University of Exeter
Department of Computer Science

Face recognition in an unconstrained environment for monitoring student attendance

Justin Nicholas Worsey

July 2016

Supervised by Professor Richard Everson

Submitted by Justin Nicholas Worsey, to the University of Exeter as a thesis for the degree of Masters by Research in Computer Science , July 2016.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) ....

Abstract

Traditional paper based attendance monitoring systems are time consuming and susceptible to both error and data loss. Where technical advances have attempted to solve the problem, they tend to improve only small portions i.e. confidence that data has been collected satisfactorily can be very high but technology can also be difficult to use, time consuming and impossible especially if the overall system is down. Camera based face recognition has the potential to resolve most monitoring problems. It is passive, easy and inexpensive to utilise; and if supported by a human safeguard can be very reliable. This thesis evaluates a strategy to monitor lecture attendance using images captured by cheap web cams in an unconstrained environment. A traditional recognition pipeline is utilised in which faces are automatically detected and aligned to a standard coordinate system before extracting Scale Invariant Feature Transform (SIFT), Local Binary Pattern (LBP) and Eigenface based features for classification.

A greedy algorithm is employed to match captured faces to reference images with faces labelled and added to the training set over time. Performance is evaluated on images captured from a small lecture series over ten weeks. It is evident that performance improves during the series as new reference material is included within the training data. This correlation demonstrates that the success of the system is determined not only by the on-going capturing process but also the quality and variability of the initial training data.

Whilst the system is capable of reasonable success, the experiments show that it also yields an unacceptably high false positive rate and cannot be used in isolation. This is primarily because the greedy nature of the algorithm allows the possibility of assigning multiple images of the same person captured in the same lecture to different students including ‘no shows’.

Acknowledgement

I would like to thank Professor Richard Everson for his assistance and direction with this research project. In particular, he sought to encourage a group of his students to participate and was responsible for the collection of source imagery which enabled me to create the real-world lecture series in which to evaluate. I would also like to thank him for encouraging me to present a related paper at the UK Computational Intelligence (UKCI) conference 2015 and all of the support he offered to enable this. In general, he was a good listener and elicited sound reasoning that assisted and directed my approach.

I would also like to acknowledge and thank the students who, not only generously agreed to participate, but provided the challenging environment for the attendance monitoring system to operate in by posing for their photographs at the beginning of each lecture. Their assistance brings a real element to the project which cannot be overstated.

Finally, I would like to thank Dr Yiming Ying who initially supervised me during this project and set me on the correct path; and of course, my beautiful wife Julia for her support throughout.

Contents

1 Introduction **14**

1.1 Outline of the Problem . . . 14

1.2 Outline of the Solution . . . 14

1.3 Contribution . . . 15

1.4 Thesis Outline . . . 15

2 Background **17**

2.1 Introduction . . . 17

2.2 Detection, Recognition and Verification . . . 17

2.3 Labelled Faces in the Wild (LFW) . . . 18

2.4 Dimensionality Reduction . . . 18

2.4.1 Principal Component Analysis . . . 19

2.4.2 Linear Discriminant Analysis . . . 21

2.5 Attendance Monitoring Prior Work . . . 22

2.6 Prior Work . . . 24

2.6.1 The Viola Jones Object Detection Algorithm . . . 24

2.6.2 Active Shape Modelling . . . 27

2.6.3 Eigenfaces and Fisherfaces . . . 28

2.6.4 Scale Invariant Feature Transform (SIFT) . . . 31

2.6.5 Local Binary Patterns (LBP) and Variants . . . 33

2.6.6 k-Nearest Neighbour (kNN) . . . 36

2.6.7 Support Vector Machines (SVM) . . . 38

2.7 Conclusion . . . 40

3 Attendance Monitoring System **41**

3.1 Introduction . . . 41

3.2 Student Attendance Monitoring . . . 41

3.3 Verification/Recognition Pipeline Definition . . . 42

3.3.1 Face Detection . . . 42

3.3.2 Alignment . . . 43

3.3.3 Feature Extraction . . . 44

3.3.4 Classification . . . 45

3.4 Recognition Algorithm . . . 45

3.5 Conclusion . . . 47

4 Experiments **49**

4.1 Introduction . . . 49

4.2 General Observations . . . 50

4.2.1 Face Detection . . . 50

4.2.2 Alignment . . . 51

4.2.3 Summary . . . 52

4.3 Evaluation of Strategies . . . 52

4.4 Labelled Faces in the Wild (LFW) Results . . . 53

4.4.1 Single Pass Results . . . 54

4.4.2 Multi Pass Results . . . 59

4.5 Lecture Series Results . . . 64

4.5.1 Single Pass Results . . . 66

4.5.2 Multi Pass Results . . . 76

4.6 Discussion . . . 86

4.7 Conclusion . . . 89

5 Conclusion **91**

Bibliography **93**

List of Figures

2.1 An example set of aligned faces taken from the deep funnelled Labelled Faces in the Wild database. 18

2.2 A simple example of the principal components of a set of two dimensional data is shown. The black line, with the greater variance is the major component. Importantly, the dimensions are orthogonal and consequently do not interfere with each other i.e. if you project the second component into the first then the resultant dot product would be zero. 20

2.3 An example of a data set which contains class information. The left hand image shows the projection of the data set onto the principal component (the image has been rotated) following PCA. This is because it contains the greatest variance in the entire set. However, the data would be better represented if it was discriminated by its class information and projected accordingly . This is shown in the right hand image which is based on using the LDA approach. 21

2.4 Three distinct types of Haar-like feature are defined and are identified by the number of adjacent rectangles they contain (two, three or four) and the relative positions of their rectangles. Each feature is available in both horizontal and vertical forms. Each distinctive area is a template used to compute contrast information when computed over the image space. 24

2.5 An example showing the pixel intensity values of the original source image and its translation to the Integral Image. Each cell value of the Integral Image contains the summation of the pixel values of the source image from the left and to the top of the cell in question, inclusive. 25

2.6 A visual representation of the use of the Integral Image to compute the total pixel intensity of an image source rectangle. 25

2.7 Two Haar-like features are shown to illustrate the strength inherent within the domain knowledge. Image taken from [Viola & Jones, 2001] 26

2.8 Viola and Jones' attentional cascade concept. Each link in the chain is used as a test for the source image. If it fails it is immediately discarded; otherwise it moves on to the next, much harder, link in the chain. 27

2.9 An example of an image that has been labelled ready for training [Cootes, 2000]. 27

2.10 Three iterations showing the active shape modelling algorithm attempting to fit the model to the image. 28

2.11 A set of principal eigenfaces derived from computing the PCA eigenvectors from the baseline Labelled Faces in the Wild aligned data set. 29

2.12 A visual method to review the variance of each eigenvector. The graph plots the eigenvalues for each dimension where the greater the variance the more relevant the associated eigenvector is to the structure of the data. . . 30

2.13 A step by step application of the weighted eigenvectors being added to the mean face in order to reform the original image of Kevin Spacey. 31

2.14 The mean face is derived from computing the average face across the entire data set. 31

2.15 SIFT octaves and scaling pyramid. The original image is convolved with a set of Gaussian functions to produce an octave which is then down sampled for scaling purposes. Image taken from [Lowe, 2004]. 32

2.16 Each pixel is compared with its neighbouring pixels to find local extrema pixels. This includes comparisons with pixels in the adjacent layers. Each extrema pixel is a potential key point. Image taken from [Lowe, 2004]. . . . 32

2.17 SIFT key points detected within the region of interest. Each key point has an associated descriptor describing its invariant context. 33

2.18 An LBP code of 11010011 derived from comparing the intensity values of each pixel against the centre pixel. The value 5, our sequence start, is greater than the middle value 4 yielding a bit value of 1 based on equation 2.21. The code is defined by comparing the pixels clockwise. 33

2.19 An encoding of the baseline LFW data set using the LBP technique. . . . 34

2.20 Local Binary Pattern (LBP) histograms of a target and candidate for subsequent distance based comparison. 34

2.21 The CS-LBP code is produced by comparing the pixel values of symmetrical pairs. e.g. if n_0 is greater than n_4 then the bit code is set to 1 otherwise it is set to 0. This figure, taken from [Heikkilä et al., 2006], shows the methods used for both CS-LBP and LBP encoding. 35

2.22 A normalised to 8-bit grey scale encoding of the baseline LFW data set using the CS-LBP technique. 36

2.23 A reconstruction of an LBP encoded face that has had PCA applied to it. The top left hand image shows the summation of the average LBP face with the dominant principal component, going left to right the other components are added until we finish with the bottom right hand reconstructed LBP version. To show a direct comparison, the image of Kevin Spacey is the same as the one used in Figure 2.13 37

2.24 An example of a kNN classification. Using the smaller circle, which shows the region occupied by 3 nearest neighbours, the green circle would be classified as part of the red triangle class. However, as the radius is increased to the dotted circle, the classification changes to reflect the new majority (blue squares). Image taken from Wikipedia. 37

2.25 A two state classification space that highlights the decision boundary that separates the classes by the maximum distance. Image taken from [Cortes & Vapnik, 1995]. 38

2.26 Two examples where it is not possible to draw an optimal hyperplane in the originating dimension. Image taken from [Cortes & Vapnik, 1995] . . . 40

3.1 The fundamentals of a face recognition pipeline. The first two stages are concerned with processing the data. Namely, filtering non-faces out of the pipeline followed by alignment. The latter stages extract the features before classification. 42

3.2 A face that has been identified using the Viola-Jones algorithm. 43

3.3 Viola-Jones redeployed to detect pairs of eyes for alignment. A simple two dimensional rotation can be applied to align the eyes. 43

3.4 An attendee whose face has been modelled using ASM and subsequently aligned. Once a face has been modelled using ASM then specific features can be examined in detail with great confidence e.g. the person's mouth. . . 44

3.5 An example of a Viola Jones detected face which is highlighted with a blue square. The image also shows the 7 nearest neighbours results for three feature extraction techniques in the three rows below (LBP, CS-LBP and SIFT respectively). Each column represents a distance ranking from best (left) to worst. In this instance SIFT clearly gives the best results as it has placed the detected face near 4 correct student matches. An attempt to align the faces by their eyes was tried and future work is needed to perfect this. 46

3.6 An example of output from the system where the top row shows the face image and the bottom row shows the students that best match the face images. In this case, 9 out of 12 students were correctly identified. 47

4.1 The angle of some of the faces within the source image prove difficult for the Viola-Jones algorithm to detect. In this image, taken from lecture 23, no faces have been detected. 50

4.2 Graph showing Viola Jones detection over 161 photographs across 23 lectures where multiple photographs were taken in each lecture. *Faces* corresponds to the maximum number of attendees at the lecture. *Expected* refers to the expected number of faces that should be detected by the algorithm. This has been determined by human inspection. *Detected Faces* and *False Positives* refer to the output from the Viola Jones algorithm. 51

4.3 Breakdown of Viola Jones detection and failure rates versus the expected rates across the 23 lectures where multiple photographs were taken in each lecture. Please refer to Figure 4.2 for the accompanying graph. 51

4.4 ASM Analysis data showing the percentage of quality fittings using an arbitrary scale ranging from 10 (Good) to below 5 (Bad) where we mark on how good the model fitted the eyes. The right hand images show some of those fittings. Three have scores of 10 with another at below 5 (bottom right image) 52

4.5 A bar chart showing the evaluation of the synthetic lecture over the k-nearest neighbour range; using the single pass approach. 54

4.6 A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 1$; using the single pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 55

4.7 A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 8$; using the single pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 56

4.8 A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 19$; using the single pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 57

4.9 Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbour range and colour coded based on the number of training set matches; using the single pass approach on the LFW synthetic data set. 58

4.10 Breakdown on a per student basis using individual methods over the entire nearest neighbour range and colour coded based on the number of training set matches; using the single pass approach on the LFW synthetic data set. 58

4.11 Example pairs of attendees to be matched. 59

4.12 A bar chart showing the evaluation of the synthetic lecture over the k-nearest neighbour range; using the multiple pass approach. 60

4.13 A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 1$; using the multiple pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 60

4.14 A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 5$; using the multiple pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 61

4.15 A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 19$; using the multiple pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 62

4.16 Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on the LFW synthetic data set. 62

4.17 Breakdown on a per student basis using individual methods over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on the LFW synthetic data set. 63

4.18 Six examples of the same attendee in varying poses highlights the difficulty faced by the system. 65

4.19 Percentage of principal components used which represent 90% of the variance for each spectral method. The line plot shows the total number of principal components found for each training set. 65

4.20 Line plot showing the first 15 eigenvalues for each of the three spectral methods whilst evaluating training set 10. The LBP and CS-LBP methods distribute the total variance within a tighter range than standard Eigenfaces method. 66

4.21 A bar chart showing the evaluation of the lecture series over the k-nearest neighbour range; using the single pass approach. 66

4.22 A bar chart showing percentage matches over three training sets with nearest neighbour $k = 1$; using the single pass approach. 67

4.23 A bar chart showing percentage matches over three training sets with nearest neighbour $k = 4$; using the single pass approach. 67

4.24 A bar chart showing percentage matches over three training sets with nearest neighbour $k = 12$; using the single pass approach. 68

4.25 A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 1$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 68

4.26 A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 4$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 69

4.27 A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 12$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 70

4.28 Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the single pass approach on lecture 12. 70

4.29 Breakdown on a per student basis using individual methods over the entire k nearest neighbour range and colour coded based on the number of training set matches; using the single pass approach on lecture 12. 71

4.30 Example output from the system where the ensemble method is used with nearest neighbourhood $k = 4$; using the single pass approach on lecture 12. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched. 71

4.31 A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 1$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 73

4.32 A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 4$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 74

4.33 A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 12$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 74

4.34 Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the single pass approach on lecture 13. 75

4.35 Breakdown on a per student basis using individual methods over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the single pass approach on lecture 13. 75

4.36 Example output from the system where the ensemble method is used with nearest neighbour $k = 4$; using the single pass approach on lecture 13. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched. 76

4.37 A bar chart showing the evaluation of the lecture series over the k-nearest neighbour range; using the multiple pass approach. 77

4.38 A bar chart showing percentage matches over three training sets with nearest neighbour $k = 1$; using the multiple pass approach. 77

4.39 A bar chart showing percentage matches over three training sets with nearest neighbour $k = 6$; using the multiple pass approach. 78

4.40 A bar chart showing percentage matches over three training sets with nearest neighbour $k = 12$; using the multiple pass approach. 78

4.41 A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 1$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 79

4.42 A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 6$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 80

4.43 A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 12$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 80

4.44 Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on lecture 12. 81

4.45 Breakdown on a per student basis using individual methods over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on lecture 12. 81

4.46 Example output from the system where the ensemble method is used with nearest neighbourhood $k = 6$; using the multiple pass approach on lecture 12. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched. 82

4.47 A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 1$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 83

4.48 A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 6$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 84

4.49 A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 12$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1. 84

4.50 Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on lecture 13. 85

4.51 Breakdown on a per student basis using individual methods over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on lecture 13. 85

4.52 Example output from the system where the ensemble method is used with nearest neighbour $k = 6$; using the multiple pass approach on lecture 13. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched. 86

1 Introduction

1.1 Outline of the Problem

Universities are interested in monitoring attendance as an indicator of student engagement for reasons including visa compliance and student welfare. Most standard attendance monitoring methods employed trade off the effort required to obtain results with their accuracy [Joseph & Zakharia, 2013]. For example, lecturers who actively take a register collect reasonable results at the expense of the available teaching time. The opposite can also be seen where audience delegation is utilised. Whilst the technique distributes the effort across the audience, it is susceptible to misrepresentation and data loss. Furthermore, this approach can be disruptive to the individuals concerned [Alia et al., 2013]. Consider the case where an uncontrolled register is passed from attendee to attendee during a critical part of the lecture.

Technological advances allow systems to employ RFID tags, magnetic strip cards and even finger print recognition but none successfully balance the effort/accuracy ratio to a satisfactory level. Certainly, technology tends to yield higher confidence in the collected data e.g. the correct ID card was swiped; but the ability to swipe another person's ID card or the time taken to queue to present one's self at a biometric sensor still present the same fundamental issues as the more traditional methods. In some ways, technology can compound the problem, especially if the system is down or difficult to use.

1.2 Outline of the Solution

Face recognition has the potential to solve this conundrum by offering a passive solution whilst maintaining a high degree of accuracy. Granted, environmental conditions such as lighting and the unassuming, yet natural, pose of the attendees make this a difficult and challenging setting to operate within. Even with such limitations, the data can be post processed off-line with human intervention to preserve the system's integrity.

Attendance monitoring using biometrics including face recognition have been previously proposed [Alia et al., 2013] [Shilwant & Karwankar, 2012] [Joseph & Zakharia, 2013] but they typically concern themselves with the overall system infrastructure whilst operating under constrained conditions in order to reduce the problems encountered by pose, alignment and multiple images of the same attendee. As this thesis will show, duplication of attendee and pose is a particular problem extending much further than two-dimensional alignment problems [Yan et al., 2010]. Without three-dimensional modelling, it is perhaps best to consider extreme pose as different profile images in order to mitigate the problem to some degree.

The University of Exeter stores lecture attendance information digitally and subsequently analyses the results. Primarily this is done to conform with current government legislation regarding international students, especially as the University is a Tier 4 sponsor license holder. Attendance, or lack of, can also be used as an early indicator for student welfare purposes. The University is not alone in having such requirements.

The vast majority of attendance data is currently collected by the University using paper based registers. Although technologies such as bar code scanners are available, they are used in a limited capacity. One prototype system used a combination of an iPad and a bar code scanner to present an image of the student together with their attendance history as part of a visual check. This system was the first to automatically add the confirmation information to the digital store. Whilst accurate, it required huge effort at the point of delivery. Due to this, the system is better suited to examination hall verification rather than as a tool to assist with the monitoring of general lecture attendance.

Accuracy and effort aside, the University is also sensitive to the needs of its students. Whilst students fully understand the need for taking registers they do not want to participate in a big brother scenario that far exceeds the original remit e.g. where CCTV surveillance monitors their every move. Furthermore, whilst it is a legal requirement to monitor international students, a consistent student experience across the board is almost certainly desirable. Consequently, any system adopted by the University must take these considerations into account. Clearly, where a face recognition system is proposed its entire purpose and functionality must be transparent at all times.

1.3 Contribution

The principal contribution of this thesis is to explore the limitations of applying standard detection and recognition methods to the problem of attendance monitoring where multiple images of the attendees have been captured. Importantly, these images have been captured in a variety of different poses and other environmental conditions. It outlines a strategy that employs a standard four stage recognition pipeline, showing results from each stage before applying the pipeline iteratively to best match attendees to those expected.

A preliminary report on this work was presented and published at the 15th UK Workshop on Computational Intelligence (UKCI) in September 2015 [Worsey & Everson, 2015].

1.4 Thesis Outline

Chapter 2 discusses the foundation building blocks necessary to build an attendance monitoring system. It suggests that the system is ultimately a recognition system given the underlying assumption that the attendees will be known prior to the lecture. In essence, the chapter is a literature review of the components necessary to enable such a recognition system to be built. It includes a review of prior attendance monitoring work.

A major theme of the chapter concerns dimensionality reduction and it will be shown that this concept forms the cornerstone of the majority of methods reviewed. Consequently a section of the chapter is given to the subject which includes the description of two of the main approaches for reduction, Principal Component Analysis and Linear Discriminant Analysis. Whilst not exhaustive, Principal Component Analysis will feature in three of the building blocks subsequently reviewed. The latter section of the chapter reviews common techniques and includes details sufficient to provide the reader with a good understanding of how they work.

Chapter 3 defines the attendance monitoring system and introduces a standard recognition pipeline that utilises the techniques outlined in Chapter 2. This includes face detection using the Viola Jones algorithm, a fast technique to extract faces from the lecture images. Whilst it does yield false positives, detection focuses the pipeline allowing computational effort to be used constructively further down the pipeline. Image alignment is also discussed as it proves useful for subsequent feature extraction in the pipeline; especially those that rely upon a localised context such as histogram based Local Binary Patterns (LBP). Finally, a simple distance metric approach is used to classify the data before a confidence measure is taken and employed.

The thesis will, in Chapter 4, evaluate the recognition system against a synthetic lecture series constructed from the Labelled Faces in the Wild (LFW) data set and against a complete lecture series. It will first consider the robustness of the Viola Jones algorithm by comparing the number of faces detected from the data against those deemed detectable via human inspection. It will then examine the effectiveness of the system's ability to align the faces into a common framework. Good alignment is essential to the system. Ultimately, the chapter will test the synthetic lecture and the real world equivalent using two approaches. In the first each student's face image is identified as the student for which there is the best match in the training data. A second multiple pass algorithm evaluates the quality of the match between each face image and all the training images. The face image with the best match is identified with the matching training image and that face image and the training images corresponding to the matched student are removed from the available training images.

The latter part of Chapter 4 forms the basis of a discussion about the system and includes area for further investigation and improvement. Chapter 5 is the conclusion.

2 Background

2.1 Introduction

This chapter is a literature review of well known techniques and resources utilised within the field of face recognition. It will begin by outlining three closely related components of the field; the detection of faces, verification and recognition. It will also introduce the use of an on-line database of face images. This is an important resource as it forms the baseline of all experiments performed.

The chapter will consider efforts to reduce dimensionality of a data set in order to avoid the curse of dimensionality. In particular, it will detail the methods of two primary concepts; Principal Component Analysis and its related Linear Discriminant Analysis. The former will be employed by several of the techniques described in the latter part of this chapter.

The main section is dedicated to a review of existing attendance monitoring research as well as a description of common techniques and their inner workings. It will define them as building blocks which will form the infrastructure of an attendance monitoring system.

Predominantly the equations used within this thesis will be based, in the appropriate context, on their use with images. The base definition of the data is of the form

$$\{(x_n, t_n)\}, n = 1 \ldots N \tag{2.1}$$

where N is the total number of images in the training set, $x_n \in \mathbb{R}^D$ is a stacked image and t_n is its associated class label. Furthermore t_n is a member of the set of labels L comprised of the identities of the available faces.

2.2 Detection, Recognition and Verification

Face detection concerns the detection of one or more human faces from within the image space. Detection acts as a filter to focus attention on regions of interest that could benefit from further scrutinization. It does not offer any knowledge of whom the face may be but does allow subsequent techniques to extract facial features. Verification and recognition are closely related classification problems that utilise the extracted features to either confirm the identification of an individual, in the case of verification, or seek to identify whom the individuals are when recognition is required.

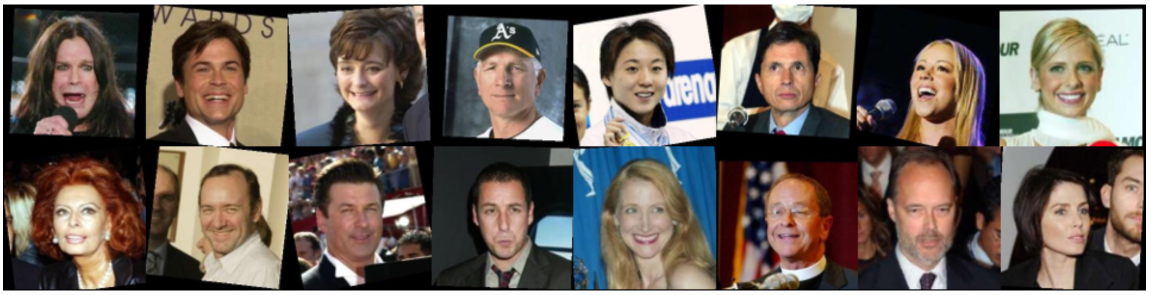


Figure 2.1: An example set of aligned faces taken from the deep funnelled Labelled Faces in the Wild database.

An attendance monitoring system resides somewhere in between verification and recognition. Whilst the list of expected attendees is known, as is the main assumption of this thesis, the system acts as a recognition system confirming the attendance accordingly. This is because the students do not make any claim to prove who they are so we are not verifying them. We simply need to recognise them.

2.3 Labelled Faces in the Wild (LFW)

Labelled Faces in the Wild (LFW) is a collection of named face images taken from the internet [Huang et al., 2007]. The database is publicly available and provides a useful resource for the training and testing of face recognition algorithms. The original data set comprises of faces that have been detected using only the Viola Jones [Viola & Jones, 2001] object detection algorithm and each face is stored in its raw and unconstrained state.

Further derivatives of the database align the face to a common set of Cartesian coordinates. In the context of this thesis, LFW images provide a benchmark for the validation of the system and so it is desirable to use pre-aligned imagery to remove issues of extreme rotation and awkward facial angles. One available data set meeting this criteria is Huang et al's deep funnelled alignment [Huang et al., 2012] set which finds a set of features within the face to use for alignment. Figure 2.1 shows a sample of the aligned data.

2.4 Dimensionality Reduction

By their very nature images have high dimensionality (width in pixels by height in pixels dimensions). Given that most classification techniques are statistical, representation of data in lower dimensions is highly desirable if we are to avoid the sparseness of information problems associated with the *curse of dimensionality* [Bishop, 1995, Chapter 1.4]. Whilst dimensionality reduction can be computationally expensive, a reduction in the amount of information subsequently processed should more than compensate.

There are numerous methods to reduce the dimensions of a data set. For example, we could select and extract features exploiting domain knowledge. Consider the image space reduction where face detection has pinpointed a potential face. Admittedly, the resultant region of interest is still likely to reside within high dimensional space so it would not be a complete solution in itself.

We could consider clustering salient features using histogram bins or similar receptacles as is evident in [Ahonen et al., 2004, pg. 471] where histogram bins containing approximately 60 uniform patterns capture the extracted features of an 8 bit local binary neighbourhood. The obvious drawback to this approach is the inverse relationship between the size of the receptacle and the information stored. Furthermore, spatial information is compromised leading to the break up of the image into smaller regional grids. Scale Invariant Feature Transform (SIFT) [Lowe, 1999] achieves dimensionality reduction by extracting features and storing them as keys with descriptors in low dimensional vectors.

2.4.1 Principal Component Analysis

Principal Component Analysis (PCA) is a technique commonly employed for dimension reduction [Jolliffe, 2002] [Shlens, 2005]. It is an unsupervised learning algorithm which performs a linear transformation of the dimensions in the original space to a space that best describes the significance of each dimension. It considers the entire data set as a whole without taking into account the labelled classes within the data set. In essence, it determines the coordinates for the data that minimises the mean square error over all linear changes of coordinate. It does this by finding the eigenvectors and eigenvalues of a related covariance matrix i.e. determining a set of vectors ordered by greatest variance. Dimensionality reduction is achieved by discarding the principal components that possess low variance. Variance within the component is proportional to the amount of information they contain suggesting that we can remove components of low variance without too much impact. However, there is no guarantee that we are not losing informative components albeit of low variance.

Furthermore, by projecting data into the computed eigenspace we can capture a set of associated weights that not only permit reconstruction of the image but has the potential to significantly compress the data in the case where a relatively small set of components describes the majority of the original data set. Consider the case of a 120 by 120 pixel image; by projecting the image into eigenspace for the first m significant components we can store an image in terms of m weights where m is less than the total number of dimensions D of the image. Section 2.6.3 defines a simple use of these weights which measures the distance between the eigenweights of two images. The weights are computed from each image's eigenvectors, which in this context, are referred to as Eigenfaces.

An outline of how to perform PCA on a set of images can be seen in the following:

1. Define X as a matrix of stacked images where each image is a column. The images are the x_n component of the set of training data from equation 2.1.

2. Compute the mean vector $\bar{\boldsymbol{x}}$ by averaging the rows of X using

$$\bar{x}_i = \frac{1}{N} \sum X_i \tag{2.2}$$

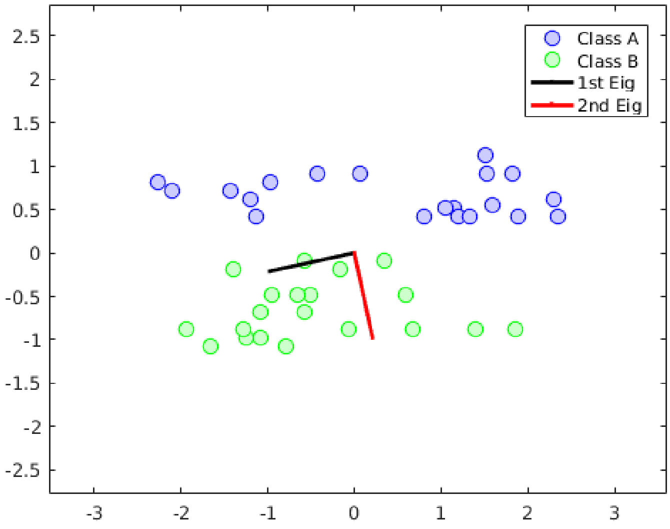


Figure 2.2: A simple example of the principal components of a set of two dimensional data is shown. The black line, with the greater variance is the major component. Importantly, the dimensions are orthogonal and consequently do not interfere with each other i.e. if you project the second component into the first then the resultant dot product would be zero.

where i is the i^{th} row of X.

3. Centre the data by subtracting the mean and placing the resultant column in a matrix A using

$$A = X - 1\bar{\boldsymbol{x}}^T \tag{2.3}$$

where 1 is a vector of 1s.

4. Compute the sample covariance matrix C by

$$C = \frac{1}{N-1} AA^T \tag{2.4}$$

Strictly speaking, we do not need to scale the matrix given that the subsequently solved eigenvectors will be identical. The scaling has been included for completeness.

