrarily dimensioned space provided the distribution of the sample data conforms to a multivariate normal distribution. A final step to reorder the estimated and original data to conform to the original order of the data is to apply the inverse permutation matrix $\boldsymbol{R}^{-1}$.

Whilst this is an indirect derivation, a proof using conditional densities directly can be found in [36].

2.4.5.1 Worked Example

Consider a distribution represented by the following samples, where some collection of artefacts with a relationship between their length and weight are being modelled:

Length	13	12	15	13	12
Weight	8	6	9	6	7

Table 2.1: Samples for a two-dimensional example

A new sample, with length 14 is observed, and a weight must be estimated based upon the known samples. First, the mean and covariance of the samples are calculated:

$$\mu=[13 \quad 7.2] \qquad \Sigma=\begin{bmatrix}1.20 & 0.92\\ 0.92 & 1.36\end{bmatrix}$$

The new sample is $X=[14 \quad x]$; we must therefore apply permutation mask [1 0], and re-order all matrices and vectors accordingly:

$$\mu_{perm}=[7.2 \quad 13] \qquad \Sigma_{perm}=\begin{bmatrix}1.36 & 0.92\\ 0.92 & 1.2\end{bmatrix} \qquad X_{perm}=[x \quad 14]$$

Referring back to Equation 36, the variables may now be substituted in thus:

$$\mu_1=7.2$$
$$\mu_2=13$$
$$P_2=14$$
$$\Sigma_{12}=0.92$$
$$\Sigma_{22}^{-1}=\frac{1}{1.2}$$

Therefore the expected weight of the new sample is

$$E(x)=7.2-\frac{0.92}{1.2}(14-13)=6.43$$

2.4.6 Compensation for sparse sampling of the training population

In order to apply the estimation technique described in section 2.2.5 the covariance matrix and mean of the face/depth space must be known. In the absence of definitively correct covariance and mean control parameters for the space, it becomes necessary to estimate from sample data. Calculation of an estimate for the population mean is simple assuming that the mean of the sample set approximates the mean of the population. Unless there is heavy bias in the training samples, this is a reasonable assumption to make. Estimation of the covariance matrix is equally straightforward assuming that sufficient samples are present. The scatter matrix, **S**, is commonly used as an estimation for covariance and is defined for the general case where $\boldsymbol{x}$ is a random variable as:

$$\boldsymbol{S}=\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T \qquad \textit{Equation 37}$$

Where N is the number of training samples and $\hat{\mu}$ is the mean of all samples. Note that this is the total scatter matrix, and that individual scatter matrices must also be defined for each class. In this instance, however, all training samples belong to the same class, and so only the total scatter matrix need be considered. It is a given that equation 37 can only yield a useful approximation in the case where N>=M where M is the number of samples needed to perform a direct estimation of the class covariance and is related to the dimensionality of the data. Whilst M is not generally determinable, it cannot be less than the number of dimensions that in-class data is composed of. It is generally considered that the number of samples used when estimating covariance from the scatter matrix should not be less than five times the dimensionality of the space. In the case of the three dimensional face models, this condition is almost guaranteed to never be met, due to both the high dimensionality of the data and the relative scarcity of training samples. As discussed, this precludes the use of PCA as technique for the estimation of the eigenvectors and hence covariance of the data. Algebraic manipulation may be employed to allow PCA to provide a result[64], but this does not address the root problem that N data samples cannot define an M dimensional space when N<M. This solution results in the calculation of several "noise" vectors which are known to have a poor relationship to the actual space. This is often considered permissible as these vectors are those with lower eigenvalues (referring back to section 2.2.4, equivalent to the vector describing the minor axis of the ellipse). However, as demonstrated, this assumption is flawed in that such dimensions can only be considered negligible for the reconstruction

task under extremely specific conditions. There is no way to guarantee the output of the sparse sample rearrangement of PCA will conform to these conditions, and so it must be considered a dangerous technique to employ at best. Dimensional reduction, the other technique which could be used to provide an estimation of at least some of the covariance matrix, is only valid if it is true that the class space is composed of a well-defined low dimensional representation mixed with higher dimensional noise. This is assumed to be the case frequently, though there is no evidence to support it. Dimensional reduction should therefore be regarded with suspicion: It is entirely possible that reduction of the dimensionality of the class is being carried out at great expense to the information content of the distribution.

Robinson[57] proposes a general solution to the estimation of covariance for multiple class systems which has the benefit of simplicity in the single-class case as only one control variable is required, as is demonstrated below. The primary thrust of the example in this paper is the use of regularised covariance estimation (RCE) for inter class discrimination, though the estimation technique is presented as useful for appearance based analysis in general.

The general form of RCE states that the best estimation for covariance is:

$$\boldsymbol{\Sigma}_i(\alpha,\beta)=\alpha\,\boldsymbol{S}_i+(1-\alpha)\,\boldsymbol{S}_{total}+\beta\,\boldsymbol{I} \qquad \textit{Equation 38}$$

Where $\boldsymbol{\Sigma}_i$ is the class covariance, $\boldsymbol{S}_i$ is the class scatter matrix, $\boldsymbol{S}_{total}$ is the total scatter matrix across all classes, $\boldsymbol{I}$ is the identity matrix and (α,β) are the regularisation parameters. Whilst α is normalised and takes a value between 0 and 1, β may vary freely in order to allow the

identity matrix to subsume the scatter matrix should the scatter matrix be sufficiently ill-defined. However, unlike the classification examples presented by Robinson, for this reconstruction application all samples belong to the same class, hence $\boldsymbol{S}_i = \boldsymbol{S}_{total}$. This is significant since equation 34 may therefore be simplified, thus removing one of the two regularisation coefficients:

$$\boldsymbol{\Sigma}(\alpha,\beta) = \alpha\,\boldsymbol{S}_{total} + \boldsymbol{S}_{total} - \alpha\,\boldsymbol{S}_{total} + \beta\,\boldsymbol{I} = \boldsymbol{S}_{total} + \beta\,\boldsymbol{I} \qquad \textit{Equation 39}$$

Therefore the estimation of the covariance matrix relies only upon one regularisation variable, β, and the scatter matrix calculated from the training samples. Other widely used schemes which may be suitable, by comparison, require the training of either two control parameters in the single class situation (in the case of Regularised Discriminant Analysis[24]) or six (in the case of LOOC1[38]), increasing the complexity and hence processing time of the training process.

It is worth taking a moment to consider the mode of operation of this regularisation. In the case where the scatter matrix is composed of sufficient training samples to adequately estimate the covariance, β should tend to zero, and hence not alter the distribution. In the extreme where the training samples have provided no useful estimation of the covariance, β should tend to infinity, turning the space into a uniform gaussian in all dimensions, rendering any estimated value equal to the mean. In essence, this variable allows the manipulation of the eccentricity of the space, increasing eccentricity (and therefore forcing any estimation closer the mean) when the distribution is less reliably defined.

S_{total} may be calculated independently of the regularisation which, as shown in equation 39, is effectively an additional step after the calculation of the scatter matrix. This allows the use of a simple but effective training technique to empirically determine the optimal value of β by reusing the training samples. Rather than a single scatter matrix, multiple matrices may be calculated by omitting each sample in turn. These samples can then be used with the corresponding scatter matrix to perform an estimation which may be compared with the ground truth (as the sample omitted and then projected is complete, with known values for all data estimated). The probe data for a sample can be passed into equation 36, using the covariance estimation from equation 39, and the result compared with the original non-probe data from the sample to yield a mean square error. Using the mean square error over all samples as a fitness value, an optimisation technique such as steepest gradient descent or even a random search may be used to determine the most effective value for β over the entire training set. Use of this value to estimate the covariance based on the scatter matrix should provide a means for the optimal estimation of data from a probe.

2.5 Conclusion

A summary of prior face structure from single image techniques has been presented. The problem has then been re-stated as a generic statistical estimation task in order to reduce the number of assumptions being made. instead incorporating the relevant properties, such as surface reflectance and the topographical commonalities between faces implicitly into a trained statistical model. A signal to noise based critique of the theory that PCA or other low dimensional parameterisation often used to limit the search space for the iterative searching required by shape-from-shading is a valid process has been given. It has been argued that low dimensional representation cannot be assumed to provide acceptable signal to noise performance in reconstruction, and that estimation of data by conditional densities without limiting the dimensionality of the space is preferable. The issue of sparse sampling has been addressed and the regularised covariance matrix estimator proposed by Robinson[57] has been suggested as a suitable technique for estimating missing data in high dimensional multivariate gaussian data and a simplification of the estimator proposed for single-class systems. It has been suggested that this provides a suitable technique for generating 3d structure from a facial image as a departure from the more usual image irradiance based techniques. It should also be noted that this is, unlike the other techniques discussed, not an iterative process. Once the estimator is calculated, an estimation may be made directly.

3. Implementation of the Estimator

To provide data for the development of a three dimensional face coding scheme, a database of captured three dimensional head models is required. In this chapter the technologies available for the capture of such data are discussed. The requirements the data must meet are presented, and third-party databases are described. The process of capturing further 3d models is also detailed. Errors during capture are highlighted, with explanations for inaccuracy in the data, and finally the process of preparing the data for analysis, including the selection of useful models and the removal of spurious artefacts in the data, is presented.

The process of training a regularised covariance estimator based upon this data is discussed, including techniques for speeding up this extremely time consuming process. The results of the training process are analysed to estimate an optimal regularisation coefficient (β). Finally, the effect of varying training set size is investigated. The purpose of this work is to train the regularised covariance estimator; the fully trained estimator will then be taken forward for analysis in the next chapter.

3.1 Data Capture, Selection and Preparation

3.1.1 Techniques of 3D data capture

3D model capture has been heavily researched over recent years, with advances in cinema and computer game technology pushing forward the commercial development of systems for the capture of human facial models alongside motion capture. The ability to record the structure of a face in 3D grants the use of "digital doubles" used extensively in films such as The Lord of the Rings trilogy, allowing the film maker to realise physically dangerous, impossible or expensive sequences. The technology has also seen the use of "real" sportsmen in games such as the FIFA football series, where a character may be seen from any angle, and the use of a high quality model of a real player adds considerably to the feeling of immersion.

The oldest and most established visual 3d capture technique is the laser scanner. There are many commercial variants of this technology, all applying slightly different algorithms and hardware, but all share a common basic principle. A laser line or point is projected onto an object, and the image of this line is recorded. The deformation of the line, or movement of a moving point, allows the shape upon which the line is projected to be calculated. By scanning the line or point across every surface of the object, sufficient information is contained within the images of the laser reflection to compute the shape of the object. This is the most accurate of all the current techniques, though it has many limitations. Not least among these is the expense of the system, and the huge quantity of raw data involved in even a small scan. More practical problems with this

system include the long capture time – with a single face model requiring the subject to remain stationary for the duration of a scan. Clearly this is both a time consuming and intrusive process and is not particularly suitable for large scale data capture. Another concern is the physical transit of the scanning head over the surface being scanned. While the head is not physically in contact with the subject, for a complete model, the head must pass over every point on the desired surface.

There are many benefits of the use of a laser scanner; whilst slow, the scanning process is capable of producing a full 3d model (rather than a three dimensional projection towards a single camera, as less intrusive techniques described below do) and is relatively resilient to lighting, with fewer errors creeping into the system from specular reflection. Specular reflections are, however, still an issue – so much so that in many industrial applications, make-up is applied to the face of the subject to minimise their occurrence. Laser scanners also perform poorly over material such as human hair – though this is again a major issue for all 3d capturing techniques due to the complex optical scattering properties of hair. For inanimate objects, the accuracy and reliability of laser scanners make them the system of choice – though even inanimate, stationary objects can be surprisingly difficult to capture, as evidenced by the length of time taken by the Digital Michaelangelo project to scan the statue of David (one month)[18].

Despite the limitations for live subjects, laser scanners are widely used in the core games and film markets. The expense, sloth and intrusiveness of the technique are of negligible importance, as the time required of the subject is minimal compared to any other technique for the generation of the end result. There are also relatively few subjects to be

scanned in any given project, and all subjects are willing participants in the process. The high accuracy and resolution produced by laser scanning often outweighs the disadvantages of the technique.

It should be noted that other non-visual 3d scanning techniques are commercially available, mainly developed for medical use, such as computed tomography (CT) and magnetic resonance imaging (MRI). These are essentially designed for imaging the internals of a soft tissue body, and cannot capture surface texture. They are also hugely expensive and (in the case of CT) extremely dangerous to the health of the subject, and can be considered to lie outside the scope of this discussion.

There are two constants of all laser scanning techniques:

- Projected structured image (the laser)
- Multiple views of the subject (the physical scanning process)

Clearly, these requirements need not necessarily be met by a laser and a scanning head, and indeed several less intrusive techniques have been developed. There is now a sliding scale of techniques, using anything from one single camera and patterned light, through multiple cameras and changing projected patterns, to video cameras with no projected lights. Despite the interest of institutions such as BBC Research[55] in the capture of 3d data from standard video, or video shot from multiple views of the same subject, the resulting models are of relatively low resolution and accuracy. Furthermore, and most significantly, no system is currently commercially available.

The only technology currently widely used as a competitor to the 3d

laser scanner involves the use of patterned light (either with or without narrow-baseline stereoscopic cameras). Several different types of capture device are available[1], and more are in development[21]. They operate upon the general principle of projecting a known pattern, and using the observed deformation of the pattern in the resulting image to calculate the geometry of the surface upon which the pattern has been projected. The advantage of this technique is that the capture device is essentially a complex camera, and is used as such. The subject is not required to be still for long periods while their face is scanned as the mode of capture is identical to that of a standard flash camera. However, this technique is often susceptible to uncontrolled ambient lighting conditions and specularities on on the target. Whilst in time these issues may be resolved, they are pertinent to the question of data collection for this work. This technique also tends to produce lower resolution models than 3d scanning.

3.1.2 The York data set

A set composed of multiple poses of some 300 subjects was collected at the University of York in collaboration with the Computer Science department using a commercially available capture device operating with a combination of stereopsis and patterned light projection, though details of operation are not publicly available. Subjects were captured in a variety of locations, some with and some without controlled lighting. As well as multiple poses, multiple images of each pose were captured. In order to standardise pose, marks were positioned on the floor radially from the subject pointing directly at the camera, and at angles of +/- 45 and 90 degrees. The subject was seated on a revolving, height adjustable chair, adjusted such that eyelines were approximated across all

subjects. Subjects were asked to rotate the chair in order to look directly forward along each of the lines in turn, and also to look at points above and below the camera, again marked consistently, whilst facing forward. Smiling, angry and partially obscured images were also taken. The capture device was capable of capturing greyscale images only, though a secondary standard colour camera was synchronised to take a colour image in many cases. This camera was located as close to the 3d device as possible, though not close enough to use the image as a texture map for the model without some registration process, as can be seen in figure 3.1

Figure 3.1: Colour image(left) and greyscale texture image (right) from a 3d capture instance

3.1.3 Third party datasets

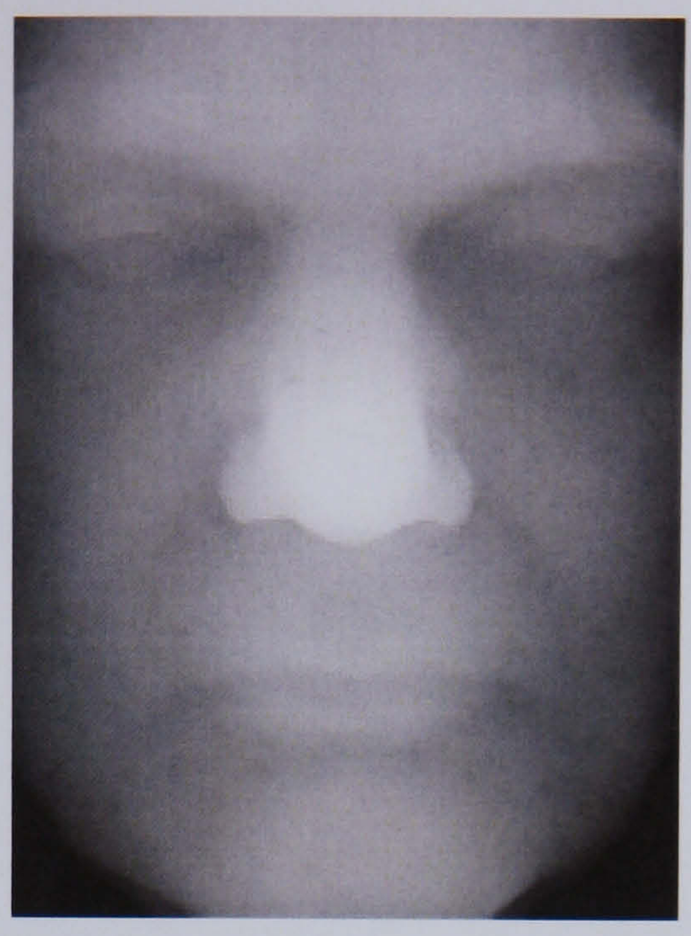


Figure 3.2: Example of a Notre-Dame depth map

A large quantity of conventional 2d images of human faces are readily available to the research community, the most famous single database being the FERET database[22,15], though a simple image search using a search engine such as google or altavista will provide further, unsorted images. This is in no small part due to the maturity of the photographic medium; photographing faces is far from remarkable. However, while techniques of 3d capture are slowly reaching maturity, no system is cheap, and none have become widely used. Sourcing third party 3d data is much more challenging, therefore, than sourcing conventional images. The most widely used – and most extensive – database generally available to the research community has been captured by the University of Notre Dame, USA, and has recently been used as the data set for the Face Recognition Grand Challenge[20] competition. As can be seen in 3.2, a depth-map representation of the 3d data, the level of detail present in these models is impressive.

3.1.4 Requirements for Three Dimensional Data

For the purposes of this work, full 3d models of the entire head are not required as the facial region is the primary area of interest. This has the benefit that excepting facial hair, the problems posed by hair to the data capture process can be ignored. Facial features must not, however, be obscured by long hair. To simplify the problem further, facial ornamentation such as spectacles should not be present in the training or testing sets. A representative spread of age, gender and race is preferable, though with the limited data available a truly representative training set is not realistically obtainable. The largest data set possible is required to best define the distribution, but all training data must meet certain minimum quality levels as well as the capture requirements described above. As noted, all 3D capture techniques degrade over regions with very high specularity or complex optical properties, leading to areas of unknown or clearly erroneously captured depth. The training data set should not contain any models with such areas, though the largest possible quantity of good data is required to minimise in order to provide the highest level of statistical significance possible.

To remove as much spurious variation in the data as possible, models should be positioned and oriented identically with respect to each other. In practise, this is a non-trivial task due to automate due to the natural variation of the human face. Both York and Notre-Dame sets ostensibly position subjects identically at capture. In reality, this is a very coarse approximation. Variation in height between subjects and the specific pose each adopts mean that whilst the strictures placed upon the subjects at capture ensure that no data required for a frontal render of the face is lost

due to occlusion they are far from ideally posed and positioned. There are several possible landmarks for use to overcome this problem; a common technique used with two dimensional data is to resize, rotate and displace the faces such that the centres of both eyes coincide for all images. As the three dimensional data is present in this case, nose-tip is a easily detected landmark for positioning of the faces.

As a frontal mask and not a full head is used, the data may be represented as a depth map, using a standard greyscale image format with each pixel representing distance rather than colour information at that point. However, unlike many truly three dimensional representation formats, a depth map is inherently quantised and bound limited. There can be no more depth quantisation levels than there are colour levels available to the image format, and the range 0 to saturation define the limits of the model in the depth axis; data beyond either 0 or saturation will be clipped to these levels. While some formats (such as the portable network graphics format) allow more quantisation levels, a very common representation across many formats is the 256 quantisation levels found in a basic Windows bitmap file. This is the standard format for representation of image data on computers and is also a convenient data size for computerised processing, being representable in a single byte.

Using this level of quantisation is desirable from the point of view of complexity; however it adds an extra task to the preparation of experimental data, as the choice of front and back clipping planes is extremely important. As all of the depth information on the face is, in the absence of evidence to the contrary, of equal importance regardless the depth, a linear quantisation scheme should be employed. As shown in

figure 3.3, if the front and back clipping planes are set too loosely then the many of the quantisation levels will be unused, with the useful data spanned by fewer levels and hence more coarsely represented. At the opposite extreme, badly centred or too narrow quantisation will result in useful data being lost in beyond the bounds of the valid quantisation levels. The central image of this figure shows optimally positioned clipping planes, with no data lost but no unused quantisation levels.

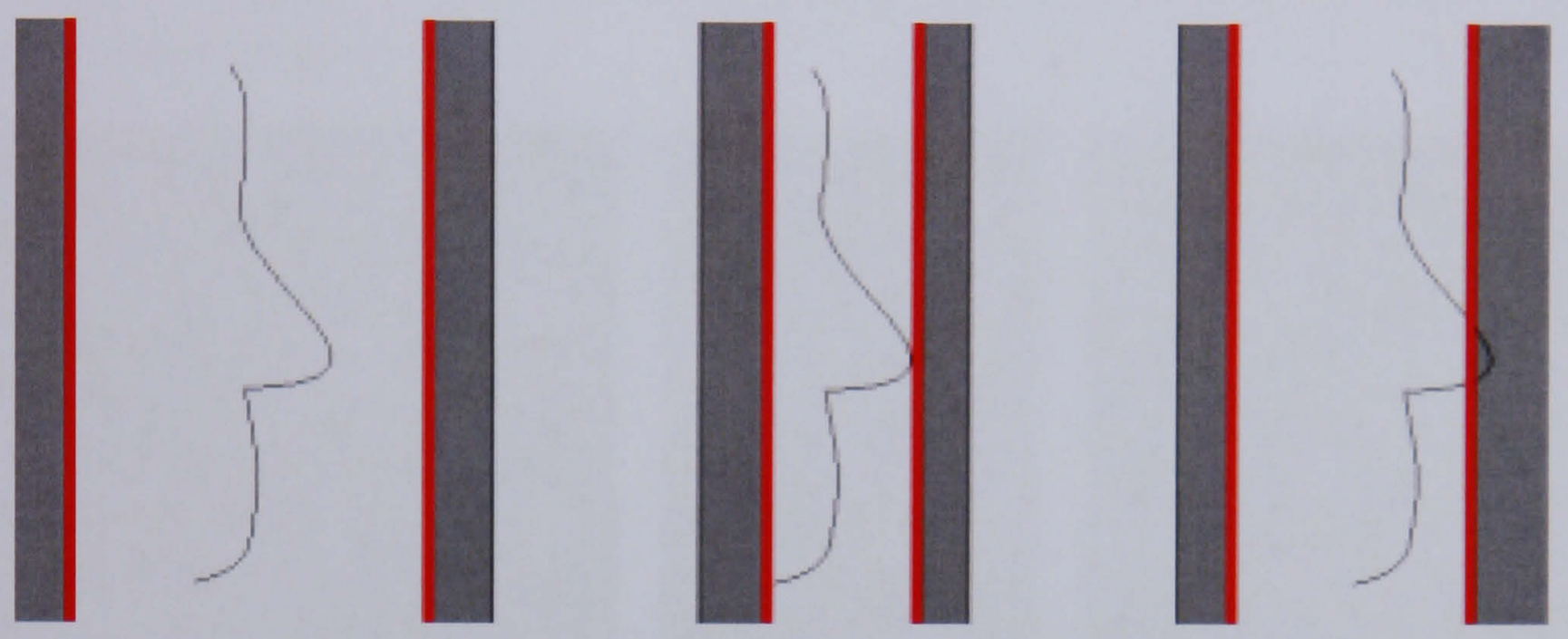


Figure 3.3: Positioning of clipping planes is critical to ensure well quantised, unclipped data

This figure slightly trivialises the task, as it assumes that there is no unwanted data (from specularity, hair, background or other non-face data) present in the model. If such data is present, it should be ignored when setting the clipping planes. Assuming the nose can be detected, it again provides a useful front clipping landmark as no facial data should be closer to the camera than the nose for a frontal image.

3.1.5 Selection of the candidate data for training

Noise present in the three dimensional data is mainly due to surface reflection effects in the physical face, primarily from specular reflection and extreme diffusion. This noise takes the form of sharp spikes, discontinuous curves and missing areas of mesh, as seen in figures 3.4, 3.5 and 3.6. These effects are concentrated particularly about the eyes and hair of the subject, as these are the most reflective and diffusing areas of the face respectively.

Figure 3.4: A point of error caused by specular reflection in the eye

Figure 3.5: Gaps in the mesh

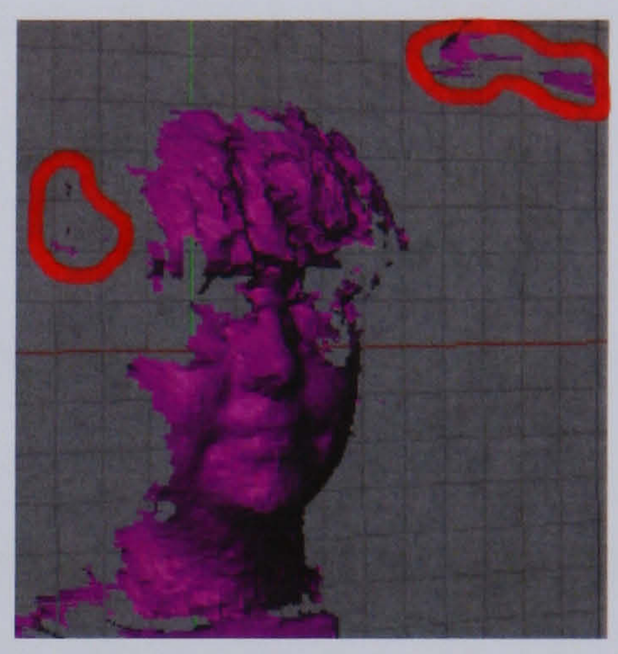


Figure 3.6: Disconnected sections of model "free floating" from the face

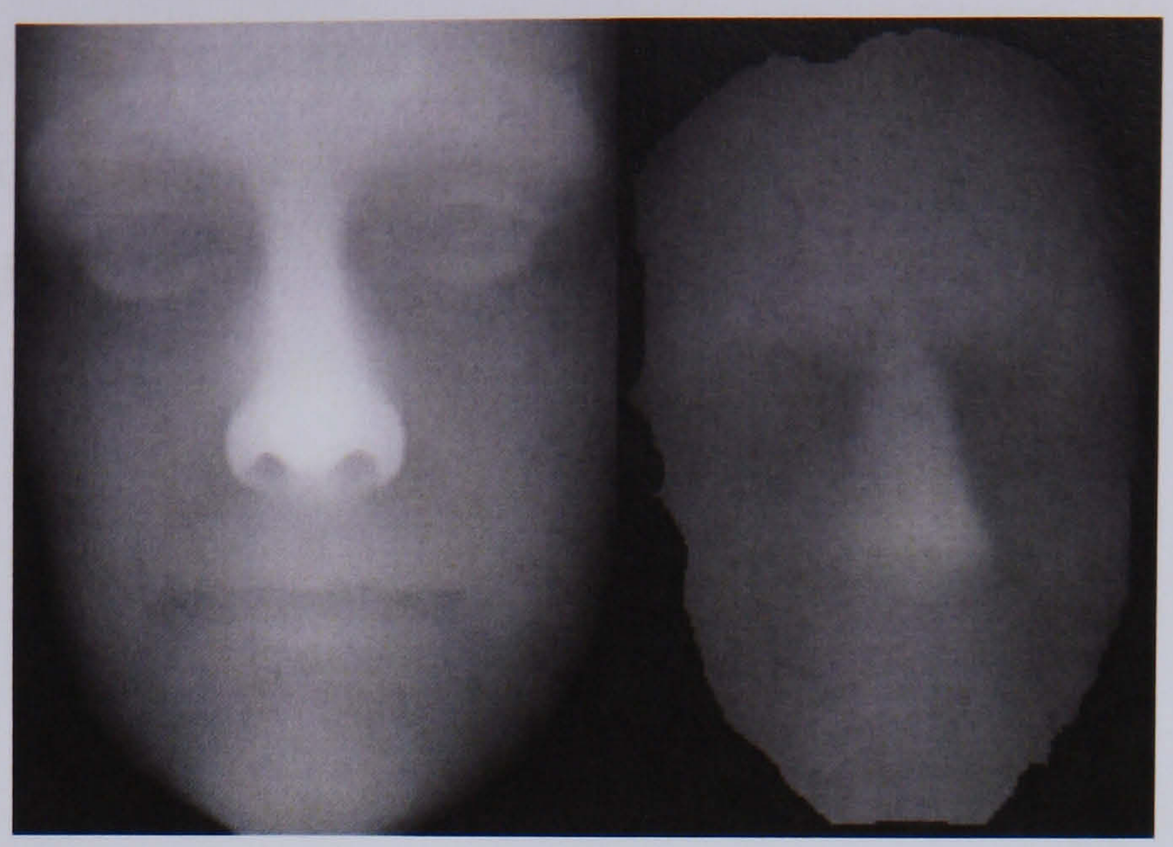


Figure 3.7: Notre Dame depth maps (left) captured with a laser scanner are of a much higher quality than the York depth maps (right) captured with a patterned light projector with shape-from-stereo

Comparison of the two depth images in figure 3.7 demonstrates that the projected light set captured at the University of York (the "York set") is of inferior quality compared to the 3d scanned set provided by the University of Notre Dame (the "Notre Dame set"). Whilst both suffer from noise and missing data, the Notre Dame set has more complete models. More significantly, the York set models are clearly of substantially lower resolution, and do not contain much of the structural detail present in the Notre Dame set. For this reason it was decided that it would be counter-productive to include the York set into the training data, but that using a smaller test set comprising only the Notre Dame set (after the removal of poor quality data) would be a better approach, as the training data used is an attempt to provide a ground truth for the system to be trained upon. It should be noted, however, that the Notre-Dame data is itself far from perfect. Whilst the exact capture technique and procedure for the Notre Dame data has not been published, there appear to be several types of noise present in the data. Inspection of the bitmaps used to texture the models

reveal that noise is not limited to the three dimensional data, but rather that there are two independent and significant sources of noise in the photographic data.

