A Novel Multispectral and 2.5D/3D Image Fusion Camera System for Enhanced Face Recognition

William Williams

A thesis presented for the degree of PhD Computer Science and Information Systems

Department of Computer Science and Information Systems

This report is substantially the result of my own work, expressed in my own words, except where explicitly indicated in the text. I give my permission for it to be submitted to the JISC Plagiarism Detection Service.

The report may be freely copied and distributed provided the source is explicitly acknowledged.

Signed:_ _

William Williams

Acknowledgement

This thesis would not have been possible without the encouragement, support and guidance of several people to whom I offer my sincere gratitude and thanks. In alphabetical order they are:

Chloe Harrison

Prof. Steve Maybank

Robin Watling

Sue Williams

Dr William G. Williams

For Peggy

Contents

Abstract

The fusion of images from the visible and long-wave infrared (thermal) portions of the spectrum produces images that have improved face recognition performance under varying lighting conditions. This is because long-wave infrared images are the result of emitted, rather than reflected, light and are therefore less sensitive to changes in ambient light. Similarly, 3D and 2.5D images have also improved face recognition under varying pose and lighting. The opacity of glass to long-wave infrared light, however, means that the presence of eyeglasses in a face image reduces the recognition performance.

This thesis presents the design and performance evaluation of a novel camera system which is capable of capturing spatially registered visible, near-infrared, long-wave infrared and 2.5D depth video images via a common optical path requiring no spatial registration between sensors beyond scaling for differences in sensor sizes. Experiments using a range of established face recognition methods and multi-class SVM classifiers show that the fused output from our camera system not only outperforms the single modality images for face recognition, but that the adaptive fusion methods used produce consistent increases in recognition accuracy under varying pose, lighting and with the presence of eyeglasses.

1 Introduction

This thesis describes the design and development of a novel multispectral camera system which is capable of capturing images in the visible (VIS) near-infrared (NIR) and long-wave infrared (LWIR) spectral bands as well as 2.5D/3D depth data (Section 3.1). To our knowledge, no camera system such as this has been developed and investigated for face recognition applications.

Two databases are generated using the camera system (Section 3.3 on page 39) under varying pose, lighting and eyeglasses. The results of extensive face recognition experiments across multiple image fusion and recognition methods are reported. The experiments explore the performance of the fused output versus single-modality images and attempt to identify an optimum method for automatic, adaptive image fusion (Section 5.3). An eyeglasses-compensation algorithm which automatically detects and synthesises areas occluded by eyeglasses in the LWIR images using information from the VIS images (Section 3.4) is also applied and the results reported in Section 5.1.

This thesis demonstrates that a manually weighted fusion technique yields improved face recognition under varied lighting conditions for a particular set of images, but that a specific set of fusion weights cannot necessarily be applied universally under any variation in lighting. We will show that in such cases, what is commonly achieved is a set of weights suitable only for a particular set of face images. Images fused using more adaptive methods produce more accurate face recognition under varied pose, lighting and with eyeglasses.

We demonstrate that images which are adaptively fused in the transform and feature spaces can be further fused with single modality images in match-score space to improve recognition accuracy and also produce more consistent recognition performance under changing lighting conditions. A method of match-score fusion using the transform fused images and feature fused images is shown to achieve higher recognition and verification rates than methods based on single modalities for all lighting modes used across a range of established face recognition methods.

Finally a feature fusion method utilising the adaptive transform fused image and fused features is described and tested using multiclass SVM classifiers. The feature vector is shown

to give a highly consistent 98% and 94% recognition rate for both lighting modes used in our experiments. From this we conclude that a multispectral camera system using an adaptive fusion algorithm in the transform and feature spaces is an effective solution to face recognition under realistic or poor lighting conditions and varying pose.