5. Compute the eigenvectors and associated eigenvalues for the covariance matrix C by solving

$$C\boldsymbol{v} = \lambda\boldsymbol{v} \tag{2.5}$$

where λ are the eigenvalues and $\boldsymbol{v}$ are the corresponding eigenvectors. The eigenvectors are orthonormal and the eigenvalues show the significance of each of the eigenvectors. Figure 2.2 provides a two dimensional example where the most significant component is clearly the direction with the greatest variance. The second

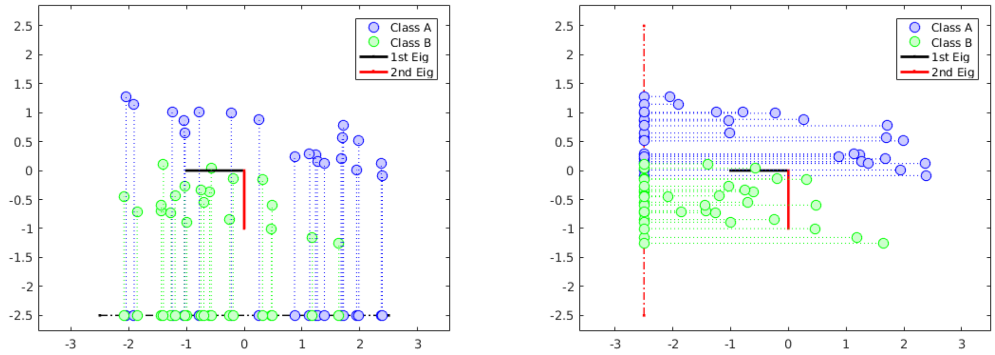


Figure 2.3: An example of a data set which contains class information. The left hand image shows the projection of the data set onto the principal component (the image has been rotated) following PCA. This is because it contains the greatest variance in the entire set. However, the data would be better represented if it was discriminated by its class information and projected accordingly . This is shown in the right hand image which is based on using the LDA approach.

component is shown perpendicular to the first.

Sirovich and Kirby [Sirovich & Kirby, 1987] show that it is feasible to reverse the AA^T multiplication in equation 2.4 which yields a covariance matrix of $N \times N$ dimensions and instead use

$$K = \frac{1}{N-1} A^T A \tag{2.6}$$

to produce a covariance matrix K of $D \times D$ dimensions where $D < N$. We can subsequently transform the eigenvectors back into the intended dimensionality due to their linear relationship across the dimensions [Sirovich & Kirby, 1987]. In doing so we have to normalise the new vectors for them to remain orthonormal.

2.4.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), also known as Fisher's linear discriminant [Bishop, 1995, Chapter 3.6] [Bishop, 2006], is a supervised learning algorithm that determines the projection that maximises class separation by finding the largest ratio of between classes to within classes scatter in a data set. Figure 2.3 shows the projection differences between PCA and LDA. Whilst PCA correctly determines the main principal component, a better representation would have been to consider the data set's class information and project, as LDA has, to a more suited dimension. LDA does not necessarily make PCA redundant, especially where a sufficiently large maxima between class means cannot be found.

An outline of how to perform LDA on a set of data which contains two classes to determine a single projection that maximises the discrimination between them can be seen below.

1. Separate the data into appropriate class matrices C_1 and C_2 where membership is determined by examining the class label of the pair (x_n, t_n) defined in equation 2.1

2. Compute the mean vector for each class $\boldsymbol{m_c}$ using

$$\boldsymbol{m_c} = \frac{1}{N_c} \sum_{x_n \in C_c} x_n \tag{2.7}$$

for each class $c \in [1, 2]$ where $N_c = |C_c|$.

3. Compute the 'within class' scatter matrix S_w by computing

$$S_w = \sum_{x_n \in C_1} (x_n - \boldsymbol{m_1})(x_n - \boldsymbol{m_1})^T + \sum_{x_n \in C_2} (x_n - \boldsymbol{m_2})(x_n - \boldsymbol{m_2})^T \tag{2.8}$$

to represent the expected covariance of the classes.

4. Compute the 'between class' scatter matrix S_b to represent the expected covariance between the mean of each class.

$$S_b = (\boldsymbol{m_2} - \boldsymbol{m_1})(\boldsymbol{m_2} - \boldsymbol{m_1})^T \tag{2.9}$$

5. Find a projection weight vector $\boldsymbol{w}$ such that the ratio of the 'between class' variance to the 'within class' variance is maximised by using

$$\boldsymbol{w_{max}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{|\boldsymbol{w}^T S_b \boldsymbol{w}|}{|\boldsymbol{w}^T S_w \boldsymbol{w}|} \tag{2.10}$$

which can be found as the eigenvectors of the generalised eigenvalue problem

$$S_b \boldsymbol{w} = \lambda S_w \boldsymbol{w} \tag{2.11}$$

corresponding to the largest eigenvalue.

In essence, LDA maximises the distance between the classes whilst minimizing the variance within them to provide the best class discriminant.

2.5 Attendance Monitoring Prior Work

Shilwant and Karwankar [Shilwant & Karwankar, 2012] devise a simple web based system with a centralised data store to allow its portability across campus. Their methodology divides the system into two distinct aspects, that of actual recognition and the post processing of the gathered information for wider system administrative purposes. They use a client side jQuery plug-in to detect faces from a continuously operating web cam before passing the captured image with face coordinates to the server for recognition. The assumption is that they accept the high computational cost of detection at the client, especially given their choice of technology and their continuous operation, in order to reduce the traffic to the server.

Ultimately, Shilwant and Karwankar deploy Eigenfaces (Section 2.6.3) as the main feature extraction technique for their recognition system. Whilst they do not explicitly state where recognition occurs (client, server or both), the data compression afforded by the use of the Principal Components Analysis (PCA, Section 2.4.1) weight coefficients, the basis of Eigenfaces, would certainly suggest that they could minimise their bandwidth by passing the weight vectors back to the server for classification instead of the entire image.

Joseph and Zakharia [Joseph & Zakharia, 2013] use a mounted and moveable camera controlled by a PIC micro-controller to capture images to be processed by a single computer using Matlab's Image Processing Toolbox. Unlike Shilwant and Karwankar, there is no consideration for bandwidth usage from the camera to the computer as they suggest images can be transferred on a frame by frame basis either wired or wirelessly. Like Shilwand Karwankar, feature extraction is achieved using Eigenfaces. Whilst they do acknowledge the possibility of detecting facial features such as eyes and the mouth, it is not clear how they actually detect faces.

Alia et al [Alia et al., 2013] present a semi-automated system where a camera captures the students' images as they enter the lecture theatre. The system relies on the actions of the lecturer including instructing the system to process the images and to confirm the recognition. The system extends to allow the lecturer to manually add undetected or unrecognised attendees before submitting a final attendance report to the wider administrative system. This is made possible using a drag and drop method. Detection is done by using the Viola Jones detection algorithm [Viola & Jones, 2001] but little is said with regards to the actual recognition method employed.

With this in mind, the final paper reviewed in this section concentrates solely on the adequate detection of faces in lecture theatres where detection must accommodate the differing face sizes of the students due to their distance from the camera. Taylor and Morris [Taylor & Morris, 2014] utilise, but are not restricted to, an ensemble of differing Viola Jones style cascades together with frontal face Local Binary Pattern (LBP, Section 2.6.5) features. Their goal is to detect as many potential faces as possible before filtering out the false positives.

One interesting component of their system is the use of PCA to define a gradient based framework over the lecture theatre in order to discard potential faces that are not within an expected size relative to their location. Once filtering has occurred the ensemble is flattened retaining a score for each unique face of how many cascades detected it. Final classification is then performed using the confidence indicated by the assigned score and by skin colour.

The authors note that system intentionally discards faces where only one cascade has detected it. This is because the number of false positives generated in such instances, outweighs the gain. In particular, they suggest that skin-coloured environments influence this phenomenon; although a review of the environmental conditions that led to the failure to detect the face by the majority of the cascades may provide useful insight.

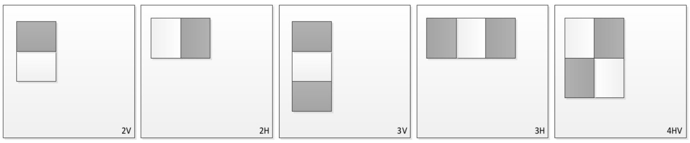


Figure 2.4: Three distinct types of Haar-like feature are defined and are identified by the number of adjacent rectangles they contain (two, three or four) and the relative positions of their rectangles. Each feature is available in both horizontal and vertical forms. Each distinctive area is a template used to compute contrast information when computed over the image space.

2.6 Prior Work

This section will describe well known techniques used within the field of face detection, verification and recognition. As building blocks, they have been ordered to reflect where they may be positioned within any recognition system.

2.6.1 The Viola Jones Object Detection Algorithm

The Viola Jones algorithm [Viola & Jones, 2001] is the de-facto standard face detection mechanism. It utilises basic and scalable geometric structures known as Haar-like features (see figure 2.4) as a template to compute the contrasting differences between distinct rectangular areas within the image space.

By using pre-defined features, domain knowledge is inherent within the model from the outset. This removes the need for any evolutionary/learning phase at the pixel level where meaningful results would be difficult to obtain, especially given a finite set of training data and time. For example, we intuitively know that the eye region is often darker than the cheeks [Viola & Jones, 2001] making it relatively easy to define a simple heuristic capable of detecting this. In fact, the boosting algorithm of Viola Jones selects such an attribute using Haar-like features. It would be a non-trivial and time intensive task for a machine learning algorithm to do likewise without this knowledge. Eon et al's work with boxlets [Eon et al., 1999] reinforces this approach by confirming pixel-level feature detection would be near impossible due to the noise in the image; although, reducing the dimensionality of the image using PCA would potentially remove noise.

A key artefact of the Viola Jones detection algorithm is the use of the Integral Image, a single pass transformation of the image space defined as

$$ii(x,y) = \sum i(x^{'},y^{'}) \,\Big|\, (x^{'} \leq x) \wedge (y^{'} \leq y) \tag{2.12}$$

where ii is the integral image, x and y form the column and row location of the summation of the pixels taken from i which is the grey-scale source image that significantly reduces the computation required to compute contrasting areas. In essence, the Integral Image is a simple summation where each cell holds the total of all the source image's pixel intensity values from the left and to the top of the current pixel, inclusive of own location.

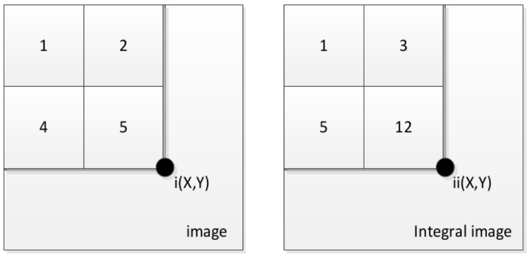


Figure 2.5: An example showing the pixel intensity values of the original source image and its translation to the Integral Image. Each cell value of the Integral Image contains the summation of the pixel values of the source image from the left and to the top of the cell in question, inclusive.

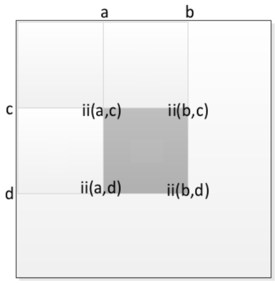


Figure 2.6: A visual representation of the use of the Integral Image to compute the total pixel intensity of an image source rectangle.

To compute the sum of pixels in a rectangular area of the image space without the use of the Integral Image would require the summation of every pixel. Equation 2.13 shows that the summation of the pixel intensity values of our target rectangle is equivalent to a simple calculation of the four corners of the Integral Image using

$$\sum_{x=a}^{b} \sum_{y=c}^{d} i(x,y) \equiv ii\,(b,d) + ii\,(a,c) - ii\,(a,d) - ii\,(b,c) \tag{2.13}$$

where i is the source image and ii is the Integral Image. This is illustrated in Figure 2.6. Furthermore, as the Haar-like features contain only adjacent rectangles of equal size, the number of cell references proportionally decrease as the number of rectangles within the feature increase.

The fundamental idea behind this algorithm is to compute the pixel intensity differences between the areas defined by the Haar-like feature to derive coarse knowledge. In the case of eye detection, a three box Haar-like feature (shown in figure 2.4) could discover potential eyes by finding areas of the image where the two outer rectangles contrast the inner rectangle significantly. By adopting the Integral Image we can compute these totals efficiently.

In the case where more than two rectangles form the feature, the summation of the areas from the same colour is calculated and used i.e. with three rectangles, two dark and one light, the sum of the areas of the dark rectangles is subtracted from the area of the light one. Importantly, Haar-like features are also scalable and it is computationally easier to

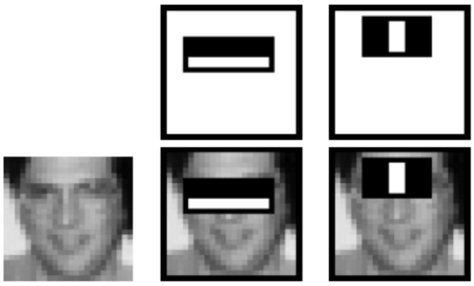


Figure 2.7: Two Haar-like features are shown to illustrate the strength inherent within the domain knowledge. Image taken from [Viola & Jones, 2001]

test scalable features than to scale the source image over many levels.

Figure 2.7 demonstrates the simplicity and strength with this approach. It should be clear from this example that the eyes are typically darker than the surrounding areas. Whilst intuitive, the system did not discover this by accident and is the result of intensive training using a modified version of Adaboost [Freund & Schapire, 1999] which selects a single feature within each weak classifier. Traditionally, Adaboost is used to select the strongest candidates from a set of weak classifiers, which perform only marginally better than 50%. It does this by putting the classifiers through a series of increasingly demanding problem solving tests and decreasing the weights of the strongest classifier at the end of each round such that the weights are inversely proportional to the strength of the classifier [Viola & Jones, 2001].

A key consideration to Viola and Jones' approach is that without the reduction achieved by training there would be circa. 45,000 possible rectangular features that could exist within the 576 pixel area of their recommended operational sub-frame, far too many to be efficient. Training only reduces the subset of features to test however. Through trial and error, Viola and Jones suggest there are no single strong features capable of detecting faces in isolation. During the latter phases of their training, most single feature classifiers only correctly detected true positives marginally better than 50%.

By chaining weak classifiers together in increasing difficulty we are able to create a set of strong classifiers. These can be tested by the third part of the Viola and Jones' detection algorithm, the attentional cascade. The cascade is a simple pruning algorithm that evaluates each element of the chain against the source image immediately discarding those that fail. Any image that successfully gets through the cascade is likely to be a face. Figure 2.8 illustrates this concept.

Another way to look at this is by considering the fact that the strongest feature selected by Adaboost training is large relative to the detection sub-frame. Features covering large areas crudely detect the attributes in which we are interested and consequently yield higher false positives rates. Given the ever-increasing level of difficulty of each stage in the chain, the cascade can afford to accommodate this whilst still remaining efficient. This is because it utilises the cheapest computational expense available at the time. For example, a large two part feature at the beginning of the chain is capable of pruning negative images quickly but tends to generate more false positives; whereas, a small four part feature consumes

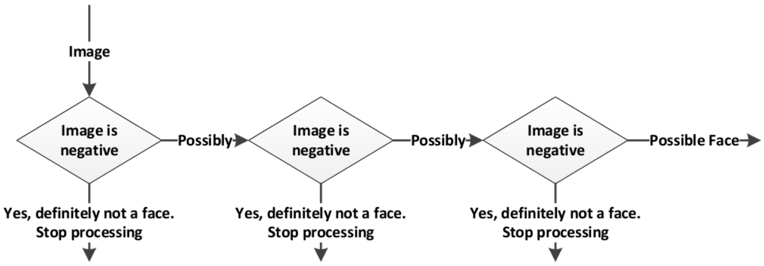


Image
Image is negative
Possibly
Image is negative
Possibly
Image is negative
Possible Face
Yes, definitely not a face. Stop processing
Yes, definitely not a face. Stop processing
Yes, definitely not a face. Stop processing

Figure 2.8: Viola and Jones' attentional cascade concept. Each link in the chain is used as a test for the source image. If it fails it is immediately discarded; otherwise it moves on to the next, much harder, link in the chain.

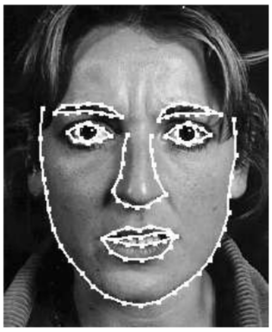


Figure 2.9: An example of an image that has been labelled ready for training [Cootes, 2000].

more resources computationally and has to be evaluated more frequently to cover the same search space, but yields a better overall level of certainty.

Once trained, the Viola Jones algorithm provides a reliable and efficient mechanism to detect faces. It exploits simple mathematical concepts to pre-process and prune the image space utilising resources proportionally to the amount of certainty the algorithm has that the image may be a face.

2.6.2 Active Shape Modelling

Active Shape Modelling (ASM) [Cootes et al., 1995] [Cootes, 2000, Chapter 7] attempts to fit a predefined set of landmarks that are be used as a deformable wire frame model of a face to the image space. ASMs are robust against scale and rotation although they require sufficient contrast to discover the correct landmarks. Training however is extremely labour intensive as it requires an operator to manually label multiple key points (or landmarks) over a large set of images. This task is made more onerous given that the number of key points must be consistent and identifiable across each and every image. Generally, the number of points used and images labelled dictate the quality of the result and should reflect the context of the model we are attempting to fit.

The ASM labelling process creates a vector of landmark pairs (x_n, y_n) per image defined as

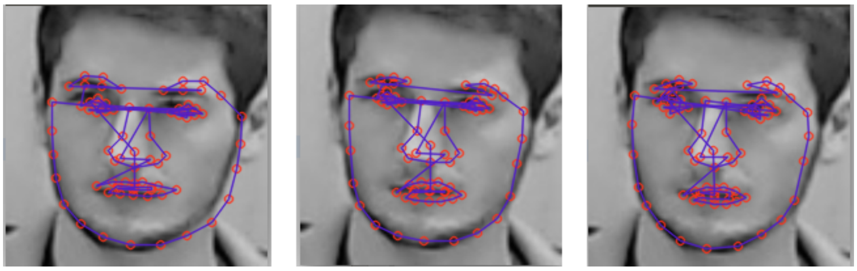


Figure 2.10: Three iterations showing the active shape modelling algorithm attempting to fit the model to the image.

$$\boldsymbol{t} = [(x_1, y_1), \ldots, (x_N, y_N)]^T \tag{2.14}$$

with each complete vector $\boldsymbol{t}$ defining the outline of the marked up face and x and y being the coordinates of each landmark. The vectors are then aligned and normalised to a common set of coordinates across all images by applying scaling, rotation and translation techniques such that we minimise the sum of residuals against a previously computed average wire frame face. Figure 2.9 shows an example of a labelled face for training.

ASM utilises PCA to reduce the dimensionality of the normalised data set to derive a set of eigenvectors describing the wire frame model in eigenspace. It can then apply weights to the eigenvectors to create scaled, rotated and transformed variants of the model in image space using

$$\boldsymbol{t} \approx \bar{\boldsymbol{t}} + V\boldsymbol{b} \tag{2.15}$$

$$\boldsymbol{b} = V^T(\boldsymbol{t} - \bar{\boldsymbol{t}}) \tag{2.16}$$

where V is the matrix whose columns are the retained principal components, $\bar{\boldsymbol{t}}$ is the mean of $\boldsymbol{t}$ and $\boldsymbol{b}$ (equation 2.16) is the set of weights varying the shape. Each shape is thus characterised by a set of coefficients in the reduced dimension PCA space.

The weight parameters are range bound to within a pre-defined number of standard deviations to ensure that the deformed shape remains broadly in line with the expected model. A simple iterative strategy using these variations can then be applied to the target image to find a suitable overlay. Each landmark of the deformed wire frame is tested for localised contrast information and the collective sum of this landmark information yields its likely fit [Chenxing, 2012]. Figure 2.10 demonstrates the deformed wire frame adjusting to the image space in order to find a suitable fit.

2.6.3 Eigenfaces and Fisherfaces

Eigenfaces [Sirovich & Kirby, 1987] [Kirby & Sirovich, 1990] [Turk & Pentland, 1991] and the subsequent Fisherfaces algorithm [Belhumeur et al., 1997] are natural progressions of

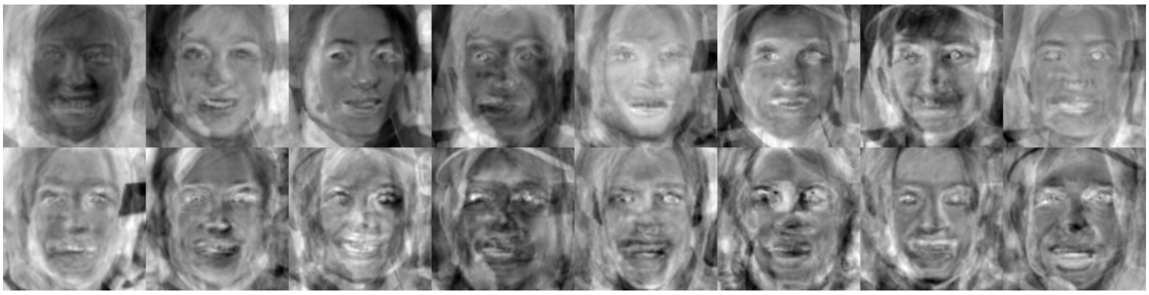


Figure 2.11: A set of principal eigenfaces derived from computing the PCA eigenvectors from the baseline Labelled Faces in the Wild aligned data set.

PCA (see Section 2.4.1) and LDA (see Section 2.4.2) respectively. During training they both reduce the dimensionality of the data set to extract variance-based features in the form of eigenvectors which can then be projected upon to derive a set of weight coefficients that describes the current aligned face in the new space e.g. Eigenfaces uses the principal components to project the aligned face into eigenspace.

In the context of image processing, PCA and by extension, the Eigenface approach, is known to be sensitive to the effects of lighting conditions. Belhumeur et al demonstrate that excluding the first three major principal components mitigates this sensitivity to some degree because these components capture illumination variation. Clearly, removal of major components risks losing useful feature descriptions in the process.

As an example, figure 2.11 contains the first 16 eigenfaces computed from the baseline 19 (resized to 120 by 120 pixel) face images taken from the Labelled Faces in the Wild aligned data set (Figure 2.1). The most significant component is shown in the top left image and decreases over the rows such that the bottom right image contains the least significant information of those components used. It is clear from the top left image that the main component has captured overall illumination information, and the second represents horizontal variations in illumination.

Figure 2.12 illustrates a graphical method to prune the dimensionality of the resultant eigenspace. It shows the eigenvalues computed from using PCA in decreasing variance where the lower the variance the less relevant the dimension is to the structure of the data. A decision can be made using

$$\frac{\sum_{m=1}^{M} \lambda_m}{\sum_{n=1}^{N} \lambda_n} > threshold \tag{2.17}$$

where the ratio of eigenvalues λ_m over the total sum of eigenvalues λ_n is compared against a threshold in order to capture the desired proportion of variance using M eigenfaces. By comparing the ratio of M potential components over N total components the desired number of components capable of exceeding the threshold can be found allowing for the remaining components, of lesser importance, to be pruned accordingly.

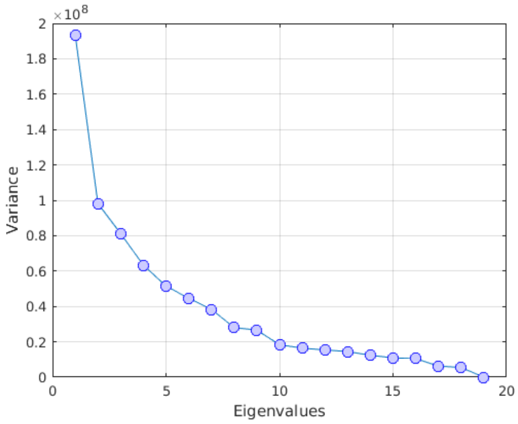


Figure 2.12: A visual method to review the variance of each eigenvector. The graph plots the eigenvalues for each dimension where the greater the variance the more relevant the associated eigenvector is to the structure of the data.

By placing the resultant eigenvectors into a column matrix V any x_n image can be projected onto the newly computed eigenspace using

$$\boldsymbol{w_n} = x_n \cdot V \tag{2.18}$$

in order to derive a vector of weight coefficients of the form

$$\hat{\boldsymbol{w}} = (w_1, \ldots, w_m)^T \in \mathbb{R}^M \tag{2.19}$$

w in equation which is a lower dimensional representation assuming M is less than the dimension D of the original images. By finding the M coefficients for all the x_n images within the training set it is then possible to project a new face onto the eigenspace and determine its position relative to the others using a simple distance metric. Furthermore, it is possible to reform an approximation to the original image by applying the weights to their respective eigenvectors, summing and reapplying the mean face by

$$\tilde{\boldsymbol{x}} = \sum_{m=1}^{M} w_m v_m + \bar{\boldsymbol{x}} \tag{2.20}$$

where M is the total number of components used. An example reconstruction is shown in 2.13 and the associated mean face $\bar{\boldsymbol{x}}$ of the training set shown in figure 2.14. The reconstruction shows $m = 1$, the primary component, as the top left hand image with the mean face added increasing row by row where the bottom right hand image is $m = 16$ plus the average face. The total number of images N used in this example is nineteen and the image space dimension D is 120 by 120 and clearly demonstrated that 16 dimensions in eigenspace accurately represents a face of 14400 dimensions.

The mean face is calculated in order to centre the data set prior to computing the eigenvectors. By subtracting the mean face we are left with a set of faces where only their

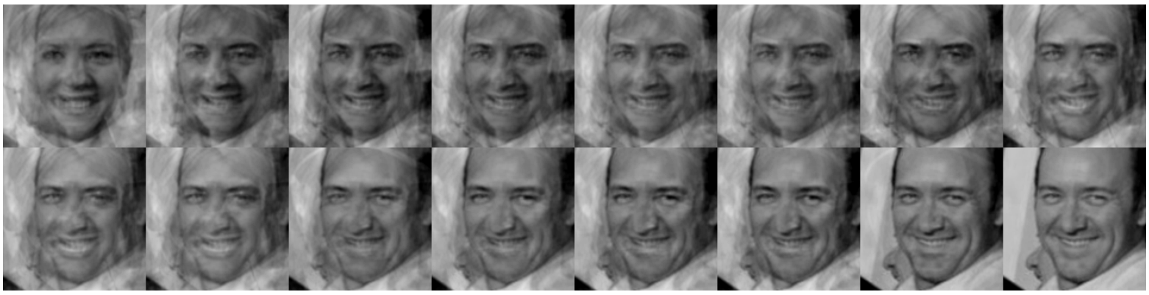


Figure 2.13: A step by step application of the weighted eigenvectors being added to the mean face in order to reform the original image of Kevin Spacey.

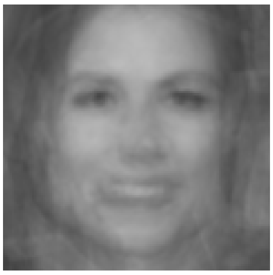


Figure 2.14: The mean face is derived from computing the average face across the entire data set.

distinct features remain. It should be clear that applying the coefficients to the vectors of the eigenspace will only rebuild our centred data and so it is necessary to add the mean face back on.

A number of proposed student attendance monitoring systems utilise Eigenfaces. Whilst not explicitly stated, they assume consistent illumination of the subjects. Joseph and Zakharia [Joseph & Zakharia, 2013] conclude that illumination was not a key factor during their experiments whilst Shilwant and Karwankar [Shilwant & Karwankar, 2012] suggest that advances in computer graphics has led to better preprocessing algorithms that can compensate for illumination variation accordingly.

2.6.4 Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) is a technique [Lowe, 1999] to extract features, which are invariant to scale and rotation, that form descriptors of the region of interest. SIFT has been widely used for the purpose of face recognition including [Luo et al., 2007], [Majumdar & Ward, 2009], [Mohamed Aly, 2006] and [Lenc & Krl, 2013]. It utilises a set of variable scale Gaussian functions that are convolved with the image to define a group of smoothed images in scale space. This grouping of scales is shown in figure 2.15. Adjacent Gaussian images in the octave are subtracted from each other to produce difference of Gaussian (DoG) images.

Furthermore, a pyramid-like scaling of octaves is produced by down sampling each convolved image. In Lowe's later work [Lowe, 2004] each level in the pyramid is computed by dropping alternate pixels. In this case, bilinear interpolation is used to scale the initial image as the first tier. Importantly, this does not fundamentally alter the main functionality

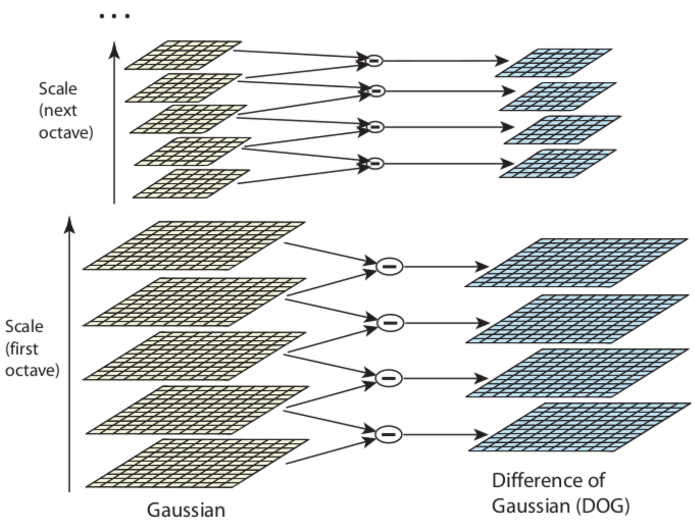


Figure 2.15: SIFT octaves and scaling pyramid. The original image is convolved with a set of Gaussian functions to produce an octave which is then down sampled for scaling purposes. Image taken from [Lowe, 2004].