Figure 3.8: The colour channels of a Notre-Dame texture image

Figure 3.8 shows an extreme example of the most problematic of these noise sources. Here, one of the full colour texture images has been divided into the constituent red, green and blue colour channels. Two issues are immediately evident; firstly the unusual colouration of the original image is due to extremely poor colour balancing at capture. The green channel is well balanced, with saturation only on areas of white clothing and small patches of high specularity. The blue channel is less well balanced, though this appears to be primarily a case of poor contrast; saturation is still not a major problem, but the entire image appears slightly washed out, with dark areas not appearing particularly dark. The red channel, by comparison, is clearly over-saturated, with large areas of the face whiting out. The result of this is that the composite colour image has a

red-purple bias, looking quite unnatural. A further problem with this data is evident when comparing the blue channel to the red and green channels: there is a significant temporal shift between blue and the other two channels. In this and several other instances the result is that the eyes are open in two channels and closed in the third, leading to coloured ghosting in the composite images (seen as a bizarre blue glow in the eyes).

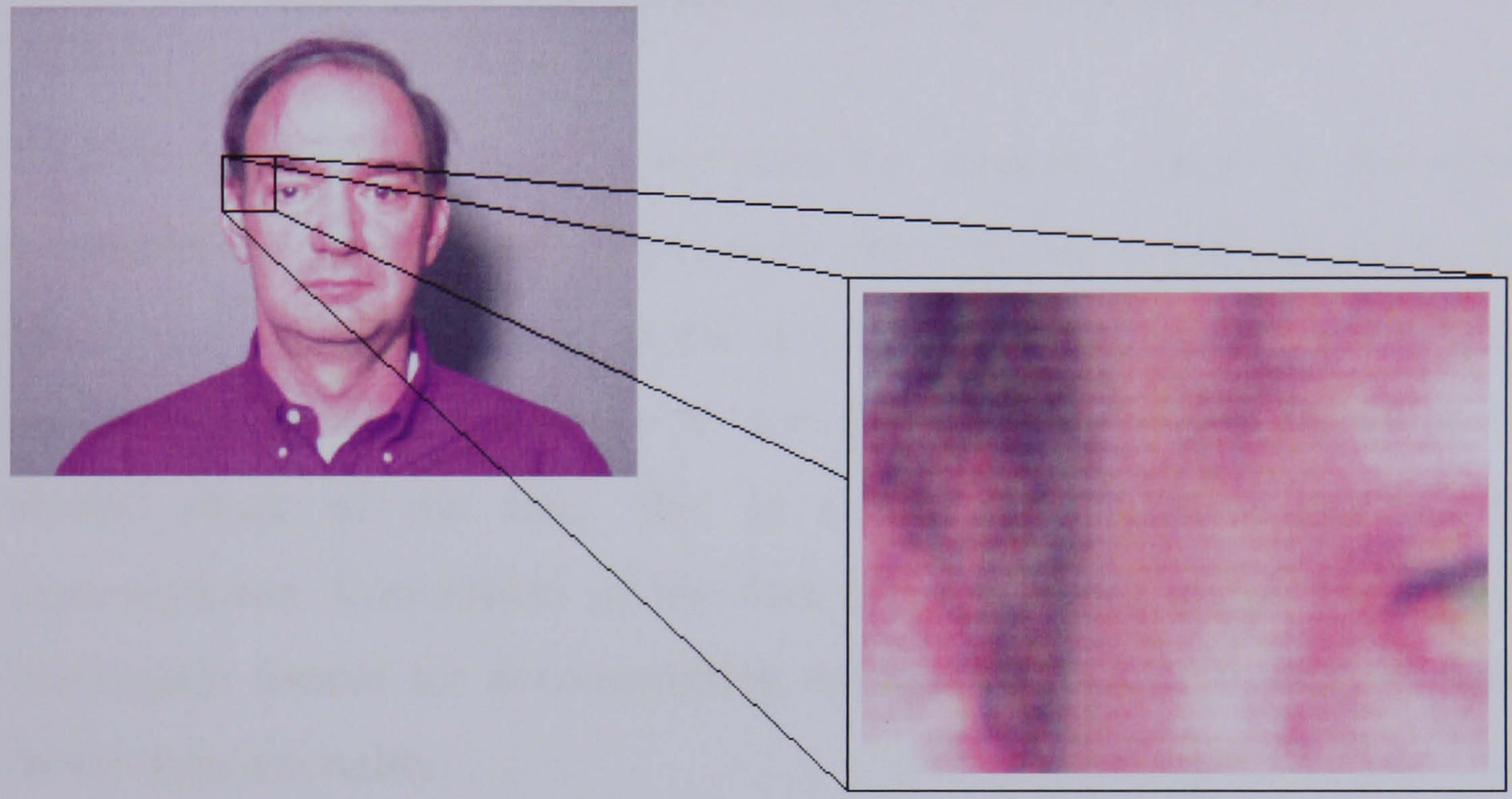


Figure 3.9: Close up of a Notre-Dame texture image reveals interlace lines, likely to be due to the use of a video rather than a stills camera. An image from a progressive scan video camera or stills camera would not exhibit these artefacts.

Figure 3.9 shows a less immediately obvious interlacing problem. Ideally, images should be captured using a stills camera to provide the highest quality data. This image appears to have been captured using some form of video camera. Very close inspection of the image reveals interlacing lines. These occur where two consecutive frames are interlaced together to produce the appearance of a higher resolution image (the camera scans odd lines on odd frames and even lines on even frames, for

example, thus doubling the effective frame rate of the camera). The implications are twofold; firstly the visible lines are of themselves spurious, and secondly rapid motion (such as blinking) can cause the two line sets to show markedly different images, introducing interlace ghosting in the composite image.

3.1.6 Noise removal, posing and rendering procedure

To provide the least noisy data for analysis, much of the noise present in this data set can – and should – be obviated by pre-processing. In addition, it should be noted that the raw 3d data takes the form of a mesh, analysis of which is complex. As the data only accurately represents a frontal mask of the face, this is a less than optimal method of representation. Conversion of the data set to a depth image provides a convenient format for downsampling the data to provide a data set with lower dimensionality.

Downsampling the data is desirable as the data available is not large. The dimensionality of the data can be considered, after conversion of the model to a depth image format, to be:

$$D = R_t \times C_t \times n_t + R_d \times C_d$$

Equation 40

where D is the dimensionality of the sample. R and C are the row and column dimensions of the images, with subscript t denoting texture and d denoting depth-map images. The texture image also may be composed of multiple colour channels, the number of which is defined as n_t. Limiting the texture information to one channel (greyscale) has the advantage of reducing the dimensionality substantially and also removing the poor

colour balance (though saturation is an issue). It is possible at this stage to remove colour channels, and build the greyscale image entirely from one of the channels, thus removing the temporal distortion apparent across channels. However, it is not known which channel most accurately maps to the 3d model and so removing channels may reduce the validity of the data. For this reason, the greyscale images are generated as the mean of the three colour channels; this is the usual technique of generating a greyscale image from a colour image and in the absence of any information about which of the channels are accurate, it is the most sensible approach. The noise in the erroneous image will be averaged with the good data; however, since the errors are evident in only limited regions of the images, and only in a minority of images.

The Notre-Dame data is composed of 2d photographs for texture and custom 3d point cloud coordinates for the 3d data, stored as text in a ".abs" file. For ease of processing, this must be converted to a depth image and noise removed. Initially, the z-coordinates of each point are read into a two dimensional array equal to the x, y extents of the point cloud. This is simplified by the data format; the Notre-Dame data uses course x, y quantisation to arrange the data into such an array in the .abs file. As the depth data will be further downsampled before processing, this quantisation is negligible. The z-data is, at this stage, stored as double data rather than quantising and clipping to the output 0-255 range. Also at this point a "valid data map" is constructed. Missing data in the .abs file is represented by extreme negative values. These can be detected and a boolean mask array created to indicate locations containing valid data.

The nose should be the point on the model closest to the camera. The depth array is scanned, and the highest point found. If this is a valid candidate, there should be approximately the same amount of data before reaching invalid data to both the left and right of the peak, and should also not appear too close to the edge of the data. An empirically determined 150 element buffer around the image is used to ensure that the peak occurs centrally enough for a face to fit into the image should a nose be detected at a given point; peaks found within this buffer are discarded before checking for centrality. If difference between valid elements to the left and right of a peak is greater than half the sum of the full distance horizontally across the model along the line upon which the peak is found (ie, the sum of elements to the left and elements to the right), then the peak is discarded. Discarded peaks are set to "invalid" on the valid data map, and all data in the immediate x, y area around the peak which is within 40 depth units (again, an empirically determined value) of the detected peak is set to -999999.00 – this being the value taken by undefined points in the .abs file. This process is repeated until a valid peak is found.

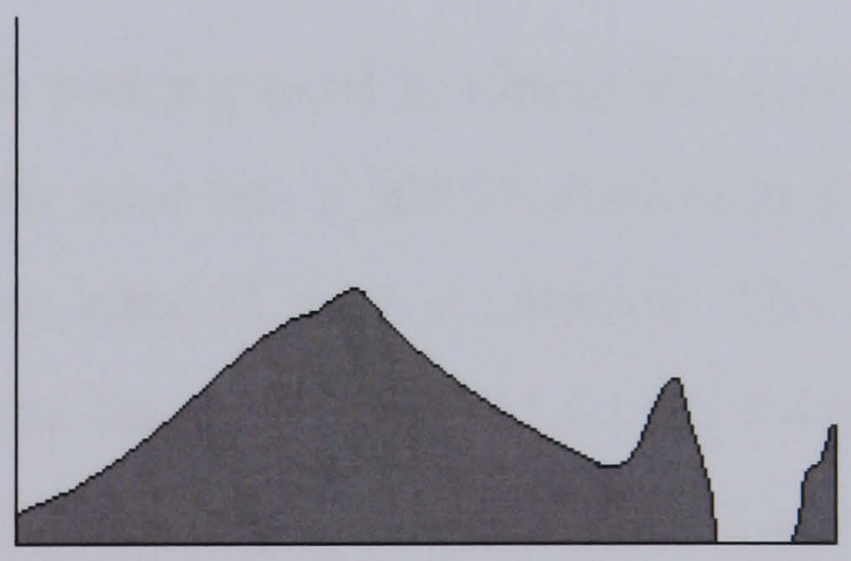


Figure 3.10: Sketch of an image histogram with an isolated peak. This peak is highly unlikely to contain useful face data, and can be quantised out, removing the wasted quantisation levels falling in the gap between it and the body of the data in the image

Other sharp positive discontinuities are detected by checking for large gaps in the histogram of the z data. If, after quantising to the output image range, the data exhibits isolated high values (such as those shown in figure 3.10) then there is a large step change in the depth information. As the human face is a smoothly contoured surface, this should never occur in valid data. Such peaks in the data will be a product of glare. The gap is defined in this process as a series of ten pixel values of which less than five have any pixels present in the image. Pixel values isolated by such gaps are considered glare peaks. Pixels taking these values in the image are set to -999999.00, and set to invalid on the valid data map. This process is repeated until no discontinuous peaks are found, with the remaining valid data re-quantised at each iteration from the original source data.

With the nose peak detected, the front clipping plane can be set to this z value. The back plane is set to 75 units behind this point, a value chosen after manual inspection of the data. All points behind this rear clipping plane are removed.

To prevent the peak picking used to detect the nose from choosing the edge of a plateau as the nose tip, a 50*50 window is centred upon the high point, and the plateau area (if any) is mapped. The centroid of this area is set as the actual nose tip. A 200 by 150 element window is centred upon this, and the data within the window quantised using the same scheme as the glare peak removal technique employs. The depth information is now in a format suitable for saving as an image file. Three passes of a three by three by three median filter are used to remove shot noise from the model, with "invalid" points set to valid where the filter causes them to be filled with data within range. The depth data is then saved as a greyscale

image.

The .abs format uses a pixel by pixel mapping from texture image to rendered model, such that each pixel in the texture image corresponds to one element in the .abs array. This is a direct one to one mapping, with no geometric distortion, and hence a 200 by 150 element window can also applied to the texture image, centred upon the same location as the nose point in the depth data, to maintain the direct relationship for texturing the depth data. The resulting texture image is converted to greyscale and deinterlaced by replacing every pixel on each odd line with the mean of the pixel values directly above and below. All texture and depth images are then resampled down to 50 by 38 pixels in order to reduce the dimensionality of the system. This dimensionality was arrived at first by resizing a sample set of facial images to three quarters, half, one quarter and one eighth the width and height, and selecting the smallest image size which still appeared based upon ad hoc experimentation to provide an image that was recognisably a face which does not give the impression of being uncomfortably small, or difficult to distinguish identity in. Given the extent of this resampling, the effects of interlace removal are completely removed; a simple half size resample would remove the artefacts and result in an image indistinguishable from one based upon an image without interlace lines. The resulting data set is of 3800 dimensions. The ghosting caused by the temporal shift in eye regions is limited, and visual inspection reveals that it is no longer readily apparent in the data. Averaging the colour channels is the standard technique to convert a colour image to a greyscale image, and of itself has no implication for the accuracy of the data.

Finally, a manual check is performed on each texture / depth pair to ensure that the data is of satisfactory quality, and images which still contain overt artefacts or missing data or which are badly posed are removed. Note that an error was made at this stage, and two badly framed images were inadvertently left in the set. The implications of this are discussed in section 3.2.1.

3.1.7 Final data set

Figure 3.11: Example of a prepared image-pair (actual pixel resolution)

This process results in a set containing 740 pairs of image, though it should be stressed that there are multiple images of many subjects in this set, and so there are images of only 166 individuals present. Of this set, 40 images were removed to create a test set leaving 700 images for training of the depth estimation algorithm. These forty were chosen such that the 35 individuals present in the test set were completely absent from the training set in order to allow fair testing of the performance of the system with completely novel data.

Even with the data reduction techniques employed, limiting the colour information to single channel greyscale, closely cropping the images to the face, representing the 3d information as depth rather than true 3d and resizing the images to reduce the number of pixels present, the dimensionality of the data is very large compared to the number of samples available. A further technique may be employed to increase the size of the sample set by exploiting a property of the human face. The human face

does not exhibit perfect bilateral symmetry and therefore a mirrored image of a face represents a subtly different face to the original image; the two faces could not belong to the same individual, but do represent valid points in face space. By mirroring each of the training images, then, the sample set size may be doubled to 1400 samples. This is still significantly smaller than the dimensionality of the space, but is at least of the same order of magnitude.

It is important to note that the training set is still severely limited by the number of individuals represented. Whilst there are 1400 training samples, these cannot be considered to be randomly distributed within the population. Rather, there are clusters of samples related to the 131 individuals. Some of the shortcomings of the set are apparent upon a cursory inspection; there are only three Afro-Caribbean subjects, none of whom appear to be particularly old. There are, in fact, few subjects that appear to be over the age of 40, and there are no children whatsoever. The test set contains a preponderance of Caucasian individuals, presumably simply due to availability to the collectors of the data set. Regions of the space generated by analysing this data set are, therefore, likely to be poorly fitted to the population where gaps in the training data occur. Effectively, the system is trained on adult Caucasian faces, and so may be expected to perform more poorly on faces of children, older adults and those of non-Caucasian origin since these will have had less weight in the generation of the regularised covariance estimator. The significance of this bias is not known; certainly all faces share roughly the same topology so it is to be expected that in general the system will still produce a non-garbage output. Experimentation with age, racial and gender specific training sets, should such data become available in sufficient quantity to be useful in such an

analysis, would be valuable. There is the potential that reverting to a multiple-class based estimator and categorising faces before training may produce improved performance. However, the fundamental similarities of faces are greater than their differences, and it is expected for this reason that whilst poor representation of a face type may produce poorer results, these results will still be a better fit than the average face.

A more significant limitation of the training set is that the lighting conditions under which all the images were collected appear to have been very similar; the system is therefore trained implicitly to operate on a lighting conditions which match the Notre Dame images. Since the control set is gathered from the same data, this could be considered to be a flaw in the test data which will provide better results than for the situation where lighting does not match. However, since the quality of the York models are so low, there are no suitable image/model pairs available to allow a quantitative evaluation of the technique other than the Notre Dame data.

3.2 Empirical estimation of the optimal regularisation coefficient

The final training set, including mirrored images, provide raw data for estimation of the regularisation coefficient independently of the test set. As discussed in chapter 2, a leave-one-out method may be used to perform this estimation. Figure 3.13 shows a flow diagram of this process. Each sample in turn is omitted from the training data, and the remaining image pairs are used to construct scatter matrix $\boldsymbol{s}$. A trial regularisation value, β_{trial} is applied to the scatter matrix as described in chapter 2. The texture

image of the omitted sample is used as a probe with this regularised matrix, and the "unknown" depth image estimated. The estimated image can be compared with the true depth data for the probe, and a mean square error value calculated, as shown in figure 3.12. β_{trial} is not varied between samples during this process. Once all samples have been left out once, and a mean square error calculated for each, an average mean square error over all images may be calculated for the current value of β_{trial} which may be considered to be its fitness value (where 0 is maximally fit, with fitness decreasing as average MSE increases). β_{trial} may be varied according to some search scheme, and the process repeated until the minimum sum mean square error value is found. Whilst at no point is the full scatter matrix generated by the training set evaluated for fitness (due to the omission of one sample at all stages of training), because the number of samples remaining in the training set is much greater than the number omitted, any given $\boldsymbol{S}_{trial}$ should tend to $\boldsymbol{S}_{total}$ (where $\boldsymbol{S}_{trial}$ represents any one-sample-omitted trial scatter matrix and $\boldsymbol{S}_{total}$ the scatter matrix generated using all training samples), hence the fitness of a given regularisation value over all $\boldsymbol{S}_{trial}$ matrices should correspond closely to the fitness of $\boldsymbol{S}_{total}$. At the point where average MSE is minimised, $\beta_{trial}=\beta_{opt}$ for the training set.

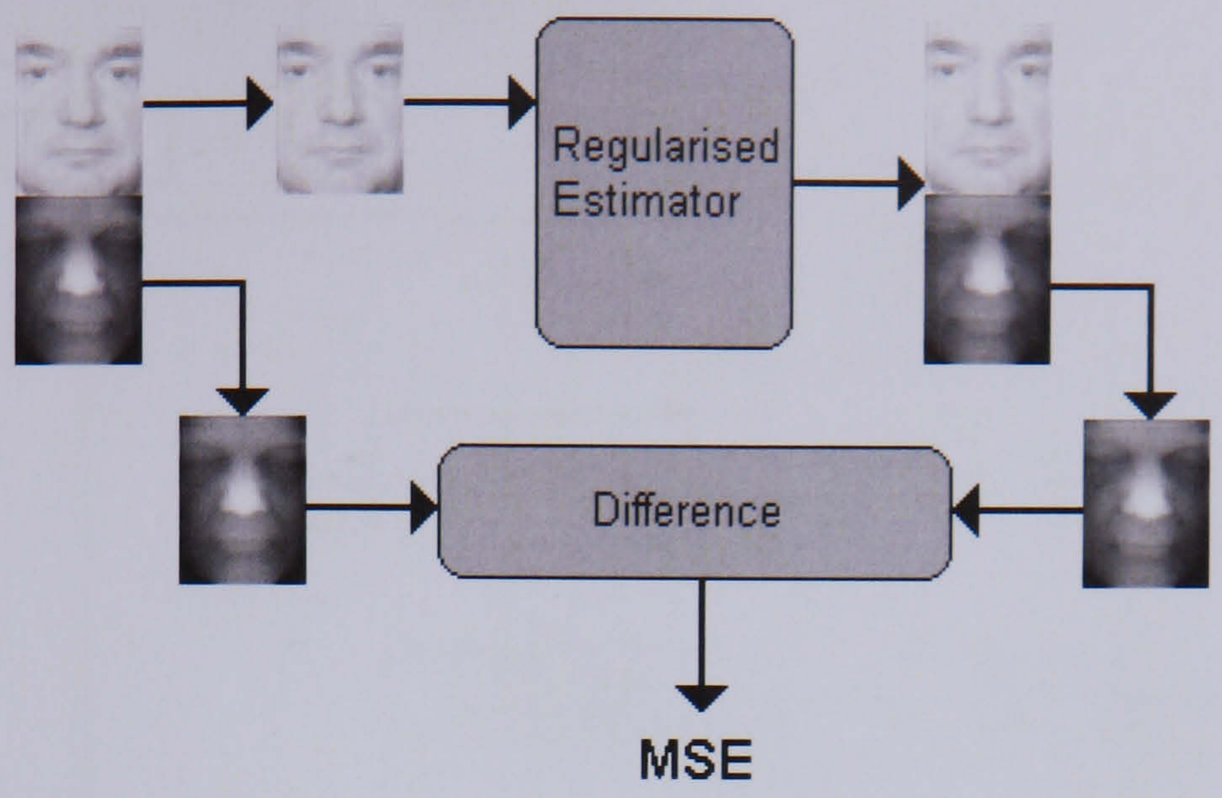


Figure 3.12:Calculating the error for training

The search technique applied is simply to equally space twenty sample points between $\beta_{trial}=1$ and $\beta_{trial}=17500$, with the upper limit chosen based upon the results of several informal preliminary tests used to establish an upper bound which apparently lay well beyond the optimal value. A curve may be plotted through the calculated average MSE for each value of β_{trial} . As $\beta_{trial}=\beta_{optimum}$ where average MSE is at a minimum, provided the curve exhibits a definite minimum the optimal regularisation coefficient may be read from the graph, assuming a simple and relatively noise free relationship exists between the two variables.

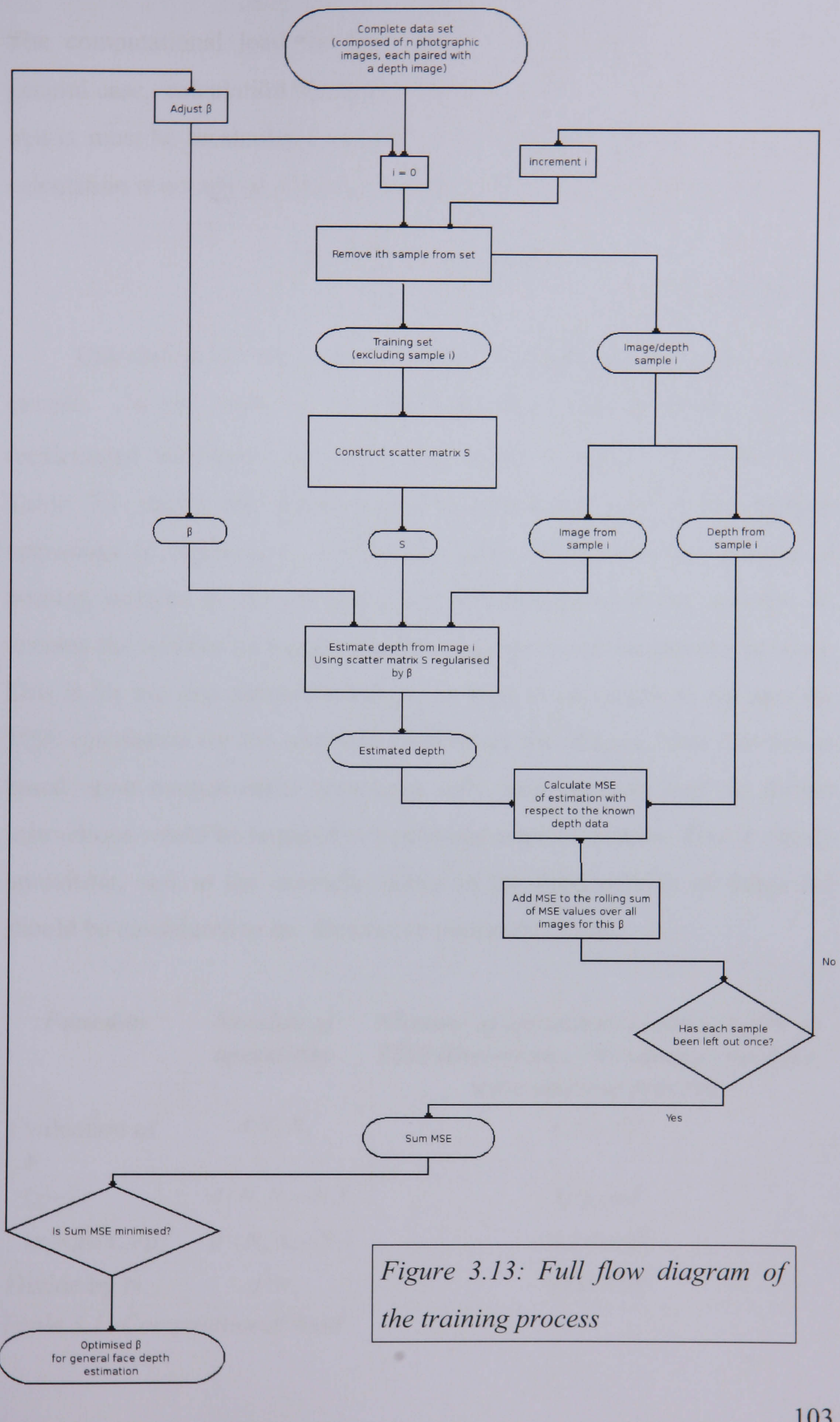


Figure 3.13: Full flow diagram of the training process

The computational load for this process, however, is extreme. In the general case, calculation the fitness for each value of β_{trial} the full scatter matrix must be recalculated once for each sample in the training set. This calculation is not trivial. Consider equation 41, as defined in Chapter 2.

$$\boldsymbol{S}=\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T$$

Equation 41

Calculation of the mean $\hat{\mu}$ must be repeated for each omitted sample. $(x_i-\hat{\mu})$ must be calculated for each sample in the set and recalculated whenever $\hat{\mu}$ varies, and hence so must $(x_i-\hat{\mu})(x_i-\hat{\mu})^T$. Table 3.1 shows the approximate computational cost of the various operations in equation 1. Within this table, N_s denotes the number of training samples in the set and d the dimensionality of the samples. N_t denotes the number of images used for the leave-one-out training process. This is for the data described half N_s as there is no reason to perform the MSE calculation for the mirrored versions of the images. Note that this is based upon mathematical operations only, and assumes that no further instructions would be required in a practical implementation. This is clearly unrealistic, and so the example values in the third column of Table 3.1 should be considered to be illustrative minimum values.

Function	***Number of operations***	***Number of operations (1400 samples of 3800 dimensions, 700 samples used for leave-one-out process)***
Evaluation of $\hat{\mu}$	dN_sN_t	3.72×10^9
$(x-\hat{\mu})$	$d(N_sN_t-N_t)$	3.72×10^9
$(x_i-\hat{\mu})(x_i-\hat{\mu})^T$	$d^2(N_sN_t-N_t)$	14.14×10^{12}
Divide by N	d^2N_t	10.11×10^9

Table 3.1: Computational load

Note that the transpose operation is ignored as this operation manifests as an output formatting (into a d square matrix) rather than any actual calculations; the internal values of the transposed and untransposed vectors are identical. It is clear from this table that the significant operation is $(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T$. In addition to the stated cost, all of these operation sets must be repeated for each value taken by β_{trial}, multiplying the number of operations in this case by a further 20, raising the total computational cost to a minimum of 0.28×10^{15} operations. On a 2Ghz processor, this would require in excess of 39 hours processing time as a minimum. Whilst this is not trivial, this is a minor consideration compared to the computational load involved in preparing the scatter matrix for use as an estimator.

Once the scatter matrix has been calculated, the regularisation process is computationally cheap as it operates only upon the d elements of $\boldsymbol{S}$ in the leading diagonal. In order to use the regularised matrix for estimation, however, a section of the matrix must be inverted, as shown in Equation 42.

$$E[\boldsymbol{X}_1]=\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{P}_2-\boldsymbol{\mu}_2) \qquad \textit{Equation 42}$$

Inversion is an extremely intensive process, with a computational load far outstripping the covariance matrix generation. Testing has shown that in total, from reading images to generating a MSE value, one image can be analysed at one regularisation value in approximately 7.3 minutes by a 2Ghz machine. The training run requires each image of 700 to be analysed 20 times, and hence a total training time on a single such platform is of the order of ten weeks. Clearly, use of a faster machine alleviates some of this problem though processing time is still extremely long.