2 Background

There are several areas of research which are relevant to this thesis. These can be subdivided as:

- Image fusion: the fusion of information from two or more images in order to produce a single image that is of greater use to an image processing task than the constituent images.
- Image fusion for face recognition: the application of image fusion specifically to face images for the purpose of recognition and verification.
- Adaptive image fusion: a method of image fusion whereby the information in the constituent images is selected for fusion by a process which adapts to variations in the constituent images.
- 2.5D/3D and multi-modal image fusion for face recognition: the fusion of images with data captured using depth sensors and other 2D images for application to face recognition and verification.
- Image fusion technology: camera systems and hardware designed for the capture and fusion of images.

A literature review and discussion of these research areas with regard to this thesis are conducted in sub-sections 2.1 to 2.5 below. A summary of the motivation and contribution of the thesis with regard to the literature review is then given in Section 2.6.

2.1 Image Fusion

Early work in fusing long-wave infrared (LWIR) and visible (VIS) sensor images was started in the late eighties by Burt and Adelson [11] and Toet et al. [84] who used Laplacian (LP) and contrast pyramid (CP) techniques respectively in order to select the important features from the constituent images. While both algorithms are designed to produce output for human observation only, it is interesting to note that, in their advance on Burt and Adelson's work, Toet et al. exploit knowledge of how the human vision system works for their fusion algorithm. This is particularly relevant to ongoing research today, twenty two years later, in cognitive image fusion [82].

Algorithms for image fusion are divided into several categories with regard to abstraction; low, mid and high. The algorithms that have been demonstrated within these categories work at signal, pixel, feature or symbolic levels where signal is at a low level, pixel and feature are mid level and symbolic methods operate at a high level of abstraction. The majority of image fusion algorithms that have been researched and developed work at the pixel level in either the spatial or a transform domain [30, 45].

Image fusion algorithms operating in the spatial domain can be very simply implemented by combining the pixel values of the constituent images using a weighted average for each image [45]. Image fusion algorithms operating in a transform domain can be implemented by applying a multi-resolution transform e.g. wavelets [45, 47] or contrast pyramids [84] and combining the lower resolution images which are produced. The fused image is then obtained by performing the inverse transform.

The initial work with contrast pyramids described in [84] creates a ratio of low pass (ROLP) pyramid for the VIS and LWIR images. By convolving an image with a weighted Gaussian filter, each level of the ROLP pyramid contains a low-pass filtered, subsampled copy of the image from the tier below it. The ratios of the low pass images at each level are then computed before a composite ROLP pyramid is created by selecting the values from the constituent pyramids. The selection criteria can be chosen with respect to the post-processing desired for the final fused output.

The use of wavelet transform coefficients for image fusion has many advantages over ROLP pyramids. They are more compact, can provide directional information and reduce redundancy as each level of resolution does not contain blocking artifacts which are common in pyramid techniques [47].

The simplest and most commonly used wavelet transforms is the Discrete Wavelet Transform (DWT) [31, 47, 56]. This is a multi-resolution transform whereby an image is decomposed into detail and average coefficients. The detail coefficient images give vertical, horizontal and diagonal edge information. The average coefficient image represents a low-pass filtered version of the original image and gives texture information. It has been shown that, when two images from different spectral modalities are decomposed using the DWT (with the obvious assumptions that they are the same size and are spatially registered), a single fused image can be created from the two constituent images by selecting or weighting a combination of coefficients from either constituent image. The final fused image is generated by using the Inverse Discrete Wavelet Transform (IDWT).

The DWT method described in [47], while producing a fused image which resulted in improved perception by humans, does have shift variant behaviour. This is inherent in most transforms. A shift in the input signal during decomposition can cause the DWT coefficients at different levels of resolution to change [42, 45, 56]. This can produce "ringing" artifacts as well as problems in preserving edge and texture detail within the fused image.