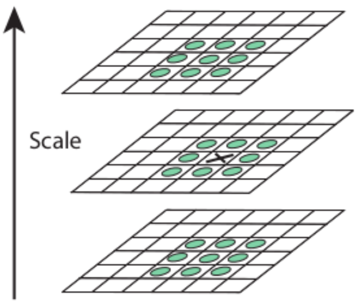


Figure 2.16: Each pixel is compared with its neighbouring pixels to find local extrema pixels. This includes comparisons with pixels in the adjacent layers. Each extrema pixel is a potential key point. Image taken from [Lowe, 2004].

of the process but provides an efficient scaling mechanism.

Once the difference of Gaussians are defined, a comparison is made with each pixel against its neighbours from a three by three neighbourhood to determine localised minima and maxima. Assuming the pixel is the extrema at its base layer, the neighbourhood is extended to include pixels in the adjacent layers of the pyramid, taking into consideration the sampling differences between the octaves. Pixels that remain extrema are potential key points that are considered to be best represented at that scale. A simple illustration of the neighbourhood can be seen in figure 2.16.

A contrast threshold test is performed on each key in an attempt to eliminate noise and provide better stability within the data set. Lowe states that it is also necessary to remove edge based keys due to the strong response found along the edges. The result is a set of keys identifying local extrema locations together with their gradient magnitude and orientation.

The SIFT keys are made invariant to translation, scaling and rotation by modelling biological equivalent functions. This is achieved by allowing a key position to to vary within

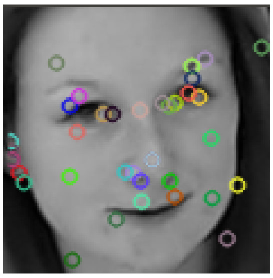
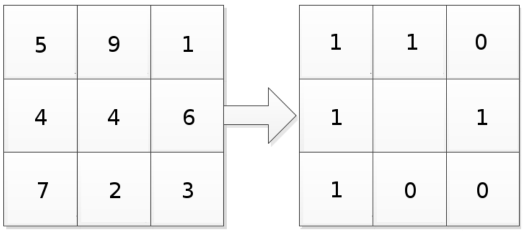


Figure 2.17: SIFT key points detected within the region of interest. Each key point has an associated descriptor describing its invariant context.



5	9	1
4	4	6
7	2	3

1	1	0
1		1
1	0	0

Figure 2.18: An LBP code of 11010011 derived from comparing the intensity values of each pixel against the centre pixel. The value 5, our sequence start, is greater than the middle value 4 yielding a bit value of 1 based on equation 2.21. The code is defined by comparing the pixels clockwise.

a small neighbourhood whilst its orientation and spatial frequency is strictly maintained. Lowe uses orientation planes based on this to form 128 histogram bins as the descriptor.

The final stage is to produce an index of the trained keys with their associated descriptors in order to match keys found in new images. SIFT uses a *best-bin-first* algorithm (based on the k-d tree algorithm) to determine the nearest neighbour as an approximation to the solution. Hough transformations are also used to cluster a series of keys into fragments that can be used to recognise objects even if they possess a large numbers of occlusions. Figure 2.17 shows a set of stable SIFT keys extracted from a face.

2.6.5 Local Binary Patterns (LBP) and Variants

Local Binary Patterns (LBP) [Ahonen et al., 2004] [Yang & Chen, 2013] are feature extractions derived by comparing pixel intensity values relative to their neighbours in image space. The technique is based on the texture mapping method developed by Wang and He [Wang & He, 1990]. They are simple to compute and are invariant to monotonic transformations of intensity given that the comparison would still hold. Figure 2.18 illustrates an encoding based on eight adjacent neighbours and produces a binary LBP code of 11010011 using

$$LBP = \sum_{i=1}^{8} \begin{cases} 0 & P_i < P_0 \\ 2^{i-1} & \text{otherwise} \end{cases} \tag{2.21}$$

where P_0 is the pixel of interest and P_i is an adjacent pixel where the pixels are labelled

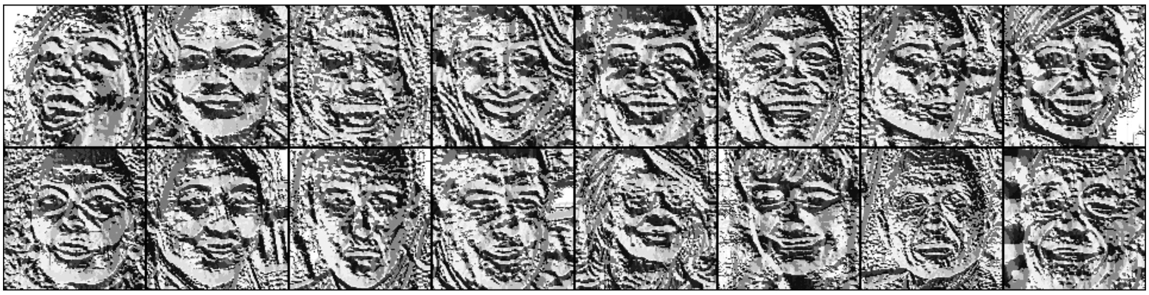


Figure 2.19: An encoding of the baseline LFW data set using the LBP technique.

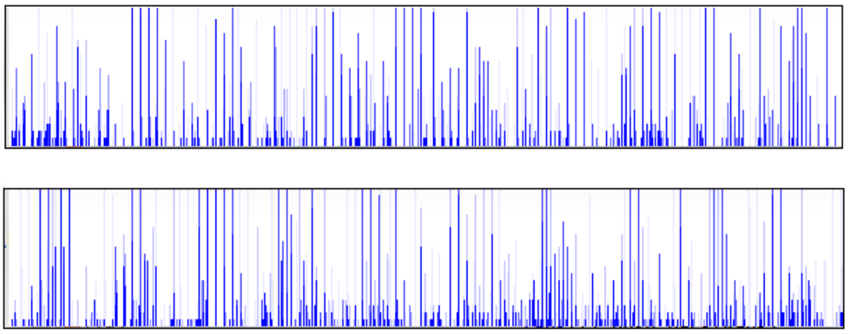


Figure 2.20: Local Binary Pattern (LBP) histograms of a target and candidate for subsequent distance based comparison.

in a clockwise direction from the top left pixel. By using bilinear interpolation we are not restricted to adjacent pixels and can choose our neighbours at any radius and select the number of neighbours that is most appropriate.

The complete LBP encodings for our baseline LFW data set can be seen in Figure 2.19 where each pixel is displayed as a grey scale value corresponding to its LBP code interpreted as an 8-bit integer. They depict obvious features and suggest that LBP is a simple noise reduction technique.

Ahonen et al [Ahonen et al., 2004] use histogram bins as feature descriptors. To preserve spatial information, the LBP codes are segregated into regions to form regional histograms which are then concatenated as a representation of the *global face*. Figure 2.20 shows two histograms, H_1 and H_2 formed from the concatenation of regions of 16 pixels. They can be compared for likeness using a variety of distance measures such as the Chi-square distance:

$$d(H_1, H_2) = \sum_i \frac{(H_1(i) - H_2(i))^2}{H_1(i)} \tag{2.22}$$

where $d(H_1, H_2)$ is the 'distance' between the two histograms and $H_n(i)$ is the i^{th} histogram bin. When $H_n(i)$ is zero the denominator is set to 1. Additional distance measures could include correlation and intersection techniques. Furthermore, by weighting local regions one can bias the comparison to areas that are more relevant to face recognition than others e.g. areas containing eyes.

Empirical evidence collected by Ahonen et al shows that most 8 bit LBP codes fall within

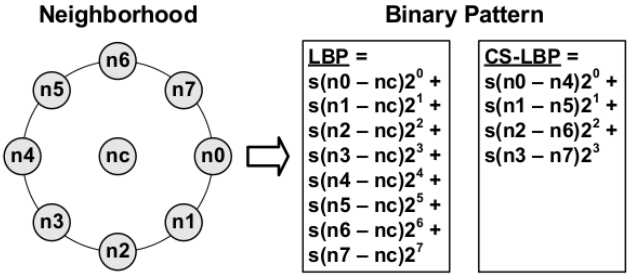


Figure 2.21: The CS-LBP code is produced by comparing the pixel values of symmetrical pairs. e.g. if n_0 is greater than n_4 then the bit code is set to 1 otherwise it is set to 0. This figure, taken from [Heikkilä et al., 2006], shows the methods used for both CS-LBP and LBP encoding.

a small subset of uniform patterns i.e. those classifications with no more than two binary transitions across the encoded pattern, such as 00000000, 10000011 and 00011100. By compartmentalising each LBP into their respective uniform pattern, with an additional compartment set aside for the remaining non-conforming patterns, a histogram of local micro-patterns can be created showing the distribution of enhanced features such as edges, spots, line ends, corners and flat areas over the entire image. Uniform patterns are also rotation invariant.

In 2006, Heikkilä et al introduced Centre Symmetric Local Binary Patterns (CS-LBP) [Heikkilä et al., 2006] as a technique that combined the strengths of both LBP and the difference of Gaussian SIFT algorithm. Instead of comparing neighbouring pixels with the central pixel, CS-LBP compares symmetrical pairs of pixels with each other using

$$CS - LBP_{N,T} = \sum_{i=0}^{(N/2)-1} s(n_i - n_{i+(N/2)})2^i \tag{2.23}$$

$$s(d) = \begin{cases} 1 & d > T \\ 0 & \text{otherwise} \end{cases} \tag{2.24}$$

where N is the number of equally spaced neighbours to be evaluated with n_i being the current pixel and $n_{i+(N/2)}$ its diagonal opposite. The threshold value T must be exceeded by the pair difference d in order for a bit value encoding of 1.

Figure 2.21 defines the encoding techniques for both CS-LBP and LBP. It shows that CS-LBP encoding is as a result of comparing diametrically opposing pixel pairs; whereas, LBP is encoded using the centre pixel for comparison. The major advantages of the CS-LBP approach are two-fold; the speed of computation is improved due to the reduction in the number of comparisons [Meena & Suruliandi, 2011] and the resultant CS-LBP code has lower dimensionality because each pixel pair is encoded as opposed to LBP which encodes each pixel. Furthermore, CS-LBP not only preserves distinctiveness but captures better gradient information.

The application of the CS-LBP technique can be seen in figure 2.22. Once more, this is an

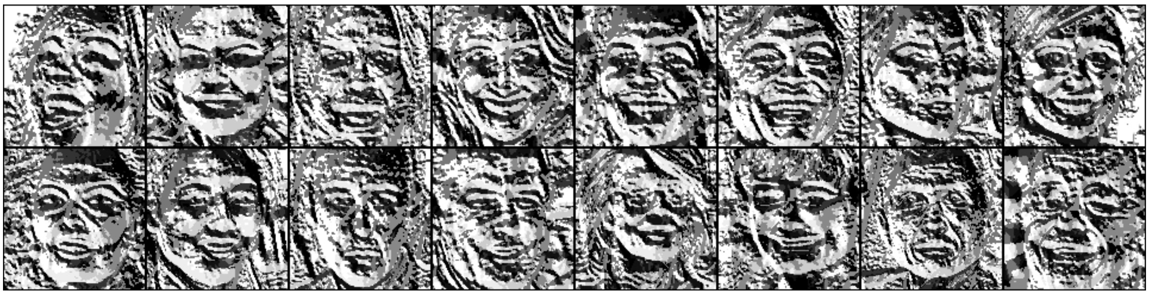


Figure 2.22: A normalised to 8-bit grey scale encoding of the baseline LFW data set using the CS-LBP technique.

extraction of features from our baseline LFW data set from figure 2.1. As with the related LBP, the resultant codes are stored in region based histograms as feature descriptors. In order to remove the influence of large gradient magnitudes the histogram bins have a 0.2 threshold applied before being renormalised to unit length [Pietikinen, 2011, Chapter 5].

Within the context of face recognition systems, Liu et al proposed a multiple pass approach that simply reapplied CS-LBP iteratively as they found a single pass computation did not contain enough facial information [Liu et al., 2011]. It was found especially true in areas of the face such as the forehead or cheeks i.e areas that are smooth. Important face areas are given the appropriate weighting and this is applied dependent on each specific pass.

Eigenfaces Applied to Both LBP and CS-LBP

Principal Component Analysis can be applied to Local Binary Pattern based techniques. Even if our goal is not to directly reduce dimensionality, projecting our LBP encoded image into eigenspace will allow us to derive a set of weights that can be used as a descriptor for comparison purposes. Furthermore, it also allows for near perfect reconstruction depending on the number of principal components we retain. Lei et al utilise LBP before projecting the encoding into eigenspace [Lei et al., 2014]. In doing so they encode both the local, using LBP, and global structures of the face (using PCA). When tested against k-nearest neighbour classifiers using either LBP or Eigenfaces based features, Lei et al demonstrate that the combined method offers significant performance improvement.

Figure 2.23 shows the reconstruction of an LBP encoded image from eigenspace using only its projected weights. This approach can also be applied to CS-LBP.

2.6.6 k-Nearest Neighbour (kNN)

The k-Nearest Neighbour classification [Webb, 2003] [Fukunaga, 2013] permits the categorisation of a datum based on its proximity to previously classified data. It is a simple strategy that assesses the datum's location in the classification space relative to the classes. If k is set to one then the class of the data point is simply assigned to the same class as its closest neighbour. Where k is greater than one the classification is performed according to the majority class type of the surrounding k neighbours. Typically, the value of k should be set in context of the classification task and on the understanding that whilst a large

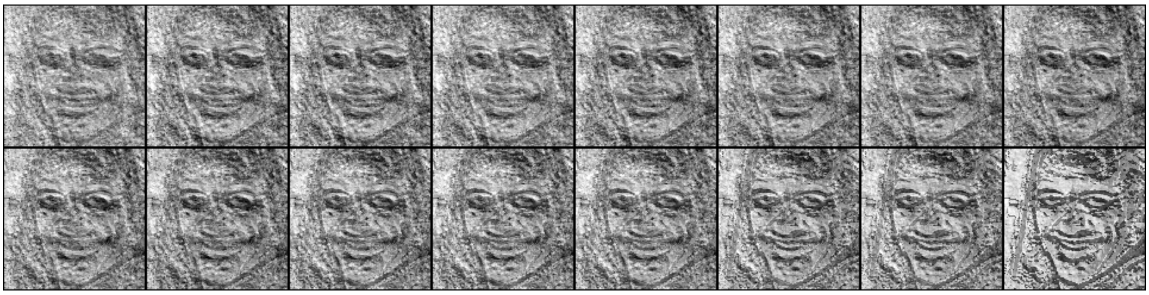


Figure 2.23: A reconstruction of an LBP encoded face that has had PCA applied to it. The top left hand image shows the summation of the average LBP face with the dominant principal component, going left to right the other components are added until we finish with the bottom right hand reconstructed LBP version. To show a direct comparison, the image of Kevin Spacey is the same as the one used in Figure 2.13

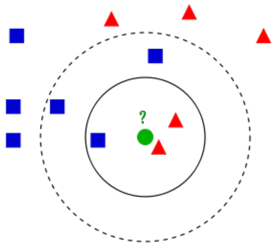


Figure 2.24: An example of a kNN classification. Using the smaller circle, which shows the region occupied by 3 nearest neighbours, the green circle would be classified as part of the red triangle class. However, as the radius is increased to the dotted circle, the classification changes to reflect the new majority (blue squares). Image taken from Wikipedia.

k removes noise it also smooths out the class boundaries. Additionally, an odd number should be selected for k to avoid tied votes especially in the case of two classes. In the experiments performed in this thesis, the distance metric used is dependent on the feature extraction technique employed e.g. Chi-Square for histogram based techniques.

Figure 2.24 shows an example classification under varying k conditions. It shows two classes, red triangles and blue squares and the location of a green circle in the classification space. Where $k = 3$ the green circle is classed as part of the red triangle class. When k is increased to five, the ratio of blue squares to red triangles increases such that the green circle is considered a member of the blue square class. To avoid bias, where one class outweighs the others because there are more data points in the group, each neighbour's vote can be weighted. This can be achieved in numerous way including computing the inverse proportion of the number of data points in the class over the total number of data points, so that if there are N_c training examples in class c, then each member of the class carries weight $\frac{1}{N_c}$.

A kNN classifier is a non-parametric learning algorithm that does not make assumptions about the probability distribution of the data and is easy to implement. However, it is sensitive to local data structures that could be erroneous, places the majority of its computational time to point of classification as opposed to within the training phase and can be saturated in higher dimensions where the datum to be tested is close to too many classes.

2.6.7 Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised learning algorithms that attempt to find an optimal hyperplane or decision boundary with the maximum separation between classes [Cortes & Vapnik, 1995] and can be used for classification. Figure 2.25 shows an optimal hyperplane partitioning two example classes, positive and negative respectively. The samples that reside closest to the boundary are the support vectors and form a minimal subset of the sample space.

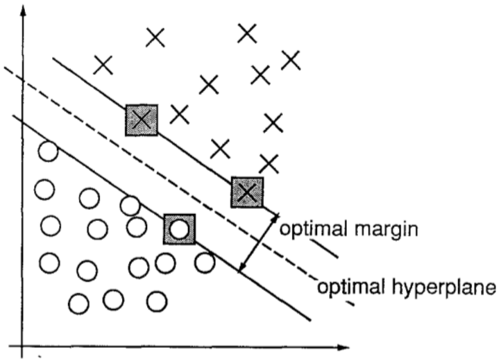


Figure 2.25: A two state classification space that highlights the decision boundary that separates the classes by the maximum distance. Image taken from [Cortes & Vapnik, 1995].

Using two classification states as an example it is possible to define a decision rule that will classify any sample as being either positive or negative. A typical rule is shown by

$$\boldsymbol{w} \cdot \boldsymbol{u} + b >= 0 \begin{cases} TRUE & \text{positive classification} \\ FALSE & \text{negative classification} \end{cases} \tag{2.25}$$

where $\boldsymbol{u}$ is our unclassified sample, $\boldsymbol{w}$ is perpendicular to the decision boundary and the scalar b defines our boundary. By projecting $\boldsymbol{u}$ onto a vector that is perpendicular to the boundary we are then able to discover which side of the boundary it resides.

Initially the vector $\boldsymbol{w}$ and corresponding scalar b which describe our maximum boundary conditions are unknown. By defining a set of constraints it is possible to use a Lagrange multiplier and solve its partial derivatives to find the extrema. The constraints necessary for this are derived using

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) - 1 >= 0 \tag{2.26}$$

and

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} + b) - 1 = 0 \tag{2.27}$$

where y_i is +1 for a positive sample x_i and -1 for a negative sample. In particular, those samples that reside at the edges closest to the boundary are those that equal zero in the constraint.

Given a positive x_+ and negative x_- sample that both reside on the boundary edges it is possible to compute the width of the boundary by projecting the difference between x_+ and x_- unto a unit normal version of $\boldsymbol{w}$. In fact, this simplifies to

$$\frac{2}{||\boldsymbol{w}||} \tag{2.28}$$

and shows that in order to maximise the width we actually need to find the minima of the magnitude of $\boldsymbol{w}$.

The Lagrangian L is described by

$$L = \frac{1}{2}||\boldsymbol{w}||^2 - \sum \alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1]_+ \tag{2.29}$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum \alpha_i y_i \boldsymbol{x}_i = 0 \tag{2.30}$$

$$\frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0 \tag{2.31}$$

[Winston, 2010] where the partial derivatives for both $\boldsymbol{w}$ and b can be solved.

Equations 2.30 and 2.31 can be re-written and highlight that the vector $\boldsymbol{w}$ is a linear sum of samples of $\boldsymbol{x}$ where the Lagrange multiplier α is not zero. Ultimately by substituting the equations of 2.32 and 2.33 back into the Lagrange we determine that the extrema is dependent on the dot product of pairs from the sample space.

$$\boldsymbol{w} = \sum \alpha_i y_i \boldsymbol{x}_i \tag{2.32}$$

$$\sum \alpha_i y_i = 0 \tag{2.33}$$

Furthermore, it is possible to find the support vectors for classification problems that do not appear to be linearly separable by applying a kernel function K to transforms the domain into a different dimension before attempting to find the optimal decision boundary.

$$K(u, v) = (u \cdot v + 1)^d \tag{2.34}$$

This can be seen in figure 2.26 which gives two examples where it is not possible to determine an optimal hyperplane without transforming the samples into another dimension. In both cases a linear kernel function (equation 2.34) has been employed where $d = 2$. In

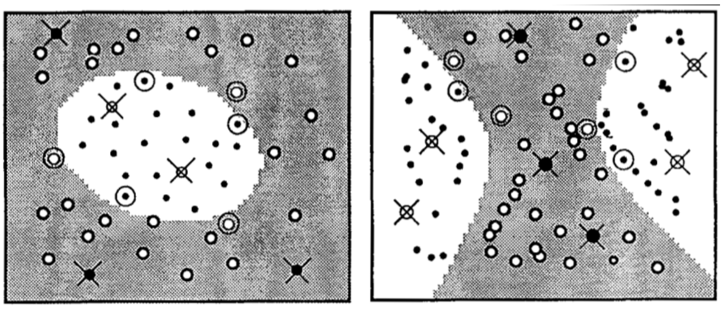


Figure 2.26: Two examples where it is not possible to draw an optimal hyperplane in the originating dimension. Image taken from [Cortes & Vapnik, 1995]

essence, the kernel function translates the dot product of sample pairs into a higher space in order to find the maximum boundary.

2.7 Conclusion

This chapter was predominantly a literature review that considered existing systems and defined the building blocks necessary for an attendance monitoring system. It began by giving an overview of the three main notions of the field (detection, verification and recognition) before enabling the reader to visualise the techniques to achieve them. With regards to face detection, it described the de-facto standard Viola Jones algorithm whilst reinforcing its position by demonstrating its ubiquitous use as the source detector in the Labelled Faces in the Wild public face database.

The chapter considered dimensionality reduction and suggested that achieving this was highly desirable. It was shown that Principal Component Analysis underpinned several of the building blocks subsequently discussed within the chapter. It was used to assist in the definition and fundamental deformation of the wire-frame model in Active Shape Modelling, in its raw form to extract the features as used in Eigenfaces and as an extension of the Local Binary Pattern family.

Dimensionality reduction remained a constant theme throughout as low dimensional embedding without significant structural loss makes efficient input into classifiers such as the k Nearest Neighbour algorithm. Other classifiers such as Support Vector Machines (SVM) and Neural Networks can be used effectively.

3 Attendance Monitoring System

3.1 Introduction

This chapter will define the requirement for monitoring student attendance using face recognition and will highlight why this technology could be well suited to the application. It will introduce and define a standard recognition pipeline which utilises the techniques outlined in Section 2.6. Specifically, it will define the four stages of the pipeline; detection, alignment, feature extraction and classification, mapping the techniques appropriately to each stage.

The latter part of the chapter will specify the proposed attendance monitoring system. It will incorporate the aforementioned pipeline and employ it to discover and process faces taken from multiple images for any lecture where the list of registered attendees are known in advance. It will include two approaches for its use: in the first each captured face image is assigned to the identity of the library image that it best matches; in the second iterative procedure, the identity of the best matching face image is assigned and that identity with all associated data excluded from further consideration.

3.2 Student Attendance Monitoring

An ideal mechanism for capturing student attendance would be where the attendees and lecturer assume passive roles. The lecturer would be able to concentrate on delivery whilst the attendees listened. Face recognition systems facilitate this to some extent by acting as a third party in the room. Theoretically, attendance monitoring should be one of the simple recognition/verification problems. This is because it is an event that can be well defined given that the possible attendees are typically known prior to the lecture. Equally, the number of attendees are usually small reducing the overall classification space considerably.

A system that incorporates face recognition is not without its challenges however. Environmental conditions, the ability of the camera to take a set of images at a suitable resolution and the poses and expressions of the attendees themselves all contribute to the difficulties the system may encounter. The experiments performed within this thesis assume that the environmental and resolution considerations are satisfied leaving the system to resolve the issues created by the attendees.

The system may also need to take a set of images and not just a single snapshot. This is to combat situations where the field of view is not sufficient to capture the entire audience

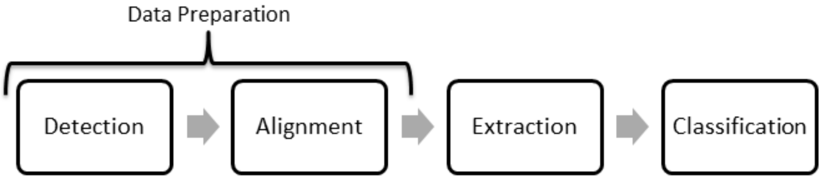


Figure 3.1: The fundamentals of a face recognition pipeline. The first two stages are concerned with processing the data. Namely, filtering non-faces out of the pipeline followed by alignment. The latter stages extract the features before classification.

at significantly high resolution and to increase the likelihood of good poses across the audience; although, the success of this cannot be guaranteed. Conversely, this further contributes to the difficulty faced by the system as it may capture the same attendee in different positions and/or poses.

Ideally, the system should be able to detect and process multiple attendees contained within a single image and do this iteratively for all available images of the lecture in question. It must attempt to match the attendees, who will be captured in an unconstrained manner and therefore in potentially differing poses, to a list of expected participants. Finally, the system should be able to cope with cases where an attendee is captured over several images, again with the potential of having a different pose or expression.

3.3 Verification/Recognition Pipeline Definition

Typical face recognition systems are based upon a four stage pipeline; see Figure 3.1. The initial stages are concerned with data filtering and preparation where the detection stage detects faces and the alignment stage aligns them, whilst the latter are used to perform the feature extraction and subsequent classification. Filtering implies an obvious fact that is worth overstating; even in the case where there are multiple faces to be detected, the faces are a small component of the overall image. It follows that any system should be designed to exploit this characteristic early on if it wishes to capitalise on this point. Coarse filtering (or basic dimensionality reduction) is achieved using a face detection algorithm to sift out unimportant areas of the space.

3.3.1 Face Detection

A face detector, placed at the beginning of the pipeline, makes an effective filter to focus the system. Figure 3.2 is a typical image that not only shows four students but also a vast amount of the image space that is not of interest. Clearly, it is desirable to remove the unwanted areas and concentrate solely on the points of interest that show promise.

One such detector is the aforementioned Viola-Jones algorithm which is described briefly in Subsection 2.6.1 [page 24]. Once trained it is able to rapidly evaluate the image space against sets of Haar-like features in order to discover promising regions of inter-

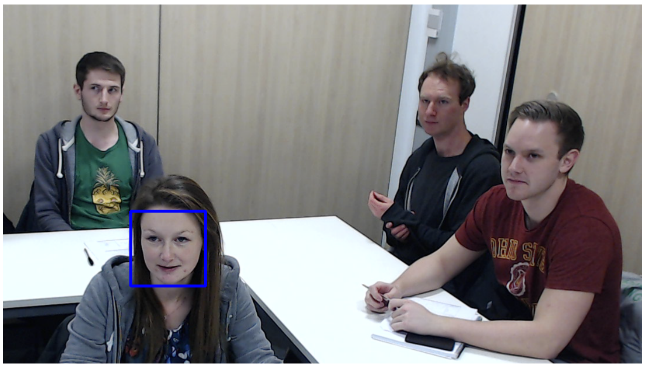


Figure 3.2: A face that has been identified using the Viola-Jones algorithm.

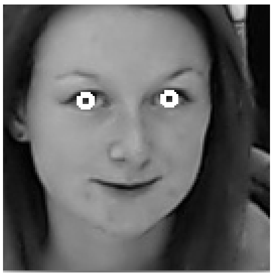


Figure 3.3: Viola-Jones redeployed to detect pairs of eyes for alignment. A simple two dimensional rotation can be applied to align the eyes.

est. In comparison with other detectors ([Rowley et al., 1998], [Schneiderman, 2000] and [Roth et al., 2000]), Viola and Jones note that the algorithm presents a higher detection rate relative to the number of false positives it yields [Viola & Jones, 2001]. Figure 3.2 illustrates the output of the detection algorithm by showing a blue box around the first of several detected faces in the image.

The Viola-Jones algorithm is not without its limitations as it was trained, in this instance, for forward facing and unobstructed faces. Consequently faces that lie outside these bounds are difficult to detect. Given the nature of a lecture room, especially where the students do not necessarily face forward, this is certainly challenging.

3.3.2 Alignment

The output from the detection phase is a set of potential faces that require subsequent alignment. Alignment is necessary because some of the feature extraction techniques used later in the system use localised information. Others, such as Scale Invariant Feature Transform (SIFT) (Subsection 2.6.4 [page 31]), are capable of feature extraction irrespective of rotation, scale or location but may still benefit from alignment due to consistency. With alignment, we simply obtain geometric information in order to move the *probable* face into the desired position at the correct angle and scale.