In order to further decrease processing time, the calculation may be parallelised across several machines, exploiting the fact the calculation of the MSE for an individual sample is independent to the calculation of MSE for all other samples. Several machines can, therefore, complete an analysis of different omitted samples simultaneously with each machine calculating the sample-exclusive scatter matrix, as shown in figure 3.14. It is more efficient, therefore, to perform all of the estimations for a given sample for each value taken by β_{trial} . The resulting MSE scores are returned to the server and stored, and a new sample specified for the client to process if required.

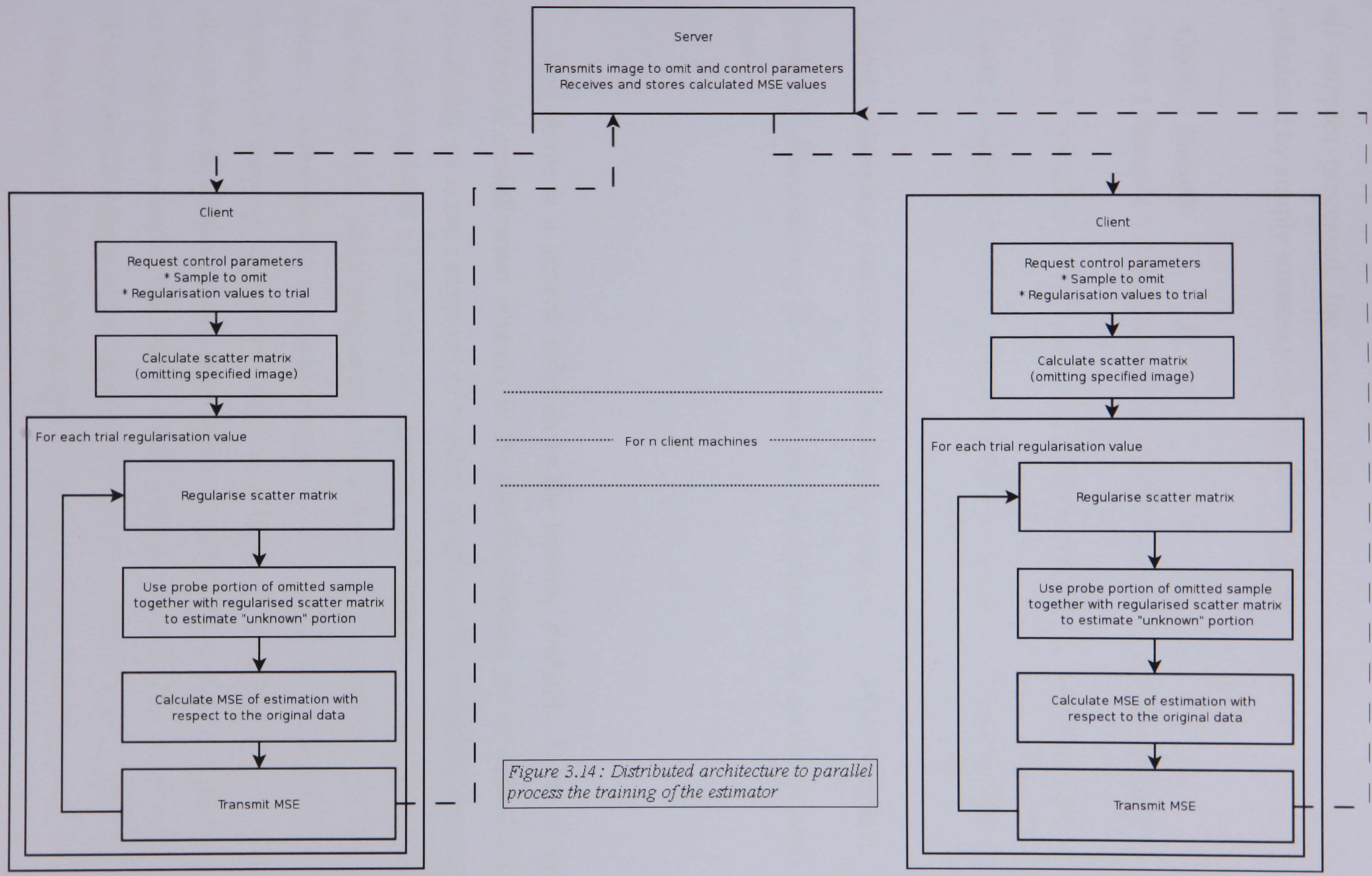


Figure 3.14 : Distributed architecture to parallel process the training of the estimator

Data received by the server can be tabulated as shown in 3.2. With all samples processed, the average MSE for each value of β may be calculated by simply summing columns of the table.

Client	***Sample***	β_1	β_2	...	β_n
Client 1	Sample 1	MSE(Sample 1, β_1)	MSE(Sample 1, β_1)	...	MSE(Sample 1, β_n)
Client 3	Sample 2	MSE(Sample 2, β_1)	MSE(Sample 2, β_2)	...	MSE(Sample 2, β_n)
Client 2	Sample 3	MSE(Sample 3, β_1)	MSE(Sample 3, β_3)	...	MSE(Sample 3, β_n)
...	...	...	...	...	...
Client *i*	Sample *n*	MSE(Sample *n*, β_1)	MSE(Sample *n*, β_2)	...	MSE(Sample *n*, β_n)

Table 3.2: Illustration of the storage and tabulation on the server as items arrive.

The above is a general solution to the training problem. There is an additional complication inherent in the specific training set described. As noted, this set is not composed of random samples of the population but of a limited number of clusters of samples, each sample within a cluster sharing the same underlying subject. 3.15 shows a large example of such a cluster, including the mirrored images used to artificially increase the number of training samples available. It is apparent from observing such a cluster that the different samples within a cluster could easily be considered to be the same sample with added noise, and it is arguable that the presence of such similar images in the training set when a sample from such a cluster is being used as the omitted sample in the leave-one-out searching process

creates an artificial situation which could skew the result of the process; a novel face presented not as a training image will not have the benefit of having closely related images used to train the regularisation images. In order to prevent the possibility of this incorrect training, rather than omitting only the sample to be estimated, all related samples should be removed from the training set.

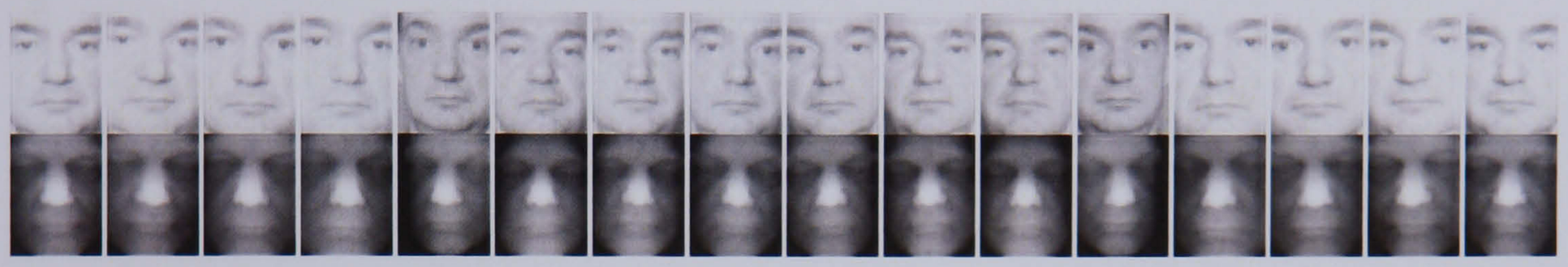


Figure 3.15: All of the instances of an individual in the training set. Not all individuals have so many images in the set.

Where reference is made to "Omitting an image", therefore, it is the case that all images of the individual have been omitted, rather than simply the one instance which is used as the source for the estimation and MSE calculation. Note also that due to the variable brightness and contrast visible in the above example images. Histogram equalisation is applied to each texture image before use either as a training image or a probe image in order to normalise the relative brightness of the image. This does not destroy any information present in the image, rather it ensures that the pixel values are spread as evenly as possible over the full range available (without differentiating pixels of the same value), thus an over exposed image and an under exposed image will be corrected to fill the full range, removing the lightness or darkness of the individual image as a whole from the analysis, and allowing variation across the image to be the dominant signal.

3.2.1 Result of training

Figure 3.16: Regularisation coefficient against mean square error; a comparison of the prematurely terminated run with the completed run

An error in the training process caused the premature termination of training. Whilst all images were used correctly to generate each scatter matrix (save the ones deliberately omitted each iteration), the leave-one-out process was completed on only the first 468 images in the set. Images were omitted in order of subject then image-of-subject, so many subjects were therefore left out of the reconstruction training completely, being present only as points in the scatter matrix for the other images. Unfortunately, this error was only detected after subsequent experiments were carried out, and resources to re-run these later experiments were not available. However, the missing training cases have been analysed and the discrepancy between the value of β used and the optimal β is small. Figure 3.16 shows the result of both the incomplete and complete training processes. As can be seen, the incomplete training result suggests an optimal value of approximately 4500, whereas the complete training result suggests a value of approximately 4100. However, due to the shallowness of the curve at the optimal point, the difference in terms of MSE performance is extremely small. This implies that the error has not significantly affected the results of the later tests. This also implies that the value of β may be estimated without such exhaustive searching in

terms of the number of images left out. The shallow curve also suggests that searching for a very accurate value for β (by use of some mathematical optimisation technique such as simulated annealing or a simplex optimiser) is not necessary; these techniques are extremely time consuming and the calculation of fitness is also extremely time consuming. With several sample points, the optimal value may simply be read from a graph with reasonable confidence that performance will not suffer.

Note the overall performance of the system appears to degrade with the complete data set. Close inspection of the results reveals that two samples are extremely poor, and should be ignored.

Figure 3.17: The bad training data found in the set

Removing the results of training specifically on these spurious samples yields markedly different performance, as shown in Figure 3.18, with an even greater shift in optimal β, which in this case is 3670. This is a larger shift, though as shown in Figure 3.19 it still yields a variation of only 0.44 MMSE. Whilst these two bad samples were not omitted from the set used to construct the full scatter matrix, their impact upon the matrix as two samples from 700 should be minimal. The erroneous value of β used in the later experiments will lead to a conservative estimation relative to an estimation performed using the lower, true optimal β; that is the estimation will be slightly over-regularised and so will tend to produce results which are closer to the mean rather than producing extreme or randomly noisy data.

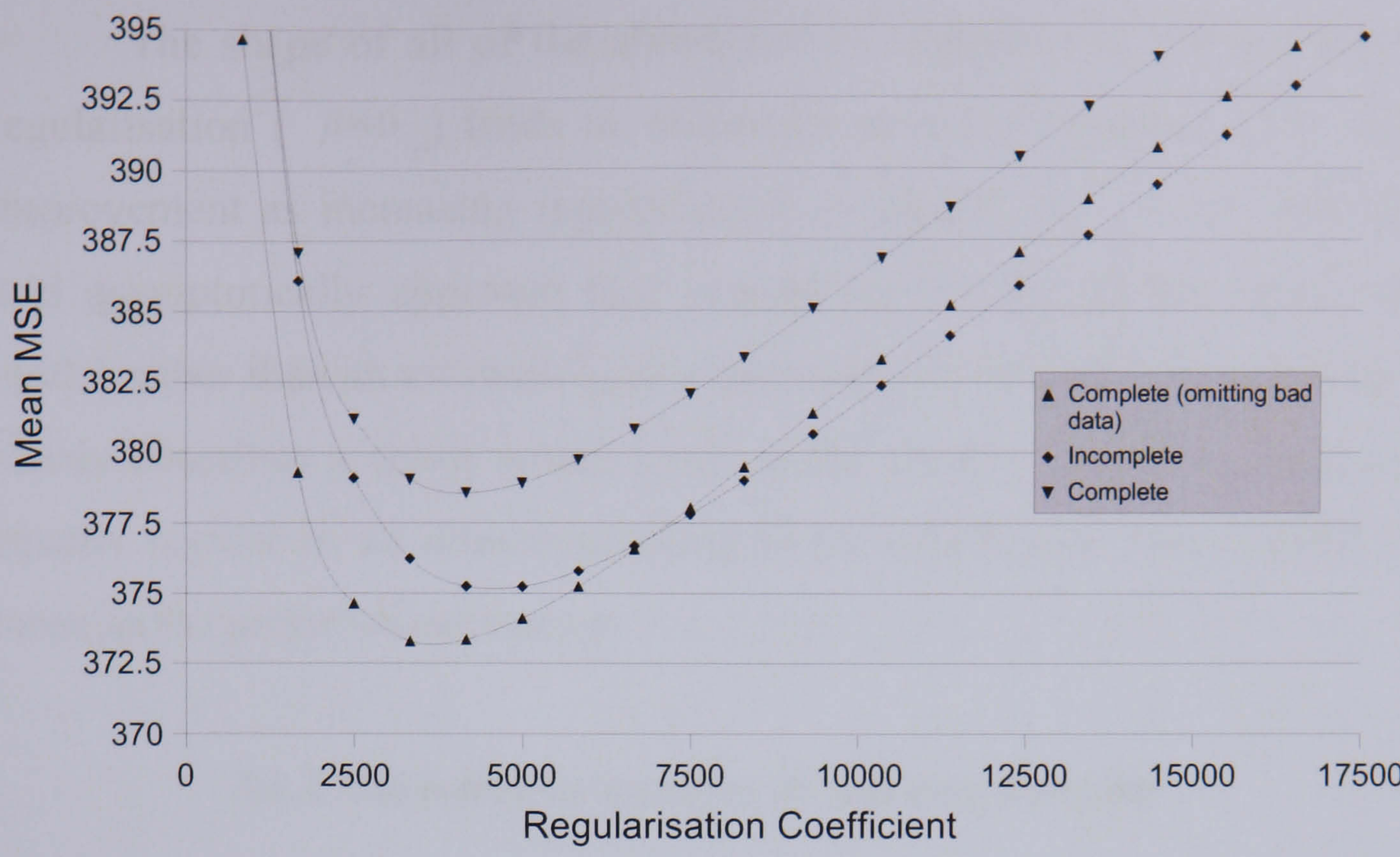


Figure 3.18: Comparison of the completed run with and without the bad data with the incomplete run from which the regularisation value was taken

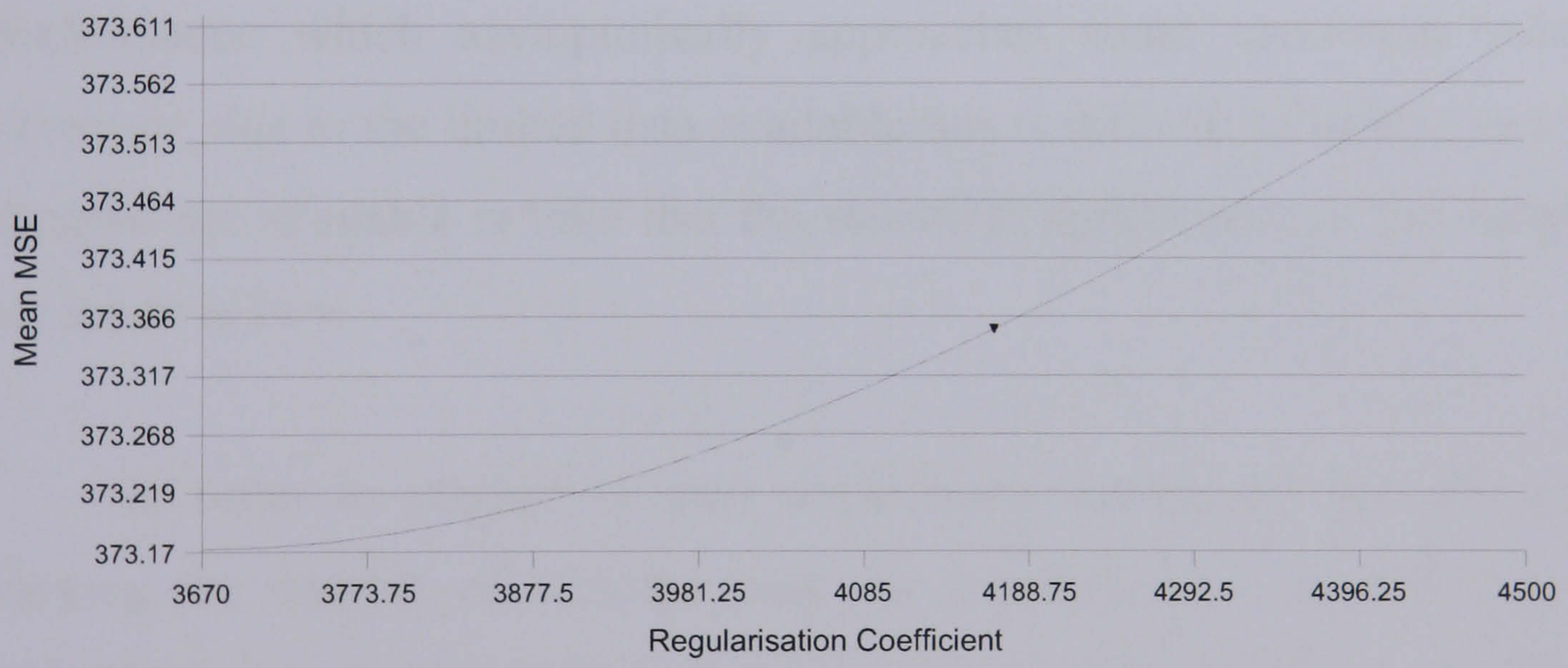


Figure 3.19: Close up on the complete (no bad data) run. The y-axis intercept is at the actual optimal value of regularisation coefficient. 4500 is the erroneous value. Note the very small variation in MSE (Y axis) over this range

The shape of all of the above curves suggests that performing no regularisation ($\beta=0$) leads to extremely poor performance, with rapid improvement as increasing regularisation is applied. As $\beta \to \infty$ the error will asymptotically approach that yielded by the use of the mean face model rather than an estimation; as regularisation increases, the covariance matrix describes a space which tends in the limit to being perfectly and equally regular in all dimensions, and hence which will always yield the mean as the preferred estimation.

3.2.2 Variation in number of training samples

Whilst the system appears to be relatively resilient to variation in the number of samples used in the leave-one-out training of β, this is not necessarily true of the number of samples used to form the underlying scatter matrix. It is postulated that increasing the number of truly independent samples used towards infinity should lead to mean performance which asymptotically approaches some maximum value. However, due to the limited data available this is difficult to verify; so few samples are available in total that the statistical significance of the sample set is low at best.

In order to attempt to gain some understanding of the effect of varying the number of samples used, the regularisation coefficient was trained on one quarter, one half, three quarters and all of the available samples. Samples were grouped such that new individuals were added as the training size increased rather than more images of the same individuals being present. Unlike the training above, the test set of twenty images was used to train the system so no leave-one-out iteration through the sample

set was required hence the scatter matrix needed only be calculated once.

The limited number of individuals available, however, severely limits this technique, since a one quarter size training set has so few individuals present in the images that the system is effectively operating near the noise floor. Training data from more individuals sufficient to allow the estimator to be trained on multiples of the original test set size, with each individual present in the image only once, would provide a clearer picture of the performance over varying set size. A more sophisticated approach could also re-run the test at each training set size in order to build a sub-set of the appropriate size from every possible combination of individuals in the full set. This would remove the effects of any unusually biased sub-set (for example, if no women happened to be present in one of the possible training sub-sets, this may detrimentally effect performance. By providing every possible combination of faces, provided the global set is representative of the population, then the error for this case will be averaged with the unusually good result of the all-women training case which would also be one of the sub sets and the middle quality results of the mixed set). The computational load of such exhaustive testing for large set sizes is considerable, but would provide a clear picture of the effect of varying training sample size upon both the optimal regularisation coefficient and the mean error performance. This could allow the estimation, rather than calculation, of the regularisation coefficient which would provide the capability for a much speedier recalculation of an approximation to the optimal covariance estimator should an application require it. More significantly, minimum average error against set size could be plotted, providing a means to identify the theoretical best performance that this technique could hope to achieve, and also reveal the point at which

additional training samples no longer contribute appreciably to the accuracy of the estimator. With the limited data available, however, this relationship is masked by the paucity of source data.

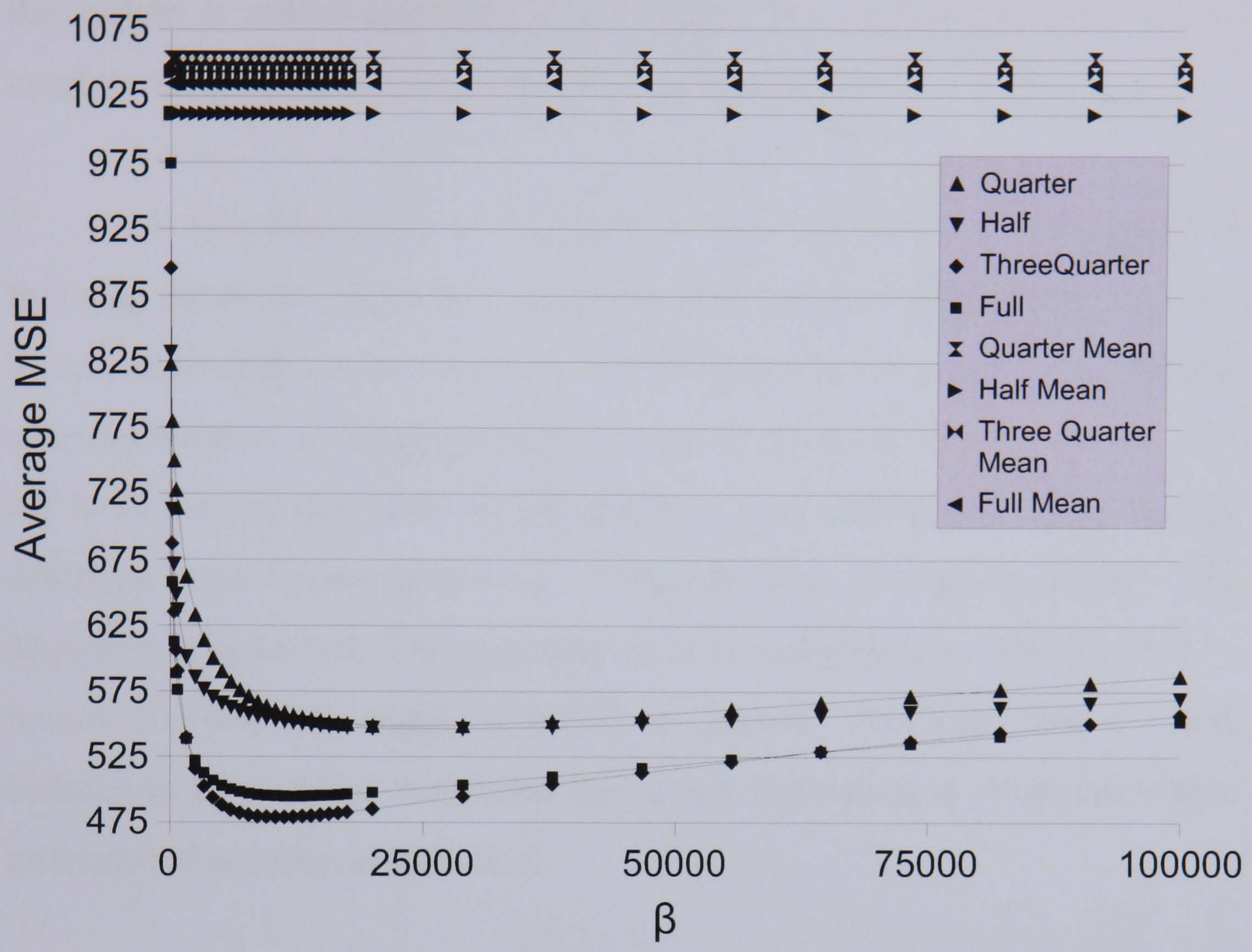


Figure 3.20: MSE against regularisation coefficient for varying training set sizes

The results shown in figure 3.20 do not show a clear improvement in every case as the training set size increases, though this can plausibly be attributed to the extremely small training set sizes involved and the relatively small test set. It is conceivable, for example, that the last quarter of images added were of individuals that are in some way less similar to the test images, thus providing a poorer estimation. Using a larger test set would also improve the accuracy of these results, since the error performance against a single face (which may be a statistical aberration) is

relatively significant in a set of only twenty images. However, it can be seen that there appears to be a trend towards better peak performance and lower regularisation coefficients as the set size increases. This suggests that the system is indeed operating at the noise floor, though this can only be confirmed by further testing with larger training sets.

It is also interesting to compare the curves with the errors calculated by using the mean (calculated using the same samples for each case as used to calculate each scatter matrix). Performance in all cases is far greater when estimation is compared with the use of the mean. The very slow roll-off from the optimal point on all curves is also interesting; whilst the the estimator must approach the use of the mean as β tends to infinity, this approach is not swift. The degradation of performance as β tends to 0 is, however, extremely rapid. It could be argued, therefore, that a high estimation of β is less dangerous than a low estimation in situations where a structured search is impractical.

3.3 Conclusion

Data capture techniques and the sourcing of data both directly and from third parties have been discussed. The quality and shortcomings of the available data has been analysed and the reasons for the choice of the Notre-Dame set as the source of training and testing data presented. The noise removal, posing and other processes required to prepare the data for analysis have been discussed. The nature of the final test and training sets, and the independence of the two, have been described. To summarise, the training set comprises 700 images of 131 individuals, and the test set comprises 40 images of 35 individuals, with no individual present in both

sets. The data in the set is composed of two images for each sample, a greyscale texture image and a greyscale depthmap image. Both images are 50 pixels in height and 38 pixels in width. Even with the dimensional reduction engendered by the use of such small, monochrome images the training set of 700 samples remains extremely small with respect to the dimensionality of the data which exists in 3800 dimensional space. This has been addressed by the use of mirroring to artificially increase the number of samples, though it is understood that this is a less satisfactory approach than the use of more genuine samples.

The implementation of the leave-one-out training process used to estimate an optimal value of β, the regularisation coefficient, including a discussion of the problems inherent in training such a computationally heavy process, has been presented. A solution to alleviating this process in the form of a parallelised training process has been described. Problems which occurred during the training process have been described, and their likely impact, analysed. A near optimal regularisation value has been estimated for use in depth estimation, $\beta = 4500$. Finally, observations on the effects of varying the training set size have been presented, though these have been severely limited by the number of training samples available.

4. Estimator Performance Analysis

4.1 Introduction

In this chapter the performance of the optimised estimator is evaluated. Experiments are described which allow the analysis of the performance of the estimator both objectively and subjectively, the results of the experiments recorded and the outcomes discussed. Evaluation by both subjective and objective measures is important; this allows the verification that as a generic technique for data estimation that the approach has merit and also allows an insight into the interaction of the technique with the human perceptual system. This is crucial to demonstrate the suitability of the method as an aid to human based recognition. Initially the raw mean square error performance of the estimator compared against ground truth data is evaluated. The performance of the estimated models relative to that of the originals and the use of a mean model is analysed in the context of human face recognition under varying conditions. Finally the use of a condition-matched estimated model is compared with a condition-mismatched true model, again in the context of human face recognition.

4.2 Test data

The basic data used is common to all of the following experiments, though rendering is an extra step required before subjective experimentation may be carried out. The regularised covariance matrix approximation described in chapter 2 and optimised in chapter 3 is used, generated from the 700 training samples. The regularisation coefficient used is 4500.

$$E[\boldsymbol{X}_1]=\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{P}_2-\boldsymbol{\mu}_2) \qquad \textit{Equation 43}$$

The two mean values in equation 1 (derived in chapter 2) are also generated using the 700 sample training set. Equation 43 may now be used to estimate the most likely depth values for a given photographic probe image. These photographic images are taken from the 40 samples from the Notre Dame data set reserved from the training process. As none of the individuals in these images was present in the training set, they are completely novel to the system. There is no chance, therefore, that the system has been unfairly trained to operate more effectively on these individuals, though as noted in section 3.1.7, the lighting conditions under which all of the images were captured are similar and so has been implicitly built into the model. This is unavoidable given the data available, and a study of the effect of this bias using good quality data captured under varying illumination conditions would be valuable. It should be borne in mind, however, that the other shape from shading techniques presented in section 2.1 all require an illuminant vector to be specified and so this situation is in effect still replicating the conditions usually assumed for the shape from shading problem. It is also noteworthy that this technique is not trained for a single illumination source – the number of sources in the

training set is irrelevant, and there is no requirement for the lighting conditions to be exactly constant. Indeed, slight variation within the training set will build tolerance to variation into the model. Whether this tolerance can be extended to large lighting variation is an open question that can only be answered through experimentation with data gathered under varying conditions.