The Dual-Tree Complex Wavelet Transform (DT-CWT) [42, 66] uses a "dual tree" of low-pass filter banks. The complex wavelet transform (CWT) has improved analytical properties in terms of reduced redundancy and shift invariance compared to the DWT as it uses complex valued scaling functions and complex valued wavelets. In a similar way to the classic Fourier transform (FT) which produces real and imaginary components of a signal (i.e. a pair of cosine and sine components that are 90° out of phase with each other), and therefore a truly analytical signal, in DT-CWT two "trees" of low-pass filter banks are designed to reproduce similar behaviour to the FT in the complex wavelet and complex scaling functions. The main advantages of this method are that it is shift invariant while also being directionally sensitive [42, 56], the

latter being particularly useful for detecting edge information within an image. The DT-CWT produces better qualitative and quantitative results than both pyramid and DWT methods when used for image fusion [56] due to the inherent shift invariance. It improves on prior attempts to produce shift invariant techniques (SIDWT) by reducing redundancy. However, it is worth noting that this improvement comes with a higher computational and system resource demand than standard DWT [45].

Not as much work has been done on techniques for image fusion in the feature or symbolic levels as has been done at the pixel level. The majority of feature level methods work by dividing the input images into regions and then applying a priority rule to select which regions from the constituent images are used in the final fused image. In [45] the detail coefficients of the DT-CWT are used to obtain texture information from the constituent image. The gradient function is then applied and combined with intensity information, thus providing information on edge locations where the resulting gradients are largest and allowing the segmentation of the image into regions of interest.

The advantages of such methods are [45, 61] a) the ability to use more intelligent fusion rules when selecting the regions to be fused, b) regions within the fused image can be highlighted by weighting them. The weight can depend on various properties of the region. c) a reduced sensitivity to signal noise d) the possibility of improved image registration by using the region features identified in the constituent images. There is also the suggestion in [45] that future sets of component images could have their fused output rapidly predicted by tracking the movements of the image regions. This is particularly relevant to real-time video surveillance in that it would drastically reduce the computational demand on the system.

The various methods of region-based fusion schemes are discussed in [45]. While more complex than transform domain schemes, the ability of feature-level fusion schemes to apply weightings to regions based on their activity, size or position relative to other regions allows the optimisation of the final fused image for its intended end use.

2.2 Image Fusion For Face Recognition

Face detection and recognition are increasingly important for security applications [94]. As camera systems and networks produce better quality images, decrease in price and generally become more widespread, the advantages of robust facial recognition systems for identification have become more and more apparent.

As mentioned above, one of the major disadvantages of sensors working in the visible range of the electromagnetic spectrum (~400-750nm) is that they are very sensitive to ambient changes in light. Face detection and identification suffer greatly when there is a decrease or shift in light source either due to natural variations (e.g. night time, shadows) or when the viewed scene is indoors and there is shuttering or obscuration of the light source. Further to this the subject's facial expression and pose relative to the camera can also affect the performance of face recognition. Sensors working in the long-wave range (~8-14μm) provide better performance under ambient light variations [74, 75, 89] as they detect emitted light coming from the subject, rather than light reflected off the subject. While LWIR images are in many ways superior to VIS images for face recognition [74] they can also be hindered by external, environmental variables. Indeed, images captured via LWIR sensors can prove problematic since any change in the temperature of the surroundings or the subject will affect the resulting image and therefore the match rate [27, 34]. Temperature changes in the subject are difficult to avoid since they can depend on any combination of metabolic processes within the subject, ambient temperature, alcohol consumption or physical exertion. In addition the presence of eyeglasses also affects the recognition performance, as glass is opaque to LWIR light. This causes eye-glasses to obscure feature information. While occlusions in the image are potentially less detrimental to recognition rates than variations in outdoor or uncontrolled conditions [34], they are obviously still a concern for face recognition applications as ~50% of the population wears some form of eye-glasses.