With appropriate training, the Viola Jones algorithm can be redeployed during this stage to detect obvious features to align. Figure 3.3 shows the effectiveness of the algorithm when trained to detect eyes. From an alignment perspective, the easiest features to align for a two dimensional image are the eyes and this is achieved by a simple rotation to

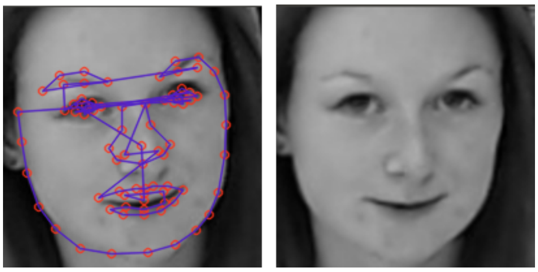


Figure 3.4: An attendee whose face has been modelled using ASM and subsequently aligned. Once a face has been modelled using ASM then specific features can be examined in detail with great confidence e.g. the person's mouth.

place the eyes on the same horizontal line. Eyes alone cannot be relied upon however, consider where a student is wearing dark glasses or their eyes are obscured by hair. Other techniques include the use of Active Shape Modelling (ASM) (Subsection 2.6.2 [page 27]) which can be used to fit a face model to the region of interest using a predefined model of the face.

Alignment stages that utilise ASMs have great potential because they *can* identity every major facial attribute. Once the model has been fitted to the target it is a trivial task to perform alignment as we have captured a rich data set of facial information. Furthermore, we can strip the background from the image to remove unwanted noise during feature extraction. The background associated with a face is a particular problem for unsupervised extraction techniques such as Eigenfaces. Figure 3.4 shows an alignment that has been performed using information derived from Active Shape Modelling.

3.3.3 Feature Extraction

Once the region has been aligned, prominent features can be extracted; not necessarily those that are obvious to our eyes. At this point the potential face resides in high dimensional space and a key requirement of feature extraction is to extract the salient key features of the face without significant loss of its structure.

Section 2.6 [page 24] gives an overview of the extraction techniques that will be used for the experiments contained within this thesis. Broadly speaking, Eigenfaces will extract directions that possess significant variance over the image, Scale Invariant Feature Transform (SIFT) will extract and cluster extrema in the image by using Differences of Gaussians filtering; and Local Binary Pattern (LBP) techniques will transform the image based on the contrast of pixel intensities.

As suggested earlier, LBP methods that extract features and represent them as a concatenation of regional histogram segments are particularly sensitive to spatial offsets and benefit from alignment. Whilst the histogram segments can be increased in size to mitigate some misalignment, it is done at the expense of the resolution of the final histogram i.e. by enlarging the area spatially we lose specific feature resolution which could impact classification. Other techniques such as SIFT and Eigenfaces are less affected by location issues. Collectively, the feature extraction techniques have been selected in order to provide a rich set of features from differing perspectives for classification.

3.3.4 Classification

During the final stage of the pipeline the features are classified. In all cases a distance metric can be computed between the training data and the target. Chi-square has been used in these experiments. Whilst the other techniques are much more direct and obvious, the key points extracted using SIFT are processed further using OpenCV's brute force match functionality which uses a Euclidean distance based k nearest neighbours to match each descriptor.

A simple technique to determine a datum's classification is to utilise the k-nearest neighbour (kNN) algorithm (Subsection 2.6.6 [page 36]) and categorise the target relative to its neighbours in the training set. The nearest neighbour algorithm simply assigns the classification of the datum by looking at the majority class that the point is closest to.

Additionally, the bias associated with the number of images per student is removed by weighting each image in inverse proportion to the number of images of that student in the training set. This is done because of the potential of a labelled student to incorrectly affect the outcome mainly because they have more images in the training set and therefore, potentially more matches. In an ideal world, there would be an equal distribution of images among identities but this is difficult to obtain due to the way in which the training set has been captured; that is, the training set is derived from each student's actual attendance and not any formal registration.

Whilst using an odd value for k in the nearest neighbourhood to specifically avoid ties, the addition of the weighting could lead to such an outcome e.g. student A has 2 images of 0.50 value and student B has 4 images of 0.25 value. In this case, the system should select the student who has the most matching images in the neighbourhood i.e. student B.

Figure 3.5 demonstrates the use of the kNN algorithm when applied on SIFT and histogram based LBP and CS-LBP. It shows the detected student and three rows of their closest matches, one row for each feature type.

3.4 Recognition Algorithm

The student recognition system, as outlined in Section 3.2, primarily uses the recognition pipeline as a mechanism to build a classification space from expected students for subsequent matching with the attendee data captured from the lecture. This section formalises the system.

1. A labelled training set $\{(x_n, t_n)\}, n = 1 \ldots N$, comprising expected attendee images is processed through the recognition pipeline up until the point of classification, that is the resultant data forms the classification space for subsequent recognition and is defined by

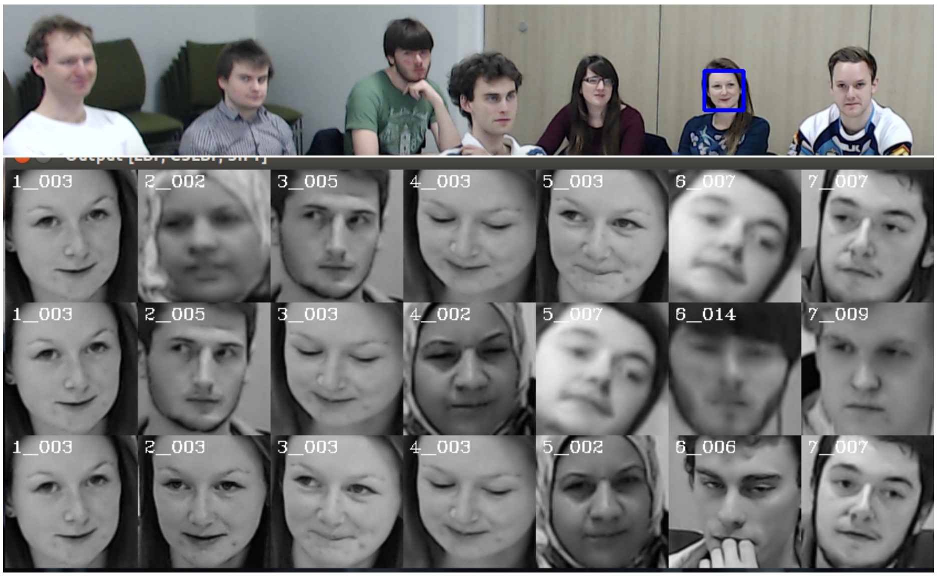


Figure 3.5: An example of a Viola Jones detected face which is highlighted with a blue square. The image also shows the 7 nearest neighbours results for three feature extraction techniques in the three rows below (LBP, CS-LBP and SIFT respectively). Each column represents a distance ranking from best (left) to worst. In this instance SIFT clearly gives the best results as it has placed the detected face near 4 correct student matches. An attempt to align the faces by their eyes was tried and future work is needed to perfect this.

$$\mathcal{D} = \{(F_n, t_n)\}_{n=1}^{N} \tag{3.1}$$

where F_n is the extracted features and t_n is the identity of the face.

Additionally a set of weights

$$\boldsymbol{b} = \{\frac{1}{N_1}, \frac{1}{N_2}, \ldots, \frac{1}{N_L}\} \tag{3.2}$$

is created to remove the bias caused by the potential of an unbalanced number of labelled images within the training set (see Section 3.3.4). These are inversely proportional to the number of labels t_n contained within the training set for each member of the set of unique labels L.

2. The images captured during the lecture are processed using the pipeline to detect faces and extract their features defining

$$\mathcal{I} = \{I_m\}_{m=1}^{M} \tag{3.3}$$

where M is the total number of faces detected and I_m is the feature set of the face to be identified. Classification is performed on each image in $\mathcal{I}$ using kNN based on training data $\mathcal{D}$ and using the biases b.

Identification of students attending a particular lecture can be done in two ways.

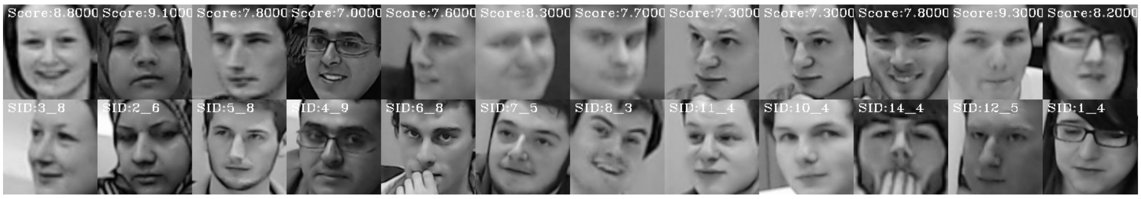


Figure 3.6: An example of output from the system where the top row shows the face image and the bottom row shows the students that best match the face images. In this case, 9 out of 12 students were correctly identified.

Single Pass Algorithm

In the first, each student's face image is identified as the student for which there is the best match in the training data and is shown as pseudo code:

Algorithm 1 Single Pass Algorithm

```
for all $I$ in $\mathcal{I}$ do
    $t_I, p_I = \text{kNN}(I, \mathcal{D}, k)$    ▷ $t_I$ is the identity and $p_I$ is the estimated probability
end for
```

This algorithm allows more than one I to be matched to a single identity. The estimated probability p_I is derived by calculating the number of identity matches within the neighbourhood as a ratio of total neighbours

Multiple Pass Algorithm

A multiple pass algorithm (see pseudo code 2) evaluates the quality of the match between each face image and all the training images. The face image with the best match is identified with the matching training image and that face image and the training images corresponding to the matched student are removed from the available training images.

Algorithm 2 Multiple Pass Algorithm

```
while $|\mathcal{I}| > 0$ do
    for all $I$ in $\mathcal{I}$ do
        $t_I, p_I = \text{kNN}(I, \mathcal{D}, k)$    ▷ $t_I$ is the identity
                                                     ▷ $p_I$ is the estimated probability
    end for
    $I^* = argmax\{p_I\}$                            ▷ Best matched image
    $t^*, p_I^* = \text{kNN}(I^*, D, k)$             ▷ Get identity for best matched image
    $\mathcal{D} = \mathcal{D} \setminus \{(F_n, t_n) | t_n = t^*\}$    ▷ Remove from training set
    $\mathcal{I} = \mathcal{I} \setminus \{I^*\}$    ▷ Remove from image set
end while
```

3.5 Conclusion

A system for performing face recognition, as an alternative attendance monitoring system, was proposed. In particular, its passive nature permits the potential for an uninterrupted lecture for both the lecturer and the attendees. In the case where the expected attendees are already enrolled, the system would only be required to search within a well defined and finite search space.

The introduction of a four stage recognition pipeline as a general solution for attendee recognition is intuitive. The pipeline detects faces from within images that possess a large amount of background noise, aligns the faces in order to best utilise the feature extraction techniques ready for classification. The latter part of the chapter defined an algorithm which fully utilised the pipeline. It suggested two approaches for its use; a single pass which simply marked off the best matches and an iterative approach that determine the best overall student match and marked off and removed the it with all of its associated data at the end of each pass to provide a decreasing set of attendees to search for.

4 Experiments

4.1 Introduction

Having defined a suitable recognition algorithm we are now in a position to test it to see how it performs under varying conditions. First, the robustness of the Viola Jones algorithm is considered by comparing the number of faces detected against those deemed detectable via human inspection. These experiments will consider those captured during a genuine lecture series. This is because our other main source of data is derived from faces already detected by the Viola Jones algorithm.

A second important step is to consider how effective the system is at aligning the faces we have detected into a common axis-aligned, centred framework. This is essential as the majority of feature extraction techniques employed within the system rely on localised information. Without good alignment, the system will simply fail to find these local features and therefore to perform to a satisfactory level.

This chapter then tests two distinct sets of data showing the results from both. The first is a synthetic 'lecture series' using famous faces taken from the Labelled Faces in the Wild (LFW) data set. It is tested using both the single pass and the multiple pass approaches. Whilst this data is already aligned it does not bypass the alignment stage of the recognition pipeline.

The final tests evaluate the system using a lecture series captured over a period of ten weeks. This data set is described in Section 4.5. The performance over the entire lecture series is tested using both the single pass and multiple pass approaches whereby in the single pass, each student's face image is identified as the student for which there is the best match in the training data. In the multiple pass, an evaluation of the quality of the match between each face image and all the training images is performed. The face image with the best match is identified with the matching training image and that face image and the training images corresponding to the matched student removed from the available training images. Given the sheer volume of result data that is available, only two lectures are studied in detail.

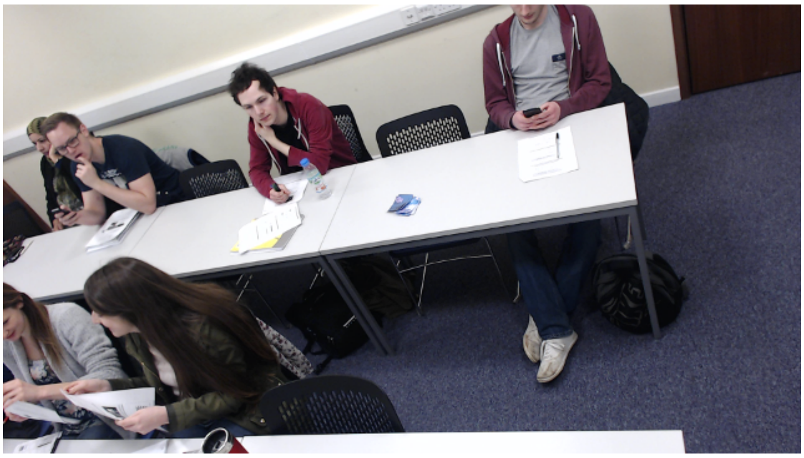


Figure 4.1: The angle of some of the faces within the source image prove difficult for the Viola-Jones algorithm to detect. In this image, taken from lecture 23, no faces have been detected.

4.2 General Observations

4.2.1 Face Detection

Given the obvious importance of face detection in the system, we begin by testing the effectiveness of the Viola Jones algorithm for accuracy. Only faces captured from the series of lectures were considered. This is because the faces within the Labelled Faces in the Wild data set were captured using the Viola Jones algorithm itself.

In total 995 faces were contained over the 161 photographs taken across the 23 lectures with a nominal class size of 12 students. Multiple photographs were taken in each lecture. From human inspection, only 61% (609) were judged to be of sufficient quality to be detected by the algorithm. In reality, 82% (499) of these faces were actually detected by the algorithm with a further 5% (54) false positive rate. Most false negatives appeared correlated to the extreme rotation caused by the hand held camera which took the face beyond the capability of the trained algorithm. An example of this can be seen in Figure 4.1 where one face is clearly visible yet has not been detected. The algorithm was trained exclusively using frontal face only and is expected to perform poorly when a non frontal face of correct rotation is presented to it.

Figure 4.2 shows a breakdown of the detected face data on a per lecture basis and is based on the table data provided in Figure 4.3. The algorithm remained broadly in line with the expectations although at 82% of what human inspection suggested. From lecture 15 onwards there is a significant difference between the faces captured in the images and those expected to be detected by the Viola Jones algorithm. This is because there is a correspondingly higher number of faces that are either blurred through motion, obscured or not facing the camera sufficiently. The significant drop off of detected faces in lecture 23 can be attributed to the extreme rotation of the images captured from the hand-held camera, as mentioned earlier and shown in Figure 4.1.

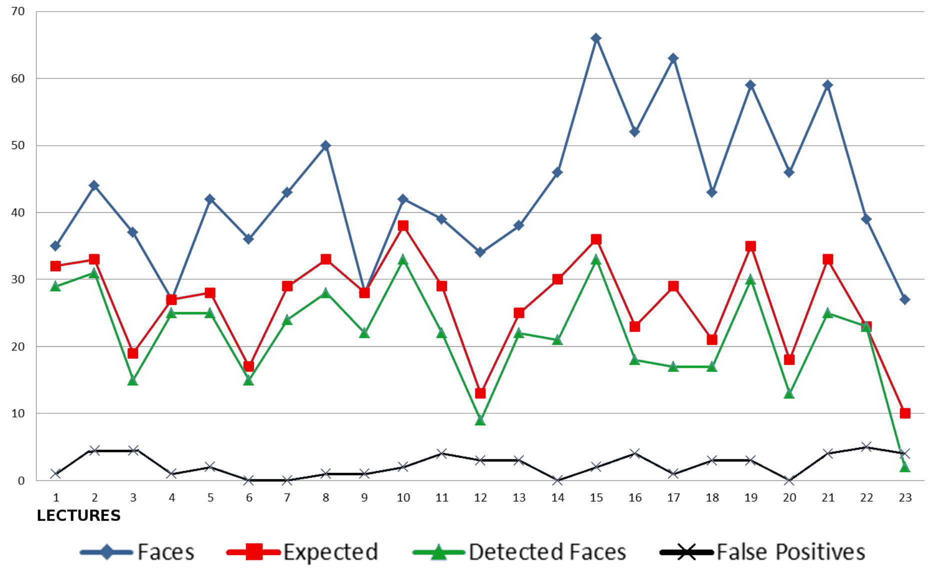
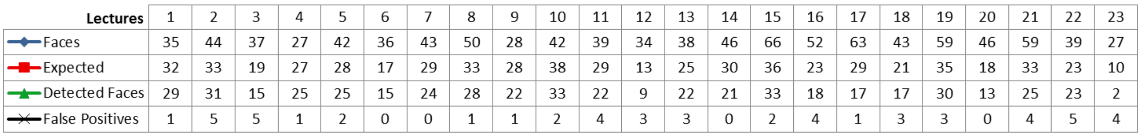


Figure 4.2: Graph showing Viola Jones detection over 161 photographs across 23 lectures where multiple photographs were taken in each lecture. *Faces* corresponds to the maximum number of attendees at the lecture. *Expected* refers to the expected number of faces that should be detected by the algorithm. This has been determined by human inspection. *Detected Faces* and *False Positives* refer to the output from the Viola Jones algorithm.



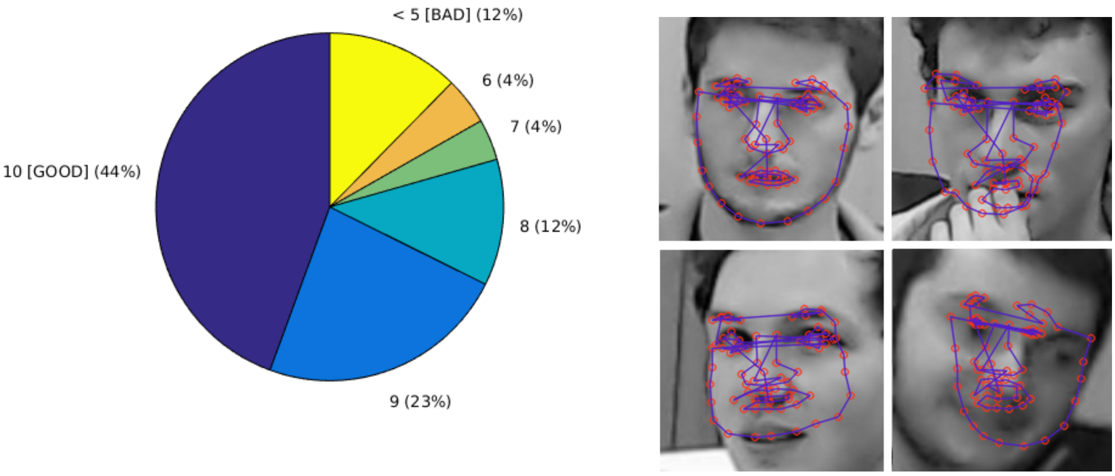
Lectures	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
◆ Faces	35	44	37	27	42	36	43	50	28	42	39	34	38	46	66	52	63	43	59	46	59	39	27
■ Expected	32	33	19	27	28	17	29	33	28	38	29	13	25	30	36	23	29	21	35	18	33	23	10
▲ Detected Faces	29	31	15	25	25	15	24	28	22	33	22	9	22	21	33	18	17	17	30	13	25	23	2
× False Positives	1	5	5	1	2	0	0	1	1	2	4	3	3	0	2	4	1	3	3	0	4	5	4

Figure 4.3: Breakdown of Viola Jones detection and failure rates versus the expected rates across the 23 lectures where multiple photographs were taken in each lecture. Please refer to Figure 4.2 for the accompanying graph.

4.2.2 Alignment

Face alignment also plays a key part in the system and, like detection, can be investigated independently from any subsequent attendance monitoring strategy. As an experiment, 340 Viola Jones detected faces were fitted using Active Shape Modelling (ASM) implemented via an openCV third party library [Chenxing, 2012]. Each face was examined by human inspection and scored between 1 and 10 to indicate the quality of the fitting of the eyes on the face. Figure 4.4a shows the percentage breakdown of this experiment.

In practice around 75% of the faces had, what could be considered as, reasonable fittings around the eyes. By reasonable I mean a score greater than 7 representing eye detection within 1 cm. The remainder were either false positives from the Viola Jones algorithm or could not be fitted correctly. One interesting observation which can be seen in Figure 4.4b is that even when the eyes were perfectly fitted, the implementation of ASM utilised could not fit the entire face correctly. In fact, the majority of faces fitted exhibited this. Whilst not performed here, a similar experiment using the Viola Jones algorithm to detect eyes could be performed and would be part of a future investigation.



(a) ASM quality scores

(b) ASM fitted examples

Figure 4.4: ASM Analysis data showing the percentage of quality fittings using an arbitrary scale ranging from 10 (Good) to below 5 (Bad) where we mark on how good the model fitted the eyes. The right hand images show some of those fittings. Three have scores of 10 with another at below 5 (bottom right image)

4.2.3 Summary

This section investigated the effectiveness of both the detection and alignment stages of the pipeline independently. It showed that the Viola Jones algorithm was successful in around 80% of cases. As training was only performed using a training set that focused on frontal face poses, and given the unconstrained nature of the environment, this success rate can be seen as successful. The alignment evaluation showed a similar set of results and the unconstrained nature and environment could be seen as a contributing factor.

4.3 Evaluation of Strategies

Having defined a pipeline based recognition algorithm (see Section 3.4), its use for solving the problem of capturing lecture attendance can be examined. It is worth noting that the problem is made much easier by the fact that not only will the attendee numbers be small but they are also known in advance. Therefore, the pipeline can be trained using a small set of targeted data. In these experiments the first lecture is used as the starting point for establishing the training set. Given that no formal enrolment has occurred it is worth recognising the limitations of this approach, especially where attendees do not attend or where the Viola Jones algorithm cannot detect their faces. In fact, in the lecture series, additional students attended some subsequent lectures and one student attempted to conceal themselves through-out.

The recognition algorithm will be employed in two ways. First, a single pass of the algorithm will be evaluated where each student will have its best match selected. A second method involves iteratively executing the algorithm removing the absolute best student match and its associated data from the pool before proceeding to determine the next best student match and so on. Furthermore, where more than one lecture has occurred, the data from previous lectures is labelled and added to the training set before the process is repeated.

The experiments evaluate the recognition algorithm using each of the following methods in isolation: LBP, CS-LBP, SIFT, Eigenfaces, Eigen LBP and Eigen CS-LBP. The spectral methods retain only an arbitrarily chosen 90% variance of the data and the LBP based methods consider only the closest radius (8 pixels). It will also evaluate the ensemble of all of these methods by using the total matching neighbours (see Section 3.4) from each method. The ensemble is simply the sum of nearest neighbours derived from the classification from all the individual methods without bias. The class with the overall majority is selected unless there is a tie, in which case, the first highest result is selected. Additionally, the experiments will be performed using a range of k-nearest neighbour between one and the total number of students and applied against each of the aforementioned methods. As each subsequent lecture is added to the training set we are able to derive the best k using Monte Carlo cross-validation [Xu & Liang, 2001] [Picard & Cook, 1984].

Finally, the experiments are performed over both a real lecture series comprising of 23 lectures and 12 students and a synthetic lecture using 19 randomly selected people from the Labelled Faces in the Wild (LFW) data set. The only selection condition for sourcing the candidates from LFW was the need to be have more than one image available for each mock student (one for the training set and one for the lecture).

4.4 Labelled Faces in the Wild (LFW) Results

A simulated lecture series was created to evaluate the recognition algorithm using well defined data. This data was taken from the Labelled Faces in the Wild deep aligned data set. Whilst previously aligned, the data was subjected to the algorithm under the same conditions as any other set.

Four lectures, three as training sets, were created which always contained all 19 attendees. Each attendee had a number of face images, typically 1 or 2 faces, made available in each lecture which were taken from the LFW data set [Huang et al., 2012]. This was done to mimic the conditions of a genuine lecture where multiple photographs could be taken capturing individual students a number of times. Nicole Kidman had a significantly higher number of face images available, totalling 19, and these were distributed over the 3 training sets with 1 added to our synthetic test lecture. The inversely proportional bias weighting should compensate for this.

The first lecture was assigned as the initial training set. To simulate the increasing number of images available as a lecture series progresses, lectures 2 and 3 were subsequently added to the training set to create training sets 2 and 3 respectively. In order to keep the results clear, only the fourth lecture is evaluated against each of the three training sets. In the following the single initial lecture is labelled ‘Training Set 1’, ‘Training Set 2’ means the initial lecture combined with lecture 2. ‘Training Set 3’ means the initial lecture combined

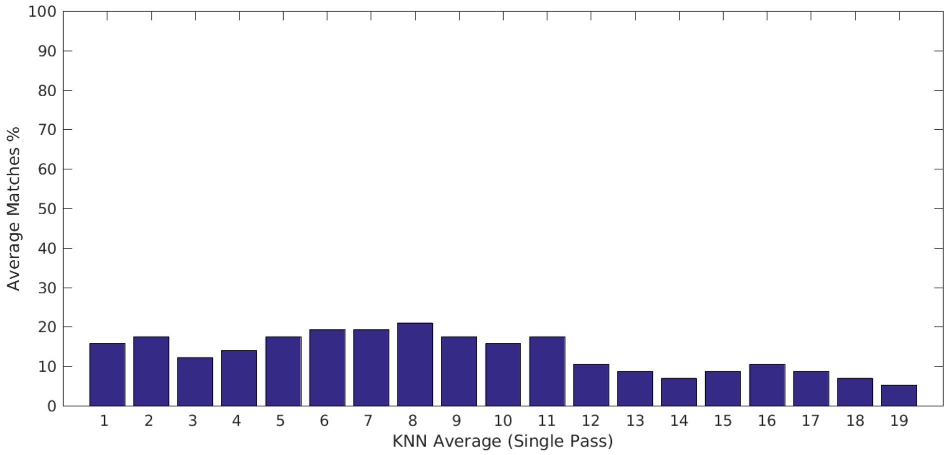


Figure 4.5: A bar chart showing the evaluation of the synthetic lecture over the k-nearest neighbour range; using the single pass approach.

with both lectures 2 and 3.

As described in Chapter 3, identification of students attending the synthetic lecture was done in two ways. In the first, each student's face image is identified as the student for which there is the best match in the training data. Importantly, this method means that not all students from the training data can be matched as a face image may never be assigned to a student if a better match for the image exists. The second selects the best match before removing the face image and the corresponding training images from the available training images. It then repeats the process until all students have been found.

4.4.1 Single Pass Results

The results of the single pass evaluation of the recognition algorithm using the synthetic LFW lecture series are presented here. The single pass method simply assigns each face image to its best student match. Figure 4.5 shows the collective (ensemble) results taken from the individual feature extraction techniques (LBP, CS-LBP, SIFT, Eigenfaces, Eigen LBP and Eigen CS-LBP) when classified using the k-nearest neighbour range from $k = 1 \ldots 19$ and applying the weights to remove the bias caused by too many images for a particular student. The ensemble is simply the sum of nearest neighbours from all of the individual methods classified using k-nearest neighbour where each method has an equal weighting. The class with the overall majority is selected. The results have been averaged over the 3 training sets and shows the respective matches for the lecture examined.

It is clear that the recognition algorithm did not verify many attendees as only just over 20% of the 19 attendees were successfully matched. From visual inspection $k = 8$ performed the best with the higher values of k producing the worst results (as low as 5%). Reviewing the results from the individual training sets, and not the average, showed that 'Training Set 1' and 'Training Set 3' also performed best when $k = 8$. 'Training Set 2' performed best when $k = 5$.

For the effort required, the nearest neighbour ($k = 1$) performed reasonably well at 15%.

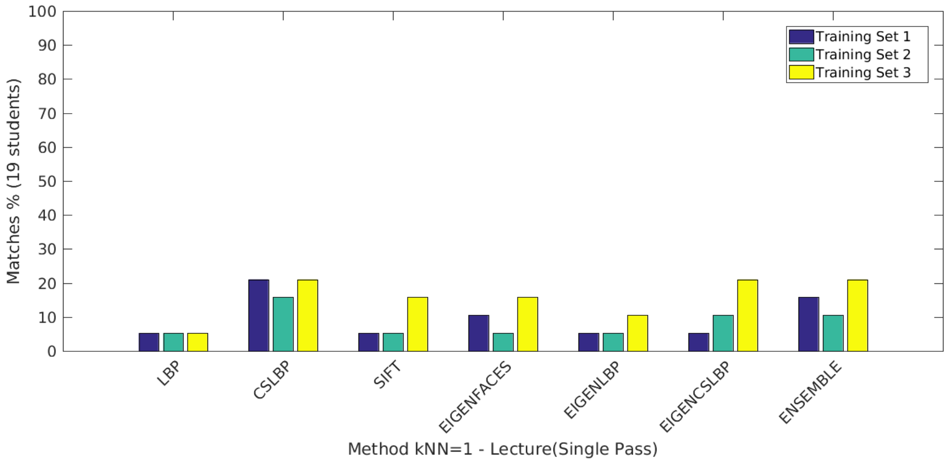


Figure 4.6: A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 1$; using the single pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

The implication here is the system is better at matching near-identical faces when k is small but these ideal faces tend to get averaged out as k increases.