These samples are split, using the photographic section as the probe, and depth models estimated. Figure 4.1 shows a selection of the estimated images produced. These are a cross-section of the test set, and demonstrate some interesting features of the system. From left to right, figure 4.1 shows the texture "probe" image, the original depth information for each sample, the estimated data, an image rendered using the texture and original depth information, an image rendered with the estimated model and finally with the mean depth over the entire training set. Note that the render engine is not raytracing in these examples, and so cast shadows are not drawn. The pose of the model relative to the camera is identical in each image, with any perceived difference due to variation in the models.

Figure 4.1: Examples of the output of the estimator. The left hand images are the original textures used as probes, the next images are the ground-truth depth maps for these images, as captured by a laser scanner. The next images are the estimated depth maps. Finally, a render with the original model, a "reconstruction" of this scene using the estimated model and the same reconstruction using the global mean face are presented for comparison. The original texture is mapped on to the models in all cases. Of interest are the second sample, showing a marked tolerance to the luminance noise in the source image, and the fifth sample, which demonstrates poor performance on a badly oriented and centred image.

Note the poor centring and orientation in the final image, and the resulting poor quality in the estimation and mean renders. The second sample, however, has clear luminance distortion in the texture image which

has relatively little impact upon the reconstruction. This is due to the fact that the system has been trained on data which has been centred in a reasonably uniform fashion. It therefore has no "knowledge" of faces which are not centred, and therefore little ability to replicate the movement required; instead the shape becomes ill-defined and incorrectly aligned with the texture. Even in the third sample, one with few similar images in the training set, the error in the estimation does not appear to the eye to be poor when rendered, though comparison of the depth map reveals more discrepancy. Compared to the mean render, however, it appears much more convincing. Faces which appear closer to the mean shape of the training set – such as image 1 – look arguably similarly convincing with both estimated and mean model used for rendering, though also could be considered to look like different individuals between the two. This apparent change of identity seems more pronounced in example 4.

This informal, qualitative discussion suggests that the human perception of such reconstructions may not necessarily agree with a raw mean square error evaluation of the success of the estimation.

4.3 Experiment 1: Mean square error performance

Mean square error allows the calculation of a non-subjective measure of performance. The disadvantages of using such a measure in image coding are well understood; fundamentally mean square error provides a poor estimation of the response of the human visual system. However, the advantages of MSE as a measure lie with its ease of calculation and

repeatability – MSE will never vary for a given image regardless of who makes the calculation. It is to be expected that a successful estimation system should produce lower MSE than the use of the mean model, as the mean is defined as the most likely model independent of the texture. If an estimation dependent upon the texture generally fails to improve upon the mean, then it is worthless.

The 40 test images were processed in exactly the same way as described in chapter 3 during the training of the estimator; that is the texture and depth components were split for each sample, a new depth component estimated and compared with the ground truth produce an error value. Figure 4.2 shows the result of this process.

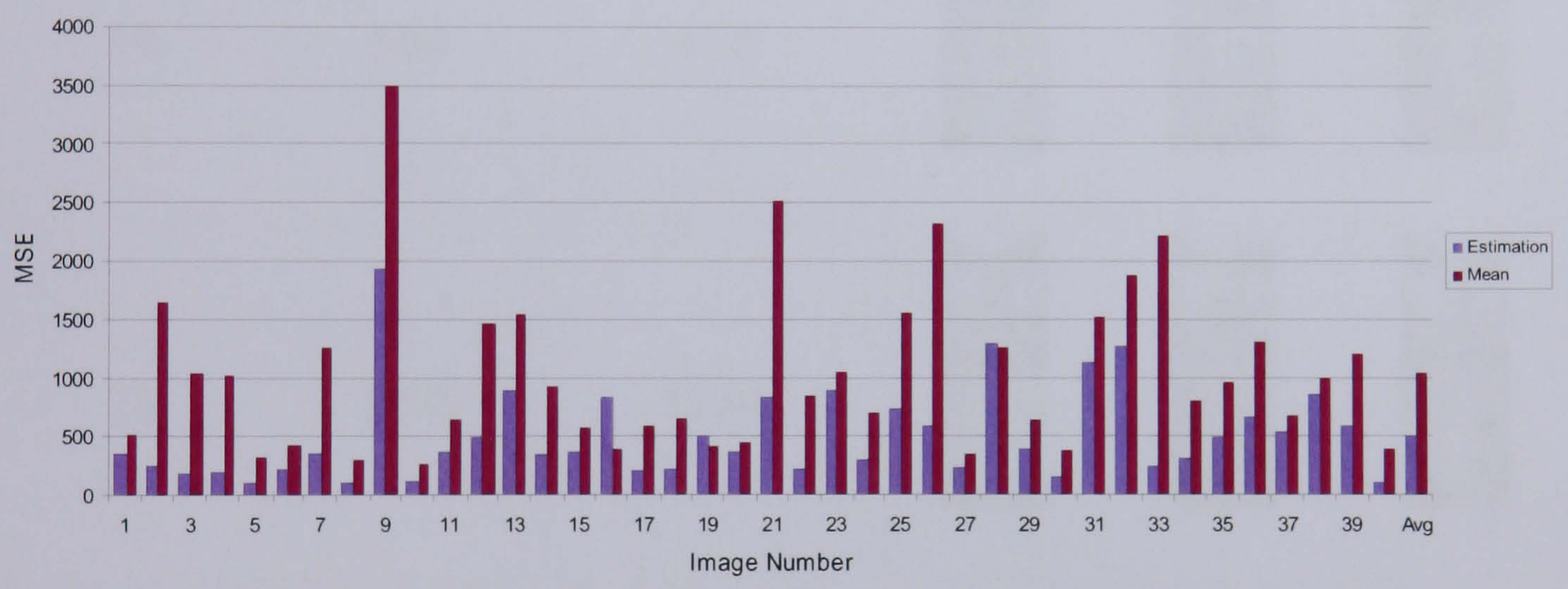


Figure 4.2: Mean square error of the estimated models compared with mean square error of the mean model, compared against the ground truth depth map for each sample.

As can be seen, in all but three cases the estimation improves upon the mean model, with an average halving of the error rate across the set (from 1034 to 506). The three samples exhibiting poorer performance are shown in table 4.1. It is arguable that the estimator was subjectively "heading in the right direction" with images 16 and 19 despite the greater

MSE – the lip and nose shape in image 16 appears to be tending towards that in the original model, and image 19 appears to exhibit a raised area around the mouth which is similar to the representation of the moustache in the original. In the case of sample 28, the variation between the estimation and the mean in error score is negligibly small.

Sample number	***Estimation MSE***	***Mean MSE***	***Original Sample***	***Estimated Image***	***Mean Image***
16	835	387			
19	504	410			
28	1288	1254			

Table 4.1: The three instances where the estimator does not perform better than the mean

Using these forty images as a sampling of the population of all non-training faces, it is possible to test whether there is a statistically significant difference between the performance of the regularised covariance estimator and a dumb mean estimation. A Students t-test where the alternative hypothesis is that the estimation result set has a lower MSE than the mean result set returns a significant difference at the 0.0001 confidence level; it is reasonable to say that the estimated result does indeed provide a

statistically significant reduction in error in the depth component.

4.4 Subjective performance

Whilst it has been shown that in pure, mathematical error terms the regularised covariance estimator provides a significant improvement, it has also been noted that the human visual system may limit the validity of this result. Subjective testing includes this system in the analysis of the performance.

4.4.1 Experiment 2: Comparison with mean and original

The first subjective experiment is designed to provide a direct comparison between human recognition with an image rendered using mean, original ground truth or estimated 3d structure. As discussed in chapter 1 the perception of identity is affected by variation in pose and lighting. The concept behind the estimation technique is that by fitting a suitably accurate model to a texture, pose and lighting conditions may be artificially matched between training and testing instances sufficiently well to eliminate the degradation in recognition observed as a result of lighting and pose change. Experiment 1 seeks to find the difference in performance when this re-posing is performed using a perfect model (the original), the population mean, and the model generated using the regularised covariance estimator.

Consider a textured 3d model of a face. The model is a depth map, and only includes the face area, not the sides of the head, and was created with the subject facing directly at the camera. If the render camera is placed directly in front of the model such that the z axis of the model is

aligned parallel to the lens normal of the camera then the three dimensional structure of the model cannot be directly observed. If in addition the render properties are set such that there is a global, diffuse ambient level (i.e., shadows are not cast) then there are no visible indirect cues to the shape of the model. Therefore, a texture applied to any model which has no discontinuities will appear unchanged between renders using different models (providing that the camera properties are set such that perspective distortion is negligible over the range of the model). Altering the position of the render camera, such that the model z-axis is no longer parallel to the lens normal allows direct observation of the shape variation in the model. Conversely providing a non-uniform directional light source and using a raytraced rendering technique to cast shadows allows a shape-from-shading perception of the variation in model shape. In the case where both of these steps are taken, both types of cue are apparent in the rendered image. Examples of these conditions are shown in table 4.2, rendered using the different model types available. The significance of this is that by careful arrangement of the render conditions, the effect of shading due to directional lighting in the form of cast shadows and the effect of direct observation of shape may be observed independently. This allows a much more complete picture of the potential benefit of the use of the estimated model.

	Control	Shape	Shade	Both
Estimated model				
Mean model				
Original model				

Table 4.2: One sample from the training set rendered in the 12 different conditions to be used in the test.

The aim of this experiment can therefore be refined using this rendering technique to compare the effect upon recognition when images rendered with a ground truth model, an average face shape model and an estimated model under conditions where direct shape deformation due to the 3d model is the only varying factor, where shape information from cast shadows is the only varying factor, and where both factors vary. Note that in all cases the subject is trained on ground truth models, and model variation only occurs at test.

4.4.2 Test procedure

24 samples were selected from the available full test set. These were selected such that no individual was represented more than once, and the samples with poorly framed and noisy original data shown in figure 4.1 were not used; after being constrained to some extent by these considerations, final selection of data was random. The 24-sample available set was not changed during the test.

Subjects were selected only on the basis that they had seen none of the data before. Beyond this, selection was not constrained; all volunteers were accepted. No record was made of age, sex or ethnic origin for reasons of privacy, though no minors took part in the experiment. Since these factors were not controlled for, there is the possibility that some bias may be present in the results. However, since 72 individuals were tested it may be considered probable that a reasonable mix of gender was achieved. The test group were drawn from a predominantly white Caucasian population, and as such the likelihood is that fewer representatives of other races were present in the data. Since individual subjects are presented with every combination of test condition during the test process, this is unlikely to have any significant effect upon the relative performance between cases.

Before beginning the experiment, each subject was shown the swatch displayed in 4.3 and asked to adjust the brightness and contrast of their monitor until the number 1 on the upper left corner and the number 16 in the lower right were both on the edge of vision or visible to ensure that all subjects were observing the same range of grey values.

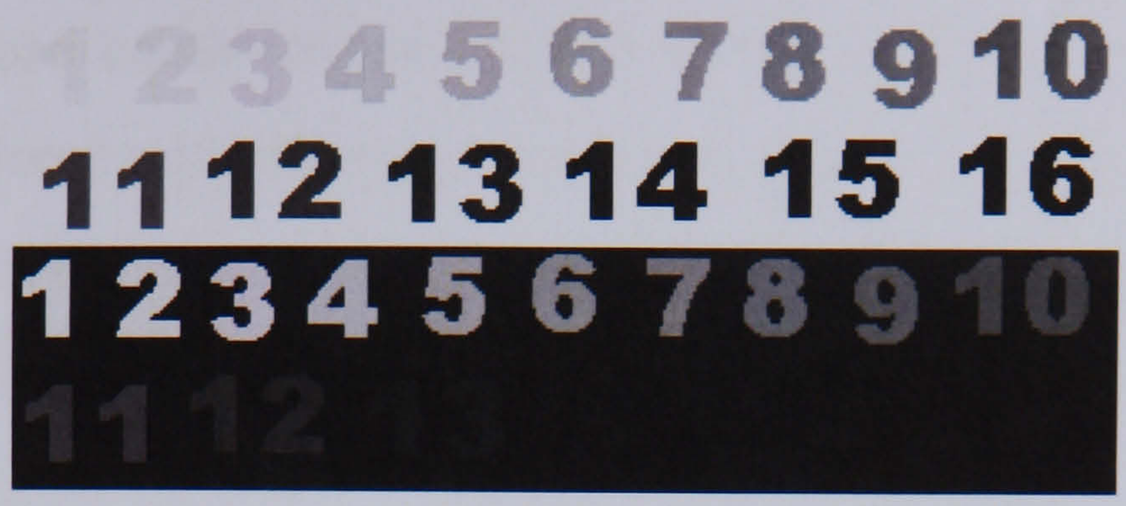


Figure 4.3: Calibration swatch

At the training stage, subjects were presented with a set of images rendered using ground truth 3d models which match the texture map image for the individual in question. In total, each subject was presented with twelve images each of a different individual, with the images selected such that after the tests had been conducted on all 72 subjects every individual present in entire test set had been presented at training an equal number of times. The test images for a given subject were equally distributed between the four render conditions (control, shape, shade, both, referring back to table 4.2). Images were presented for five seconds each, with only one image displayed upon the screen at a time. A three second mid-grey and then a three second fixation point image used to separate each image. The order in which the 12 images were presented was randomised at run-time. All images were set up using Blender and rendered with the YAFRay raytracing plugin. Reflectance properties were set up as in the experiment described in section 1.3 for the image requiring cast shadows. Shadow casting was turned off for the images requiring that shadows be absent.

In the test stage each subject was exposed to 24 images, 12 representing the same individuals as the training images presented and 12

novel distractors, again with a mid grey and fixation image shown between each stimulus. Of the repeated images, all were rendered under the same lighting and pose conditions as their training counterpart. However, one third were rendered with the mean model, one third with the original model and one third with the estimated model. Every condition represented in table 4.2 was, therefore, presented once as a true repeated stimulus to each subject (albeit each condition being represented by a different sample). The twelve distractors, again each representing a different individual, also comprised one render for each of the 12 combinations of render condition and model type. Over all subjects, each sample was represented an equal number of times as distractor and as valid stimulus. The order in which the images were presented was again randomised at run-time. Subjects were asked to respond by pressing the spacebar if they believed they had seen a face which was present in the training set, and any other key if they believed a face was novel, and were told that accuracy was more important than speed.

4.4.3 Results

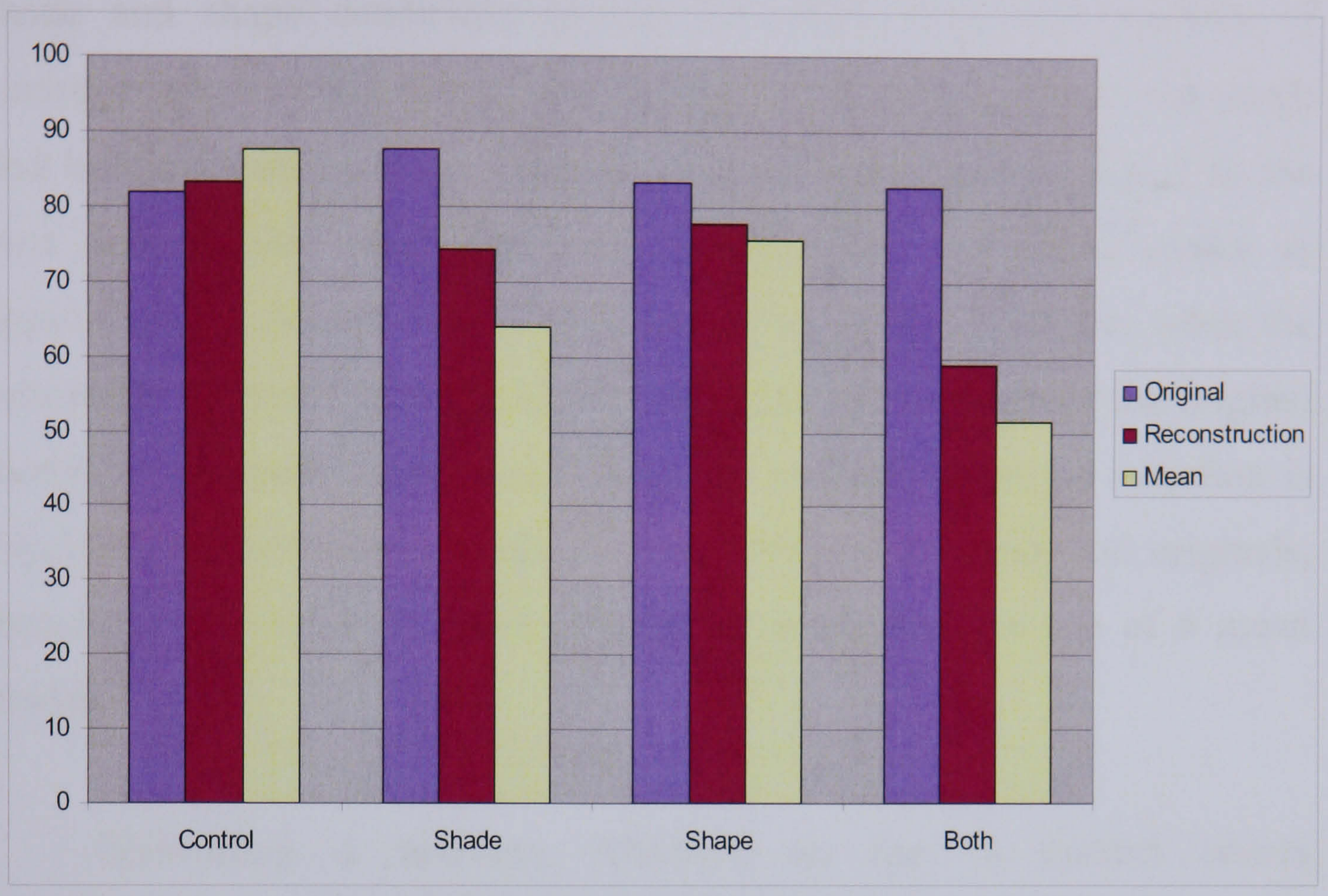


Figure 4.4: % correct decisions categorised by model type and depth cues available

Figure 4.4 shows the percentage of correct decisions made in each condition. Note that unusually high score for the mean model in the control case, and also the better performance in the shade case of the original models. Both of these are unexpected, as the original model renders do not vary between training and testing cases at all, and any variation between model types in the control case is limited to a few extra pixels of background information around the jaw on some images rendered with the mean model – these occur because the extent of the mean model exceeds the extent of the original model around the jaw, depending upon the shape of the face. It is probable that this is a saturation effect – in all render cases original to original matching varies between just above 80% and just

below 90%. The shape render condition shows less variation between model types than the shade and both conditions. This may be because the shade and shape conditions do not introduce equivalent amounts of variation; an exact like for like comparison is difficult as the viewing angle and lighting conditions are fundamentally different variable types. In the case where both vary, performance with with the mean model is approximately chance, with a 9% increase in correct decisions when the estimated "reconstruction" model is used. Both fall far short of the original model. In all render conditions except the control, where no variation is expected, the estimated models perform between the mean and originals, suggesting that they do assist recognition more than the use of a mean model.

Performing a two-way ANOVA on the % correct scores demonstrates a signficant difference between the 3d model types and also a significant difference between the render conditions at the 95% confidence level. However, the interaction F-value falls below the critical value at this confidence level, suggesting that there is some interaction between render condition and model type. If the control render condition is omitted from the analysis, however, then a significant difference in interaction is apparent at the 95% level (i.e., there is not an interaction between render condition and model type), suggesting that this is a feature of the saturation effect in the control condition, since all model types when rendered under this condition produce essentially identical testing images to their respective training images. Continuing the analysis with the control render condition omitted, application of a Tukeys post-hoc test demonstrates a significant difference between all model types but no significant difference between the shade and shape render conditions at the the 95% level. Both

shade and shape are found to be significantly different to the "both vary" condition.

To summarise, this demonstrates that:-

- Each of the model types used produce genuinely different error rates, the original ground truth models for each face providing the best performance, and the mean face model providing the worst. Whilst the estimated model produces a significantly poorer performance than the ground truth model, it is also significantly better to use the estimated 3d model to render images for human recognition than it is to use a mean 3d model.

- There is no significant difference between directly observed 3d shape changes and shape changes inferred from shading under these conditions. However, providing an image where both shading patterns and direct observation reveals changes in facial surface shape, performance drops significantly.

Performing a signal detection test upon the data yields the following values of discriminability index (d'), where 0 is a signal which is not detected and positive results are detected increasingly reliably.

d'	Control	Shade	Shape	Both
Original	1.85	2.31	1.94	1.88
Estimation	1.94	1.33	1.57	0.46
Mean	2.31	0.74	1.46	0.09

Table 4.3: Signal detection test results

This again shows consistently poorer performance when the mean model is used. Note that d' in the mean model / both cues present case is very close to 0 (i.e., the signal is not detected at all). The increase in d' comparing the mean model and estimated model cases where only shading cues are present is substantial, with the discriminability nearly doubling. The increase in the "both" case is also dramatic in these terms, with the estimation proving over five times more effective than the mean, though both are still in absolute terms quite poor.

4.4.4 Experiment 3: Pose and lighting matched reconstruction versus pose and lighting varied original

The final experiment again incorporates the subjective human perceptual process. The aim of this experiment is to directly compare the use of the estimated model to match pose and lighting with no use of any correction whatsoever. The latter is the current standard case in almost all genuine applications; this experiment serves to evaluate the benefit (if any) of applying estimation and condition matching over the de-facto standard. Subjects are presented in the training stage with images rendered with the original model in exactly the same way as in the experiment described in 4.4.1. However in this case, all models are rendered under the same condition; the camera at a fixed angle to the z-axis of the depth map, with directional lighting and cast shadows rendered. 12 images are used by each subject for training, with the same samples used as were selected for the previous experiment. In total 24 training images are available in the training pool, each of a different individual. Each training image is presented an equal number of times across all subjects, never more than

once to each individual. Again, a mid-grey and fixation image are presented to the subject between each training stimulus, with the training images presented for five seconds each. In the testing stage, subjects are presented with 24 images, 12 of individuals present in the training set shown to the subject, and 12 distractors. Half of both of these for each subject are rendered using the original model, the render camera aligned with the z-axis with extremely diffuse ambient illumination from directly behind the camera (so no shadows are cast by the face). The other half of the test set are rendered using an estimated 3d model, with the render camera position and lighting conditions duplicating those in the training images exactly. Table 4.4 shows the three images used in the test for one sample – note that only one of the two testing images would be used for a given subject, and the training image used for only half of the subjects. Over all subjects, every sample was used both as distractor and true repetition an equal number of times, and also represented in each testing render condition an equal number of times. Individuals were never present in more than one image in a given test run. Subjects were asked to decide whether each testing image shown represented a new individual or someone they had seen in the training set.

Training Original Model	**Testing** Original Model	**Testing** Estimated Model
		

Table 4.4: The three renders of an individual used in the test

Conceptually, this is replicating the situation where an image is captured in some environment with known pose and lighting conditions (the training image), but is being compared with an image captured in quite different conditions – a passport photo for example. The original model testing case represents this use of a poorly matched still photograph. The estimated image, with the render conditions matched to the original, is an attempt to use the regularised covariance estimator to manipulate the photograph in such a way that the effect of the mismatched conditions may be minimised.

4.4.5 Results

The correct decision rates produced by this test are shown in table 4.5. As can be seen, mismatched conditions result in an approximately 8% lower correct decision rate compared to matched conditions with an estimated underlying 3d model.

	Mismatched	Estimated, Matched
% Correct	56	64

Table 4.5: Percentage correct decisions

Figure 4.5 shows these percentages broken down by response type. The mismatched conditions (rendered using the original model) show a greater negative bias than the reconstruction of the original image using the estimated model. This suggests that the mismatched conditions degrade performance pessimistically; that is, the subjects tend to fail to recognise faces rather than mis-recognising distractors.

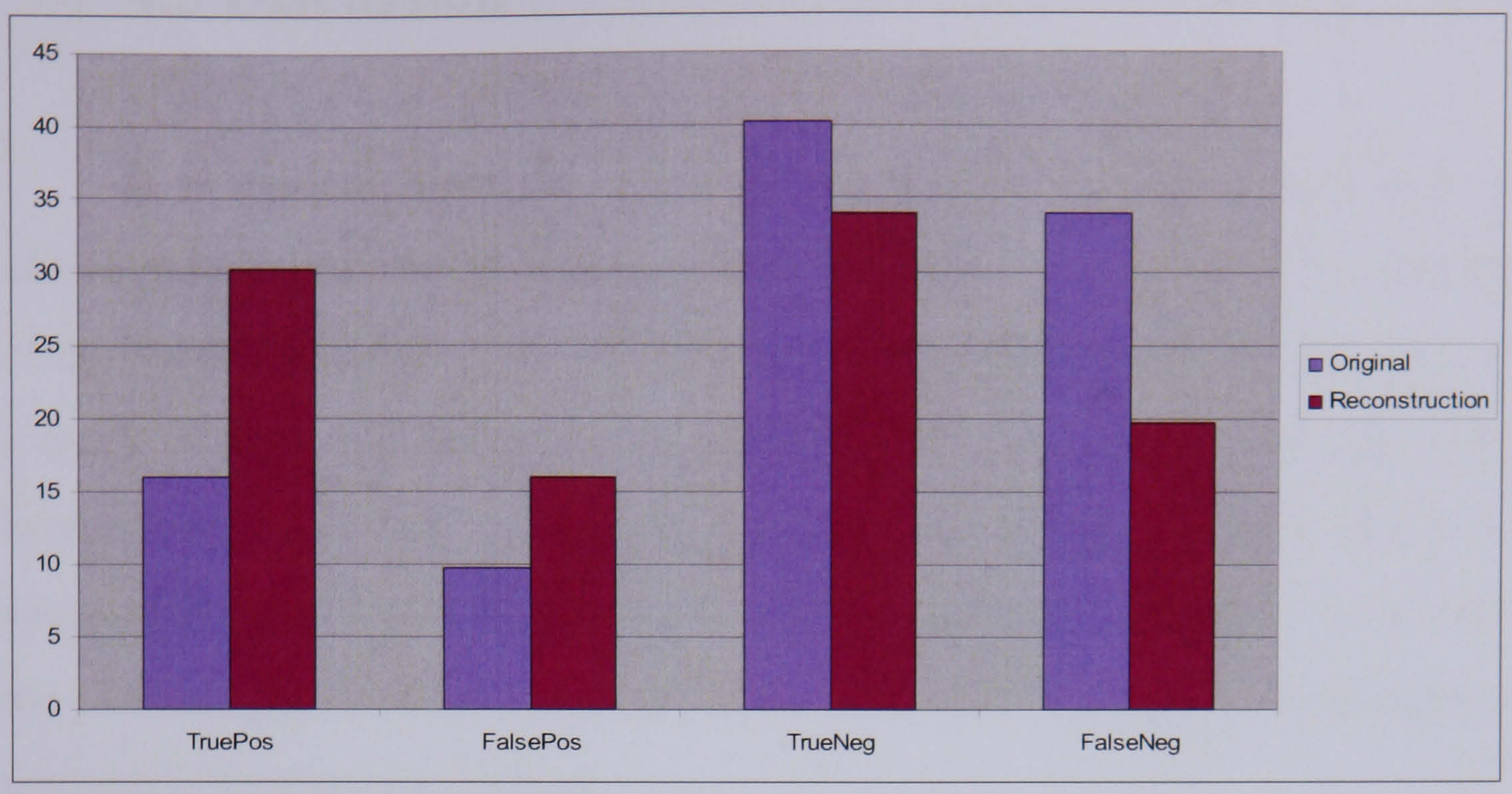


Figure 4.5: Percentage of results by response and model type

Performing a signal detection analysis reveals that whilst subjects are capable of discriminating the true images from the noise of the distractors in both cases, in neither is the performance spectacular. With the estimated reconstruction, however, discriminability is substantially raised over the mismatched case.

	Mismatched	Estimated, matched
Signal Detection	0.39	0.73

Table 4.6: d' for the two cases

A Students t-test reveals the probability that the mismatched and estimated matched responses share the same distribution to be 0.0367; therefore it can be concluded that at the 5% significance level there is an improvement in performance gained by matching the lighting and pose conditions using an estimated model.

4.5 Discussion

It is evident from the result of experiment 1 that the estimator is effective at improving upon the population mean in general, with a halving of the mean square error and vanishingly small probability that the result is a quirk of the samples, as shown by the result of the t-test. However, this improvement is purely in terms of mathematical accuracy, and it does not necessarily follow that human perception is similarly improved. However, the results from experiment 2 support the idea that the improvement carries through to human recognition.

It is evident that in like-for-like render conditions, the estimated models produce higher accuracy than the mean model in this test set. This improvement is very slight in some cases, though statistically significant.

It is important to remember that comparing the direct shape observation and shape from lighting render conditions is not as clear cut as it may first seem; lighting set-up and viewing angle are fundamentally different and it is not possible to say that the lighting condition used is "equal" in some way to the viewing angle used. It is to some extent not relevant to this work, however, as there is a systemic constraint inherent in the model: It has been produced using a depth map. It is impossible to vary the viewing angle as much as the lighting conditions without rotating the model so far that most of the useful information is obscured, and the main feature observed is the unrealistic detail at the very edge of the model. What should be understood, then, is that *given the limitations imposed by the use of a depth image* the estimated model provides a significant increase in performance from shading cues, and no significant benefit from

shape cues. In addition to this, experimental results suggest that in terms of the effect upon human recognition, the variation revealed to the subjects by both the shading cues and direct shape cues is approximately the same, since performance between these two sets did not significantly vary.