Due to the advantages and disadvantages of the two sensor modalities it was found that under more realistic experimental conditions i.e. with the subject moving or changing expression or out of alignment with the camera, neither the VIS or LWIR images were superior to the other with regard to face recognition [88]. In an attempt to overcome these limitations the fusion of multispectral face images was found to produce images that provide a higher performance for

face recognition than either single modality image on its own. As a result, there has been an increasing amount of research focused on multispectral or multisensor image fusion algorithms that produce an output image optimised for use with various face recognition algorithms [6, 12, 14, 31, 41, 44, 69, 71, 88].

In [31] a DWT fusion method is used to decompose the images into approximation coefficients and detail coefficients using Haar wavelets, before weighting the coefficients and combining them. Finally, the fused image is generated by applying the inverse DWT (IDWT) to the set of fused coefficient images. The coefficient weights applied during the fusion are determined by the quality of the respective images for face recognition. If, for example, there is a large reduction in light, the VIS image will lose a lot of the information required for face recognition. In this case the weights are adjusted such that the fused coefficients are weighted towards the LWIR than the VIS coefficients. The optimal values of these weights under different lighting conditions are fixed and set a priori through experimentation with a particular face image database.

In [31] the fixed-weight method applied to the LWIR and VIS images (0.3 and 0.7 respectively) was reported to produce a higher success rate for face recognition than simply averaging the DWT coefficients (i.e. weights of 0.5 and 0.5) although problems remained when using weighted DWT fusion in the presence of eye glasses. Because LWIR is blocked by glass, eye glasses appear as black ellipses in the LWIR image. This can severely decrease the success rate of any face recognition algorithm applied to the final fused output image. If the weighting in the final fused image is in favour of the LWIR image, the success rate may be decreased still further. In [44] a similar fusion method to [31] is used. However, an ellipse fitting algorithm is employed to locate and then replace any eye glasses found within the LWIR image with a generic eye template. The template is generated by averaging all the LWIR images with no eyeglasses that are available in the test database. This is shown to produce large improvements in the success rate in the first match (39% and 33% increases across two databases) and the first ten matches (16.8% and 6.7% across two databases). However, injecting "non-scene" information is not a desirable solution to the problem of eye glasses in LWIR images as in a real-world environment an ambient temperature change could cause a difference in the contrast between the replacement eye data and the rest of the face. It is also assumed that a large repository of LWIR images is

available for creating the averaged data. Further to this, while it is not explicitly discussed, the ellipse fitting, eyeglass removal and eye template superimposition, which requires rotation and resizing of the template may not lend themselves to a real-time application.

An alternative method to weighted averages of the DWT coefficients is image-based fusion using some selection criteria to pick the coefficients to be used from each constituent image. An example of this method is described in [71] where genetic algorithms (GA's) are used for making the selection of coefficients during image fusion. The chromosome in the genetic algorithm is encoded as a bit string, with each bit representing a wavelet coefficient in the fused image. The values of these bits are then used to select the image coefficient to be used in the final fused image i.e. 0 = LWIR, 1 =VIS. Once the GA has converged on a solution, the image fusion is completed by applying the optimised chromosome as a binary mask to the VIS and LWIR coefficients. The final fused image is then produced using the IDWT as before.

While genetic algorithms are an interesting approach that provides more variability than static weighted averages and also proves more robust to the presence of eye glasses and ambient light variations, they do present some problems. The inherent problems of GA's are that they a) don't always provide the best coefficient selections for fusion and b) have a variable rate of performance depending on the random value used to seed the algorithm.

In [71] a GA is used for fusing face images by first selecting the parent chromosome and applying it to the LWIR and VIS image coefficients. The inverse transform is then applied to the fused coefficient image and the final fused image is passed to a face recognition algorithm which provides a recognition score. The chromosome is then recombined and mutated as per the GA, which gives a new chromosome which, in turn, is used to produce another fused image and recognition score. The recognition scores give a measure of "fitness" which allows the GA to identify which chromosome produces fused images with best face recognition performance. The process is then iterated until a convergence criteria is met i.e. a minimum recognition score is achieved. This carries a large computational cost and renders the proposed application of genetic algorithms, in this specific manner, totally unsuitable for real-time image fusion. Further to this, as the chromosome is optimised to a sepcific set of training images there is a risk of overfitting

the fusion to a particular database.