Reviewing the results by method over each training set where nearest neighbourhood $k = 1, 8$ and 19, figures 4.6, 4.7 and 4.8 show that the trend for matches increases in relation to the amount of available faces in the training set. i.e. the more faces to use for recognition the greater the number of correct matches. On average an additional labelled face is added for each attendee per training set.

Given that the alignment of each face is dependent upon the reliability of both the Viola Jones algorithm (to detect eyes) and the ASM fitting for all methods other than SIFT it is possible to get a sense of how well the alignment actually worked. In all cases CS-LBP outperformed the other feature extraction methods including SIFT which was not dependent upon alignment. Furthermore, CS-LBP performed best when $k = 1$ suggesting that not only was there good alignment but also a closeness of face images. CS-LBP performed significantly better overall than the related LBP feature extraction technique. This suggests that CS-LBP's improved ability to capture gradient information and lower dimensionality (see Section 2.6.5) has assisted. Additionally, the features extracted for the LBP are based on abstracted uniform patterns (e.g. edges, spots, lines; see Section 2.6.5) whereas the CS-LBP is a straight feature extraction over neighbouring pixels.

All other methods perform to a similar level as each other and offer a basic sanity check across these methods. Whilst not overwhelming, the ensemble, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, shows great strength in all cases reviewed, performing only marginally worse than the CS-LBP feature in isolation. Given that all single feature extraction methods have equal weighting in the ensemble it cannot be the case that it is heavily biased by the CS-LBP result but instead suggests that the collective knowledge of all methods tends to produce better results.

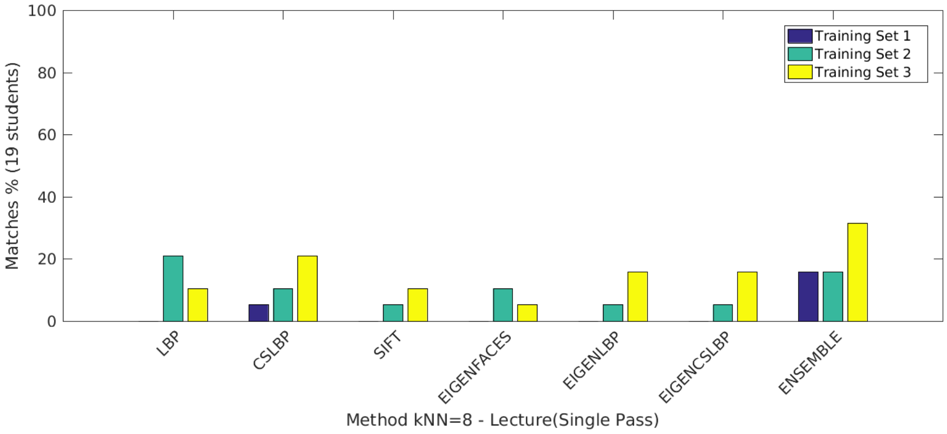


Figure 4.7: A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 8$; using the single pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

Figures 4.9 and 4.10 consider the results on a per student basis. Again, these are broken down by method and compared over the k-nearest neighbour range. The colour bar indicates the number of correct training set matches that have been made for the student.

It is noticeable that some students tend to match more than others. Reviewing the attendee match data shown in these figures against the faces in the training set reveals why some attendees fare better than others; some attendees are much more similar to the training set they are attempted to be matched with. Whilst this is unavoidable, it is worth pointing out that in a genuine lecture series we expect less variations than our sample LFW data set due to the time scale differences between the collection of the LFW images. i.e. the LFW data has been collected over a number of years.

Figure 4.11 illustrates this by showing three attendees; James Cunningham (student 7), Nicole Kidman (student 11) and Patricia Clarkson (student 13). It is much harder to see the similarities between the two images of James Cunningham than the others. Furthermore, Nicole Kidman has 18 faces in the training data set where James Cunningham only has one to compare with although the bias has been mitigated and should not affect the outcome. One final observation is that some students were not successfully matched at all. This is because the system is attempting to find the best student match for the face image and has incorrectly selected the wrong student.

As expected the ensemble method (see Figure 4.9), which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, has reflected the stronger matches from the individual methods (see Figure 4.10). Not only that but they tend to be matched over the majority of the k range at least once. Unsurprisingly, student 7 (James Cunningham) has no matches using any of the methods. What is apparent is the loss of student 11 (Nicole Kidman) from the ensemble's results. At least 3 of the methods (CS-LBP, EIGEN-LBP and EIGEN-CS-LBP) had matched her yet she was absent from the collective result. This demonstrates two things. First, given

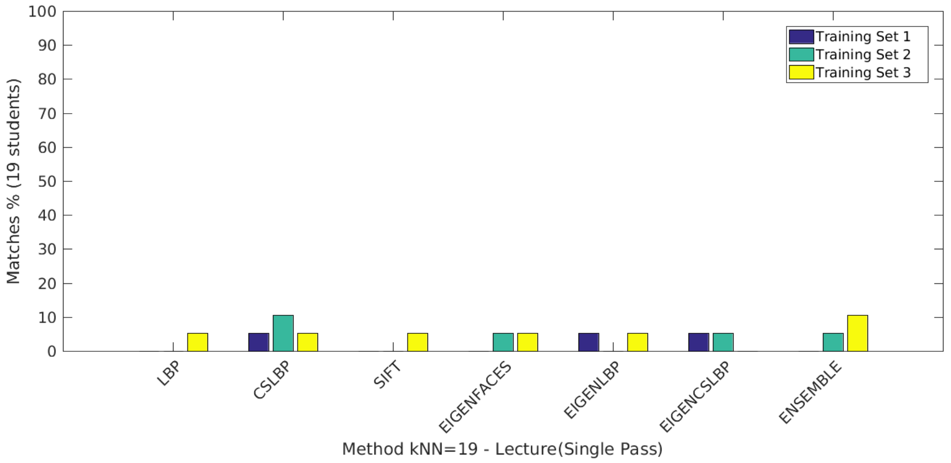


Figure 4.8: A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 19$; using the single pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

her overwhelming representation within the training data, it shows that the inversely proportional weighting scheme has successfully down-weighted the bias that can arise from the differences in the number of available training faces for the individual attendees. Furthermore, it suggests that the collective sum of all matches for her, especially from those methods which had not correctly matched, favoured another candidate.

Conclusion

The single pass evaluation of the LFW synthetic lecture was performed to identify each face image as the student for which there is the best match in the training data. Overall, the performance of the system performed poorly as it was only able to successfully match 20% of the students at best. This could be because of the time-scale for which the images were collected. In some cases these appear to be taken over years where in others, those that tended to be successful, over shorter periods (including minutes).

For this experiment it was clear that the CS-LBP method outperformed the other methods including its related LBP method whilst the ensemble performed strongly. The ensemble considers the information from each method without bias and so was not heavily influenced by the CS-LBP method.

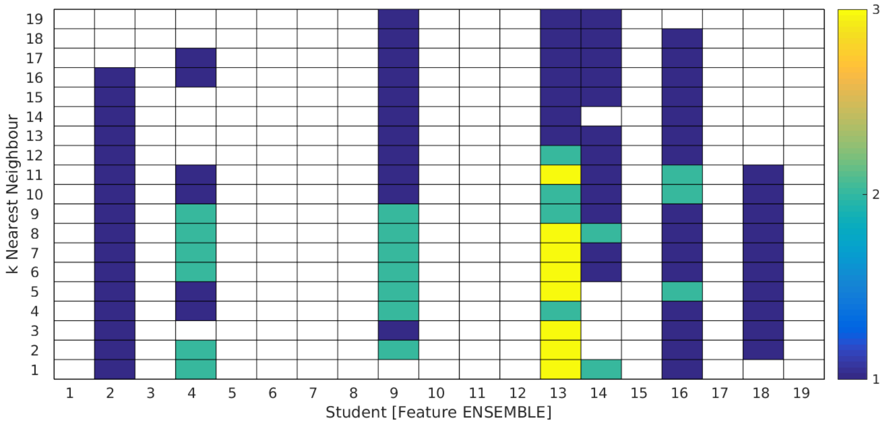
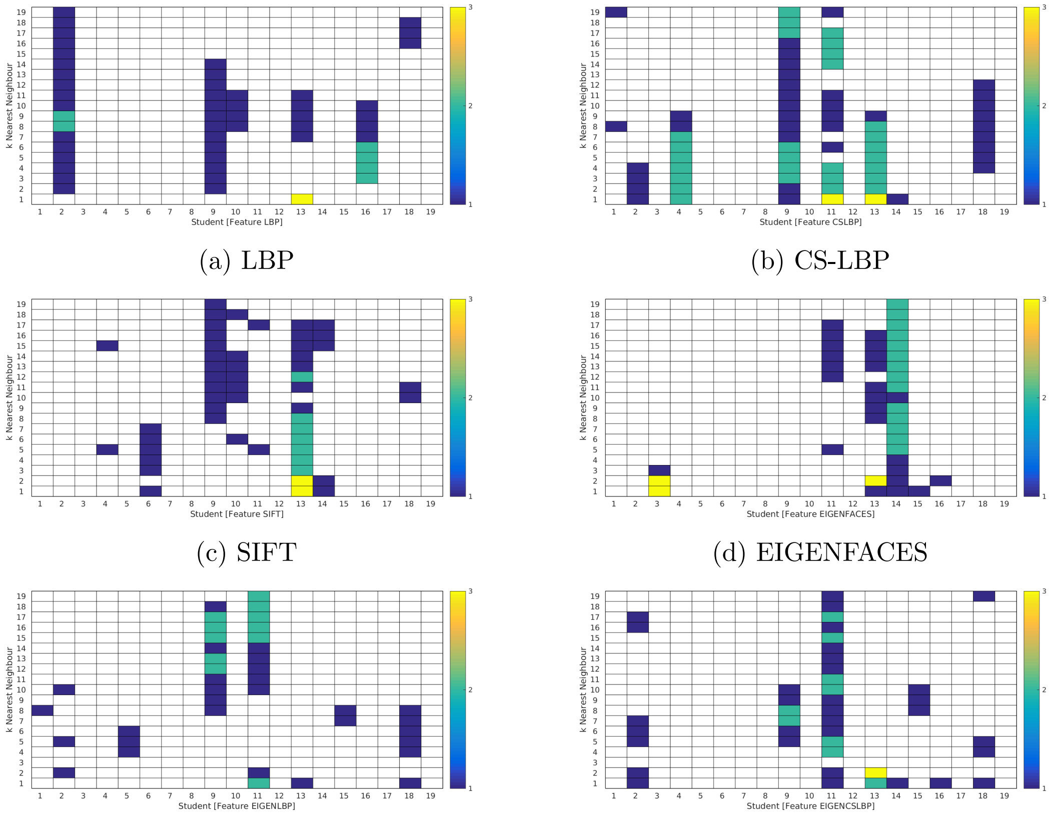


Figure 4.9: Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbour range and colour coded based on the number of training set matches; using the single pass approach on the LFW synthetic data set.



(a) LBP

(b) CS-LBP

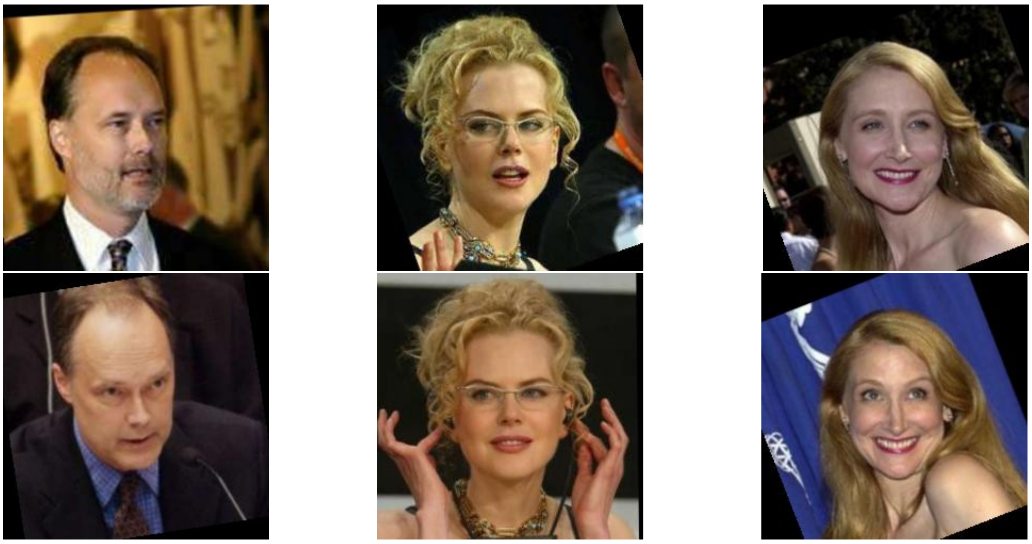
(c) SIFT

(d) EIGENFACES

(e) EIGEN LBP

(f) EIGEN CS-LBP

Figure 4.10: Breakdown on a per student basis using individual methods over the entire nearest neighbour range and colour coded based on the number of training set matches; using the single pass approach on the LFW synthetic data set.



(a) James Cunningham (b) Nicole Kidman (c) Patricia Clarkson

Figure 4.11: Example pairs of attendees to be matched.

4.4.2 Multi Pass Results

The results of the multiple pass evaluation of the recognition algorithm using the synthetic LFW lecture series are presented here. As with the single pass experiment, Figure 4.12 shows the evaluation of the synthetic lecture using the collective results taken from the individual feature extraction techniques (LBP, CS-LBP, SIFT, Eigenfaces, Eigen LBP and Eigen CS-LBP) when classified using the k-nearest neighbour range from $k = 1 \ldots 19$ and applying our inversely proportional bias weights. Again, the results have been averaged over the 3 training sets and shows the respective matches for the lecture. The fundamental difference is that the algorithm has been modified to iteratively select the best candidate before removing that target face and the student from the training set data until all faces have been matched (see Section 3.4).

The multiple pass approach consistently performed better across the board. Whilst an improvement can be seen it still performed poorly matching a maximum of only 30% of the 19 attendees. From visual inspection $k = 4$ and 5 performed the best with the higher values of k producing the worst results (as low as 8%). Again, for the relatively small effort required, the nearest neighbour ($k = 1$) performed reasonably well at 25%. Reviewing the results from the individual training sets, and not the average, showed that 'Training Set 1' performed best when $k = 1$, 'Training Set 2' performed best when $k = 4$ and 'Training Set 3' performed best when $k = 5$.

Reviewing the results by method where $k = 1, 5$ and 19 over each training set figures 4.13, 4.14 and 4.15 show that the trend for matches increases, as expected, in relation to the amount of available faces in the training set. Like the single pass, an additional labelled face is added for each attendee per training set on average.

Unlike the single pass approach there is no feature extraction technique which consistently outperforms the others. Again, the ensemble, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, performed

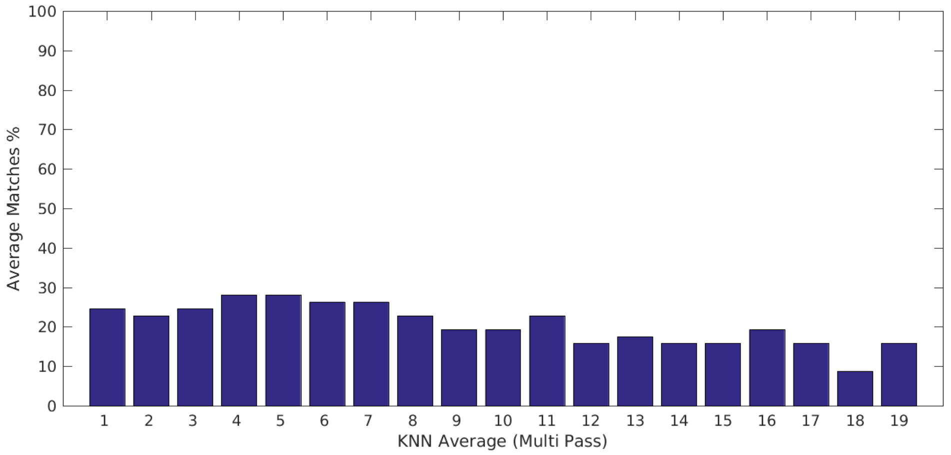


Figure 4.12: A bar chart showing the evaluation of the synthetic lecture over the k-nearest neighbour range; using the multiple pass approach.

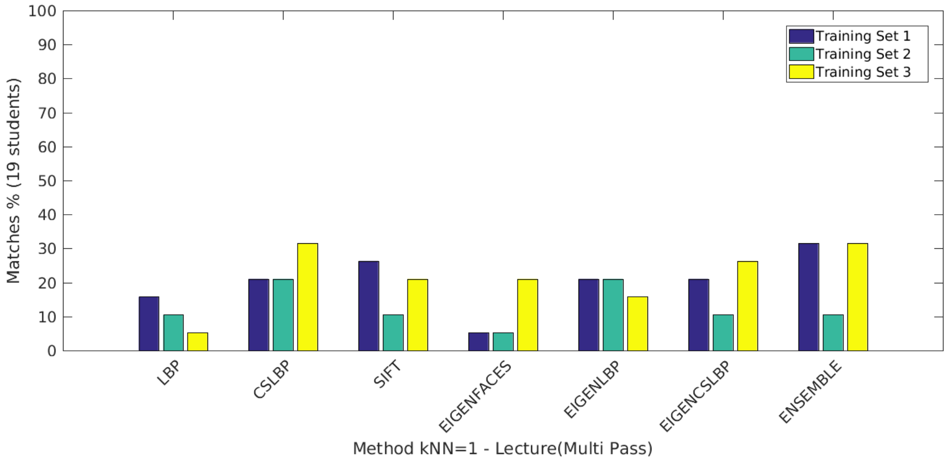


Figure 4.13: A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 1$; using the multiple pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

only marginally worse than the best performing feature classification in isolation.

Figures 4.16 and 4.17 consider the results on a per student basis. In general, the multiple pass approach produced better results than the single approach equivalent and this can be clearly seen with the corresponding increase in the number of students matched.

The multiple pass method finds the best match between a face image from the training data. It removes this optimal face image from the test data and all of the associated student data before iteratively performing this until all students have been matched to a face image. In the single pass method we discussed the low likelihood of matching student 7 (James Cunningham) and find that every method has matched him at least once. This is because once all of the more likely students have been found the pool of matching those that remain is much reduced and can afford to be coarser. This is evident where our better candidates tend to show strongly across the nearest neighbourhood range whereas

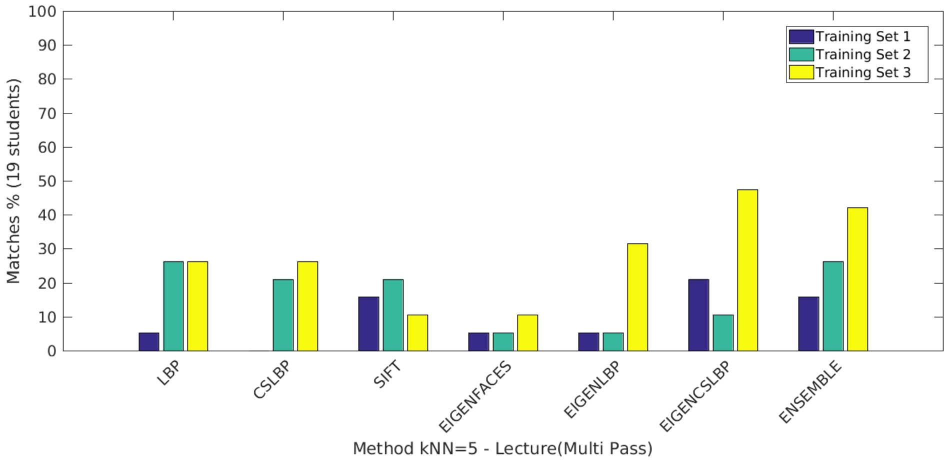


Figure 4.14: A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 5$; using the multiple pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

the weaker candidates are sporadic.

Conclusion

The multiple pass evaluation of the LFW synthetic lecture was performed. Unlike the single pass approach, this algorithm systematically removed the strongest match between a face image and student in the training data before iteratively doing this for all students. Whilst there was no clear leader in the feature extraction methods used over the nearest neighbourhood classification, they all performed consistently better than the single pass approach.

Furthermore, it was apparent that by removing the associated data of the stronger candidates we were able to remove some of the distraction that was affecting the classification of the weaker student matches.

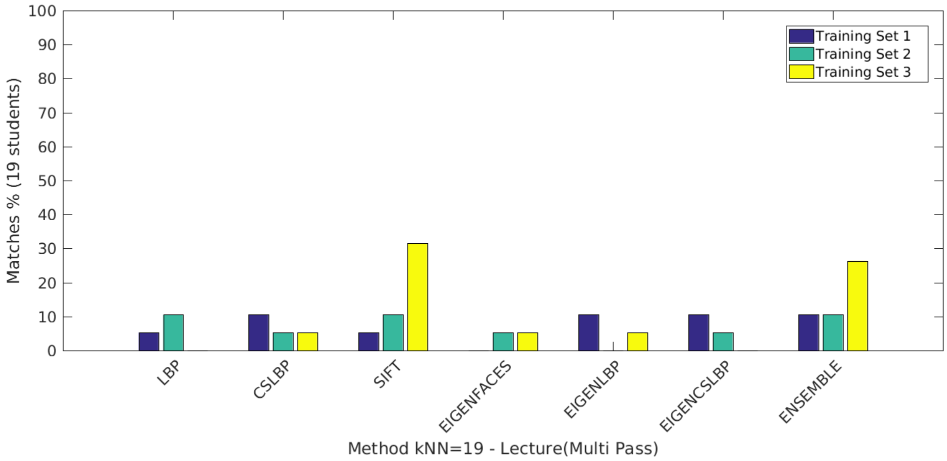


Figure 4.15: A bar chart showing the evaluation of the synthetic LFW lecture series in terms of the method employed over each training set where nearest neighbourhood $k = 19$; using the multiple pass approach on the synthetic lecture series. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

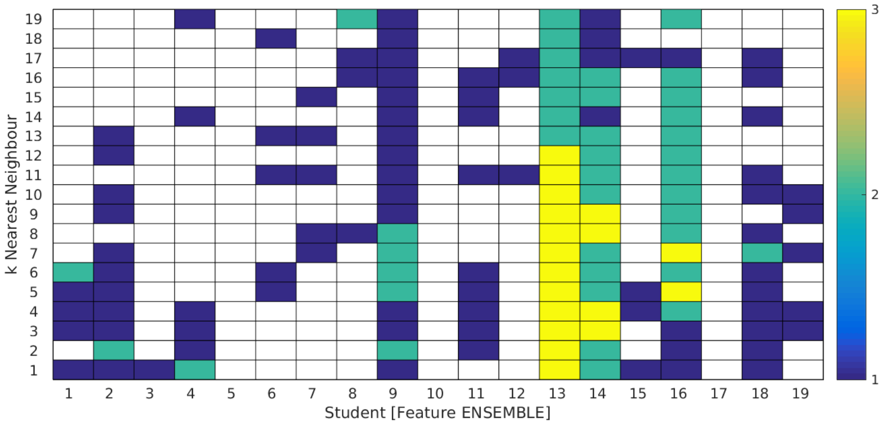
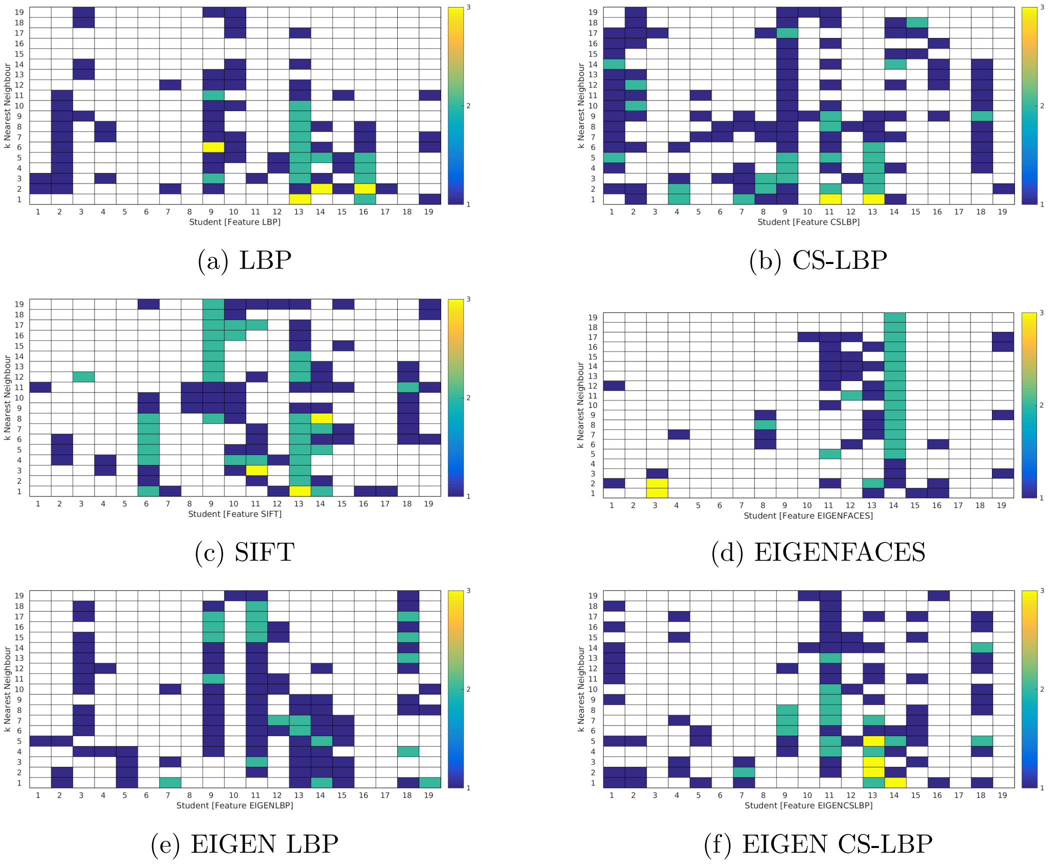


Figure 4.16: Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on the LFW synthetic data set.



(a) LBP

(b) CS-LBP

(c) SIFT

(d) EIGENFACES

(e) EIGEN LBP

(f) EIGEN CS-LBP

Figure 4.17: Breakdown on a per student basis using individual methods over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on the LFW synthetic data set.

4.5 Lecture Series Results

A lecture series comprising 23 lectures was recorded over a period of ten weeks and assessed using the recognition algorithm. Between 4 and 10 (average 7) images were captured using a hand-held web cam taken by the lecturer at the start of each lecture; most attendees appeared in one or two of the captured images. Almost certainly, each of the individual attendees was either in a different physical location in the lecture theatre as they moved to their seat or had a different expression, perhaps intentional, or pose within each photograph. Figure 4.18 shows one such student in a variety of poses; albeit over a series of lectures. Furthermore, the data set was imperfect as some students were not facing the camera whilst the photographs were being taken. The unconstrained environment makes this a challenging experiment.

Unlike the Labelled Faces in the Wild equivalent, where the result of the Viola Jones algorithm was already determined and where the faces had already been aligned, there was no guarantee that any attendee from any of the lectures would be detected or that they would ever be in a pose similar enough to be classified appropriately. Section 4.2.1 reviewed the imagery and concluded that only 50% of the faces captured were actually detected by the system; although this is only a little worse than what was deemed detectable via human inspection (circa. 60%). One other minor difference between the two experiments is the fact that there multiple faces could be detected from each image whereas the LFW version dealt with a single face per image.

On review, 12 students are assessed using 22 lectures over the ten week period. The 23rd lecture was discarded because detection performed poorly (see Section 4.2.1 for further details). As with the synthetic Labelled Faces in the Wild experiment, the single and multiple pass strategies are evaluated. Again, the system was initially trained using images captured from the first lecture and faces from subsequent lectures (named 1 to 21) were labelled and incorporated into the training set as the experiment proceeded. Each lecture is assessed against the ever increasing training data.

The spectral based methods used in the system have been constrained to use 90% of the variance using the cut off technique described in equation 2.17. Figure 4.19 shows the percentage of principal components that represent 90% of the variance found using PCA when applied to the directly to the face image, the encoded LBP image and the encoded CS-LBP image. The total number of principal components is shown as an increasing line and matches the total number of face images used. This is because of the formation of the covariance matrix using $A^T A$ as suggested by Kirby and Sirovich [Sirovich & Kirby, 1987] (see Section 2.4.1). Consequently, the same number of principal components are available for all spectral methods.

Figure 4.19 shows that as the number of principal components found increases the percentage of principal components that represent 90% of the variance decreases. Partly, this is because the initial number of principal components is low and achieving high degree of variance requires a relatively high proportion of them. The main reason is that the variance within the face images will naturally plateau and will therefore be less affected

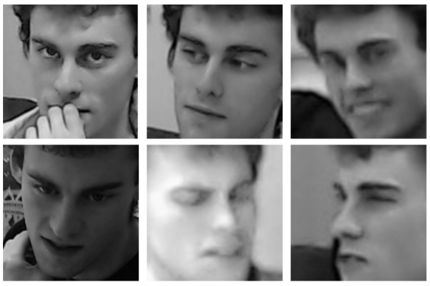


Figure 4.18: Six examples of the same attendee in varying poses highlights the difficulty faced by the system.