The case where both cast shadows and direct shape information are present is also complex. It could be expected, from the relatively small drop in performance observed for both mean and estimated models with respect to the original in the shape-only case that the "both" case would yield results very similar to the "shade" case, with the effect of direct shape providing minimal contribution. This is not the case, however. The performance with the mean model drops almost to the chance floor, with the estimator faring less than 10% better. The combined effect of direct shape and cast shadows is clearly not simply additive. The 59% accuracy rate of the estimated models could be considered disappointing, but if the 82% correct decision rate garnered when the original models are used is considered to be the performance ceiling, and the 51% rate from the use of the mean considered to be the floor, then the use of the estimated model yields an improvement of 28% of the available full scale.

The small system training set used to generate and optimise the regularised covariance estimator must be considered; with such a small set, the regularisation process will produce an estimator which will tend to mean more than a more well-trained estimator (as suggested in chapter 2). Further, any individuals in the test set which have facial features which are not well represented in the system training set will be poorly served by the estimator. The reduction of the geometric dimensions of the images required to constrain the dimensionality of the face space to allow the

estimator a chance of generating a reasonable description of the space is not ideal for subjective testing; all source images were limited to 38 * 50 pixels before rendering; much detail is therefore guaranteed to be missing from the images. The 24 images used to test performance also make up a limited set; should further data become available then more extensive testing may allow a reassessment of the extent of the performance difference between the mean and estimated models.

The final experiment is less ambitious. It seeks only to address the initial starting point for this work: *can an estimation of a three dimensional model from a photographic image improve human recognition when used to match pose and lighting conditions with some novel image*. The result of this is a clear positive; the reconstructions of the original image using an estimated model are recognised more reliably than the pose and lighting mismatched images. Interestingly, the reconstruction case here is a direct repetition of the both-estimated case in experiment 2, yet yields a substantially higher correct decision rate. The mismatched images are recognised with a slightly higher accuracy than the both-mean case in experiment 1. Whilst a direct comparison of this latter pair has not been carried out, the implication is that the use of a mean model will not provide much benefit, whereas the estimated model demonstrably does.

4.6 Conclusion

The results of estimation of depth for novel images have been discussed qualitatively, and a critique of unusual cases presented. Performance has been evaluated both using pure mean square error and subjective human trials. Mean square error performance demonstrated a large improvement over the population mean as an estimate for the depth map. Human performance in recognition tests based on images rendered with the estimated 3d models has been shown to be significantly worse than for images rendered with ground truth 3d models, though significantly better than images rendered using an average facial shape. The performance has been discussed with suggestions for further testing and improvement of the estimator (principally by the addition of further training samples). A trial of the basic concept – that an estimated model rendered under matching conditions is superior for human recognition to a non-estimated image with mismatched lighting and pose conditions – has been presented, with favourable results, suggesting that the technique may impart a benefit to human recognition in real applications.

5. Further Work

Whilst the estimator has proven successful, there are a variety of avenues to explore which may improve the system further. As noted in Chapter 3, varying the size of the data set used to create the scatter matrix and then train the regularisation coefficient has a significant effect upon the performance of the estimator. However, due to the limited data set available structured testing of this dependency is not possible. Should further 3d data become available, calculating the average error performance over some common testing set such as that used in Chapter 4 with controlled variation in the number of samples available for training should provide insight into the response of the system to increasing the training set. It could be expected that such an experiment could characterise the system sufficiently well to produce an approximate peak performance curve, minimum MSE against training samples available. A similar curve could be created describing subjective test performance as the number of system training samples vary – this may be expected to follow the MSE curve, but this is by no means guaranteed. Such a test run would, however, require a large number of test subjects.

A criticism that could be made of all the evaluation experiments reported in this document is that though the individuals in the test set of data were not present in any of the system training images, all the data was captured at the same source with identical equipment; this is a system that

performs well with University of Notre Dame data, but what of completely unrelated images? As shown in figure 5.1, reconstruction of depth models from completely novel images captured in completely novel environments is certainly possible, but without ground truths to compare with, evaluation of performance is difficult. The York data set is relatively poor, with very little detail present in the 3d models and not well suited to use as a benchmark: A system performing well may score poorly when compared with such low quality models.

Figure 5.1. Dudley Moore[34], Sean Bean[34] and Uma Thurman[34] (left) rendered in 3D (right)

Another area for investigation lies in the use of two (rather than one) regularisation coefficients. The system as implemented assumes all data within the sample is of the same type, and so should be regularised to the same extent. There is, however, a clear divide between texture image and depth image. The two are clearly different; there is no convincing reason why it should be assumed that the two halves of the data should share the same regularisation coefficient, rendering the covariance estimator thus:

$$\boldsymbol{\Sigma}(\alpha,\beta)=\boldsymbol{S}_{total}+\begin{bmatrix}\beta_1\\ \dots\\ \beta_2\end{bmatrix}\boldsymbol{I}$$

Equation 44

β_1 and β_2 would both require optimisation during the training stage, increasing the complexity of the optimisation task greatly; some structured optimisation algorithm would be required to perform this task. With the

fitness function already defined as the MSE calculation, this is not a difficult task. A further class-based distinction may be made for the subject of the images; it may be that performance may be improved by the categorisation of faces into age, gender or racial classes.

Another area of investigation which may prove fruitful is the fusion of this technique with other shape from shading methods discussed in section 2.1. for example, is is noted that the Prados and Faugeros[54] technique is relatively successful but produces models which still do not appear natural. Combination of the two techniques may yield a more natural model in the situation where a facial mugshot is being analysed.

There are a further two points to note about the subjective tests; firstly the simulated nature of the data and secondly the omission of any automated pose and light estimation technique.

The first of these issues is relatively straightforward: The concept behind the work is to re-render an image taken from one photograph to match more closely an image in another. The tests carried out, however, all have identical texture information and at no point are the rendered models compared to photographic images. Carrying such tests out is an obvious next stage in the subjective testing of the system. The latter is more complex; it has been assumed throughout the work that the lighting and pose condition to be matched to is known. This will rarely if ever be the case outside the laboratory. Whilst there is an argument that a user could estimate the relevant pose and lighting settings to approximate a scene, ideally some automated technique would perform this step. This is beyond the scope of the work presented, but a natural extension to it when

considering the further development of the technique as a tool.

It can be envisaged that the estimation technique could be used in live recognition situations; for example at passport check points. Here, the passport could be handed to the operator who would scan the photograph. A model can be estimated rapidly due to the direct nature of the estimation technique. Either by manual adjustment, some automatic process or by prior measurement of the conditions the photograph can then be re-rendered to match the observed lighting conditions, facilitating more accurate verification of identity. This is dependent upon render time, and current photo-real rendering techniques have hefty computational time overheads. A plausible alternative scenario which would be more tolerant of this is the situation where some closed-circuit television or surveillance still of a suspect is being analysed by the police. A mugshot of the suspect can be scanned, a model estimated, and then the image re-rendered to approximate the conditions in the original security data.

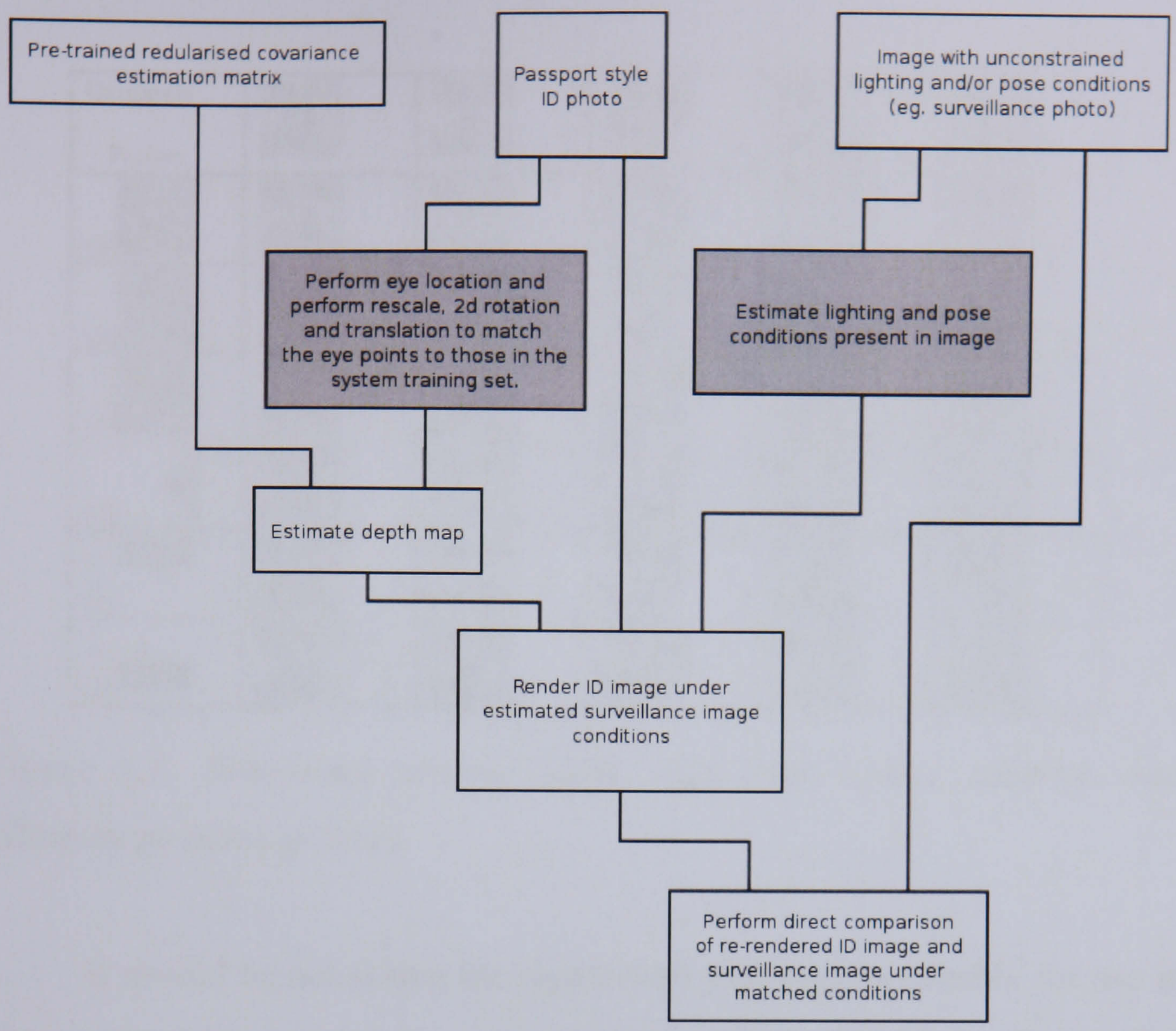


Figure 5.2: Proposed system for enhanced human recognition or identity verification. The greyed boxes represent sub-systems not addressed in this work; they may be carried out by hand or some automated technique applied. Eye location algorithms are readily available commercially[32, 46]*, and are extremely swift in operation.*

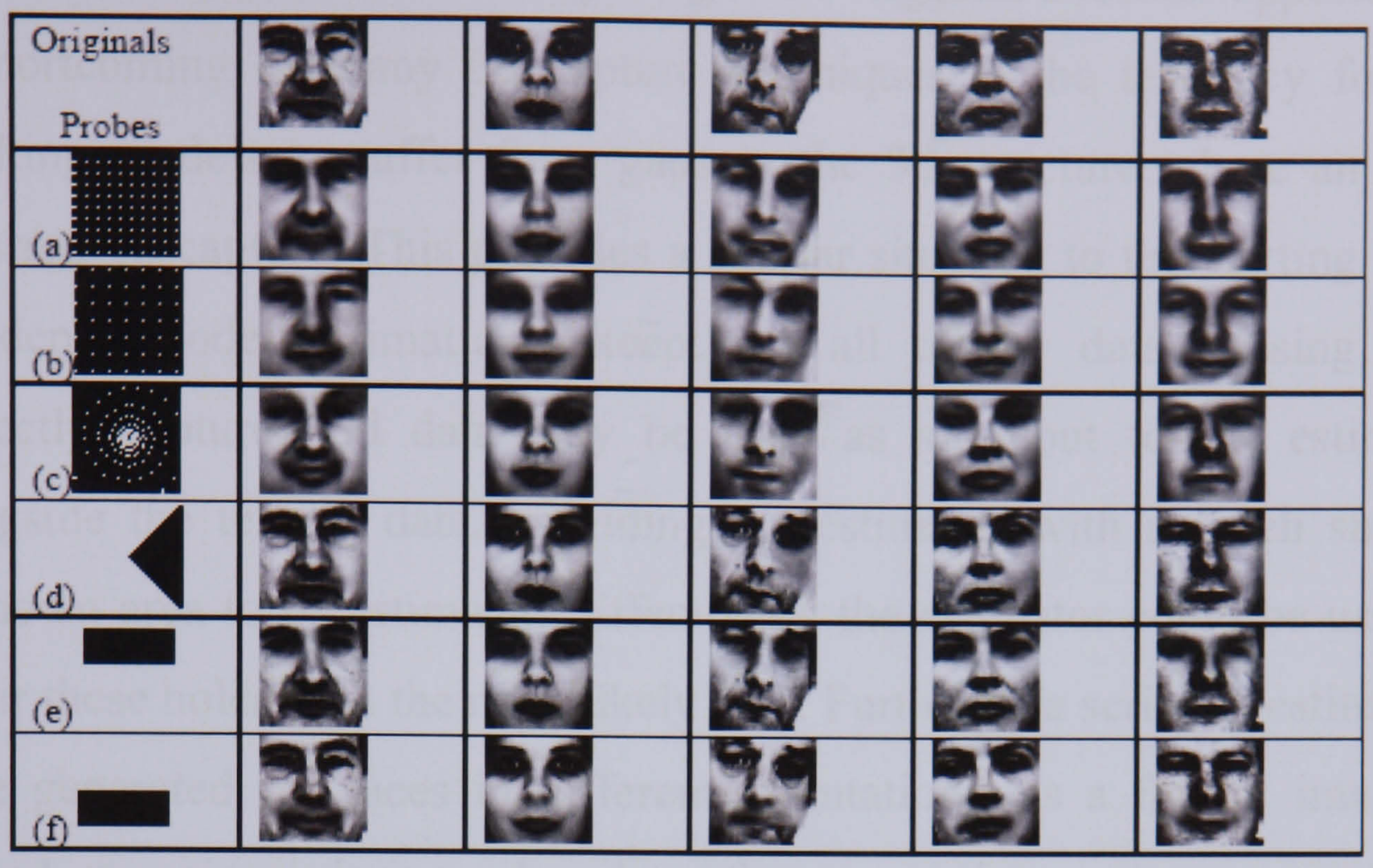


Figure 5.3: Estimating missing image data from sparse samples, and filling large missing areas

It should be noted that the regularised estimator is suitable for use in many situations beyond the reconstruction of depth images. Figure 5.3 shows an example of this; here facial images have been masked using the probes in the leftmost column, with black pixels in the mask removed from the image, and non-zero mask pixels representing known data. The system here has been trained on 2325 38 square images of faces, with none of the five images shown in figure 5.3 present. Even with the sparsest of input samples, a reconstruction of the original image is always possible to some extent, with large missing areas plausibly filled with geometrically correct but neutral data. The results of both the structured tests and these illustrative examples suggest that the application of a regularised covariance estimator to the wider field of image based analysis – and further to any system with high dimensional data conforming approximately to a multivariate gaussian – is worth pursuing.

This estimation of missing image data suggests a further opportunity. A shortcoming of many 3d capture techniques is the tendency for the resulting models to suffer from gaps in the 3d structure where an error occurred on capture. This provides a similar situation to the starting point for depth model estimation, except not all of the data missing. The correctly captured 3d data may be used as an input to the estimator alongside the texture data, providing the estimator with a much smaller unknown area to be estimated. Effectively, the estimator could be used to repair these holes with the most likely data. Further, if a series of estimators were generated for faces in different orientations, as a frontal image is rotated exposing missing edge data, the appropriate estimator could be applied to fill these areas, increasing the range of poses over which the model is of use, since these gaps and edges are unsightly and distracting to the human eye.

6. Conclusion

The poor performance of humans at the recognition task under varying lighting and pose conditions has been discussed and demonstrated, with lighting variation causing a more significant degradation of performance over the range of render conditions tested – this range being limited by the properties of the depth map "mask" model of a face, precluding extreme viewing angles but not large variation in lighting.. The use of an estimated 3d model to allow the re-posing and re-lighting of an image to match the testing conditions, thus removing much of the pose and lighting variation, has been proposed. The reliance of shape from shading techniques on unrealistic assumptions has been discussed, and the use of a statistically trained estimator as an alternative has been suggested, and an argument against the dimensionality reduction techniques generally applied when image-based analysis of facial data is carried out is presented. Estimation by conditional densities using a regularised scatter matrix as a covariance estimation has therefore been proposed as a workable alternative to shape-from-shading based on image irradiance. The training process for such an estimator has been described in detail, and an optimised estimation transform generated. The performance of the estimator has been tested for both mean error and subjective performance. In terms of pure error scores relative to the ground truths, the estimator yields approximately half the error rate when compared with the use of the mean model as an estimate. The subjective performance is less definitive: where

only cast shadows are available as shape cues subjects recognise images correctly significantly more reliably. When shape cues are absent, and direct shape cues (visible through viewing the model at an angle to full-frontal) are present, the estimation provides a negligible increase in performance. In the case where both cues are present, however, the estimated models provide a large increase in performance, though the variance of the subjective results mean that the distribution cannot be considered significantly different to the responses garnered using the mean model. However, performing recognition using the estimated model to match pose and lighting conditions has been shown to provide a significant improvement in performance over the use of an image with mismatched pose and lighting conditions. Given the earlier finding that lighting appeared, over the range available, more significant to human recognition performance, these results are consistent with what could be expected, though it is speculated that the severely limited training set and small image size result in much poorer performance – and hence less differentiation from the performance of the mean – than the system would be capable of were more data available. Further work to characterise the performance of the estimator as training set size varies will be valuable in confirming this and evaluating the full potential of the system.

Finally, it must be noted that the estimator described in this document has the potential for a distinct speed advantage over the traditional shape-from-shading technique. Shape-from-shading is an iterative process, with a significant time overhead as the estimated model is adjusted to match the observed image. Unlike other work in the area, which aims to improve performance by constraining the search space, this technique is a fundamentally different approach. Whilst training of the

estimator is extremely time consuming, this need only be performed once. Any estimation carried out is then a direct matrix multiplication of the image data with the estimator, and as such could be performed with great efficiency on a device such as a sufficiently large FPGA.

6.1 Contributions of this thesis

Primary contributions:-

- The influence of pose and lighting upon human face recognition has been assessed specifically over the range of conditions that a frontal mask model of a face may plausibly be re-posed and re-lit. This has revealed clear evidence that the variation of pose and lighting within these ranges has a severe impact upon recognition rates, with the variation in lighting conditions being the more significant component. This suggests that the generation of this type of mask model has the potential to allow re-rendering of an image to substantially improve reliability of recognition.

- The unsuitability of traditional shape-from-shading techniques has been discussed, and estimation based purely upon a trained statistical model proposed as an alternative. Dimensional reduction has been shown to be an unsafe technique to apply. This thesis has proposed that a regularised covariance matrix is a suitable estimator for depth from image intensity given sufficient training samples. Given the sparsely sampled nature of the problem, RCE[57] has been proposed as a suitable covariance estimator, with a simplification based upon the single-class nature of the face data described.

- The method developed has been assessed through subjective experimentation and shown to be of benefit to the human recognition of faces where lighting and pose conditions are at variance.

- A complete system to allow the enhanced identity verification of humans across varying lighting and pose conditions within images has been proposed. The sub-systems of this process involve application of existing methods or human input; the estimation technique introduced, developed and verified in this thesis allows the construction of a system, as it provides a practical solution to the depth estimation problem.

Secondary contributions:-

- A distributed optimisation technique for the generation of a regularised covariance estimator has been devised

- A large set of three dimensional facial models has been collected in collaboration with Tom Heseltine of the Department of Computer Science, University of York

- Both the York data set and the Notre-Dame data set (used for the Face Recognition grand challenge) have been analysed, with flaws in the data presented and methods for noise removal discussed.

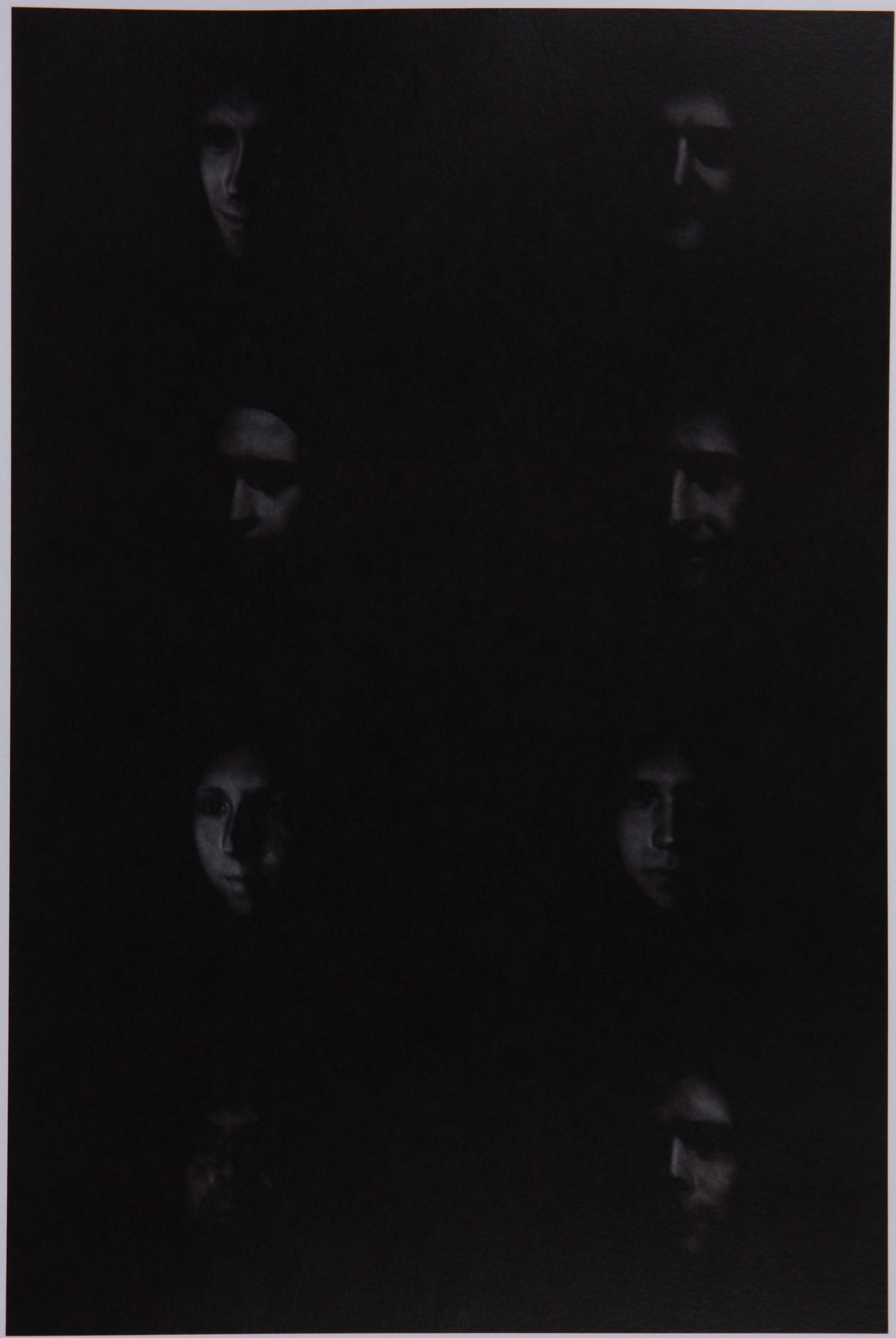
Appendix A: Example Subjective Test Results

A.1. Example test sequence and result from lighting / pose experiment detailed in chapter 1

Subject Test ID: 10
Training Images:
(In order of presentation to the subject)

Test Images:

(In order of presentation to the subject)



Results:

Test Number 10	20 Images Displayed from set of 20	Images shown for 5 seconds	Set 2 Used

Summary:

Correct Identifications:	Same Lighting: 8, Different Lighting: 4,	Total: 12
False Positives:		Total: 4
True Negatives:		Total: 16
False Negatives:	Same Lighting: 2, Different Lighting: 6,	Total: 8

Raw Results:

Images presented (in order):

317-0002	271-0001	144-0004	117-0005	139-0004	250-0001
131-0004	110-0002	324-0004	120-0001	114-0004	272-0005
135-0005	122-0002	141-0002	113-0001	147-0005	138-0001
247-0005	125-0002				

Stimulus	Image	Response Time (s)	Response	Lighting
0	319-0001	17.32	Unrecognised correctly	N/A
1	111-0004	21.70	Recognised incorrectly	N/A
2	110-0004	3.98	Unrecognised incorrectly	Different
3	232-0004	1.49	Recognised incorrectly	N/A
4	123-0005	3.42	Unrecognised correctly	N/A
5	112-0002	6.82	Unrecognised correctly	N/A
6	143-0002	6.25	Recognised incorrectly	N/A
7	269-0002	12.54	Unrecognised correctly	N/A
8	141-0004	1.15	Recognised correctly	Different
9	317-0005	1.65	Recognised correctly	Same
10	131-0002	1.54	Recognised correctly	Different
11	132-0001	3.14	Unrecognised correctly	N/A
12	138-0005	3.26	Recognised correctly	Different
13	122-0004	6.49	Unrecognised incorrectly	Different
14	125-0005	3.25	Recognised correctly	Same
15	325-0004	2.99	Unrecognised correctly	N/A
16	113-0004	1.15	Recognised correctly	Same
17	130-0001	4.70	Unrecognised correctly	N/A
18	135-0001	2.11	Unrecognised incorrectly	Different
19	129-0002	2.73	Unrecognised correctly	N/A

20	121-0004	3.85	Unrecognised correctly	N/A
21	120-0005	1.13	Unrecognised incorrectly	Different
22	139-0001	2.07	Recognised correctly	Same
23	115-0001	2.70	Unrecognised correctly	N/A
24	140-0004	3.45	Unrecognised correctly	N/A
25	146-0002	0.85	Unrecognised correctly	N/A
26	277-0005	7.22	Unrecognised correctly	N/A
27	136-0005	4.77	Unrecognised correctly	N/A
28	250-0004	2.97	Recognised correctly	Same
29	324-0001	5.64	Unrecognised incorrectly	Same
30	117-0001	2.03	Recognised correctly	Different
31	144-0002	5.59	Unrecognised incorrectly	Different
32	272-0002	2.23	Recognised correctly	Same
33	118-0005	6.65	Unrecognised correctly	N/A
34	114-0002	7.77	Unrecognised incorrectly	Different
35	329-0002	5.14	Unrecognised correctly	N/A
36	262-0004	1.84	Recognised incorrectly	N/A
37	247-0002	2.33	Recognised correctly	Same
38	271-0004	1.30	Recognised correctly	Same
39	147-0002	6.91	Unrecognised incorrectly	Same

A.2. Example test sequence and result from Chapter 4 Experiment 2

Subject Test ID: 41, **Training Images** *(In order of presentation to the subject)*:

Test Images *(In order of presentation to the subject)*:

Results:

Test Number			
41	12 Images Displayed from set of 12	Images shown for 5 seconds	Set 1 Used

Summary:

ORIGINAL	Control	Shade	Shape	Both
TruePos	0	0	0	0
FalsePos	0	0	0	0
TrueNeg	1	1	1	1
FalseNeg	1	1	1	1
RECONSTRUCTED	Control	Shade	Shape	Both
TruePos	0	0	0	0
FalsePos	0	1	0	0
TrueNeg	1	0	1	1
FalseNeg	1	1	1	1
MEAN	Control	Shade	Shape	Both
TruePos	0	1	0	0
FalsePos	0	0	0	1
TrueNeg	1	1	1	0
FalseNeg	1	0	1	1

Raw Results:

Images presented (in order):

16-3-1-1 3-3-2-2 1-3-2-1 20-3-2-2 9-3-2-1 13-3-1-1
15-3-1-2 22-3-1-2 17-3-1-2 7-3-2-2 10-3-1-1 5-3-2-1