In [18] a fusion method is proposed using the non-separable wavelet frame transform (NWFT) which is similar to the DT-CWT method discussed above and in [42]. The non-separable wavelet transform (NWT) method treats the image as an area as opposed to rows and columns in DWT. By altering the filter coefficients to the NWT (low-pass and high-pass filters), the NWFT is obtained. As explained in [42], this means the NWFT is shift invariant and has an improved directional property. However, after the NWFT is obtained, a coefficient selection method is still required in order to obtain a final fused image. In [18], a maximum absolute value rule is applied as the selection criteria. The results showed an above 90% recognition performance for experiments conducted with variations in illumination, eyeglasses and expression. However, the recognition performance was reduced to 84.85% when all three variables were changed simultaneously. Under the same conditions, a DWT method (using an absolute maximum selection rule) produced a 75.76% recognition performance. When we compare these results to those found in [31] in which weighting the coefficient selection of a DWT fusion method produces a recognition performance of 95.84% (improved from 90.31% using an absolute maximum selection rule), it suggests that fusion using the NWFT may not produce particularly large increases in performances compared to other transforms reported in the literature.

Feature-based image fusion uses fusion selections based on features identified within the image [69]. The image fusion algorithm described in [71] effectively treats the selection of these features as a pattern recognition problem in which GA's are employed to find the optimum set of coefficients for fusion. In [69] a post-image fusion process is used whereby a supervised learning method (specifically a Support Vector Machine or SVM) is used to select feature vectors from a fused image for face recognition. While it is noted that the initial fusion method using DWT gives a low enough complexity for it to be used in real-time, there is no similar analysis of the complexity of the SVM post-processing stage. This approach to image fusion, when applied to multispectral situations, appears to give the best results when using VIS and near-infrared (NIR) which is reflected light. The results from [69] show an error rate of 3.18% when the scheme is applied to VIS and LWIR. In comparison, the results in [31] obtained by fusing weighted DWT coefficients, without the computational demand of the secondary feature fusion via SVM, have

an error rate of 4.16%.

An alternative to fusing the face images in image or feature space is match-score space fusion whereby the match scores for the images in each spectral modality are calculated separately and the similarity scores from each modality then fused by a weighted average. This has been shown in applications to some mulitspectral image databases to produce better recognition results than image or feature fusion[12]. An example of this method can be seen in [25] where match score fusion is applied to a set of VIS, NIR and LWIR face images using a manually adapted set of score fusion weights to achieve a 95% recognition rate under varying light and expression, although it is worth noting that the face images used had no variation in pose and did not feature eyeglasses. The application of match score fusion to an already fused image set is investigated in [70] whereby the VIS and LWIR image pairs for each subject in a database are first fused in the wavelet transform domain. Two sets of features, namely the 2D log Polar Gabor and Local Binary Pattern (LBP) features, are then extracted from the fused image set and the recognition algorithm applied before fusing the resulting match scores. It was shown that this hybrid method of image and match score fusion gave a verification rate of 98.08% at a false acceptance rate (FAR) of 0.01% when applied to the Equinox database [70]. It is worth noting, however, that only verification results are reported so no conclusion can be drawn for recognition performance.