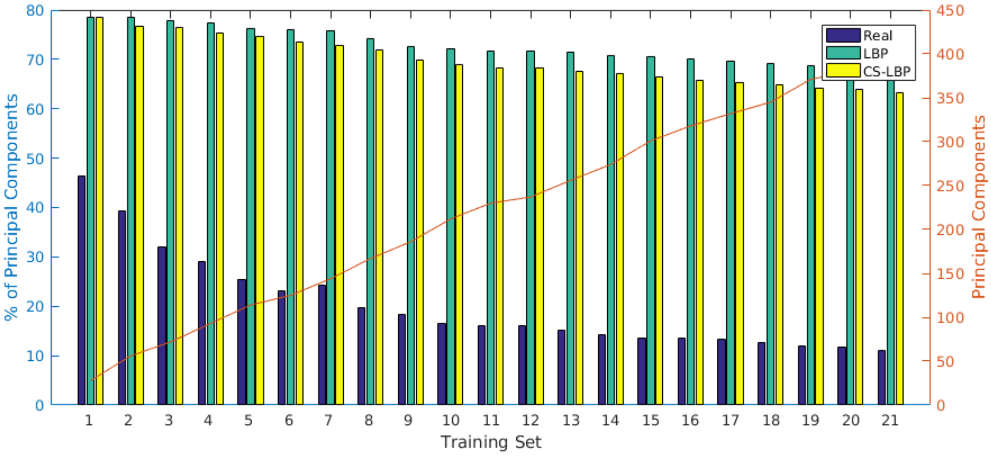


Figure 4.19: Percentage of principal components used which represent 90% of the variance for each spectral method. The line plot shows the total number of principal components found for each training set.

by the increase of face images.

Notably, capturing 90% of the variance using either of the Local Binary Pattern spectral methods requires most of the principal components, decreasing only slightly as the number of principal components increases. This is in stark contrast to the applying PCA directly to the face image. The reason for this is that the local binary pattern encoding has already reduced the dimensionality of the original face image and, whilst all spectral methods find the same number of principal components, the variance is spread more evenly with the encoded methods.

Figure 4.20 shows the first 15 eigenvalues of the three methods from training set 10. It is clear that the early eigenvalues from the real face image reflect much more of the variance of the data set than the other methods and confirms our suspicion that the variance is distributed more evenly over the encoded methods, albeit, with little variance.

Given an large increase in the amount of data to review, versus the single synthetic lecture from the LFW experiment, only lectures 12 and 13, representing a lecture that the system is expected to perform badly with (lecture 12) due to poor Viola Jones detection (see figure 4.2) and one it is expected to do well (lecture 13), are analysed in this chapter.

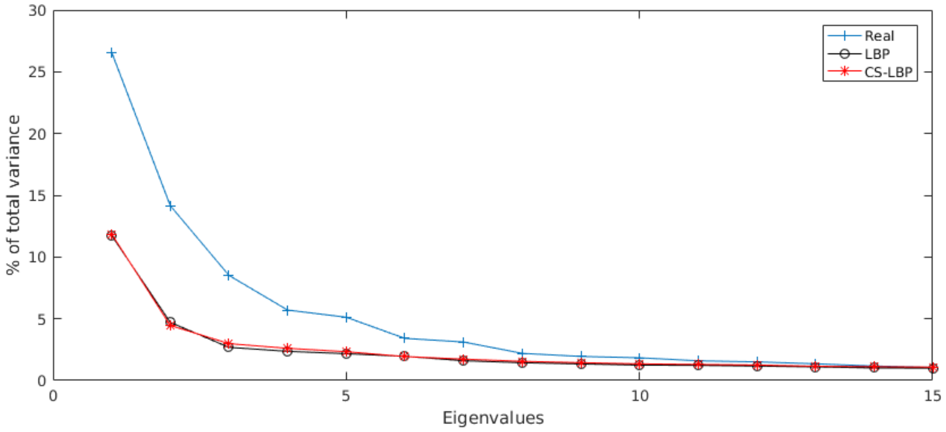


Figure 4.20: Line plot showing the first 15 eigenvalues for each of the three spectral methods whilst evaluating training set 10. The LBP and CS-LBP methods distribute the total variance within a tighter range than standard Eigenfaces method.

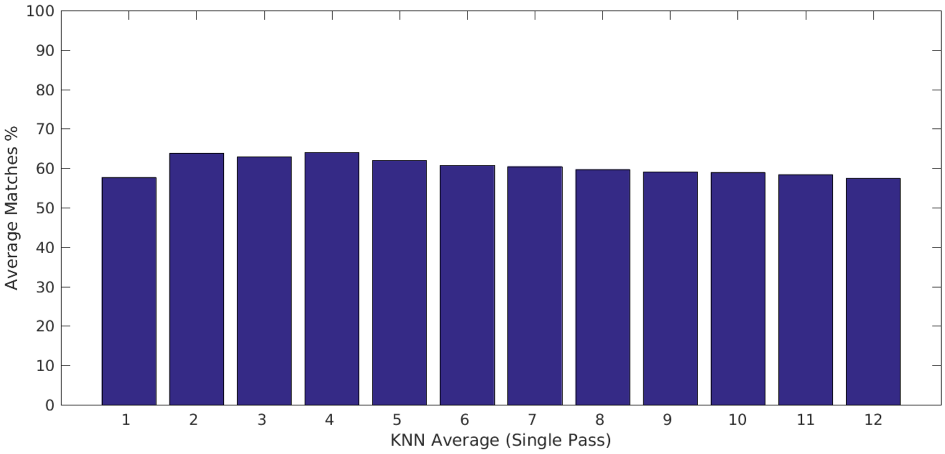


Figure 4.21: A bar chart showing the evaluation of the lecture series over the k-nearest neighbour range; using the single pass approach.

4.5.1 Single Pass Results

The results of the single pass evaluation of the recognition algorithm using the lecture series are presented here. The single pass method simply assigns each face image to its best student match. Figure 4.21 shows the collective (ensemble) results taken from the individual feature extraction techniques (LBP, CS-LBP, SIFT, Eigenfaces, Eigen LBP and Eigen CS-LBP) when classified using the k-nearest neighbour range from $k = 1 \ldots 12$ (where $L = 12$ represents the total students monitored) and applying the weights to remove the bias caused by too many images for a particular student. The ensemble is simply the sum of nearest neighbours from all of the individual methods classified using k-nearest neighbour where each method has an equal weighting. The class with the overall majority is selected. The results have been averaged over the all of 21 training sets and shows the respective matches over the lecture series.

It is clear that the recognition algorithm performed reasonably well matching around 60%

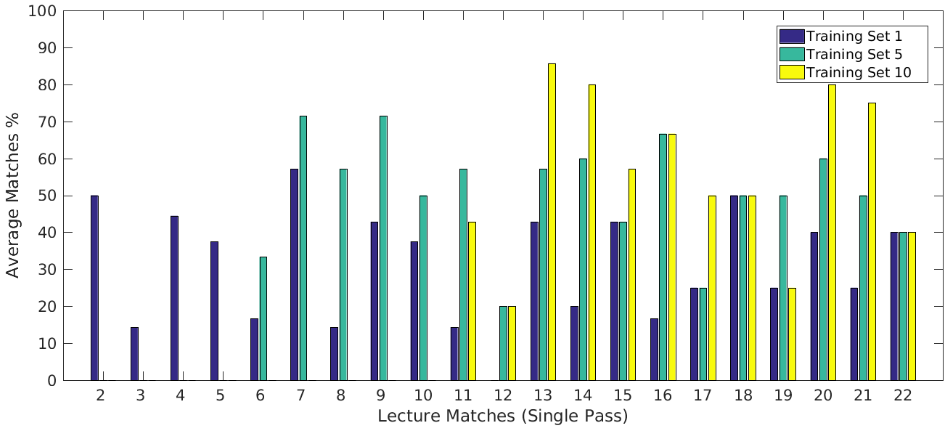


Figure 4.22: A bar chart showing percentage matches over three training sets with nearest neighbour $k = 1$; using the single pass approach.

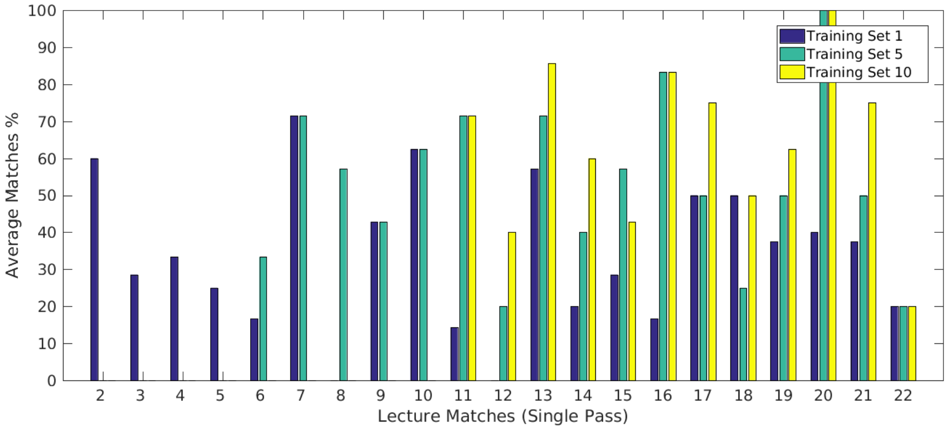


Figure 4.23: A bar chart showing percentage matches over three training sets with nearest neighbour $k = 4$; using the single pass approach.

of the 12 attendees. From visual inspection $k = 2$ and 4 performed the best with the higher values of k producing the worst results, albeit only marginally lower. Equally, the nearest neighbour ($k = 1$) performed reasonably well at 58%. Again, the system appears good at matching near-identical faces when k is small but, unlike our LFW experiments, this does not reduce much over the nearest neighbourhood range. This is because the quality of the matches is clearly better.

Reviewing the results of the ensemble method over three training sets comprised of lectures 1, 5 and 10 where $k = 1, 4$ and 12, figures 4.22, 4.23 and 4.24 show the respective matches on a lecture by lecture basis. From these, it is clear that lecture 12 performed poorly and lecture 13 performed reasonably well. These are the lectures that will be reviewed in this section.

Lecture 12 Review (Single Pass Approach)

Breaking the results for lecture 12 down by method where $k = 1, 4$ and 12, figures 4.25, 4.26 and 4.27 show the system performing recognition over each of the prior 11 training

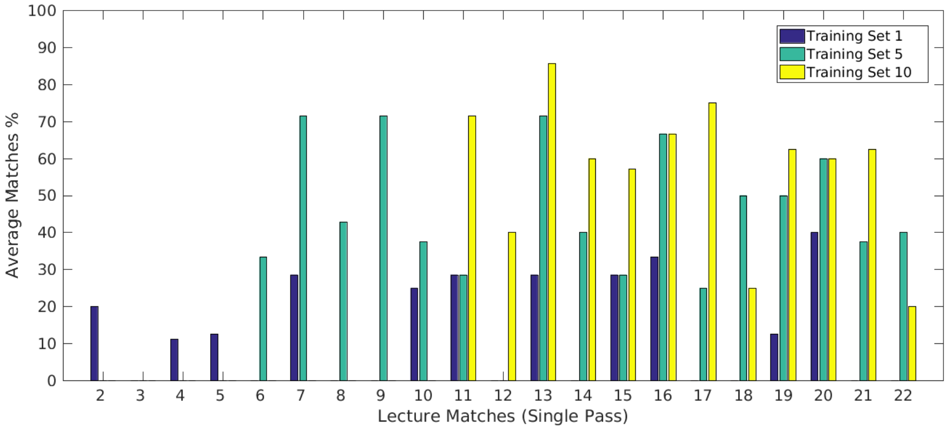


Figure 4.24: A bar chart showing percentage matches over three training sets with nearest neighbour $k = 12$; using the single pass approach.

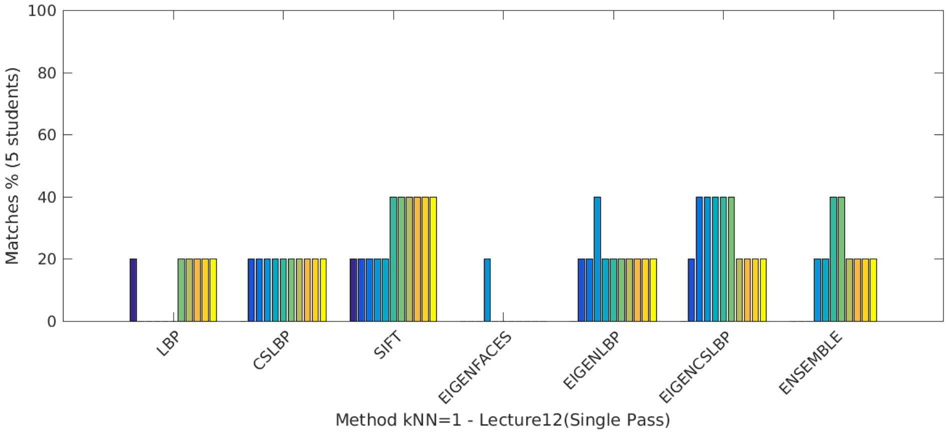


Figure 4.25: A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 1$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

sets on a per method basis. The figures reveal that the system only successfully matched a maximum of 2 of the 5 expected to be detected students. CS-LBP and SIFT performed the best but it is noticeable that SIFT was unable to match across all training sets tested especially when $k = 4$ which was previously stated as one of the two ks to offer the best overall match potential.

The difference between the CS-LBP method and its closely related LBP method is also apparent. Both extract a similar feature set from the aligned face images. The main difference is concern with the implementation of each of these feature extraction methods. LBP extracts features and compartmentalises them into uniform patterns (e.g. edges, spots, lines) whereas CS-LBP simply extracts them without further abstraction.

In this instance $k = 1$ and 4 produced an overall stronger performance than where $k = 12$ but does not appear to show an upward trend of matches in relation to the amount of available faces in the training set. This is because the training sets are composed from previous lectures and only the students from those lectures are labelled. i.e. there is no

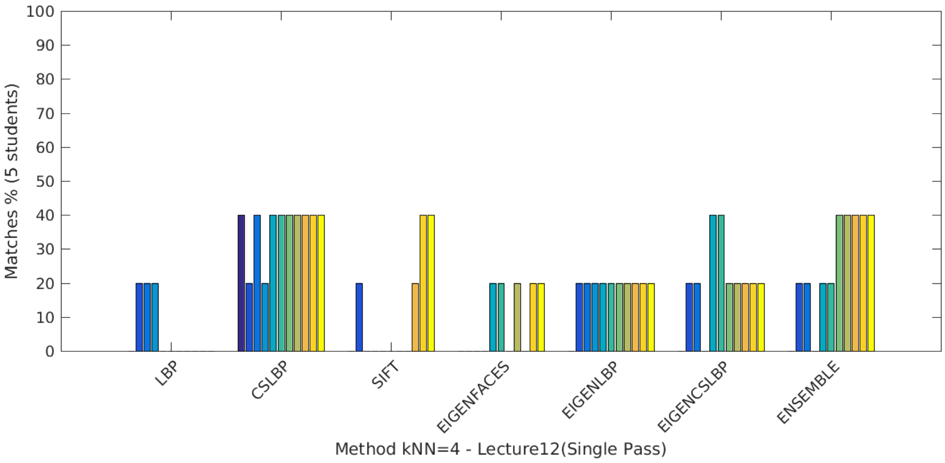


Figure 4.26: A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 4$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

certainty that a student will attend any subsequent lecture so the strength of the training data is not linear for any one student. Furthermore, poor alignment of the face images taken from lecture 12 will make matching from previous lectures difficult.

Figures 4.28 and 4.29 consider the results on a per student basis. Again, these are broken down by method and compared over the k-nearest neighbour range. The colour bar indicates the number of correct training set matches that have been made for the student.

It is noticeable that some students tend to match more than others. In particular students identified as SID 1 and SID 12 match frequently. In the system all students are labelled using student IDs (SID) because the data was collected anonymously. They are certainly the ones correctly matched in the majority of methods employed. Whilst not conclusive, a review of the images reveals that these two students tended to maintain a similar pose through-out the lecture series.

Figure 4.30 shows example output from the system where the ensemble method has been used and nearest neighbour $k = 4$. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched. It is clear to see that the first and last students are a good match. As stated earlier, this is reflected in the number of matches we witnessed using any of the feature extraction methods. Only SIFT is able to match more students frequently and this difference is most likely explained by the poor alignment as illustrated in Figure 4.30.

The remaining observation from the example output is the fact that more students are presented as being matched (correctly or not) than the 5 students human inspection believed would have been matched by the system. The reason for this is because the system assumes there will be 12 attendees. From analysis it has deemed that it was able to match 7 students from the face images taken.

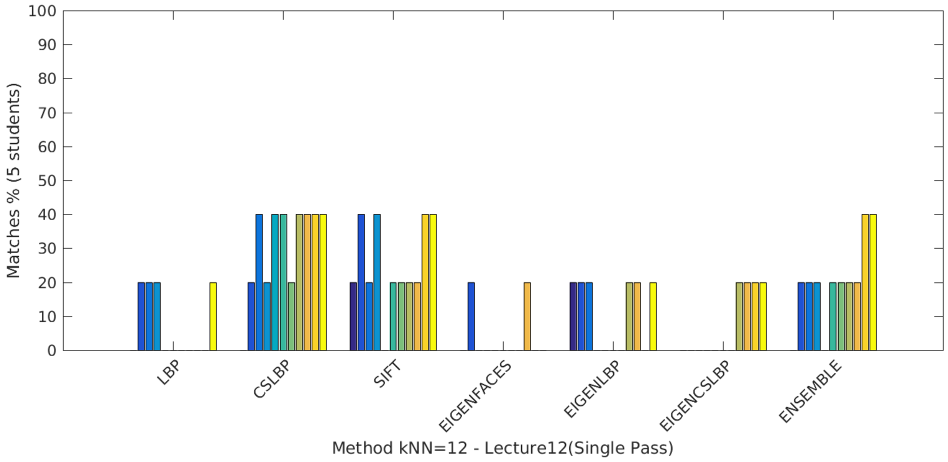


Figure 4.27: A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 12$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

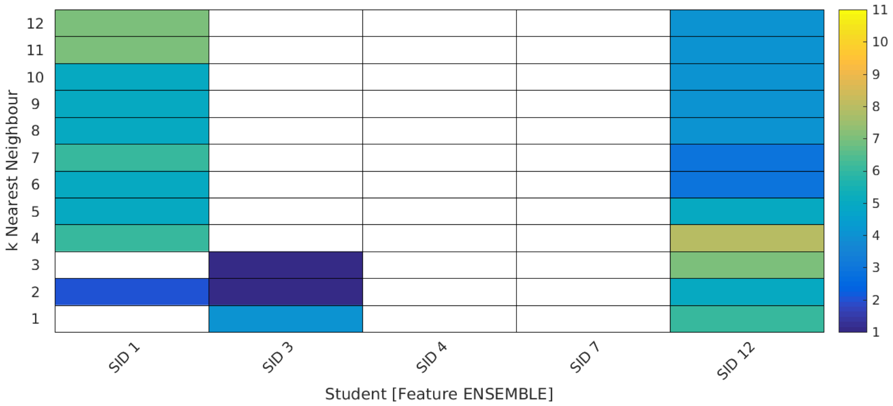
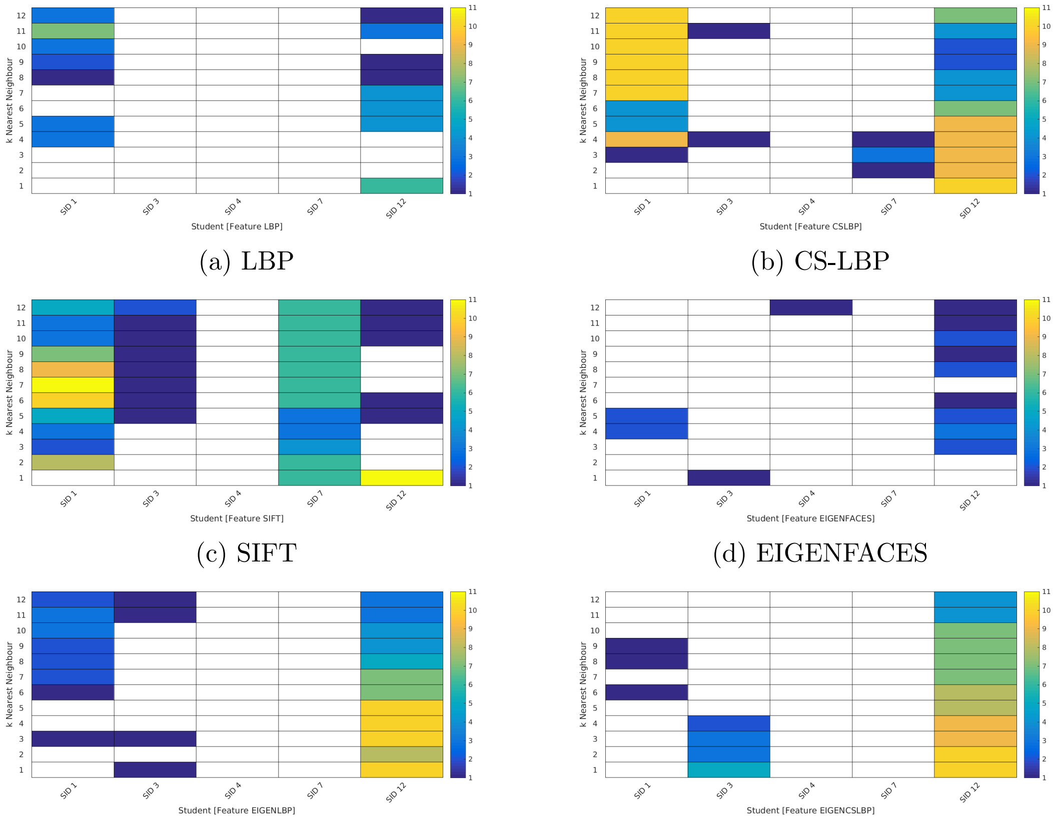


Figure 4.28: Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the single pass approach on lecture 12.



(a) LBP

(b) CS-LBP

(c) SIFT

(d) EIGENFACES

(e) EIGEN LBP

(f) EIGEN CS-LBP

Figure 4.29: Breakdown on a per student basis using individual methods over the entire k nearest neighbour range and colour coded based on the number of training set matches; using the single pass approach on lecture 12.

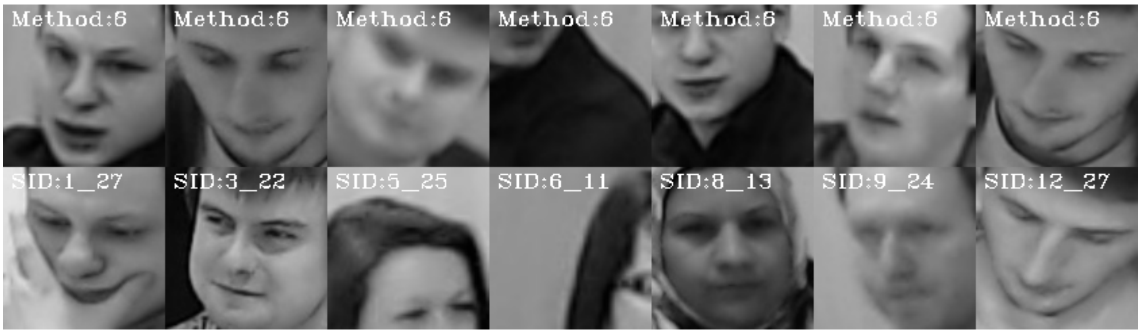


Figure 4.30: Example output from the system where the ensemble method is used with nearest neighbourhood $k = 4$; using the single pass approach on lecture 12. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched.

Lecture 13 Review (Single Pass Approach)

Breaking the results for lecture 13 down by method where $k = 1, 4$ and 12, figures 4.31, 4.32 and 4.33 show the system performing recognition over each of the prior 11 training sets on a per method basis. The figures reveal that the system successfully matched all of the students although averaged just under 5 of the 7 students with a minimum of 1 student across the board. Again, CS-LBP and SIFT performed well but the other independent methods performed well considering this was a single pass evaluation of the students. Noticeably the collective knowledge of the ensemble, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, outperformed all other methods employed.

Again, the difference between the CS-LBP method and its closely related LBP method is also apparent. Both extract a similar feature set from the aligned face images and, in this case, performed reasonably well when tested in their spectral equivalents (Eigen LBP and Eigen CS-LBP). This is possibly because the dimensionality of this domain has been reduced to only consider 90% of the variance of the data. The implication being that LBP is suffering from noise. Furthermore, the standard Eigenface spectral method performed worst.

In this instance $k = 1$ and 4 produced an overall stronger performance than where $k = 12$ and does appear to show an upward trend of matches in relation to the amount of available faces in the training set. Given that the training sets are composed from previous lectures it follows that more frequently attending students are captured and/or the lecture contains faces that are better aligned.

Figures 4.34 and 4.35 consider the results on a per student basis. Again, these are broken down by method and compared over the k-nearest neighbour range. The colour bar indicates the number of correct training set matches that have been made for the student.

In this instance, most of the students are correctly matched frequently with exception of the student identified as SID 1. Figure 4.36 shows example output from the system where the ensemble method has been used and $k = 4$ is used for the nearest neighbourhood classification. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched. Visually, it is easy to see why the face image from the lecture is matched to SID 1 given their similarities. This is especially true when you compare this face image with the correct image. This can be seen by reviewing the face image incorrectly matched with SID 9 in Figure 4.36.

As with lecture 12, the example output presents more students as being matched (correctly or not) than the 7 students human inspection believed would have been matched by the system. Again, this is because the system assumes there will be 12 attendees. From analysis it has deemed that it was able to match 9 students from the face images taken.

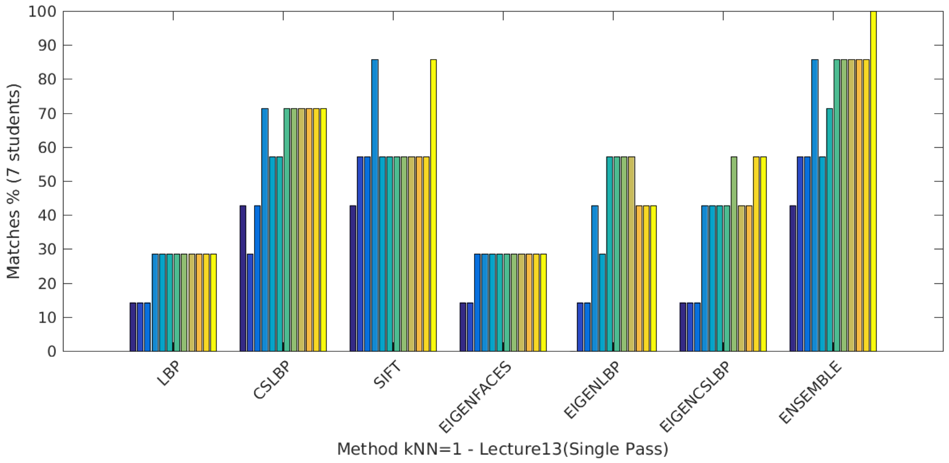


Figure 4.31: A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 1$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

Conclusion

The single pass evaluation of the lecture series was performed to identify each face image as the student for which there is the best match in the training data. Overall, the performance of the system performed well successfully matching 60% of the students. This improvement over the results from the synthetic LFW lecture could be explained by a variety of reasons. First, the time scale between each lecture is less than a week and the students tend not to change so quickly. Additionally, the environmental conditions of the lectures is similar including the lighting and the lecture locations.

From the two lectures evaluated it was clear that lecture 12 performed badly relative to the above statistic as it only successfully matched around 2 of the 5 students expected. The most apparent explanation for this was poorly aligned data but the lack of captured face images also proved difficult for the system to handle. Lecture 13 had much more available face images and performed better. It was also seen that one student had difficulty being matched to an attendee of the lecture because another attendee resembled that student more closely that the face image of the actual student.