Stimulus	Image	Response Time (s)	Response	Condition	Model Type
0	11-2-2-2	0.27	Recognised incorrectly	Shape + Shading	Mean
1	17-3-1-2	0.10	Unrecognised incorrectly	Shape	Original
2	22-2-1-2	0.04	Unrecognised incorrectly	Shape	Mean
3	9-2-2-1	0.02	Recognised correctly	Shading	Mean
4	2-1-1-1	0.09	Unrecognised correctly	Control	Estimation
5	12-1-2-1	0.11	Recognised incorrectly	Shading	Estimation
6	4-1-1-2	0.06	Unrecognised correctly	Shape	Estimation
7	18-2-2-1	0.08	Unrecognised correctly	Shading	Mean
8	14-1-2-2	0.07	Unrecognised correctly	Shape + Shading	Estimation
9	23-3-2-1	0.07	Unrecognised correctly	Shading	Original
10	15-1-1-2	0.02	Unrecognised incorrectly	Shape	Estimation
11	20-2-2-2	0.03	Unrecognised incorrectly	Shape + Shading	Mean
12	10-2-1-1	0.07	Unrecognised incorrectly	Control	Mean
13	13-1-1-1	0.00	Unrecognised incorrectly	Control	Estimation
14	5-3-2-1	0.64	Unrecognised incorrectly	Shading	Original

15	3-1-2-2	0.13	Unrecognised incorrectly	Shape + Shading	Estimation
16	1-1-2-1	0.10	Unrecognised incorrectly	Shading	Estimation
17	16-3-1-1	0.02	Unrecognised incorrectly	Control	Original
18	19-2-1-1	0.04	Unrecognised correctly	Control	Mean
19	6-3-1-1	0.00	Unrecognised correctly	Control	Original
20	8-3-1-2	1.16	Unrecognised correctly	Shape	Original
21	21-3-2-2	0.11	Unrecognised correctly	Shape + Shading	Original
22	7-3-2-2	0.05	Unrecognised incorrectly	Shape + Shading	Original
23	24-2-1-2	0.08	Unrecognised correctly	Shape	Mean

A.3. Example test sequence and result from Chapter 4 Experiment 3

Subject Test ID: 7

Training Images *(In order of presentation to the subject)*:

Test Images *(In order of presentation to the subject)*:

Results:

Test Number 7 | 12 Images Displayed from set of 12 | Images shown for 5 seconds | Set 3 Used

Summary:

Analysis	Original	Reconstruction	Total
True Positive	1	1	2
False Positive	1	1	2
True Negative	5	5	10
False Negative	5	5	6

Raw Results:

Images presented (in order):

8 14 24 11 18 21 2 6 12 19 23 4

Stimulus	Image	Response Time (s)	Response	Model Type
0	19-2	1.81	Unrecognised incorrectly	Original
1	2-1	0.61	Recognised correctly	Estimation
2	6-1	0.95	Unrecognised incorrectly	Estimation
3	5-1	0.68	Unrecognised correctly	Estimation
4	18-2	0.79	Unrecognised incorrectly	Original
5	20-2	1.65	Unrecognised correctly	Original
6	1-1	0.93	Recognised incorrectly	Estimation
7	13-1	1.74	Unrecognised correctly	Estimation
8	23-1	2.20	Unrecognised incorrectly	Estimation
9	9-2	1.00	Recognised incorrectly	Original
10	7-2	1.92	Unrecognised correctly	Original
11	8-2	2.21	Unrecognised incorrectly	Original
12	24-2	0.78	Unrecognised incorrectly	Original
13	17-2	0.80	Unrecognised correctly	Original
14	21-2	1.25	Recognised correctly	Original
15	15-1	0.93	Unrecognised correctly	Estimation
16	4-1	0.71	Unrecognised incorrectly	Estimation
17	22-2	0.75	Unrecognised correctly	Original
18	12-1	0.82	Unrecognised incorrectly	Estimation
19	11-2	0.94	Unrecognised incorrectly	Original
20	10-2	0.79	Unrecognised correctly	Original
21	14-1	0.76	Unrecognised incorrectly	Estimation
22	3-1	0.68	Unrecognised correctly	Estimation
23	16-1	0.59	Unrecognised correctly	Estimation

Appendix B: Test set depth estimation results

Sample	*Original*	*Estimation (β=4500)*	*Mean*	*MSE (Estimation)*	*MSE (Mean)*
1				358	514
2				252	1651
3				181	1035
4				198	1015
5				97	320

Sample	*Original*	*Estimation (β=4500)*	*Mean*	*MSE (Estimation)*	*MSE (Mean)*
6				221	427
7				358	1261
8				105	293
9				1932	3502
10				115	260
11				361	614

Sample	*Original*	*Estimation (β=4500)*	*Mean*	*MSE (Estimation)*	*MSE (Mean)*
12				488	1464
13				888	1544
14				345	929
15				370	571
16				834	387
17				210	582

Sample	*Original*	*Estimation (β=4500)*	*Mean*	*MSE (Estimation)*	*MSE (Mean)*
18				218	656
19				503	409
20				369	440
21				832	2518
22				213	845
23				890	1025

Sample	*Original*	*Estimation (β=4500)*	*Mean*	*MSE (Estimation)*	*MSE (Mean)*
24				302	693
25				735	1551
26				487	2318
27				225	338
28				1288	1254
29				394	641

Sample	*Original*	*Estimation (β=4500)*	*Mean*	*MSE (Estimation)*	*MSE (Mean)*
30				154	381
31				1129	1517
32				1268	1871
33				244	2219
34				311	794
35				488	961

Sample	*Original*	*Estimation (β=4500)*	*Mean*	*MSE (Estimation)*	*MSE (Mean)*
36				665	1304
37				537	669
38				861	997
39				586	1198
40				106	386

Appendix C: Published Papers

The first of these publications has been included to provide an insight into the starting point of the thesis as discussed in the introduction. Whilst the central work has been the estimation of depth from images, this has been a consequence of initial work into the coding (compression) of 3d models reported in the paper. An extension of the work discussed was the idea that as the estimation portion of a coding scheme (typically used in a predictive scheme prior to the encoding of an error signal) improves, the error signal requiring entropy coding tends to zero – and hence the system tends towards zero-bit coding – i.e., pure estimation.

Paper 1: J Hyde and J Robinson – Coding 3d facial models for mugshot applications, *Proceedings of Video, Vision and Graphics 2003, pp127-135, Bath, July 2003*

Paper 2: J Hyde and J Robinson – Estimation of face depths by conditional densities, *British Machine Vision Conference 2005, Vol. 2, pp609-618, Oxford, September 2005*

Video, Vision and Graphics (2003)

Coding 3D facial models for mugshot applications

John Robinson and Justen Hyde

Department of Electronics, University of York, York, United Kingdom

Abstract

Three-dimensional information about a human face may have some correlation with the colour information present in its flat texture image. In order to maximise the available information for human identification of faces, a variety of coding schemes based on Binary Tree Predictive Coding 5 (BTPC5) are proposed and evaluated against similar schemes applied to the JPEG coder. The results of these schemes are presented quantatively with some discussion of the subjective results.

1. Introduction

Three-dimensional mugshots – that is, facial images represented with colour and depth – may be more reliable for identity checks by humans than conventional 2D photos which lack much of the shape information used by humans in face recognition[1]. The extent to which implied depth information is included in sketch drawings by shading, and the common use of three-quarters views in portraiture supports this theory. Work aimed towards the computerised generation of cartoons of human faces has also shown a reliance on the preservation of shape from shading information[2,3]. With a 3D head model and appropriate rendering and manipulation tools, a user can obtain full face, profile and three-quarters views or any other angle desired. Arguably, this addition of explicit depth information provides cues to expose impersonation. We are interested in assessing the effectiveness of 3D mugshots, and in developing efficient methods for their display (e.g., animation of head movement requiring no direct user control of view – essentially providing a video of a moving head). Other work has attempted to reconstruct three dimensional data from two dimensional images[4], which can be considered (albeit indirectly) a method of coding. This approach assumes that no explicit three dimensional data is available at the source. In this paper, however, we report a method for coding such 3D pictures efficiently. We are aiming to represent the colour and geometry of the face in very few bits, so that such information can be encoded on printed media. Ultimately we seek to represent all the information for the generation of pictures like those in figure 1 in a two-dimensional glyph code.

We describe our approach to this problem together with promising initial results.

Figure 1: *Multiple views of a face can be obtained with no extra transmission of information*

2. Approach

2.1 Simplification of model – 3d mesh to depthmap

Coding a full three dimensional model of a head involves the manipulation of a huge amount of data. However, the face only occupies one side of the head and contains most of the recognition information. This allows us to treat 3D face coding as a 2.5D problem. The three dimensional model of the face can be reconstructed from depth information only – i.e., by placing the camera directly in front of the model, and taking a snapshot of the distance back from the camera plane at each point on the model, a "depthmap" is created which can be later used to recover a 3D model of the captured projection. This results in a four component image of a face.

Figure 2: *Colour and depth maps of the full face projection of a 3D facial model*

2.2 Coding

The 2.5D image has four components - red, green, blue and depth. Our approach is to exploit dependencies between the channels to reduce the amount of data to be coded. Almost all picture coding schemes apply a colour transform which serves to reduce the correlation between components. Most often this is simply a conversion from RGB to luminance and two chrominance channels, and usually, after transformation, the components are coded independently. We, however, propose to use "Binary Tree Predictive Coding (BTPC)"[5,6], because it provides two-level of inter-component coding that can be extended naturally from the conventional three colour channels to four channels. In BTPC the components are first transformed with the statistically-optimal Karhunen-Loeve Transform (KLT). The highest-energy component is coded first, then successive components are coded relative to the previous ones. This is possible in BTPC because it is a predictive scheme: a poor prediction in an early component can be modified to a better prediction for later components if they have the same local shape properties. In the experiments below we compare this approach with a JPEG-derived alternative which also uses the KLT for pre-processing but does not include the second stage of dependent component coding.

The KLT can also be applied to an ensemble of head images where the feature vector for decorrelation is not the four-component pel, but the whole image (i.e. a 4xNxM-dimensional vector for images of N rows and M columns). This is similar to the Principal Components Analysis approach that has been widely applied for automatic face detection and recognition[7,8]. We have not adopted a full KLT at the image level because, as yet, our training and test sets are very small, however, we are able to remove some of the correlation between faces, simply by subtracting the mean head image from particular inputs before coding. Our experiments below test the different alternatives for this.

3. Method

3.1 Data capture

The data set used has been captured using a 3-d camera which records both texture data (using a standard digital camera) and geometric information (using a stereo vision system enhanced by light pattern projection). The data was captured by Tom Heseltine of the University of York Computer Science Department. The camera records a three dimensional image of the subject in the same way as a normal camera records a two dimensional image (i.e., the subject stands before the camera for a second whilst a photograph is taken). Whilst this is less accurate than scanning the person, it is also much less time consuming and intrusive.

3.2 Coding schemes applied

Figures 3 and 4 show the four different types of coding we have investigated. In each case the KLT is used to preprocess the colour and depth information and in each case BTPC is used as the final coding engine. The KLT converts four components of colour and depth into four channels, the most significant of which is quantised most accurately, and the least significant of which is quantised the most coarsely. The KLT is integrated into BTPC, but has been adapted for the four-channel situation. We have also used similar structures with JPEG as the coder, and in these cases have used a separate KLT stage.

3.3 The four coder configurations

To test whether the inclusion of depth information in the KLT process significantly alters the distribution of colour information in the transformed channels, we performed tests where an unmodified depth map was coded alongside KLT transformed colour components, where the least energy colour component is subject to the most extreme quantization. Test coders 1 and 3 use this method of depth information inclusion. Test coders 2 and 4 treat the depth channel no differently to the colour channels.

Features such as eyes are regions of high interest, but their colour information may be swamped by the less significant but more common flesh tones. To retain this information, difference from mean depth and colour images are

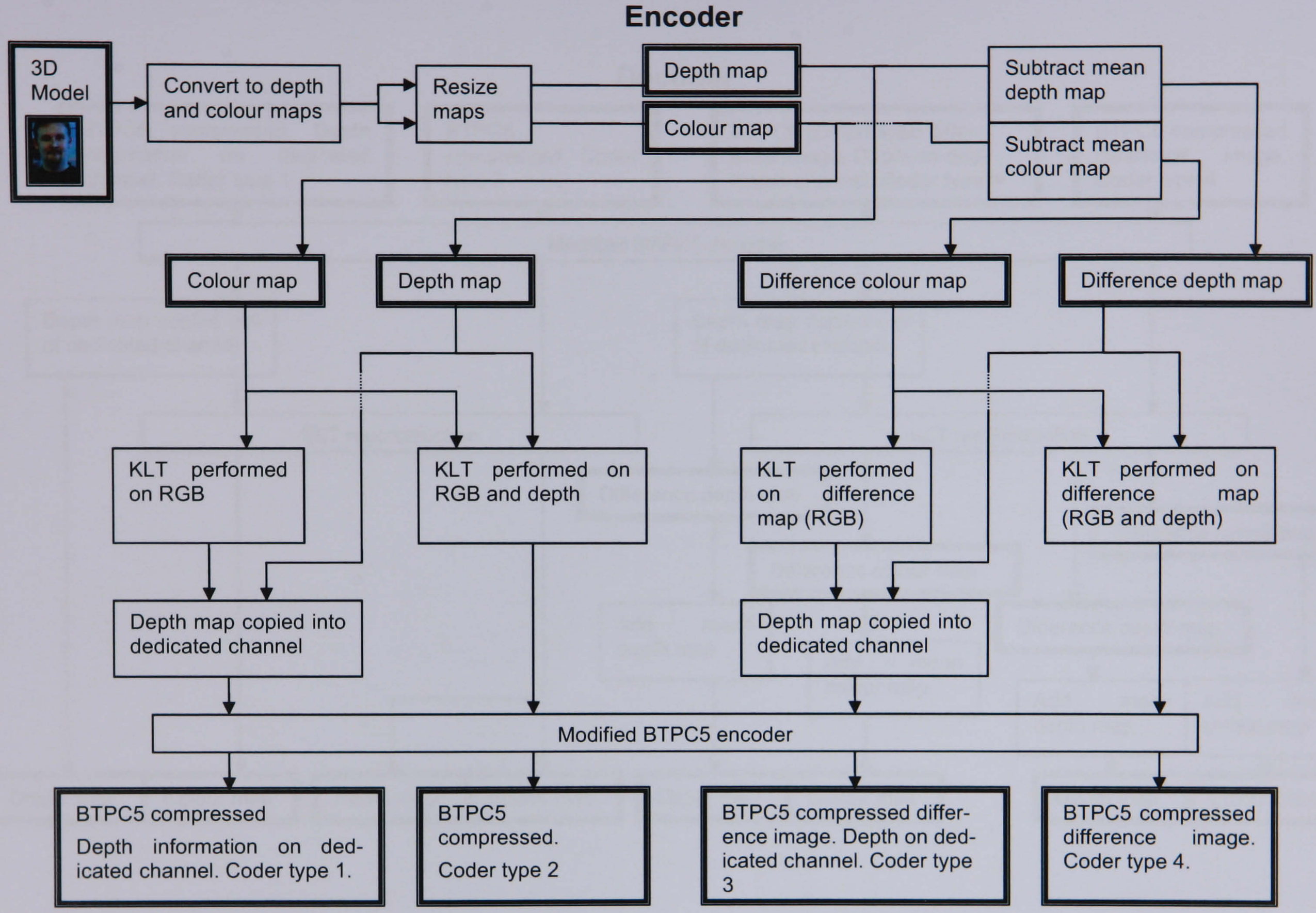


Figure 3: *Schematic diagram for four encoder alternatives*

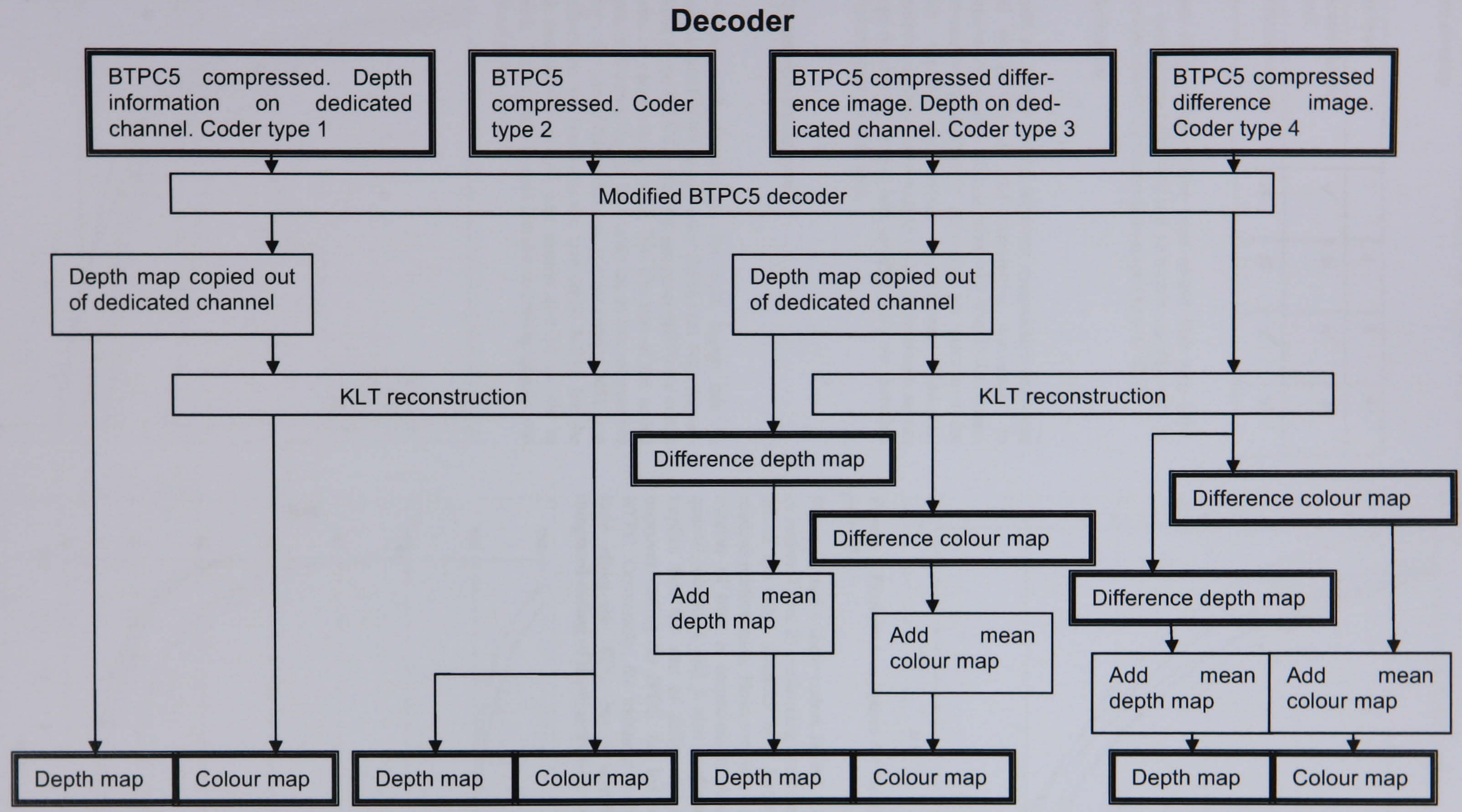


Figure 4: *Schematic diagram for four decoder alternatives*

generated and coded rather than the actual images. Coders 3 and 4 use difference images, whilst 1 and 2 code the picture directly.

Coder type	1	2	3	4
Dedicated depth channel	✓	✗	✓	✗
Difference images	✗	✗	✓	✓

Images of three sizes have been tested: Full size (250, 168), quarter size (125, 84) and sixteenth size (62, 42). In the results section we concentrate on the first of these.

4. Evaluation

As with all image coding schemes, quantitative measures (MSE or PSNR) do not necessarily correspond to subjective test results. This is especially true if the system is envisaged as ultimately providing an extra tool for human identity recognition applications. However, quantitative analysis allows us to make comparisons across a large number of images and qualities and we therefore present summary results first.

4.1 Quantative Evaluation

The success of compression for each image can be evaluated based mean square error (MSE) (or Peak Signal-to-Noise Ratio (PSNR)) and bits per pixel (BPP) (or coded file size or compression ratio). The file size of the coded images, and hence the bit rate, must be as low as possible to allow the use of the system on printed media. MSE can be calculated by summing the individual MSEs for the depth and texture images, and should also be as low as possible. The source images for the following graphs can be found in section 4.2.

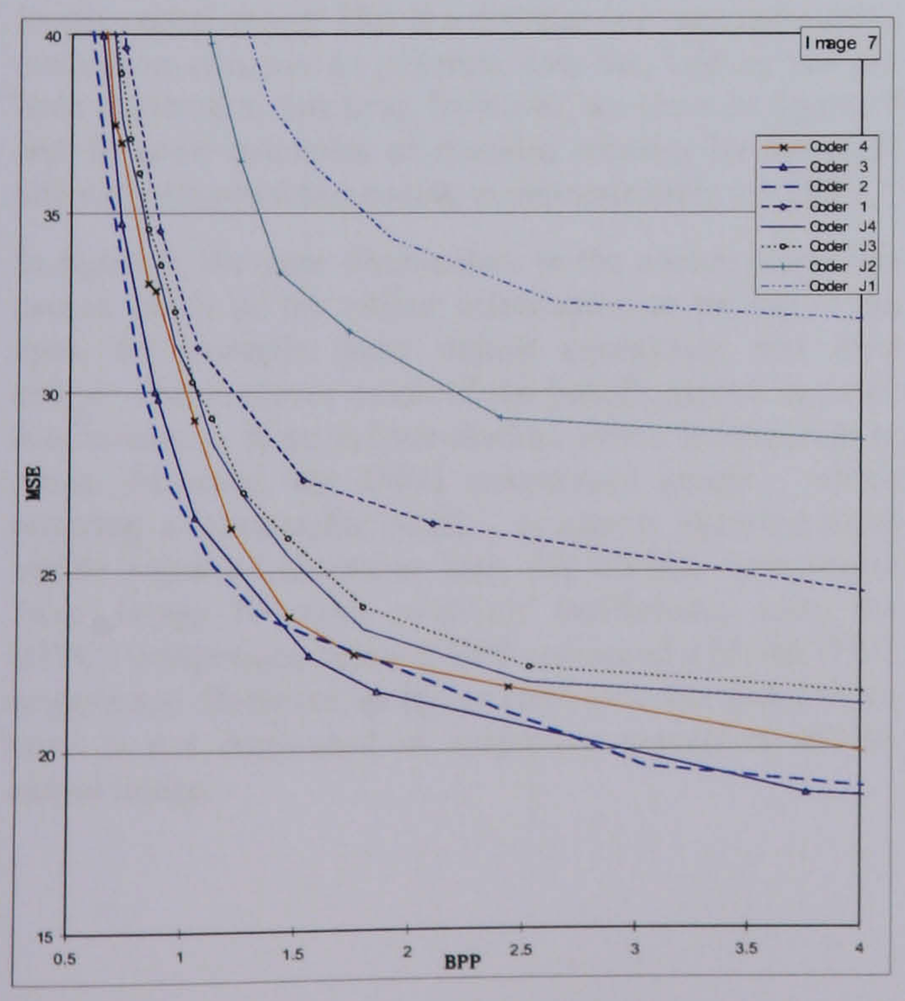


Figure 5 *Performance figures for the eight coders on input face 7.*

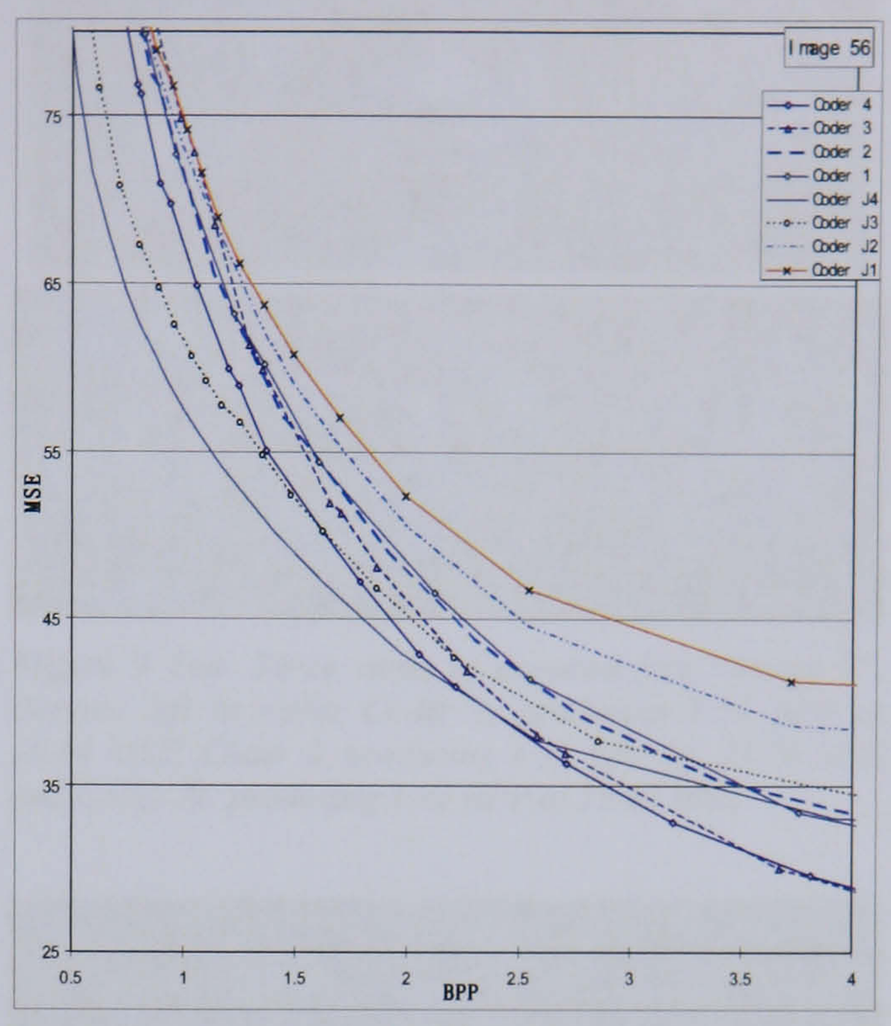


Figure 6 *Performance figures for the eight coders on input face 56.*

Figures 5 and 6 show coders J1 and J2 (the JPEG versions of coders 1 and 2) performing consistently poorly, though all of the other schemes have large variation on their relative performance. However, when the mean of a large number of tests is inspected, it becomes clear that, in general, coders 1 and 3 also perform very poorly. This implies that the use of difference images appreciably improves coding for JPEG, but has less significance for BTPC. Conversely, the independent coding of depth has little effect for JPEG, but is significantly worse than integrated coding of depth and colour in BTPC.