2.3 Adaptive Image Fusion

Research into image processing using the DWT has shown that the low-pass, average coefficients, while being the most suitable for face recognition applications [67, 68] are very sensitive to changes in illumination [55, 67]. In comparison, the highpass detail coefficients capture texture and geometric details and are more resilient to changes in illumination, but edge and corner information can still be lost under extreme lighting variations [1] and changes in expression [67]. These variations can be compensated by applying a set of static weights, however image fusion can unintentionally become over-optimised for the particular test database being used. That is to say, the resulting algorithm can fail if the database is changed. The risk of this overfitting to a particular database is clear and indeed, it has been found that fusion methods which have previously been reported to perform well under controlled conditions do not do well in uncontrolled or outdoor environments [38]. The aim, therefore, is to find a method of multispectral image fusion which not only allows accurate face recognition and verification but can do so under widely varying or previously unobserved conditions e.g. lighting,pose, expression. This is identified in the survey literature [38, 97] as a promising area for the future of face recognition .

For an image fusion routine to avoid this overfitting it must be, to some extent, adaptive to the type and quality of the component images. Image quality measures have been an active area of image processing research for the last two decades and are well documented in the literature [24, 26, 63, 87]. For the purpose of image fusion, however, the majority of the research has concerned the fusion of images used for surveillance (VIS and NIR images), medical (CT and MRI images) or multi-focus applications. These lend themselves well to image quality measures as there is an *a posteriori* comparison to be made between the final fused image and the component images after which the fusion algorithm can be adjusted in order to optimise the output. Obviously for face recognition this post-fusion comparison is not available. We do not know if the fusion of two or more face images has been successful in terms of improved recognition until we compute similarity scores across training and testing sets. In a real-world application the fusion system would not know if a face submitted for verification is valid or an impostor and any attempt to iteratively optimise the fusion based on the resulting verification score risks increasing the rates of false acceptance and false rejection.

For face recognition applications, therefore, it would be advantageous to detect any quality problems in the component images and either process them prior to fusion, or adjust the fusion routine to minimise their impact on recognition accuracy. In [68] it is found that if all of the images in a VIS modality face database are normalised by histogram equalisation (HE) prior to performing recognition tests, the recognition accuracy is reduced in many cases. Their research showed that adaptively applying HE only to images of low luminance quality gave improved recognition accuracy. To do this, probe images were measured for a global luminance distortion and HE applied if they were below a threshold.

In [67] a set of VIS face images are processed using the DWT and the sensitivity of different subbands to changes in lighting and expression are measured. Having identified that the approximation subband provides the most accurate recognition rate under uniform lighting, but the low-highpass (LH) subband is the least sensitive to variations in lighting, a "multi-stream" recognition method is proposed by the authors. The match score for each subband of the probe DWT image is calculated with those in the gallery database and the match scores are weighted and combined. The authors concluded that adaptive fusion weights were necessary to compensate for variations in the probe image lighting.

For non-adaptive image fusion in the transform domain such as DWT, the aim is to preserve the discriminatory features of the approximate (low frequency) and detail (high-frequency) subbands while simultaneously minimising the transference of any noise to the final fused image. For an adaptive image fusion method the aim is to be able to measure and identify a probe image with poor illumination and reduce the weighting of the VIS low frequency coefficients in the final fused image. Similarly for adaptive fusion in the feature and match-score spaces, it would be advantageous to detect any variations in lighting and down-weight the affected features or match scores in order to improve the accuracy of the face recognition.

Recently, the trend for 'deep learning' techniques have been applied to feature extraction for image fusion [49, 95]. The application of a Deep Neural Network (DNN) for this purpose is attractive as the features which minimise the intra-class variability and maximise the inter-class variability are selected automatically by the DNN once. However, a large number of samples

are required in order to produce a trained network which is capable of generalising. This would require thousands of samples per class i.e. face images per subject, and for our application of multispectral image fusion we are limited by the size of our database (see Section 3.3). There are also recently researched flaws within DNNs showing that very small changes in pixel values can change the output classification of a DNN [54, 79].

2.4 2.5D/3D and Multimodal Image Fusion for Face Recognition

Over the past decade the increasing availability and quality of low-cost depth sensors has ensured the rapid growth of research into the application of 3D information for face recognition [2, 7, 36, 46, 52, 53, 59, 81, 85]. The ability to capture a face image in 3D intuitively suggests a solution to the problem of pose variation and, if the depth sensor used works in the IR, it is also argued that 3D face recognition can be made insensitive to lighting variations [8, 50].