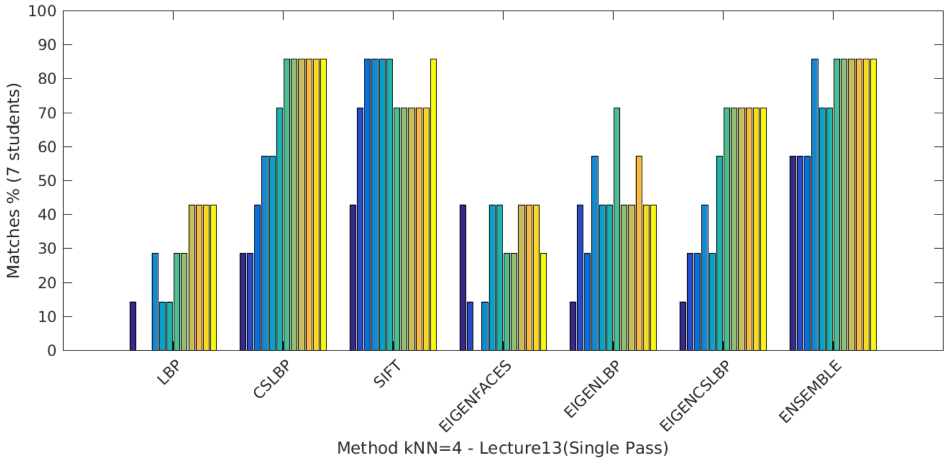


Figure 4.32: A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 4$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

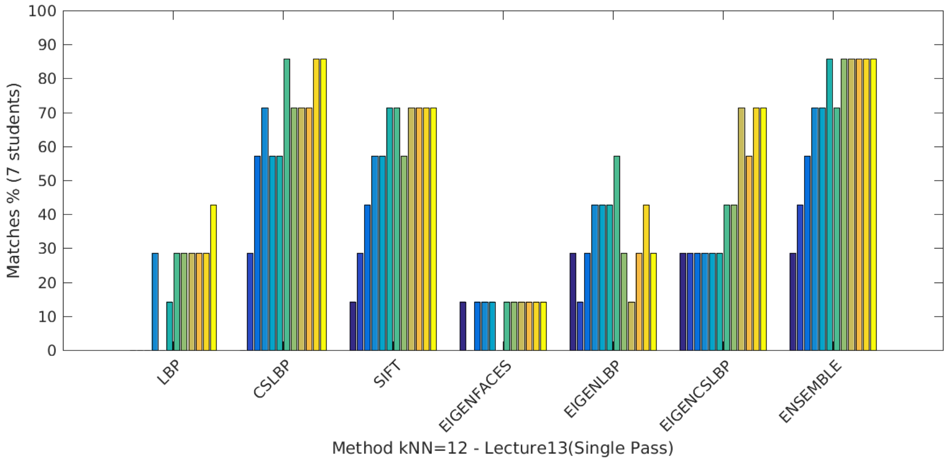


Figure 4.33: A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 12$; using the single pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

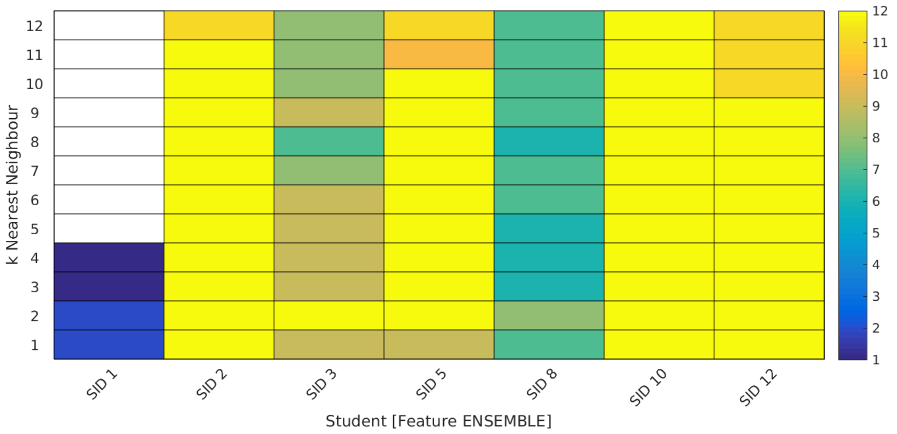
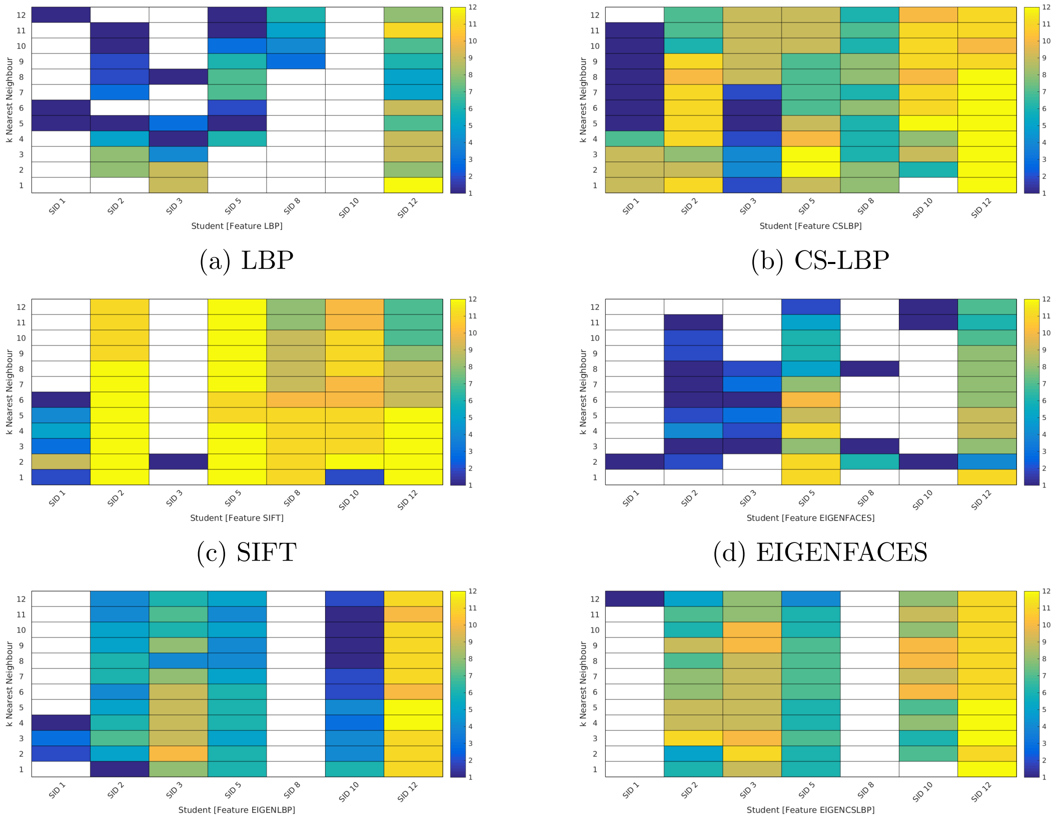


Figure 4.34: Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the single pass approach on lecture 13.



(a) LBP

(b) CS-LBP

(c) SIFT

(d) EIGENFACES

(e) EIGEN LBP

(f) EIGEN CS-LBP

Figure 4.35: Breakdown on a per student basis using individual methods over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the single pass approach on lecture 13.

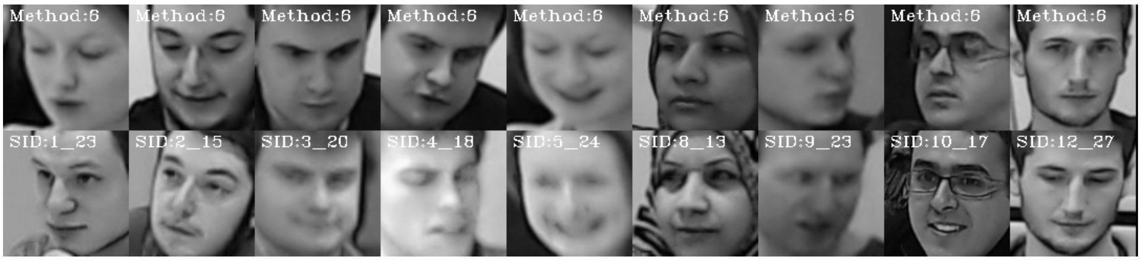


Figure 4.36: Example output from the system where the ensemble method is used with nearest neighbour $k = 4$; using the single pass approach on lecture 13. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched.

4.5.2 Multi Pass Results

The results of the multiple pass evaluation of the recognition algorithm using the lecture series are presented here. As with the single pass experiment, Figure 4.37 shows the evaluation of the lecture series using the collective results taken from the individual feature extraction techniques (LBP, CS-LBP, SIFT, Eigenfaces, Eigen LBP and Eigen CS-LBP) when classified using the k-nearest neighbour range from $k = 1 \ldots 12$ and applying our inversely proportional bias weights. Again, the results have been averaged over the entire 21 training sets and shows the respective matches for the lecture. The fundamental difference is that the algorithm has been modified to iteratively select the best candidate before removing that target face and the student from the training set data until all faces have been matched (see Section 3.4).

The multiple pass approach performs a little better than the single pass equivalent approach successfully matching around 65% of the 12 attendees whereas the single pass approach matched 60%. From visual inspection $k = 6$ performed the best although not significantly better than any of the others. The nearest neighbour ($k = 1$) performed reasonably well at 63%.

Reviewing the results of the ensemble method over three training sets (1, 5 and 10) where $k = 1, 6$ and 12, figures 4.38, 4.39 and 4.40 show the respective matches on a lecture by lecture basis. From these, it is clear that even with the multiple pass approach, lecture 12 performed badly and lecture 13 performed reasonably well suggesting the technical issues discussed in the review using the single pass approach affect both systems.

Lecture 12 Review (Multiple Pass Approach)

Breaking the results for lecture 12 down by method where $k = 1, 6$ and 12, figures 4.41, 4.42 and 4.43 show the system performing recognition over each of the prior 11 training sets on a per method basis. The figures reveal that the system improved its performance over the single pass approach by successfully matching a maximum of 4 of the 5 expected to be detected students. Although this was not consistent across all methods.

SIFT performed the best but was closely followed by the LBP based methods. The ensemble method, which selects the class with the majority from the summation of the

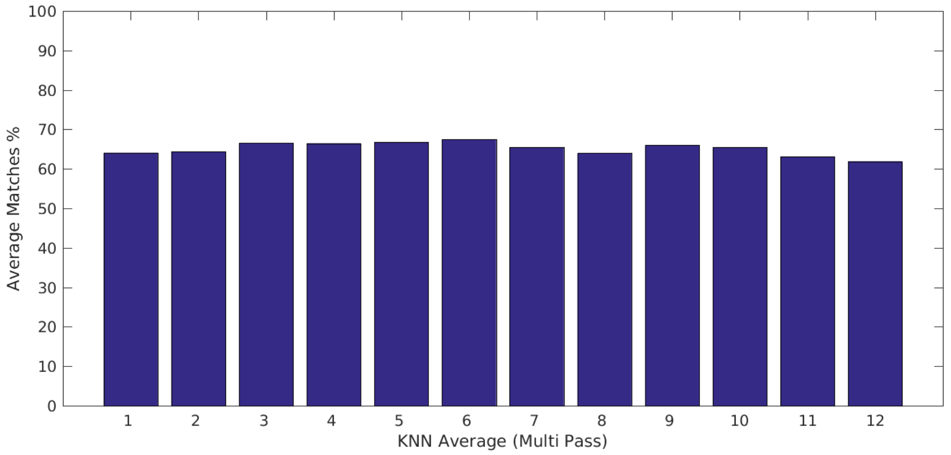


Figure 4.37: A bar chart showing the evaluation of the lecture series over the k-nearest neighbour range; using the multiple pass approach.

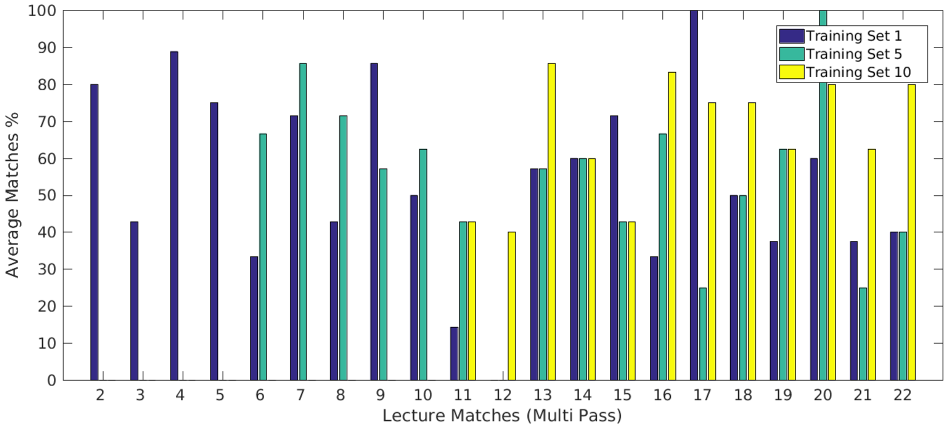


Figure 4.38: A bar chart showing percentage matches over three training sets with nearest neighbour $k = 1$; using the multiple pass approach.

nearest neighbours of all of the individual methods, did not perform to the level of these methods. The system is designed to use the results of the individual methods equally in the ensemble and so it follows that the strength of SIFT is diluted in that instance.

Unlike the previous results, $k = 12$ produced the best results when compared with $k = 1$ and 6. This unforeseen effect is likely because of the way the multiple pass approach works. As the algorithm progresses the best student matches and all corresponding data is removed from the training data. This means that they cannot interfere with the results of any subsequent nearest neighbour classification. In all likelihood, when $k = 12$ there are not enough classification nodes still available towards the end of the multiple pass to make the classification impartial.

Figures 4.44 and 4.45 consider the results on a per student basis. Again, these are broken down by method and compared over the k-nearest neighbour range. The colour bar indicates the number of correct training set matches that have been made for the student. In line with our expectations, based on the review on the single pass approach of lecture 12, it is noticeable that the students identified as SID 1 and SID 12 match much more

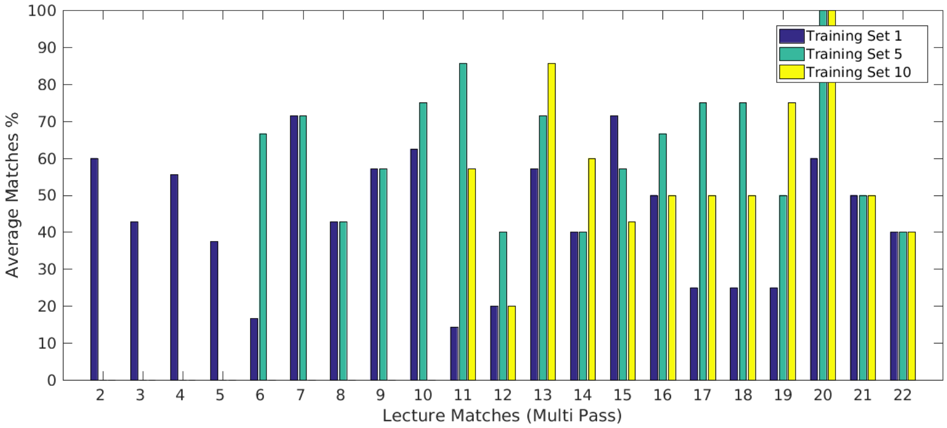


Figure 4.39: A bar chart showing percentage matches over three training sets with nearest neighbour $k = 6$; using the multiple pass approach.

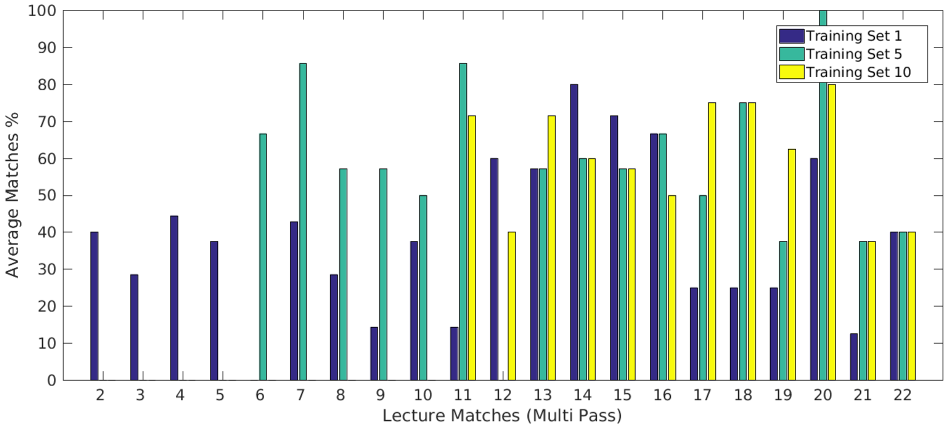


Figure 4.40: A bar chart showing percentage matches over three training sets with nearest neighbour $k = 12$; using the multiple pass approach.

frequently than the others. Other students do now get matched and this can be attributed to having removed the stronger students (SID 1 and SID 12) from the pool of available data.

Figure 4.46 shows example output from the system where the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, has been used and nearest neighbour $k = 6$. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched. Again, poor alignment of the face images taken from lecture 12 is evident and will make matching from previous lectures difficult.

It is noticeable that only 9 of the total 12 students have been matched (correctly or not). In the single pass system we anticipated that the system would cluster face images around stronger student matches and that could result in less than 12 student matches, especially given that the 12 students may not have attended; in this case only 5 did attend. However, the multiple pass approach attempts to match each face image to a student until all students have been found. The explanation is simple; of the 5 students that attended only 9 face images were captured. The system simply ran out of data to process. Clearly,

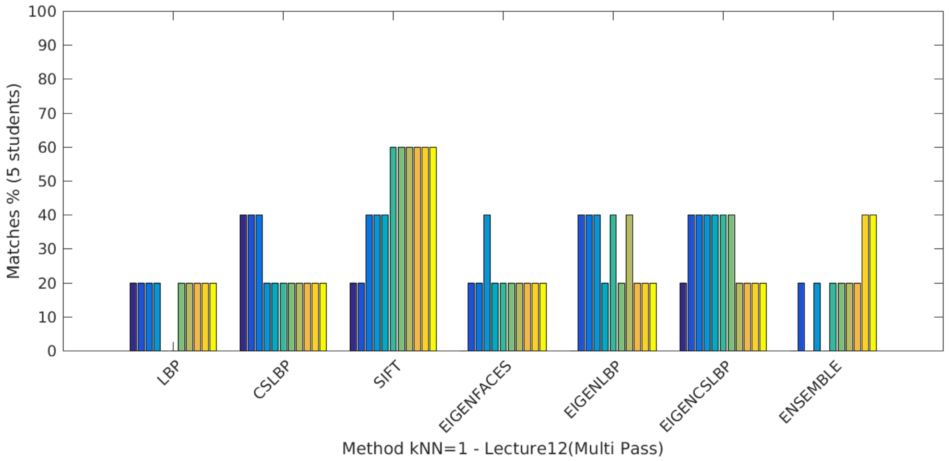


Figure 4.41: A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 1$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

the fact that the system is trying to match 12 students when only 5 actually attended (or were deemed to be captured sufficiently) is cause for concern and will be discussed in Section 4.6.

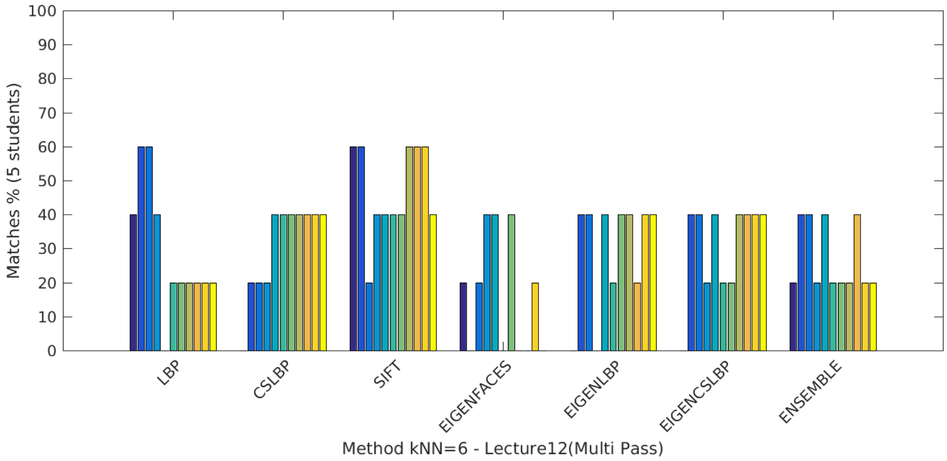


Figure 4.42: A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 6$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

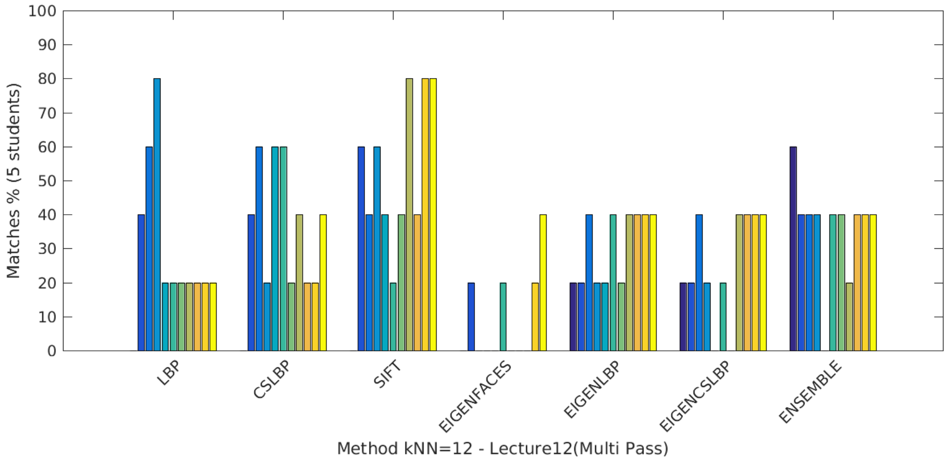


Figure 4.43: A bar chart showing the evaluation of the lecture 12 in terms of the method employed over each training set with nearest neighbour $k = 12$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

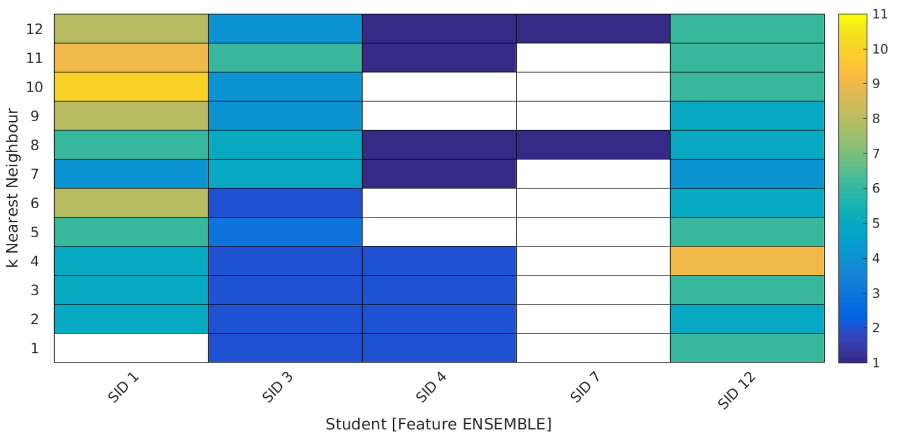
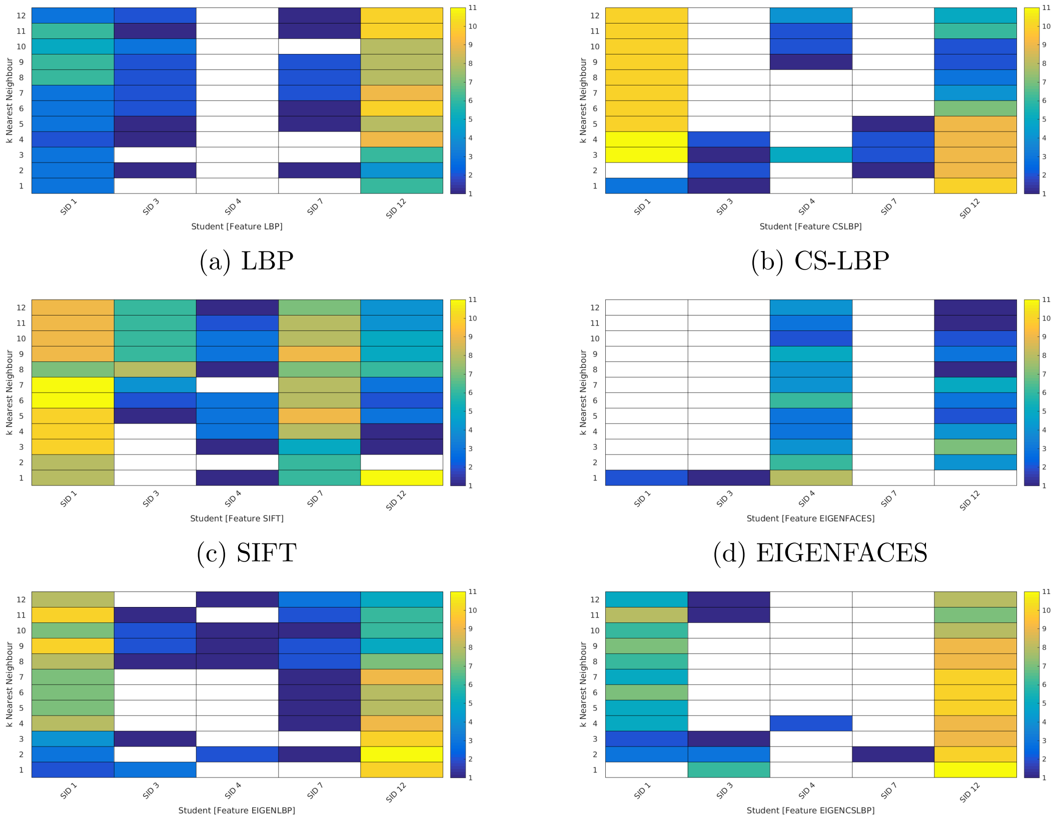


Figure 4.44: Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on lecture 12.



(a) LBP

(b) CS-LBP

(c) SIFT

(d) EIGENFACES

(e) EIGEN LBP

(f) EIGEN CS-LBP

Figure 4.45: Breakdown on a per student basis using individual methods over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on lecture 12.

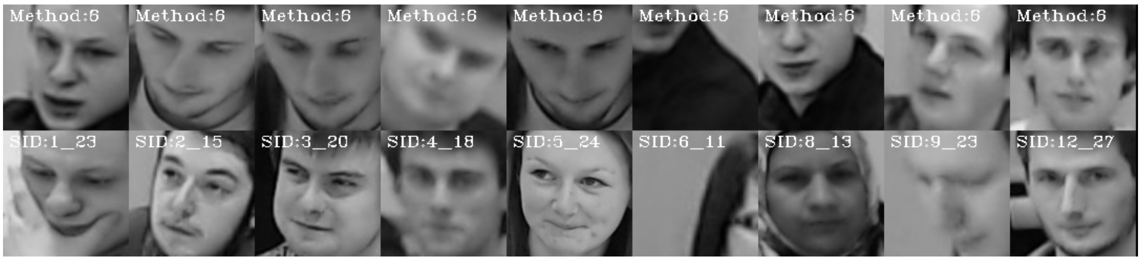


Figure 4.46: Example output from the system where the ensemble method is used with nearest neighbourhood $k = 6$; using the multiple pass approach on lecture 12. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched.

Lecture 13 Review (Single Pass Approach)

Breaking the results for lecture 13 down by method where $k = 1, 6$ and 12, figures 4.47, 4.48 and 4.49 show the system performing recognition over each of the prior 11 training sets on a per method basis. The figures reveal that the system successfully matched all of the students with a consistently strong matching of students across the board. Again, CS-LBP and SIFT performed the best but the other independent methods performed well and positively tracked the new data added to the training sets. The ensemble produced a strong response especially when $k = 1$.

Again, the difference between the CS-LBP method and its closely related LBP method is also apparent. Both extract a similar feature set from the aligned face images and, in this case, performed well when tested in the spectral domain. Overall the standard LBP method performed worst.

In this instance $k = 6$ produced an overall stronger performance than when k was either 1 or 12 and all methods appear to show an upward trend of matches in relation to the amount of available faces in the training set. The sparse neighbourhood problem considered in the analysis of lecture 12 is not seen in the analysis of lecture 13. It was stated that as the stronger students, together with their associated data, are removed from the training data that the nearest neighbourhood becomes sparsely populated with only a few classification candidates to choose from. It is possible that lecture 13 is affected by this but has not been so fortunate to correct classify the face image to the student as much as it could in lecture 12.

Figures 4.50 and 4.51 consider the results on a per student basis. Again, these are broken down by method and compared over the k-nearest neighbour range. The colour bar indicates the number of correct training set matches that have been made for the student.

In this instance, all of the students are correctly matched frequently with some students being matched much more frequently than others. This is to be expected given that the available training data may favour individual students e.g. a collection of similar poses for the student in question. The students identified as SID 10 and SID 12 are good examples of this and can be seen in Figure 4.52 which shows example output from the system where the ensemble method has been used and nearest neighbour $k = 6$. The top row of the

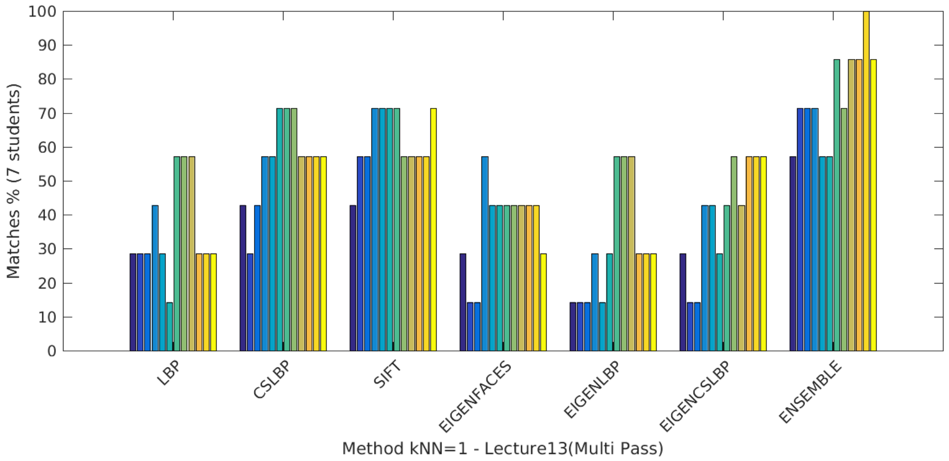


Figure 4.47: A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 1$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched. Furthermore, these two student performed well in the single pass approach version of the system.