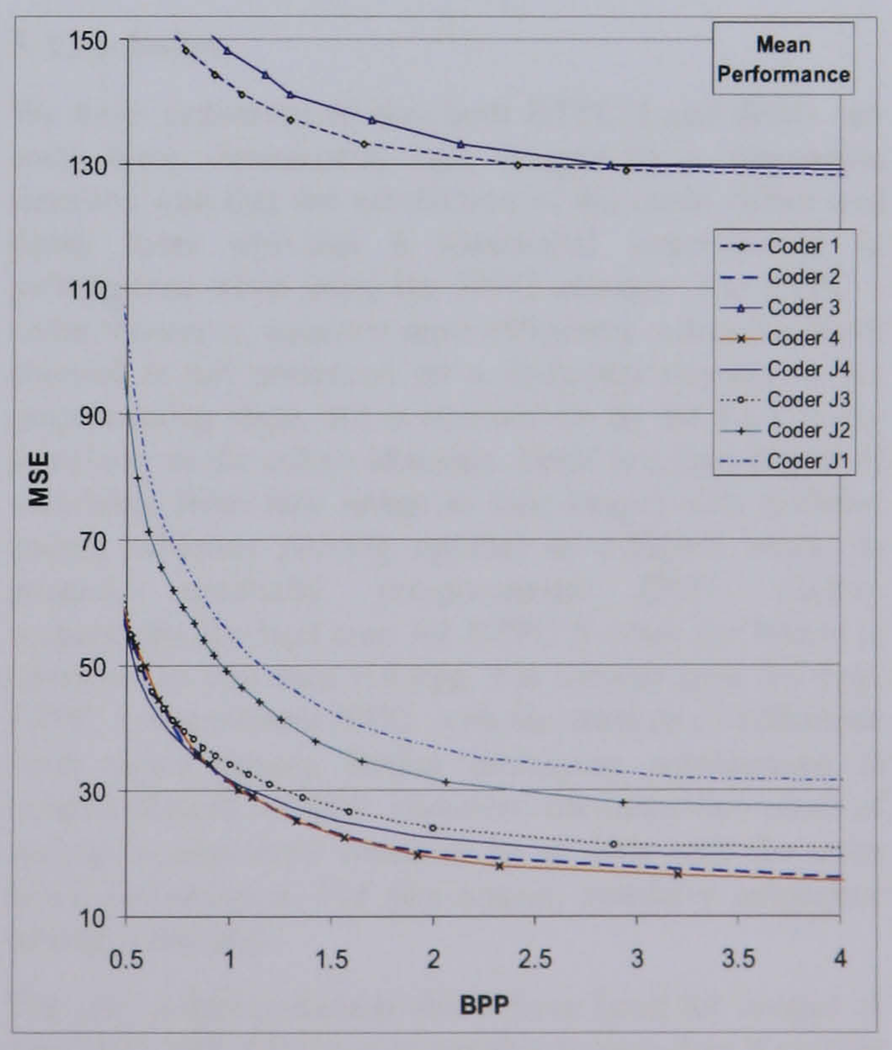


Figure 7. *Mean performance over entire test set.*

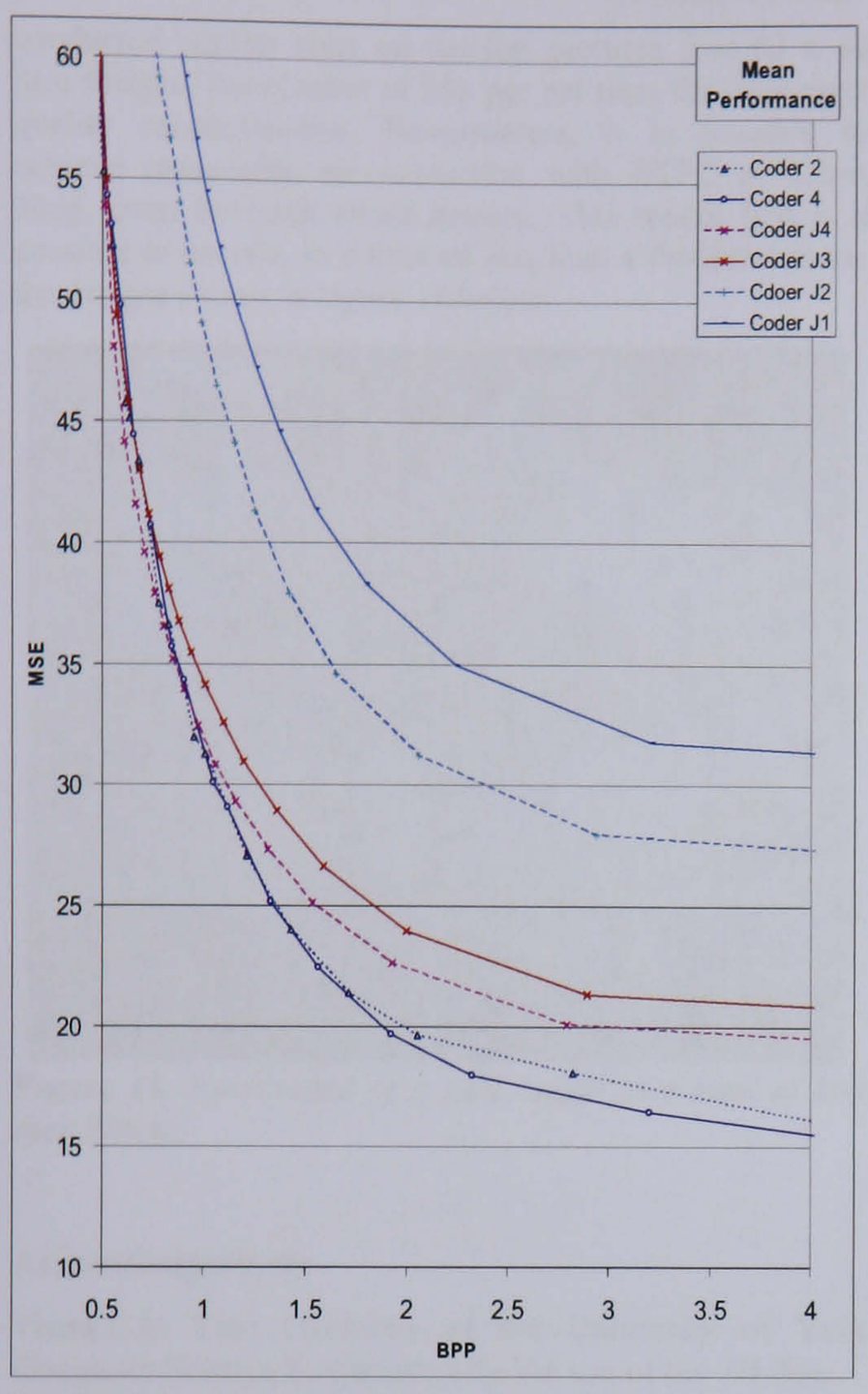


Figure 8. *Mean performance over entire test set. Expanded view of MSEs below 60.*

4.2 Subjective Evaluation

Ultimately, the system must be evaluated against file size and human recognisability rather than MSE, which is itself a poor estimator of the perceptual degradation present in a lossily coded image. This is a difficult and time consuming evaluation criterion to generate data for, and so has not been explored at this time. However, we show in figures 9 and 10 some examples of decoded pictures for the eight alternative approaches coding to approximately 1.5 BPP.

In figure 9, the poor illumination in the source image has caused much of the colour information to be lost – the eyes, for example, have almost completely lost their colour. This removes much of the benefit gained in mean subtraction, as there is little distinct colour information to retain. However, the JPEG compressed image – whilst retaining a comparable MSE – is clearly showing more visible signs of distortion then the BTPC5 compressed faces. Image 56 codes relatively inefficiently using the BTPC5 compressor below 2 BPP compared with the JPEG compressor. However, as figure 10 shows, this quantitative error is not duplicated in subjective perception of the output image.

Figure 9 *Top: Three views of uncoded face "Image 7". Bottom, left to right: Coder 3, producing 1.53 BPP at 28.84 MSE, Coder 2, producing 1.35 BPP at 23.76 MSE and Coder J4, producing 1.47 BPP at 25.47 MSE*

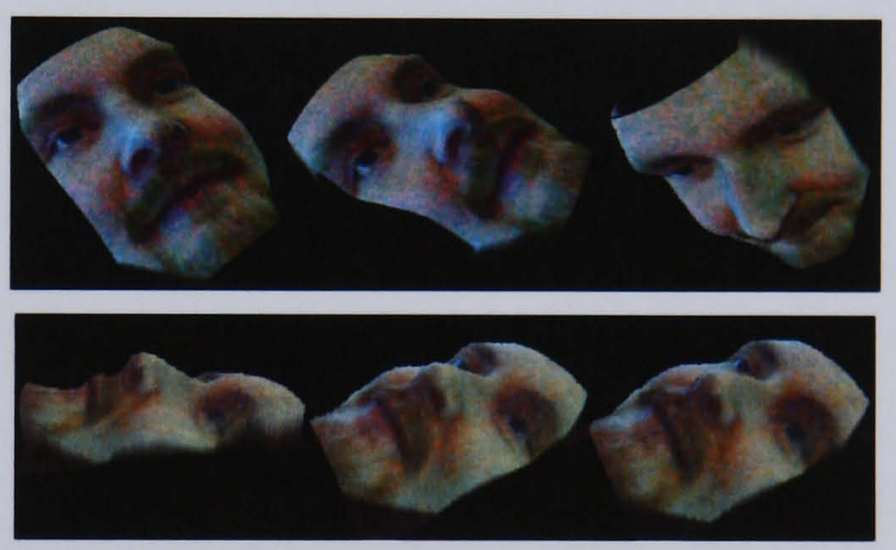


Figure 10 *Top: Three views of uncoded face "Image 56". Bottom, left to right, Coder J4 (1.40 BPP,50.07 MSE) , J3 (1.48 BPP, 52.37 MSE) and 4 (1.61 BPP, 50.25 MSE).*

5. Conclusion

We have demonstrated that both BTPC 5 and JPEG can code three dimensional face images to a reasonable standard, and that the subtraction of the mean colour and depth faces provides a substantial improvement in performance when using the JPEG encoder. The BTPC 5 coder, however, operates most efficiently when the depth channel is not preserved on a dedicated channel in the preprocessing stage, but is operated on by the KLT in the same way as the colour channels. There is a high degree of variability from face image to face image, with different coding schemes proving optimal in different cases. In general, optimally pre-processed JPEG slightly outperforms the best case for BTPC 5 when the bitrate is restricted to less than 0.9 bpp. For bitrates over 0.9 bpp, BTPC 5 outperforms JPEG, with the creation of difference from mean images further increasing performance at bitrates of over 1.7 BPP. However, the subjective result of coding is sometimes found to be at odds with the strict MAE performance. For this reason, extensive subjective testing is required.

The test results presented above have been for images of size 250 x 168. Clearly it is possible to save data if smaller images are adequate for recognition. We have therefore

conducted similar tests on smaller pictures. For 62 x 42 face images, the number of bits per pel rises for equivalent quality reconstruction. Nevertheless, it is possible to achieve reasonable reconstruction with BTPC at below 3bpp, even for such small images. This means that is it possible to encode, in a total of less than a thousand bytes, the images shown in figure 11 below.

Figure 11. *Four views of a face coded in a total of less than 1kbyte.*

Acknowledgements

Thanks to Tom Heseltine of the University of York Computer Science Department for the use of the 3D data.

References

[1] V Bruce and A Young, In the Eye of the Beholder, *Oxford University Press.* 1998

[2] D E Pearson and J Robinson – "Visual communication at very low data rates", *Proceedings of the IEEE*, **73** (4), pp 795-812, April 1985

[3] D E Pearson, E Hanna, K Martinez – "Computer-generated cartoons", in *Images and Understanding,* 1990.

[4] T Vetter – "Synthesis of novel views from a single face image", *International Journal of Computer Vision 1998*

[5] J A Robinson, "Efficient General-Purpose Image Compression with Binary Tree Predictive Coding", *IEEE Trans on Image Processing,* **6** (4), April 1997, pp 601-607.

[6] J A Robinson, "Exploiting Local Colour Dependencies in Binary Tree Predictive Coding", *Proceedings of Visual Information Engineering*, VIE 2003, June 2003.

[7] M Kirby, L Sirovich "Application of the Karhunen-Loeve Procedure for the characterization of human faces," *IEEE Trans on Pattern Analysis and Machine Intelligence,* **12** (1), pp 103-108, 1990.

[8] M Turk, A Pentland, "Eigenfaces for recognition", *Journal of Cognitive Neuroscience,* **3** (1), pp 71-86, 1991.

Estimation of face depths by conditional densities

John A Robinson, Justen R Hyde
Department of Electronics, University of York, YO10 5DD
jar11@ohm.york.ac.uk

Abstract

The expected value of missing data in a sample taken from a multivariate normal probability distribution is the mean of the conditional distribution of the missing dimensions given the known dimensions. We explain the derivation of this result, demonstrate its application to face image processing, then use it in a new method for recovering shape from image data. The context of our work is the use of 3D facial models to aid in recognition of human faces by humans. We explain the requirement for such models and review the practical possibilities for encoding depth information alongside photographs in identity documents like passports. The best alternative is to derive depths automatically from the photos, as this requires no side information. We show experimentally that conditional density estimation provides accurate face depth recovery, without recourse to explicit modelling of surface shape.

1 Introduction

In this paper we are principally concerned with the analysis of images of human faces. These may either be 2D images, where an array of pixels represents greyscale, or 2.5D (or, informally, 3D) images, where a second array represents depths. In common with other appearance-based approaches (e.g. [1]-[4]), we stack all the measurements (greyscales and depths) for an image into an n-dimensional column vector, which we consider as a sample point in a multidimensional space. Unlike most other appearance-based approaches, we do not assume that "face space" is a low-dimensional subspace of "image space". Rather, we model faces as an n-dimensional normal distribution, with the parameters derived from training data. With this representation, conditional distributions can be used to estimate any number of missing measurements in a sample. Section 2 explains the theory, highlights some properties of the estimation, then shows how it can be applied to image reconstruction from partial data. All examples are faces, but the generalization to other domains is briefly considered.

In section 3 we introduce the application with which we are concerned: the recovery of depth information from face photos.

Because the number of samples available for training is less than n and therefore far too low for the scatter matrix of training samples to characterize the covariance matrix, we use a regularized covariance estimation method [5]. This mixes the scatter matrix and the identity matrix according to a reconstruction criterion, as described in section 4, which also contains the results of experiments to find the appropriate mixing parameter.

Section 5 reports experiments on the recovery of depth data from photos of faces.

2 Conditional distributions for data estimation

2.1 Theory

Consider the random vector $\boldsymbol{X}$, distributed as $N_n(\mu,\Sigma)$. Suppose we have a particular sample from the distribution $\boldsymbol{P}$, which we call the "probe", in which some of the measurements are

known and some are not. A binary vector $\boldsymbol{M}$ (or "mask") may be used to indicate which dimensions in $\boldsymbol{P}$ are known values, and a permutation matrix $\boldsymbol{R}$ can then be defined which will reorder $\boldsymbol{M}$ into a column of q zeroes followed by p $(= n - q)$ ones.

When $\boldsymbol{R}$ is applied to $\boldsymbol{P}$, it moves all the unknown values to the top:

$$P_{perm} = RP = \begin{bmatrix} P_1 \\ \hline P_2 \end{bmatrix} \tag{1}$$

where P_1 and P_2 are column vectors of dimensionality q and p respectively (P_1's values are undefined). Similarly,

$$X_{perm} = RX = \begin{bmatrix} X_1 \\ \hline X_2 \end{bmatrix} \tag{2}$$

We also define $\mu_{perm} = R\mu = \begin{bmatrix} \mu_1 \\ \hline \mu_2 \end{bmatrix}$ and $\Sigma_{perm} = R\Sigma R = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ (3)

I.e. μ_{perm} is the mean vector reordered to match $\boldsymbol{RP}$, the reordered probe vector. Similarly Σ_{perm} is the covariance matrix with rows and columns appropriately reordered.

We now follow [6] (page 170) and define a matrix A with submatrices and dimensions as shown:

$$\underset{(n\times n)}{A} = \begin{bmatrix} \underset{(q\times q)}{I} & \underset{(q\times p)}{-\Sigma_{12}\Sigma_{22}^{-1}} \\ \hline \underset{(p\times q)}{0} & \underset{(p\times p)}{I} \end{bmatrix} \tag{4}$$

Now the random vector

$$A(X_{perm} - \mu_{perm}) = A\begin{bmatrix} X_1 - \mu_1 \\ \hline X_2 - \mu_2 \end{bmatrix} = \begin{bmatrix} X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ \hline X_2 - \mu_2 \end{bmatrix} \tag{5}$$

is a linear transformation of the normally-distributed random vector X_{perm} and so itself is normally distributed with mean $E\left[A(X_{perm} - \mu_{perm})\right] = AE\left[(X_{perm} - \mu_{perm})\right] = 0$ and covariance matrix $A\Sigma_{perm}A^T$ where the T denotes transpose.

Using the identities $\Sigma_{22}^T = \Sigma_{22}, \Sigma_{12}^T = \Sigma_{21}$ (because of the symmetry of covariance matrices), we can calculate equation (6):

$$A\Sigma_{perm}A^T = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ \hline 0 & I \end{bmatrix}\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{bmatrix}\begin{bmatrix} I & 0 \\ \hline (-\Sigma_{12}\Sigma_{22}^{-1})^T & I \end{bmatrix} = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ \hline 0 & \Sigma_{22} \end{bmatrix}$$

Equation (6) above is the motivation for choosing A as we did: when the covariance matrix is calculated, the top right and bottom left corners turn out to be 0 submatrices, meaning that $X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ and $X_2 - \mu_2$ have zero covariance and are therefore independent. We can therefore consider the quantity $X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ as a distinct q x q multivariate normal distribution. When X_2 takes the value P_2, the random variable becomes $X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(P_2 - \mu_2)$. As shown above, $E\left[A(X_{perm} - \mu_{perm})\right] = AE\left[(X_{perm} - \mu_{perm})\right] = 0$, so $E[X_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(P_2 - \mu_2)] = 0$. But $\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(P_2 - \mu_2)$ is a constant, so the mean of X_1, i.e. the expected value of the missing data we want to fill in, is given by:

$$E[X_1] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(P_2 - \mu_2) \quad (7)$$

We therefore have a direct method for estimating P_1 from P_2. All that then remains is to apply $\boldsymbol{R}^{-1}$ to P_{perm} to recover the full image corresponding to the probe.

The above derivation is indirect. It is also possible to construct a proof that uses the densities directly ([6] pp 217-218). Also note that, as well as the mean, the covariance matrix of the estimated data has been derived as $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. This could be used to explore the principal components of variation, i.e. the modes in which the real values are likely to differ from the estimate. We do not pursue use of the conditional density covariance further here.

2.2 Geometric interpretation of the estimation.

Figure 1 illustrates the estimation of a missing value in two dimensions. The geometric interpretations it suggests apply in *n*-dimensional space, namely: (a) The use of the expectation of the conditional density is equivalent to finding the point where the known subspace is tangent to an equivariance contour, or, equivalently, finding the minimum Mahalanobis distance from the known subspace to the class mean. (b) Given a particular known subspace (i.e. a particular distribution of pixels in the probe), the derivation of the missing values is a linear transformation. (c) It is important to have a full *n*-dimensional covariance matrix for the distribution. Were the method to be applied in a principal component subspace, for example, the missing values would be estimated by reflecting in the principal components. This will lead to large errors when the retained principal components are near parallel to missing value axes.

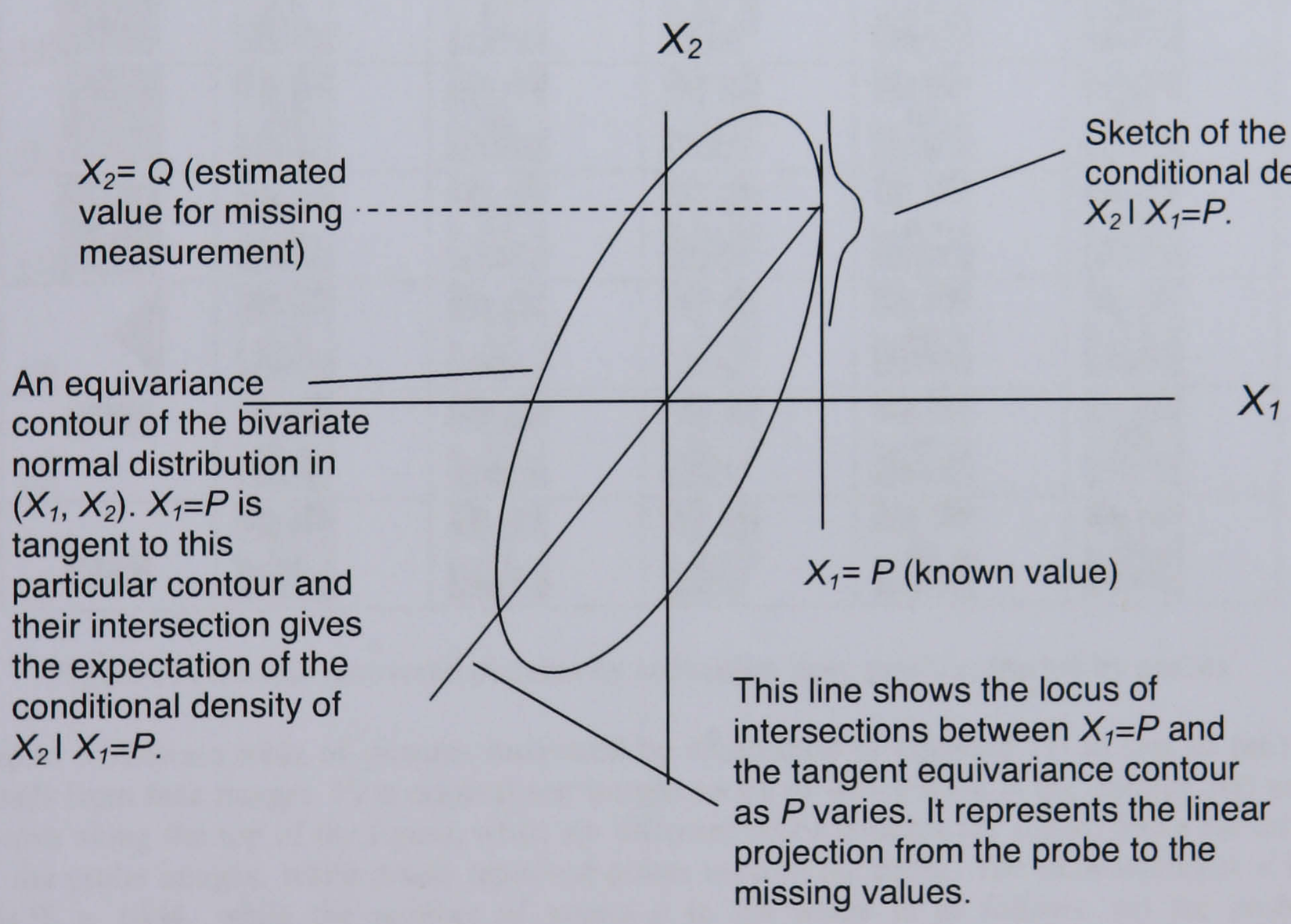


Figure 1: Illustration of estimation for a 2D distribution

2.3 Applying conditional estimation to the recovery of missing data in face images

Although the focus of this paper is the recovery of depths from greylevels, we briefly illustrate the use of conditional density estimation for filling in missing or damaged data in images.

Figure 2 shows the mean of a training set of 2325 38x38 face images together with the individual pixel variances (i.e. the diagonal of the scatter matrix). The faces are modelled as a multivariate normal distribution using the sample mean as maximum likelihood estimate of the distribution mean, and with the covariance matrix estimated as described in Section 4 below (and, more fully, in [5]).

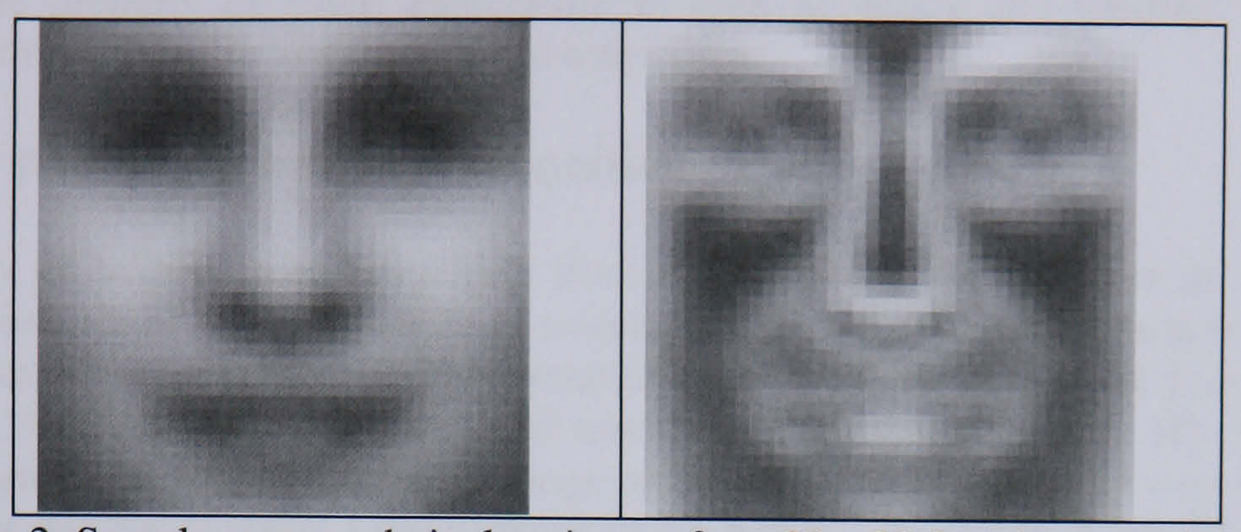


Figure 2: Sample mean and pixel variances for a 38 x 38 face-image training set

Figure 3: Example recovered pictures by estimation from pixels extracted by probes

Figure 3 shows a table of pictures recovered by application of equation (7) to sets of probe pixels from face images. Five original test images (none of which were in the training set) are shown along the top of the figure, while six different probe patterns are shown down the left. In the probe images, white pixels represent points used in the probe. The dimensionality n is 38x38 = 1444, while the number of points p in the probe is as follows: (a) the probe subsamples the image at regular intervals, $p = 100 = n/14$, (b) the probe subsamples the image at coarse regular intervals, $p = 49 = n/29$, (c) the probe is a fovea-like logarithmic spiral, $p = 104 = n/14$, (d) the probe consists of all the face except the right-hand triangle, $p = 1083 = 3n/4$, (e), the probe consists of the whole face except the eyes, $p = 1015 = 7n/10$, (f) the probe consists of the whole face except the mouth region, $p = 1015 = 7n/10$. Note that the known data in every original probe is taken directly from a test image and is not filtered.

Figure 3 shows that recovery is possible no matter what the probe. Each additional pixel in the probe provides additional data about how the recovered picture should diverge from the

mean, but the divergence is conservative. For example, in row (b), the recovered faces are all plausible though blurred, and reflect the lighting and pose of the original. So far as identity and expression are concerned, however, they tend towards the mean, as would be expected given the dimensionality requirements for identification and expression analysis [7]. Reconstructions improve with increasing probe size, though wherever there are large areas missing, the fill-in again has the property of being appropriate in both geometry and shading and therefore plausible, but neutral so far as expression is concerned.

3 3D photos for human recognition

The accuracy of human face recognition from single photographic images as on passports, driving licenses, etc., is surprisingly low, as demonstrated strikingly by Kemp *et al* [8]. Here, six subjects in the environment of a supermarket were asked to be alert for fraudulent use of credit cards. A test set of volunteers used credit cards with photographic ID. 35 percent of cards with significant variation between image and bearer were accepted, as were 64 percent of cards with similar but different faces to those of the bearer. 14 percent of non-fraudulent cards were rejected.

Part of the reason for this poor performance is believed to be that in many practical cases the faces being recognised are not well known to the observer [9]. This is significant as with only one viewpoint to compare with a testing view, the observer's perception of the ground truth (the person in the photo) is heavily reliant upon the environmental conditions when the 2D image was taken – lighting, pose, and the presence of any transient features (spectacles, hats, etc) at the time. This is not such a significant problem when the face is already known by the observer, as familiar faces are recognised more accurately than unfamiliar faces under most circumstances. Familiarity appears to compensate for distracting influences that could be considered external. Two of these are significant for our purposes:

1. As a face becomes more familiar, the observer becomes more adept at remembering, and therefore using as recognition cues, the "internal" features of the face – that is, features such as eye separation, mouth width, etc., which are not necessarily immediately apparent, but which are highly distinctive between individuals. Less familiar faces are recognised by using more obvious, large scale features, such as hairstyle, facial hair, spectacles, etc, which are readily apparent, highly distinctive when compared to the average face, and also potentially highly misleading as they are both easily changed and do not distinguish between – for example – all people with the same hair colour and style very effectively.
2. As a face becomes familiar, more views of the face are stored and associated with the identity. In this model, transient features (such as hairstyle and facial ornaments) also become less significant, though for different reasons. Here, the presence or absence of, say, spectacles, becomes irrelevant as the observer associates both states to the person in question. The more significant element of this recognition process is the ability of the observer to recognise the subject under varying lighting conditions and in different poses as images of the subject in increasing numbers of different poses and conditions are stored in the associative memory of the observer. Troje and Kersten [10] show the significance of this process effectively. Here, subjects are presented with a selection of familiar faces (close colleagues) from a variety of angles along with an image of their own face (which they are familiar with in the frontal view) in profile. The speed at which they recognised and correctly named the faces was recorded, and showed while other familiar faces could be recognised equally quickly independent of view, the subject's own face was consistently recognised more slowly in profile than in frontal view. Troje concludes that this is because although we are familiar with our own faces in the frontal view (as from our reflection), we rarely see our own face in profile, and so are less primed to recognise it.

It has long been understood that familiar and unfamiliar faces, as subsets of familiar and unfamiliar objects, are recognised so differently that completely different memory processes are used – familiar using semantic memory, and unfamiliar using episodic memory. This makes it difficult to draw conclusions about one type of recognition from the other, as both

have distinct and very different properties [11]. However, we can approach the problem from a different angle: we now have a list of external distractors which make unfamiliar face recognition rates poorer than familiar recognition rates. If these distractors could be in some way removed, then recognition should be less difficult.

A further, very significant, interpretation of the Troje study [10] is that recognition is based on a collection of views, rather than a mental 3D model of a face. If a 3D mental model were used, recognising a familiar face from different angles should have no impact upon performance, whereas recognition based on a collection of views (akin conceptually to searching through a photo album of many views of a candidate before deciding upon recognition) will degrade when a familiar face is presented in a novel orientation. Therefore, the use of a virtual 3D model should provide functionality not available implicitly to the human recognition process.

A natural extension of the study by Troje, and indeed the concept of pose as a problem for unfamiliar face recognition, is to posit that if a face is unfamiliar then pose, lighting and external feature changes will have a disproportionate impact on the observer's identification of the face. Several studies [12,13] have suggested that changes in pose have a detrimental effect upon recognition though Troje and Bulthoff [14] suggest that the training view provided is significant, and the testing view is unimportant – i.e., with a good training view, any change in pose becomes unimportant. [15] demonstrates that changes in illumination direction can hinder identification, while Liu and Chaudhuri [16] support the view that facial structure is determined principally from shading – not stereopsis – which suggests that lighting has a strong effect upon our perception of face shape. Also of note is the finding [17] that perspective distortion caused by camera distance can have a severe effect upon recognition.