The two most common methods for estimating depth are structured light and time-of-flight (TOF). In structured light sensors a known NIR laser pattern is projected onto a scene which is viewed by an NIR camera operating at the same wavelength as the laser. Any object or set of objects in the scene reflects the points of the laser pattern back into the NIR camera. If the laser pattern as measured by the NIR camera is known when the pattern is projected onto a flat plane at a known distance, then any deformation of the pattern due to an object can be used to estimate depth. Due to the nature of this configuration the pairing of the NIR camera to the NIR laser projector is critical and specific to each structured light sensor. Calibration is carried out at the point of manufacture and built into hardware within the unit.

In contrast, a TOF sensor uses the known speed of light to calculate distances based on the time taken for an emitted photon to be projected into the scene and reflected back. In order to do this the TOF sensor has a NIR emitter which emits high frequency pulses of NIR light onto a scene. In order for the NIR camera to discriminate between the probe laser light and other light sources the probe laser is either modulated using a radio frequency (RF) or synced with a high-speed electronic shutter within the NIR camera.

A depth sensor will generally use a single IR wavelength laser source to probe and capture a scene; thus large changes in visible light that cause inaccuracies in VIS face recognition have only a small effect on the depth sensor only. However, to state that that 3D depth sensors are totally insensitive to illumination variations is not accurate.

Most artificial light sources emit radiation in the NIR waveband where most depth sensors operate. Also, terrestrial sunlight can be considered to be a broad-band from ~300nm to

>2000nm and thus natural light variations will also include the NIR waveband (750-1400nm). While depth sensors do attempt to prevent external noise from interfering, a depth sensor must detect and measure very small changes in a single wavelength of light with a high dynamic range. Thus areas of a subject's face with high reflectivity will invariably cause "blooms" in the detected NIR image, which can translate into inaccurate depth measurements. Indeed in [8] the "myth" of 3D image illumination invariance is discussed for this very reason.

A modern depth sensor will generally produce depth information in two modalities:

- a set of 3D points, or "point cloud", recorded in a three dimensional array
- a 2D intensity image where the gray scale value of each pixel is related to a depth value within a scene. Commonly referred to as a 2.5D, "range" or "depth" image.

There has been increasing interest in and, indeed, successful demonstrations of face recognition applied to either the 3D data [2, 46] or 2.5D data [36, 52] alone. However, early research reporting high recognition performances in 3D alone have proved to be too optimistic when the size of the database increased from tens to hundreds of subjects, or when more challenging poses or lighting were used during capture [8, 97]. With most modern depth sensors producing both a 2D colour and depth image, the application of 2D and 2.5D/3D fusion has also been investigated in the research literature. The most common approach has been to apply a recognition algorithm to each modality separately and then apply a match-score fusion [8, 16, 81, 85, 92]. This has been shown to outperform a single modality approach.

There has also been research conducted into face recognition using recovered depth images from multiple 2D cameras, as opposed to a dedicated depth sensor. In [92] a photometric stereo camera system consisting of four cameras is constructed and used to automatically capture face images of subjects as they pass in front. Depth images are reconstructed from the four images and the recognition scores for the 3D and 2D images are fused in match-score space. The results show that the match-score fusion of 2D images with depth images recovered from photometric stereo also produces an increased recognition accuracy compared to using a single modality.