Unlike the results of lecture 12, all 12 students have been matched (correctly or not). Given that only 7 students were deemed to have been attended then this presents an issue and is discussed further in Section 4.6. Again, the explanation for the matching of the 12 students is straight-forward. 19 face images were captured in this lecture and so all 12 students could be assigned one of those faces.

Conclusion

The multiple pass evaluation of the lecture series was performed to identify each face image as the student for which there is the best match in the training data. Unlike the single pass approach, this algorithm systematically removed the strongest match between a face image and student in the training data before iteratively doing this for all students. In the case of lecture 12 we faced two related problems. First, there were alignment issues which adversely affected the results. Additionally, there were only 9 face images captured of the 5 students which presented a challenge in itself.

Lecture 13 clearly performed better than lecture 12. In this instance, lecture 13 had captured 19 face images so could operate within a deeper pool of resources.

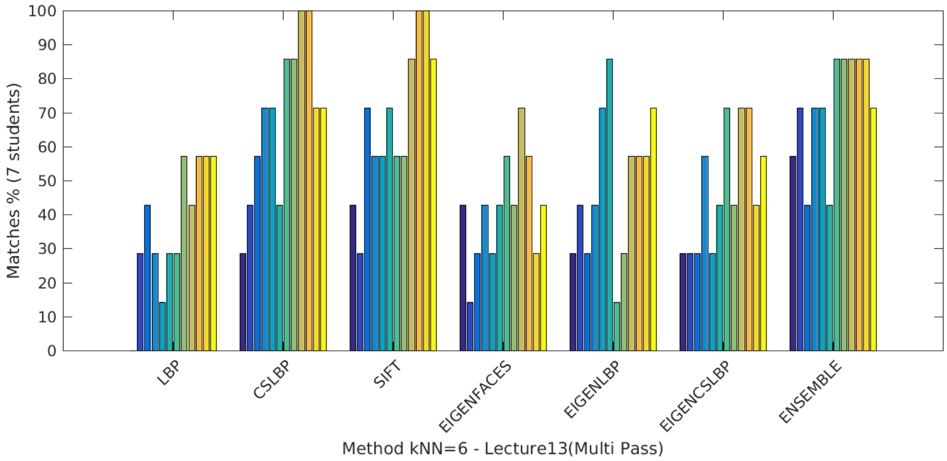


Figure 4.48: A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 6$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

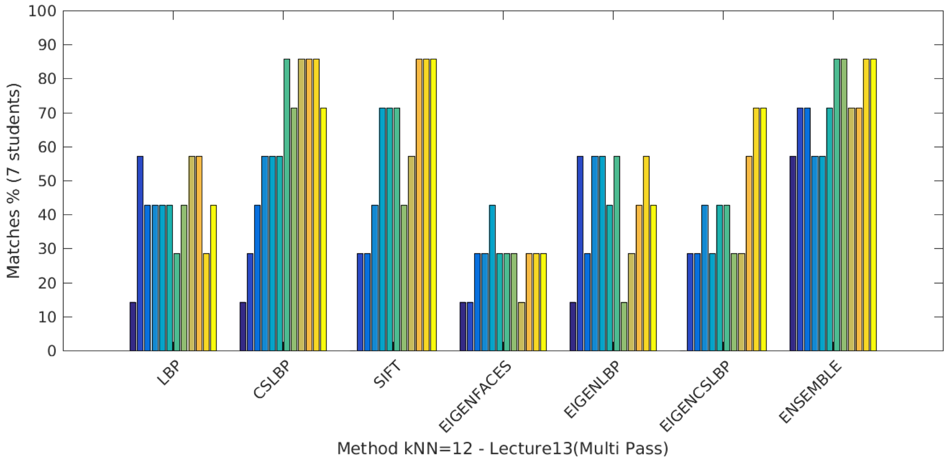


Figure 4.49: A bar chart showing the evaluation of the lecture 13 in terms of the method employed over each training set with nearest neighbour $k = 12$; using the multiple pass approach. Each bar (column) indicates an individual training set's results for the method; beginning with the first bar being training set 1.

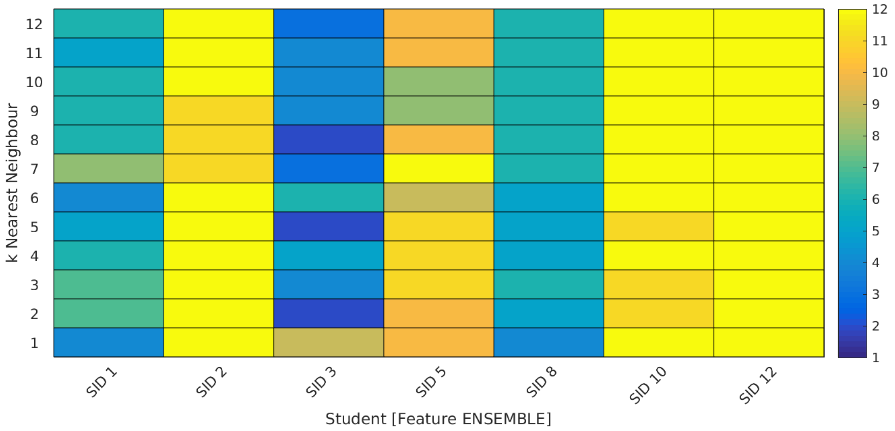
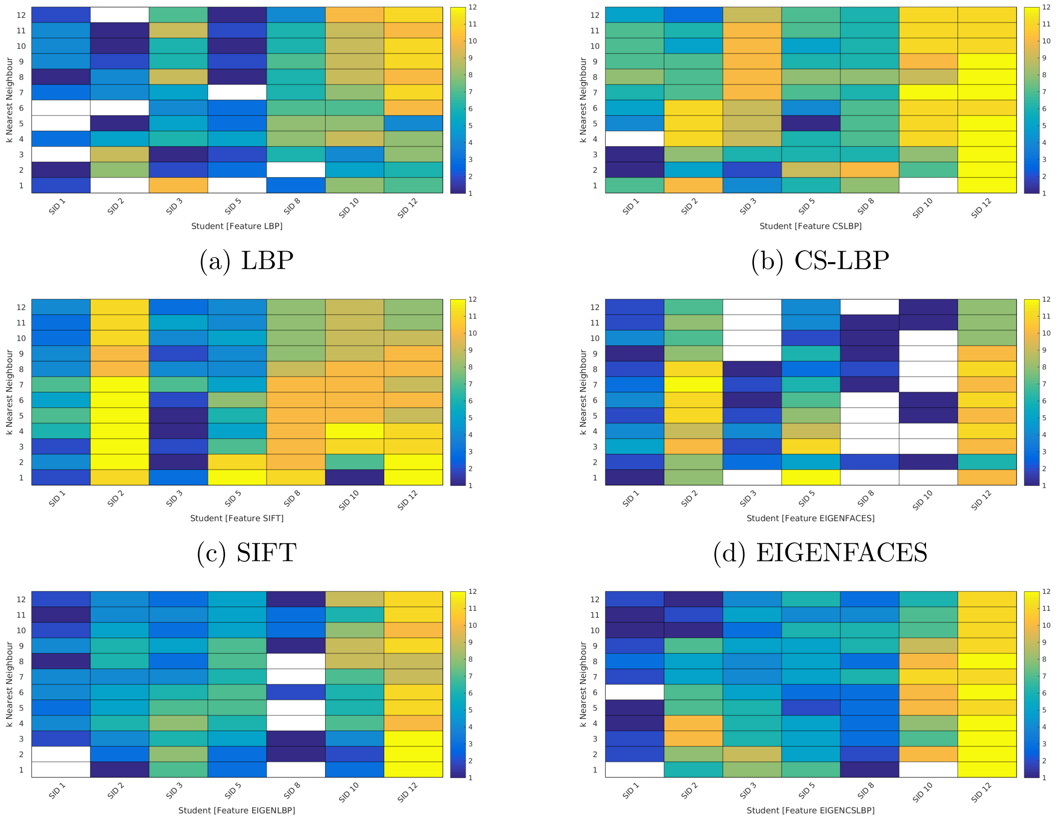


Figure 4.50: Breakdown on a per student basis using the ensemble method, which selects the class with the majority from the summation of the nearest neighbours of all of the individual methods, over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on lecture 13.



(a) LBP

(b) CS-LBP

(c) SIFT

(d) EIGENFACES

(e) EIGEN LBP

(f) EIGEN CS-LBP

Figure 4.51: Breakdown on a per student basis using individual methods over the entire nearest neighbourhood range and colour coded based on the number of training set matches; using the multiple pass approach on lecture 13.

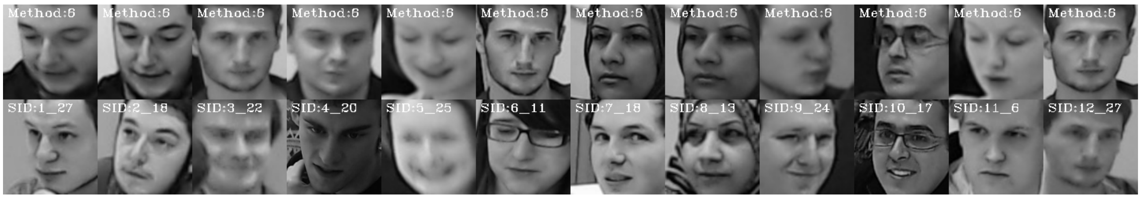


Figure 4.52: Example output from the system where the ensemble method is used with nearest neighbour $k = 6$; using the multiple pass approach on lecture 13. The top row of the figure represents the face images captured at the lecture and the bottom row shows the student to which they were matched.

4.6 Discussion

Research showed that it is highly desirable to consider dimensionality reduction within the overall strategy of any classification system, mainly to concentrate on key facets of data together with the removal of noise. Principal Component Analysis (PCA) lends itself to such an application and was employed in a number of different scenarios including assisting Active Shape Modelling (ASM) to fit face models and as a key feature extraction technique for methods such as Eigenfaces.

The experiments reviewed in this chapter showed that the spectral methods (Eigenfaces, Eigen LBP and Eigen CS-LBP) performed reasonably well but never out performed either SIFT or CS-LBP methods. There are a number of explanations for why this may be the case including that the experiments retained only the principal components that formed 90% of the variance (see Section 4.5). In fact, in the example training set used in Figure 4.20 it was clear that the major differences in variance dissipated were captured in the first 15 eigenvalues suggesting that our 90% constraint was too much especially where the local binary based methods were concerned. Future work could, as with the range tested for the nearest neighbourhood classification, evaluate a range of retained variances from a coarse to fine representation. Clearly, this range testing principle could also be extended to test the radius and members of the local binary pattern methods as these methods were constrained to their nearest pixels.

Principal Component Analysis based methods are also known to be sensitive to the background of the image [Turk & Pentland, 1991]. Removal of the background from the face image would therefore be an obvious goal. Arguably, good face fitting using ASM used in the alignment stage of the pipeline would allow us to do this because we would be able to completely identify the face. However, these experiments revealed that whilst ASM was very good at finding eyes for alignment, it failed to fit the entire face. Whilst PCA is also used within ASM; in this instance the technique is not affected by background as it is only concerned with a set of labelled coordinates defining the shape model. Instead, the problem is more likely to be an issue of poor contrasting at the key points of the face. Another factor is the choice of labelled wire frames in the system's training set.

Analysis of the Viola Jones algorithm showed that it underperformed human inspection by approximately 20%. The most obvious explanation was found to be the rotation of the camera relative to the attendees. Furthermore, the detection stage was hindered by the multitude of different poses that included those that were not frontal face poses. Clearly, the choice of training data for the algorithm plays a key role in its ability to detect faces. In this case the readily available training set from OpenCV was used that was specifically trained for frontal face poses although alternatives were available and could be used in future work.

Whilst it is relatively simple to rotate images in two dimensions to allow better detection, the preferred solution would be to maintain the camera in a fixed position such that rotations are unnecessary. Attempting to capture good poses could be achieved by asking the attendees to look forward during a registration event and the system would definitely benefit by insisting on this during a formal enrolment session. Whilst this is convenient for the system, it does start to question the passive nature of face recognition systems advocated so highly by this thesis. It is probable that all attendees will assume the correct pose at some point during the lecture and it is possible that a video stream could capture this and restore the passive behaviour of the system. Especially if deployed in a fixed position. Further work could examine how feasible this would be in practice.

The recognition system employed a simple classification technique in the form of k-nearest neighbourhood (kNN). The nearest neighbourhood strategy appeared to perform satisfactorily, producing a reasonable set of results. For the effort required, simply using the nearest neighbour for classification proved fruitful yielding results only slightly worse than the best k value tested. The implication of this is that the system identifies similar faces easily; something that should not surprise.

Another aspect of classification that requires a re-examination is the bias equation employed by the system. A simple method of inversely proportionally weighting the images in the training set is potentially flawed where there is a large disparity of available images combined with a small set of training set data. Consider the case where one attendee A has twenty available images in the training set where another attendee B has only one. If we were to weigh A's images as one twentieth then classification would potentially incorrectly favour B under circumstances where B's single image is in the neighbourhood. e.g. nineteen images of A would equal 0.95 whereas one image of B would equal 1.0. As the number of neighbours increases we are likely to witness this if the training set data size is small. Whilst not presented in this thesis, early experiments performed on the synthetic lecture, without the bias, highlighted the problems of classification where one student has too many images available. Most face images were incorrectly matched as Nicole Kidman because she had 6 times more labelled images available in the training set.

Overall, two distinct strategies were considered for the attendance monitoring system. In the first each student's face image is identified as the student for which there is the best match in the training data and is shown as pseudo code in Section 1. Whereas, a second multiple pass algorithm evaluates the quality of the match between each face image and all the training images. The face image with the best match is identified with the matching training image and that face image and the training images corresponding to the matched student are removed from the available training images.

Computationally, there was a negligible increase in the resources required for the multiple pass approach over the single pass. This is because the majority of resources used were consumed up to and including the feature extraction stage of the pipeline, which both strategies had to perform. Irrespective of the strategy employed, the system consumed more resources as the data in the training set increased because it had more comparisons to perform. Furthermore, the time taken to determine the principal components and associated average face increased approximately proportionally to the number of faces in the training set N^3. This is unlike the other methods which performed feature extractions independently for each face image added to the data. Admittedly, training set processing is not required to be performed in real time.

What did become apparent by evaluating these two strategies was the lack of a distinct winner. Whilst the multiple pass approach (see Figure 4.12) produced slightly better results than the single pass approach (see Figure 4.5) when testing the synthetic lecture they both performed poorly as they correctly matched under 30% of the expected attendees. It appeared that public images taken over varying time-scales in multiple poses could be the reason for such a performance issue although we cannot discount the impact of the varying backgrounds associated with these images. By contrast, the real lecture series performed significantly better and reinforced the idea that shorter time-scales may improve matching success as there was a lower chance of changes to the individuals. Again, the distinction between the two approaches is not clear with the multiple pass approach (see Figure 4.37) slightly ahead of the single pass approach (see Figure 4.21) successfully matching 65% and 60% on average, respectively.

What was noticeable was how the two strategies performed with limited data. The system detected 9 face images from the images captured for lecture 12. From human inspection, only 5 students were actually captured whereas the system was expecting 12. The single pass approach attempted to match 7 candidates whilst the multiple pass approach matched all 9 face images. Similarly, the system detected 19 face images from the images captured during lecture 13 where, after human inspection, only 7 students were classed as actually being captured. In this instance, the single pass approach attempted to match 9 students whereas the multiple matched all 12 students. In other words, the multiple pass system always attempts to match up to the expected number of attendees or until it exhausts all available face images in the process. Clearly, when the number of attendees is lower than expected there will be an unacceptable number of false positives in the multiple pass approach.

One potential solution for this is to look for matches between face images taken in the same lecture in order to identify groups of images all belonging to the same person before removing the identified student and all associated data. Another possibility would be to attempt to cluster the face images based on similarities between themselves or even based on their location in a similar way to how Taylor and Morris [Taylor & Morris, 2014] flatten their detected faces but across multiple images. Again, further investigation is required and this would certainly be improved if a fixed position camera was deployed.

In this set of experiments the true identity of the faces in the training sets are known through-out. In reality, the training sets are built dynamically using the recognition results from previously tested lectures. Without human inspection errors can creep into the system especially where a student who did not attend the first lecture subsequently attends. This clearly demonstrates the need for formal registration events even if only using their student ID photographs. Furthermore, therecognition system has to overcome the natural challenges of face pose, expression and the conditions of the lecture theatre without exacerbating the problem posed by the multiple pass approach. Alia et al's [Alia et al., 2013] attendance monitoring system allowed the lecturer to intervene and correct recognition errors prior to insertion into the wider administration system, something that could be considered in this case.

It was also observed that the trend for matches when evaluated over the entire k-nearest neighbour range peaked early into the experiments for both the synthetic lecture and the single pass lecture series. The exception being the multiple pass approach reviewed over lectures 12 and 13. Again, this is an unforeseen characteristic of the multiple pass system when presented with limited data. As the system progresses it removes the best student matches and their associated data. This means that the pool of available neighbours decreases and so the classification is limited in its selection. In the case of the multiple pass approach applied to the synthetic lecture, it simply did not match enough students to be affected by this sparseness.

Finally, the simplistic nearest neighbourhood classification method was employed for these experiments. Future work could involve expanding the classification system to include Support Vector Machines (SVM), neural networks and others.

4.7 Conclusion

This chapter described a series of experiments using our recognition system. It began by considering the effectiveness of the detection stage and showed that Viola Jones algorithm detected circa 80% of what was deemed detectable by human inspection. An explanation for this difference was offered which suggested that the rotation of the camera relative to the audience was beyond the bounds to the training set used to train the Viola Jones algorithm and that the algorithm also performed badly when the rotation was too extreme.

The alignment stage of the pipeline was also tested in isolation. In particular the fitting of a model to each face using Active Shape Modelling was verified and it was concluded that around 75% of all faces tested fitted well when inspection concerned itself with looking at how the ASM algorithm fitted eyes but performed poorly when reviewing its ability to fit the entire face. Given that the alignment stage of the recognition algorithm uses only the eyes the observed failing of the ASM algorithm to fit the entire model to the face was deemed irrelevant in this context.

The evaluation of the synthetic lecture series using faces taken from the Labelled Faces in the Wild data set proved a disappointment matching only a handful of attendees out of a total of 19 students. Whilst disappointing, it suggested that photographs taken over longer periods of time could result in a loss of similarity. Nicole Kidman is a good example where her appearance has radically changed over the years. Fortunately, in Nicole Kidman's case there is a sufficiently large pool of training data to withstand such changes of appearance. If this theory was to hold we would expect to, and did, see better matches using the lecture series because it only spanned ten weeks and allows for continuously updating of the training data with time.

Our final set of tests concerned the real lecture series. Overall, the success rate of the system was significantly better than the synthetic lecture but highlighted problems that the synthetic set could never have. One such example is seen clearly when the number of attendees that actually attended is fewer than expected. The multiple pass approach matched every expected attendee to the available face images captured in the lecture without hesitation. This resulted in an undesirably high number of false positives especially in the case where there was a low lecture turn out.

5 Conclusion

The problem of verifying/determining lecture attendance is well known and is concerned with balancing the effort required to capture the results against the their reliability. This thesis outlined the two extremes of this; a lecturer whose time was consumed by taking a register and the case where the attendees signed a register based on trust. It also considered technological advances but suggested that such advances did not necessarily reduce the effort required, especially if the system was difficult to use. The introduction of the thesis suggested a motivation for universities to capture attendance, namely as international student visa sponsors they have an obligation to prove student engagement.

The thesis introduced the idea of using face recognition as a mechanism to register attendees, accepting that this would be a difficult environment for it to operate within due to its unconstrained nature. Chapter 2 described the building blocks that could be used as the foundation of such a system and provided a literature review of these components and existing systems. It emphasised the importance of dimensionality reduction, suggesting that this was common to most of the building blocks. It also suggested that the ultimate system would be a recognition system rather than a verification based system because, whilst they are known in advance, the attendees do not make any attempt to prove who they are.

Chapter 3 defined the attendance monitoring system and introduced a standard recognition pipeline which incorporated many of the techniques described in Chapter 2. These included using the de facto standard Viola Jones algorithm for detection, Active Shape Modelling (ASM) to assist in the alignment of the face; and a variety of feature extraction techniques including Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBP) and Eigenfaces. Finally, two basic strategies were introduced. In the first each student's face image is identified as the student for which there is the best match in the training data. A second multiple pass algorithm considered the quality of the match between each face image and all the training images. The face image with the best match is identified with the matching training image and that face image and the training images corresponding to the matched student are removed from the available training images.

The thesis evaluated the attendance monitoring system against a synthetic lecture taken from the Labelled Faces in the Wild (LFW) data set and against a complete lecture series. It first considered the effectiveness of the detection algorithm showing a reasonable, but not perfect, success when compared against those deemed detectable via human inspection. It then examined the system's ability to align the faces before running and presenting the results from a series of tests executed on both the synthetic and genuine lecture data using both the single and multiple pass approaches.

Ultimately, analysis of a series of 23 lectures over a ten week period was performed using the attendance monitoring system. Each previous lecture was labelled and added to the training set on the assumption that an improvement to the positive matches would be seen. As the results show, face recognition systems have the potential to correctly identify lecture attendees. The technology is easy to use and requires little or no effort at the point of lecture delivery. It is certainly capable of improving the ratio between effort employed and reliability of the results.

Whilst face pose, expression and the environment conditions of the lecture theatre all challenge the system several improvements could be made to mitigate this. First, a fixed forward-facing camera would remove some of the detection and alignment issues faced by the system especially if the students were asked to look forward for a registration event. Additionally, capturing many more images during the lecture would assist in better matching as there would be a wider pool of face images to utilise.

Furthermore, a review of matching strategies could be performed to find less error prone mechanisms. Whilst, the multiple pass system would appear to be robust it was clearly demonstrated that it is an inappropriate matching strategy as it yielded a high false positive rate because it matched all expected student without hesitation. Consequently, human intervention is almost certainly required as a fail safe. One final improvement could be the introduction of better classifiers including Support Vector Machines and neural networks; or an ensemble of these.

In its current form, the student attendance monitoring system is crude and at the early stages of development. For this reason it is not a system suitable for general use. However, with the improvements mentioned above it could become a viable system.

Bibliography

[Ahonen et al., 2004] Ahonen, T., Hadid, A., & Pietikäinen, M. (2004). *Face Recognition with Local Binary Patterns*, (pp. 469–481). Springer Berlin Heidelberg: Berlin, Heidelberg.

[Alia et al., 2013] Alia, M., Tamimi, A., & Al-Allaf, O. (2013). Integrated System For Monitoring And Recognizing Students During Class Session. *AIRCCs: International Journal Of Multimedia & Its Applications (IJMA)*, 5(6), 45–52.

[Belhumeur et al., 1997] Belhumeur, P., Hespanha, J., & Kriegman, D. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7), 711–720.

[Bishop, 1995] Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Advanced Texts in Econometrics. Clarendon Press.

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

[Chenxing, 2012] Chenxing (2012). ASM implementation in OpenCV. `https://github.com/cxcxcxcx/asmlib-opencv`.

[Cootes, 2000] Cootes, T. (2000). Model-Based Methods in Analysis of Biomedical Images. In *Image Processing and Analysis* (pp. 223–248).

[Cootes et al., 1995] Cootes, T., Taylor, C., Cooper, D., & J.Graham (1995). Active Shape Models - Their Training and Application. *Computer Vision and Image Understanding*, (pp. 38–59).

[Cortes & Vapnik, 1995] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

[Eon et al., 1999] Eon, P. S., Simard, P. Y., Haffner, P., & Lecun, Y. (1999). Boxlets: a Fast Convolution Algorithm for Signal Processing and Neural Networks. In *Advances in Neural Information Processing Systems* (pp. 571–577).: MIT Press.

[Freund & Schapire, 1999] Freund, Y. & Schapire, R. E. (1999). A Short Introduction to Boosting. *Japanese Society for Artificial Intelligence*, (pp. 771–780).

[Fukunaga, 2013] Fukunaga, K. (2013). *Introduction to Statistical Pattern Recognition*. Computer science and scientific computing. Elsevier Science.

[Heikkilä et al., 2006] Heikkilä, M., Pietikinen, M., & Schmid, C. (2006). Description of Interest Regions with Center-Symmetric Local Binary Patterns. In P. Kalra & S. Peleg

(Eds.), *Computer Vision, Graphics and Image Processing*, volume 4338 of *Lecture Notes in Computer Science* (pp. 58–69). Springer Berlin Heidelberg.

[Huang et al., 2012] Huang, G. B., Mattar, M., Lee, H., & Learned-Miller, E. (2012). Learning to Align from Scratch. In *NIPS*.

[Huang et al., 2007] Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report 07-49, University of Massachusetts, Amherst.

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer.

[Joseph & Zakharia, 2013] Joseph, J. & Zakharia, K. (2013). Automatic Attendance Management System Using Face Recognition.

[Kirby & Sirovich, 1990] Kirby, M. & Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1), 103–108.

[Lei et al., 2014] Lei, L., Kim, D.-H., Park, W., & Ko, S. (2014). Face recognition using LBP Eigenfaces. *IEICE Transactions on Information and Systems*, (pp. 1930–1932).

[Lenc & Krl, 2013] Lenc, L. & Krl, P. (2013). Matching Methods for Automatic Face Recognition using SIFT.

[Liu et al., 2011] Liu, C., Lu, J., & Li, L. (2011). Three-level face features for face recognition based on center-symmetric Local Binary Pattern. In *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on*, volume 4 (pp. 394–398).

[Lowe, 1999] Lowe, D. G. (1999). Object Recognition from Local Scale-Invariant Features. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2* Washington, DC, USA: IEEE Computer Society.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

[Luo et al., 2007] Luo, J., Ma, Y., Takikawa, E., Lao, S., Kawade, M., & Lu, B. L. (2007). Person-Specific SIFT Features for Face Recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 2 (pp. II–593–II–596).

[Majumdar & Ward, 2009] Majumdar, A. & Ward, R. (2009). Discriminative SIFT features for face recognition.

[Meena & Suruliandi, 2011] Meena, K. & Suruliandi, A. (2011). Local binary patterns and its variants for face recognition. In *Recent Trends in Information Technology (ICRTIT), 2011 International Conference on* (pp. 782–786).

[Mohamed Aly, 2006] Mohamed Aly (2006). Face Recognition using SIFT Features.

[Picard & Cook, 1984] Picard, R. R. & Cook, R. D. (1984). Cross-Validation of Regression Models. *Journal of the American Statistical Association*, 79(387), 575–583.

[Pietikinen, 2011] Pietikinen, M. (2011). *Computer vision using local binary patterns.* London New York: Springer.

[Roth et al., 2000] Roth, D., Yang, M., & Ahuja, N. (2000). A SNoW-Based Face Detector. In *Advances in Neural Information Processing Systems 12* (pp. 855–861).: MIT Press.

[Rowley et al., 1998] Rowley, H. A., Baluja, S., & Kanade, T. (1998). Neural Network-Based Face Detection. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 20(1), 23–38.

[Schneiderman, 2000] Schneiderman, H. (2000). A statistical approach to 3d object detection applied to faces and cars. In *PhD Thesis* (pp. 0–6).

[Shilwant & Karwankar, 2012] Shilwant, D. S. & Karwankar, A. (2012). Student Monitoring By Face Recognition System. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, 2(2), 24.

[Shlens, 2005] Shlens, J. (2005). A tutorial on Principal Component Analysis. In *Systems Neurobiology Laboratory, Salk Institute for Biological Studies.*

[Sirovich & Kirby, 1987] Sirovich, L. & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A*, 4(3), 519–524.

[Taylor & Morris, 2014] Taylor, M. & Morris, T. (2014). Enhanced Face Detection: An Adaptive Cascade-Mixture Approach for Large-Scale Detection. In *British Machine Vision Conference (BMVC).*

[Turk & Pentland, 1991] Turk, M. & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3, 71–86.

[Viola & Jones, 2001] Viola, P. & Jones, M. (2001). Robust Real-time Object Detection. In *International Journal of Computer Vision.*

[Wang & He, 1990] Wang, L. & He, D.-C. (1990). Texture classification using texture spectrum. *Pattern Recognition*, 23(8), 905 – 910.

[Webb, 2003] Webb, A. (2003). *Statistical Pattern Recognition.* Wiley InterScience electronic collection. Wiley.

[Winston, 2010] Winston, P. (2010). 6.034 Artificial Intelligence. *Massachusetts Institute of Technology: MIT OpenCourseWare.*

[Worsey & Everson, 2015] Worsey, J. & Everson, R. (2015). Facial recognition in an unconstrained environment for the application of monitoring student attendance. In *2015 15th UK Workshop on Computational Intelligence (UKCI) - ACCEPTED* (pp. 1–7).

[Xu & Liang, 2001] Xu, Q.-S. & Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1), 1 – 11.

[Yan et al., 2010] Yan, S., Wang, H., Liu, J., Tang, X., & Huang, T. S. (2010). Misalignment-robust face recognition. *Image Processing, IEEE Transactions on*, 19(4), 1087–1096.

[Yang & Chen, 2013] Yang, B. & Chen, S. (2013). A comparative study on local binary pattern (LBP) based face recognition: LBP histogram versus LBP image. *Neurocomputing*, 120, 365 – 379. Image Feature Detection and Description.