In subjective tests we have shown that both lighting and pose changes do indeed have an impact on fully textured 3D facial masks, lighting being greater but pose being far from insignificant. To summarise briefly, test subjects were exposed to a set of twenty training faces, in a random order, posed either facing the camera or looking off at approximately twenty degrees, for five seconds each. Lighting for each image varied between sources to the lower left or upper right of the face. The same twenty faces, all shown with either the alternate pose, lighting or both relative to the training image were then shown to the subjects mixed with twenty novel faces, rendered under the same range of conditions. After testing, a distinctiveness survey was used to remove unusually distinctive faces from the results. As can be seen in table 1, both pose and lighting changes pose a significant problem for recognition. Further, with a compound change in both lighting and pose correct identification drops to less than 60% after normalising for the positive or negative bias of each test subject.

	Correct	False Negative	Normalised Correct	Normalised False Negative
Same Lighting, Different Pose	74.50%	25.50%	77.38%	23.44%
Different Lighting, Same Pose	60.00%	40.00%	72.07%	33.41%
Different Lighting, Different Pose	52.88%	47.13%	59.20%	43.52%

Table 1: Recognition of faces in different poses and lighting conditions

We conclude, therefore, that adjusting pose and/or lighting can significantly degrade the ability of an observer to correctly identify an unfamiliar face. If a full, accurate 3D model of the face in question were provided, this degradation may be ameliorated by mapping the initial 2D image onto the model, re-positioning and re-lighting the model to conform to any given environmental test conditions and re-rendering a new 2D image. For example, a face caught on camera in profile may be more accurately compared with a passport photo by applying the full frontal passport image to the 3D model of the passport owner's head and rotating to a profile view, re-illuminating to match the conditions in the test scene. Further manipulation, such as

the addition or removal of spectacles, beard, hats, etc, are also possible as is the adjustment of perspective distortion upon the model. While this does not confer any form of familiarity with a novel face, which would be the ideal method of improving recognition, we suggest that the use of a well-fitted 3D model would allow us to remove many of the distracting factors that cause significant degradation to unfamiliar face recognition, in essence trying to boost the performance of unfamiliar recognition by providing some of the empirically apparent functionality of the familiar face recognition process.

People do not carry laser-scanned 3D models of their own heads but only 2D photographs. Hyde and Robinson [18] suggest that there is a useful correlation between the greyscale information in a standard passport style photo and a depth map of the face which can be exploited in an application-specific coding scheme. We propose that the depth information for a face could be estimated from the greyscale information present in a 2D image with sufficient accuracy to provide a better 3D model than the mean face, and allow the re-posing and re-illumination of a 2D image. In the conditional density estimation process described in section 2, we have a mechanism for doing this. Training images consisting of greyscales and depths will be used to estimate a regularized normal model for faces. Greyscale faces will then be used as probes to recover all the missing depth dimensions.

4 Covariance matrix estimation

Robinson [5] reviews methods for covariance matrix estimation in the context of face classification and detection, then proposes an estimator of the form:

$$\mathbf{\Sigma}_i(\alpha,\beta) = \alpha\mathbf{S}_i + (1-\alpha)\mathbf{S}_{total} + \beta\mathbf{I} \tag{8}$$

where $\mathbf{\Sigma}_i(\alpha,\beta)$ is the estimated coefficient matrix for class i, $\mathbf{S}_i$ is the scatter matrix derived from a subset of the training samples of class i, $\mathbf{S}_{total}$ is the scatter matrix of all available training samples over all classes and α,β are regularization parameters estimated by classifying the remainder of the training samples for class i and choosing the α,β that give best performance. The estimator in [5] includes further regularization parameters γ_i which adjust the volumes of each class, again by maximizing classification performance. In our context, we have a single class so $\mathbf{S}_i = \mathbf{S}_{total}$, and α and γ_i have no effect. The estimator therefore reduces to

$$\mathbf{\Sigma}_i(\beta) = \mathbf{S}_i + \beta\mathbf{I} \tag{9}$$

In contrast to earlier regularization methods (e.g. [19,20]), the estimator of [5] optimizes its parameters according to the application. Therefore during training we consider the estimation of missing data as a recovery problem and search for the β that produces the average best estimate during a training phase using sample images with known depths.

The data available to us for the 3D face estimation comprises some 740 laser scanned images from the University of Notre Dame biometrics database. This database contains multiple images of some subjects in slightly different poses. In order to provide fair test images, 40 were removed from the set. These 40 were selected such that no other images of any of the subjects in the control set were present in the remaining 700 training images. The 3D models were automatically converted into depth map images by filtering the images to remove noise caused by surface reflectance peaks which can cause highly inaccurate readings from a laser scanner, and then centring the point closest to the camera. It can be assumed that all images are facing forwards, so this nearest point is always the nose. This is verified by checking that the image has approximately the same amount of non-zero data to both the left and right of the peak. A set-size sample window is set about this central peak, cropping the model and texture information to the face portion of the images only. A backplane is then set by scanning through all models to determine the nearest backplane which does not result in data loss. The texture and depth maps are then generated using a rendering library (OpenGL) to return depth and colour information for specified sample points.

As so few training samples are available relative to the size of the space of the problem, we have resized the images to 38 by 50 pixel greyscale texture and depth maps, limiting the dimensionality of the space to 3800 dimensions. All training images are mirrored to double the number of training samples available to 1400.

Optimization was performed by sweeping through values for β from 1 to 17500. The covariance matrix was computed using a "leave-one-out" method – that is, omitting one of the training images, computing the covariance with the remainder, regularising using the current control value, calculating the projected depth model and hence a mean square error. As several subjects in the set are represented in multiple images, all instances of a given test subject were omitted from the training set when testing using an image of that subject. As shown in figure 4, a clear minimum is evident in the results, at approximately 4500.

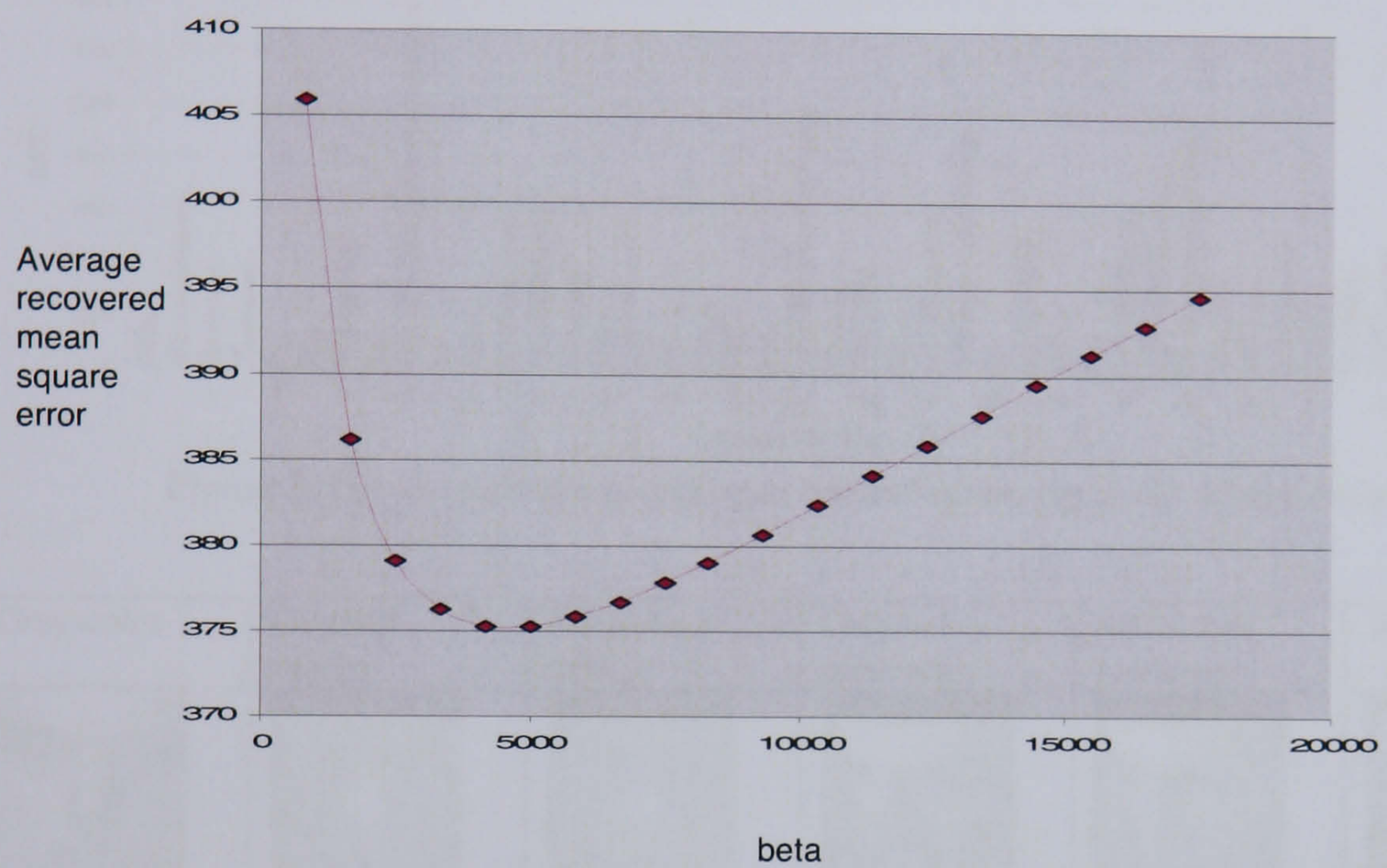


Figure 4. Average reconstruction error vs parameter β during training

In order to better understand the impact of varying the size of the training data, we have applied a simplified method of parameter estimation to cases where the scatter matrix was estimated from 175, 350, 525 and 700 images. In this experiment, we simply evaluated the optimal β for the forty test images. Table 2 shows the results:

Number of training images	175	350	525	700
Optimal β for the test set	20000	28889	11329	12210
Average MSE over the test set at this β	549	546	479	495

Table 2. Optimal parameter values for restricted training sets

In each case we also constructed the equivalent of figure 4. The graph's shape is always the same: there is an initial rapid improvement in performance as the parameter value is increased. This then reaches a minimum with a very shallow curve, and then degrades towards the standard deviation of the training set (i.e. the average distance between the training samples and their mean). It will asymptotically approach this value, which corresponds to the use of the mean depth for recovery, as β tends to infinity. While the estimation improves over the mean in all cases, improvement is clearly greater with more training samples. The β yielding the optimal performance decreases as the number of samples increases.

5 Recovery of depths from test images

Application of an estimation transform using the optimal β value of 4500 determined in section 4 to the forty test images not used in training yields results which consistently outperform the use of the mean depth image. Figure 5 shows that recovery via the conditional density provides on average an estimation with less than half the MSE of the mean depth. As shown in figure 6, the estimation provides a satisfactory model for use in a 3D render of the face. The top four rows of figure 6 are representative examples, while the bottom row shows the worst case (image 9 in figure 5). The original depths in this image were distorted and that has contributed to the poor MSE, but the reconstruction is also subjectively poor, probably because of the lack of images of similar framing and pose in the training set.

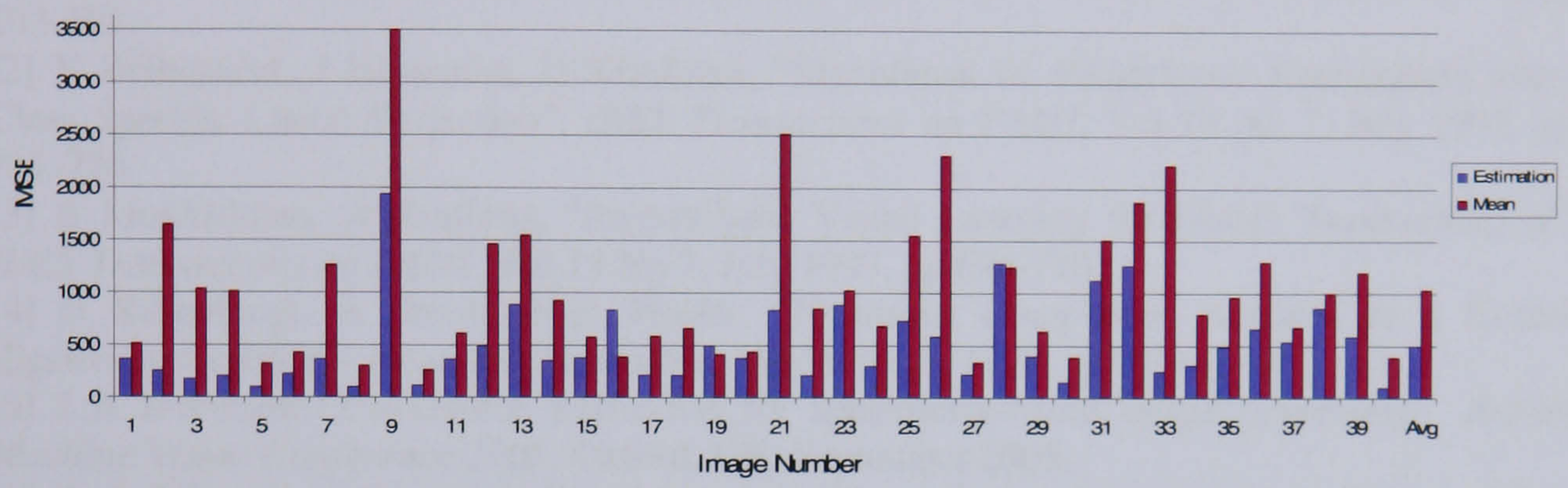


Figure 5. Depth recovery accuracy as Mean Square Error for 40 test images

Figure 6. Example depth reconstructions

6 Conclusions

We have applied conditional distribution estimation to the recovery of missing data in images. In particular, we have shown how accurate estimates of face depths are recoverable from face photos. In future work we will use the depth data to relight the photographic information, and test the extent to which this aids human recognition.

References

[1] M Kirby, L Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces", *Transactions on PAMI*, Vol 12, No 1, January 1990, pp 103-108.

[2] V Belhumeur, J Hespanha, D Kriefman, "Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection", *IEEE Transactions on PAMI*, Vol 19 No 7, July 1997, pp 711-720.

[3] B Moghaddam, A Pentland, "Probabilistic Visual Learning for Object Representation", *IEEE Transactions on PAMI*, Vol 19 No 7, July 1997, pp 696-710.

[4] B Schoelkopt, A Smola, K-R Muller, "Nonlinear Component Analysis as a Kernal Eigenvalue Problem", *Neural Computation*, Vol 10 No 5, 1998, pp1299-1319.

[5] J A Robinson "Covariance estimation for appearance-based image processing", *British Machine Vision Conference 2005*, Oxford, UK, September 2005.

[6] R A Johnson, D W Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Upper Saddle River, NJ, 1998.

[7] P S Penev, L Sirovich, "The Global Dimensionality of Face Space", *Fourth IEEE Internat Conf Automatic Face and Gesture Recognition,* Grenoble, France, March 26-30 2000.

[8] R Kemp, N Towell, G Pike, "When Seeing should not be Believing: Photographs, Credit Cards and Fraud", *Applied Cognitive Psychology*, Vol 11, 1997, pp 211-222

[9] P J B Hancock, V Bruce, A M Burton, "Recognition of unfamiliar faces", *Trends in Cognitive Sciences*, Vol 4, 2000, pp 330-337

[10] N F Troje, D Kersten, "Viewpoint dependent recognition of familiar faces", *Perception*, Vol 28 February 1999, pp483-487

[11] E Tulving, *Elements of episodic memory,* Clarendon Press, 1985

[12] F L Krouse, "Effects of pose, pose change and delay on face recognition performance", *Journal of Applied Psychology,* Vol 66, pp 651 - 654

[13] V Bruce, T Valentine, A D Baddeley, "The basis for the ¾ view advantage in face recognition", *Applied Cognitive Psychology,* Vol 1, 1987, pp 109-120

[14] N F Troje, H H Bülthoff, "Face Recognition Under Varying Poses: The Role of Texture and Shape", *Vision Research*, Vol 36, No 12, 1996, pp 1761-1771

[15] H Hill, V Bruce, "The effects of lighting on the perception of facial surfaces", *Journal of Experimental Psychology: Human Perception and Performance,* Vol 22, 1996, pp 986 - 1004

[16] C H Liu, C A Collin, A Chaudhuri, "Does face recognition rely on encoding of 3-D surface? Examining the role of shape-from-shading and shape-from-stereo", *Perception,* Vol 29, 2000, pp 729-743

[17] C H Liu, A Chaudhuri, "Face recognition with perspective transformation", *Vision Research*, Vol 43, 2003, pp 2393-2402

[18] J R Hyde, J A Robinson, "Coding 3D facial models for mugshot applications", *Vision, Video and Graphics,* 2003, pp 127-133

[19] J H Friedman, "Regularized Discriminant Analysis", *Journal of the American Statistical Association*, Vol 84 No 405, March 1989, pp 165-175.

[20] B-C Kuo, D A Landgrebe, "A Covariance Estimator For Small Sample Size Classification Problems and its Application to Feature Extraction", *IEEE Transactions on Geoscience and Remote Sensing*, Vol 40, No 4, April 2002, pp 814-819.

References

1. A4 Vision 3d face capture cameras

http://www.a4vision.com/5_overview.html#sml

2. J.J. Atick, P.A. Griffin and A.N. Redlich – "Statistical approach to shape from shading: reconstruction of 3D face surfaces from single 2d images", *Neural Computation, Vol. 8, No. 6, pp 1321-1340, 1996*

3. P. Barkowitz and J.C. Brigham - "Recognition of faces: Own-race bias, incentive and time delay", *Journal of Applied Social Psychology, Vol. 12, pp 255-268, 1982*

4. P.N. Belhumeur, D.J. Kriegman, and A.L. Yuille - "The bas-relief ambiguity", *International J. Comput. Vision, Vol. 35, No. 1, pp 33–44, 1999*

5. M. Brooks and B.K.P. Horn - "Shape and source from shading", *Proceedings of the International Joint Conference on Artificial Intelligence, pp 932-936, 1985*

6. V. Bruce- "Changing faces: visual and non-visual coding processes in face recognition", *British Journal of Psychology, Vol. 73, pp 105-116, 1982*

7. A.J. Bruce, K.W. Beard, S. Tedford, M.J. Harman, and K. Tedford – "African-American and Caucasian-American recognition and likeability responses to African-American and Caucasian-American faces", *Journal of General Psychology, Vol. 124, pp 143-156, 1997*

8. V. Bruce, P. Healey, M. Burton, T. Doyle, A. Coombes and A. Linney – "Recognising Facial Surfaces", *Perception, Vol. 20, No. 6, pp 755-769, 1991*

9. V. Bruce and H. Hill – "Effects of lighting on the perception of facial surfaces", *Journal of Experimental Psychology: Human Perception and Performance, Vol. 22, No.4, pp 986-1004, 1996*

10. V. Bruce, T. Valentine and A.D. Baddeley – "The basis of the three quarter view advantage in face recognition", *Applied Cognitive Psychology, Vol. 1, pp 109-120, 1987*

11. V. Bruce and A. Young – "In the Eye of the Beholder", *Oxford University Press, 1998*

12. D.M. Burt and D.I. Perret – "Perception of age in adult Caucasian male faces: computer graphic manipulation of shape and colour information", *Proceedings of the Royal Society of London, Series B, Vol. 259, pp 137-143, 1995*

13. A.M. Burton, S. Wilson, M. Cowan and V. Bruce – "Face recognition in poor quality video", *Psychological Science, Vol. 10, No. 3, pp 243–248, 1999*

14. M. Castelan and E.R. Hancock - "Acquiring height data from a single image of a face using local shape indicators", *Computer Vision and Image Understanding, Vol. 103, No. 1, pp 64–79, 2006*

15. Colour FERET database
http://www.itl.nist.gov/iad/humanid/colorferet/home.html

16. G. Davies, H. Ellis and J. Shepherd – "Face recognition accuracy as a function of mode of representation", *Journal of Applied Psychology, Vol. 63, No.2, pp 180-187, 1978*

17. J. van Diggelen - "A photometric investigation of the slopes and heights of the ranges of hills in the Maria of the moon", *Bull. Astron. Inst. Netherlands, vol. 11, pp 283-289, 1951*

18. The Digital Michaelangelo Project
http://graphics.stanford.edu/projects/mich/

19. J-D. Durou, M. Falcone, and M. Sagona - "A survey of numerical methods for shape from shading", *Technical Report 2004-2-R, IRIT, 2004*

20. Face Recognition Grand Challenge http://www.frvt.org/FRGC/

21. Fast 3D Scan device, Sheffield Hallam University
http://www.shu.ac.uk/research/meri/gmpr/projects/fast_3D_scan.html

22. FERET database http://www.itl.nist.gov/iad/humanid/feret/

23. R.T. Frankot and T. Chellapa - "A Method for enforcing integrability in shape from shading algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, pp 439-451, 1988*

24. J.H. Friedman - "Regularized Discriminant Analysis", *Journal of the American Statistical Association, Vol. 84, No. 405, pp 165-175, 1989*

25. F. Galton -" Regression towards mediocrity in hereditary stature", *Journal of the Anthropological Institute Vol 15 pp 246-263, 1886*

26. M.J.C. van Gemert, S.L. Jacques, H.J.C.M. Sterenborg, and W.M. Star - "Skin optics", *IEEE Transactions on Biomedical Engineering, Vol. 36, No. 12, pp 1146–1154, 1989*

27. P.J.B. Hancock, V. Bruce and A.M. Burton – "Recognition of unfamiliar faces", *Trends in Cognitive Sciences, Vol. 4, No.9, pp 330-337, 2001*

28. B.K.P. Horn - "Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View". *PhD thesis, Massachusetts Institute of Technology, 1970*

29. B.K.P.Horn - "Height and gradient from shading", *International Journal of Computer Vision, Vol. 5, No. 1, pp 37-75, 1989*

30. J.R. Hyde and J.A. Robinson – "Coding 3d facial models for mugshot applications", *Proceedings of Video, Vision and Graphics, pp 127-135, 2003*

31. J.R. Hyde and J.A. Robinson – Estimation of face depths by conditional densities, *Proceedings of the British Machine Vision Conference, Vol. 2, pp 609-618, 2005*

32. Identix face recognition system http://www.identix.com/

33. K. Ikeuchi and B.K.P. Horn - "Numerical shape from shading and occluding boundaries", *Artificial Intelligence, Vol. 17, No. 1-3, pp 141-184, 1981*

34.Image of Dudley Moore sourced from
http://www.mughsots.com/Celebrity/
Images of Uma Thurman and Sean Bean sourced from
http://www.imdb.com/

35. T. Jebara, K. Russell and A. Pentland - "Mixture of Eigenfeatures for Real-Time Structure from Texture", *Proceedings of the Sixth International Conference on Computer Vision, pp 128-135, 1998*

36. R.A. Johnson and D.W. Wichern - "Applied Multivariate Statistical Analysis", *Prentice-Hall, Upper Saddle River, NJ, 1998.*

37. R. Kemp, N. Towell and G. Pike – "When seeing should not be believing: Photographs, credit cards and fraud", *Applied Cognitive Psychology, Vol. 11, pp 211-222, 1997*

38. B-C. Kuo, D.A. Landgrebe - "A Covariance Estimator For Small Sample Size Classification Problems and its Application to Feature Extraction", *IEEE Transactions on Geoscience and Remote Sensing, Vol. 40, No. 4, pp 814-819, 2002*

39. D.S. Lindsay, P.C. Jack Jr., M.A. Christian – "Other Race Face Perception", *Journal of Applied Psychology, Vol. 76, No. 4, 587-589, 1991*

40. P.L. Lions, E. Rouy, and A. Tourin - "Shape-from-Shading, Viscosity Solutions and Edges", *Numerische Mathematik, Vol. 64, No. 3, 323-353, 1993.*

41. L.Z. McArthur and K. Apatow – "Impressions of baby-faced adults", *Social Cognition, Vol. 2, No. 4, pp 315-342, 1984*

42. B. Moghaddam, J. Lee, H. Pfister and R. Machiraju – "Model-Based 3D Face Capture with Shape-from-Silhouettes", *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures 2003, pp 20-27, 2003*

43. J.M. Montepare and L.Z. McArthur – "The influence of facial characteristics on children's age perceptions", *Journal of Experimental Child Psychology, Vol. 42, No. 3, pp 303-314, 1986*

44. Y. Moses, S. Ullman and S. Edelman – "Generalization to novel images in upright and inverted faces", *Perception, Vol. 25, No. 4, pp 443-461, 1995*

45. D. Nandy and J. Ben-Arie – "Shape from recognition and learning: Recovery of 3d face shapes", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp 2002-2007, 1999*

46. OmniPerception face recognition system
http://www.omniperception.com/

47. A.J. O'Toole, T. Price, T. Vetter, J. C. Bartlett and V. Blanz – "3D shape and 2D surface textures of human faces: The role of "averages" in atractiveness and age", *Image and Vision Computing, Vol. 18, No. 1, pp 9-19, 1999*

48. A.J. O'Toole, T. Vetter and V. Blanz – "Three dimensional shape and two dimensional surface reflectance contributions to face recognition: an application of three-dimensional morphing", *Vision Research, Vol. 39, pp 3145-3155, 1998*

49. A.J. O'Toole, T. Vetter and V. Blanz – "Three-dimensional caricatures of human heads: Distinctiveness and the perception of facial age", *Perception, Vol. 26, No. 6, pp 719-732, 1997*

50. K.E. Patterson and A.D. Baddeley – "When face recognition fails", *Journal of Experimental Psychology: Human Learning and Memory, Vol. 3, No.4, pp 406-417, 1977*

51. A.P. Pentland - "Local shading analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, No. 2, pp 170-187, 1984.*

52. E. Prados and O. Faugeras - "A rigorous and realistic shape from shading method and some of its applications". *Technical Report RR-5133, INRIA, 2004*

53. E. Prados and O. Faugeras - "Unifying approaches and removing unrealistic assumptions in shape from shading: Mathematics can help". *Proceedings of the 8th European Conference on Computer Vision, Vol. 3024, pp 141–154, 2004*

54. E. Prados and O. Faugeras - "Shape from shading: A well posed problem?", *Proceedings of CVPR, Vol 2, pp 870-877, 2005*

55. The Prometheus Project
http://www.bbc.co.uk/rd/projects/prometheus/index.shtml

56. T. Rindfleisch - "Photometric method for lunar topography", *Photogrammetric Eng, vol. 32, pp 262-276, 1966*

57. J.A. Robinson – "Covariance matrix estimation for appearance-based face image processing", *Proceedings of the British Machine Vision Conference, Vol. 1, pp 389-398, 2005*

58. E. Rouy and A. Tourin - "A viscosity solutions approach to shape from shading", *SIAM Journal of Numerical Analysis vol. 29 no. 3 pp 867-864, 1992*

59. K.B. Russell – "Eigenheads for reconstruction", *SB Thesis, MIT, 1997*

60. M. Sonka, V. Hlavac and R. Boyle – "Image Processing, Analysis and Machine Vision", *PWS publishing , 1999*

61. L. Strub and J. Robinson – "Automated facial conformation for model-based videophone coding", *Proceedings of the 1995 IEEE International Conference on Image Processing, Vol. 2, pp 587-590, 1995*

62. The Taming of Smeagol, The Two Towers: Extended edition, New Line Cinema, 2002

63. N.F. Troje and D. Kersten – "Viewpoint dependent recognition of familiar faces", *Perception, Vol. 28, No. 4, pp 483-487, 1999*

64. M. Turk and A. Pentland - "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience, Vol. 3, No.1, pp 71-86, 1991*

65. M. Turk and A. Pentland – "Face recognition using eigenfaces", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 586-590, 1991*

66. T. Vetter and V. Blanz – "Estimating coloured 3D face models from single images: An example based approach", *Proceedings of the 5th European Conference on Computer Vision, Vol. 2, pp 499-513, 1998*

67. United Kingdom Passport Agency photographic standards, http://www.ukpa.gov.uk/press_050804.asp cache available through google http://66.102.9.104/search?q=cache:iq7XnTF5oHQJ:ukpa.gov.uk/press_050804.asp+&hl=en&gl=uk&ct=clnk&cd=1&client=firefox-a

68. T. Valentine – "A unified account of the effects of distinctiveness, inversion and race in face recognition", *Quarterly Journal of Experimental Psychology, Vol. 43A, pp 161-204, 1991*

69. T. Valentine and V. Bruce – "The effects of distinctiveness in recognising and classifying faces", *Perception, Vol. 15, No. 5 pp 525-535, 1986*

70. Y. Su, W. Wang, K. Xu, and C. Jiang - "The optical properties of skin", *Proceedings of the SPIE, Vol. 4916, pp 299–304, 2002*

71. R. Zhang, P.S. Tsai, J.E. Cryer, and M. Shah - "Shape-from-shading: a survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 8, pp 690–706, 1999*

72. Q. Zheng and R. Chellappa - "Estimation of illuminant direction, albedo, and shape from shading", *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 7, pp 680–702, 1991*