While there has been an increasing interest in the fusion of 2.5D and 3D face data from one depth sensor, comparatively little work as been done on the fusion of 2.5D/3D data with

2D intensity images captured in VIS, and NIR/LWIR spectral modalities. This is noted in the recent survey literature where the authors conclude:

"...the work on 3-D + IR and visual+IR+3-D is comparatively rare. That is mainly because 1) visual images carry face texture information that is particularly useful for face recognition; 2) visual images are easy to acquire and process; 3) no devices currently exist that can capture faces with the three modalities synchronously." [97]

2.5 Image Fusion Technology

There has been a large discrepancy between the extent of research into novel image fusion techniques and viable methods for real-world deployment. This is noted in the image fusion technology review literature:

"...much of the literature to date on the topic of image fusion has not drawn attention to system-level issues or practical matters associated with making a real-time image fusion system a genuine reality, preferring to explore ever more complex or obscure methods of combining pixels."[73]

As sensor technologies have become more compact and cheaper, reliable real-time image fusion systems have recently become viable, although the complexity of the fusion algorithms employed is still limited by the computational resources available [70]. There exists a multitude of papers that, while mathematically interesting, propose fusion algorithms that are too complex to be employed in a real-time scenario, [70, 71]. To this end, when we refer to 'Image Fusion Technology', we are focusing on hardware, software and algorithm design developments that have a real-time system as the primary goal.The various process involved in a generic image fusion system are identified as image registration, image pre-processing, image fusion and image post-processing [73].

Unexpectedly the choice of fusion algorithm used is not the greatest factor affecting the quality of the system's output. The registration of the two images, such that they are spatially matched to one another prior to fusion, has the largest effect with regard to output quality [73]. Indeed, in [40] it was found during comparisons in recognition rates using VIS and LWIR images that the LWIR images were very sensitive to errors in eye-position registration. Further research has shown that it is difficult to reliably identify salient facial features such as eyes in LWIR images [27]. The advantages, therefore, of a multi-spectral camera system with a common optical path are considerable [23, 73] as the registration of the separate modalies is inherent at the point of image capture. This means that the computational demand on the fusion system's resources is massively reduced and a manual registration stage during image fusion is unnecessary.

The initial work in image fusion of LWIR and VIS images in [84] utilised a system that obtained images via two sensors over the same optical path. This was achieved by using a dichroic germanium mirror, or "beam-splitter" at 45° to the incident light. The mirror splits the incoming light into the constituent wavebands relevant to the sensors used. The LWIR sensor is placed in-line with the incident light, behind the germanium dichroic mirror. In this arrangement, the mirror acts as a long wave pass filter for the LWIR sensor. The VIS sensor is placed at 45° to the LWIR sensor, such that it is exposed to the reflected, visible-range light from the dichroic mirror. Registration between the two sensors was then carried out by placing an array of lights within the scene and adjusting the camera optics in order to "overlay" the lights in the component images.

Germanium, whilst having a relatively low transmittance (T%) of ~45% in the range of 2-12µm still provides enough light in the appropriate wavebands to produce an image in the LWIR sensor, however the signal is reduced and the resulting images are of rather poor quality. The LWIR sensor also uses a germanium lens which reduces the transmitted light by a further 45%, as can be observed in the Figures in [84]. Similarly, the reduction in signal to the VIS sensor due to the fact that there is a less than 100% reflectance (R%) from the dichroic mirror, results in a poor quality image.

The resolution of the image can be increased by the use of a single crystal mirror, as opposed to the multicrystaline mirror used in [84], as the advances in crystal growth technology over the past twenty years now allow. This would increase the level of detail observed in the LWIR image from that obtained in [84], however the overall reduction in detail due to this particular dichroic mirror method is still unacceptable, especially if the resulting images are to be used for feature extraction in face recognition in the post-processing stage.

Development on the dichroic or "beam-splitting" method has advanced from this initial stage of using only germanium, to a more highly engineered solution. With the development of coating technology, particularly dual magnetron reactive sputtering, it is now possible to design and produce "custom" optical components consisting of hundreds of layers of material coatings that will transmit and reflect light within desired wavebands at high transmittance (T%) and

reflectance (R%) to maintain image quality. It is possible to create a coated optical glass that will give a much better performance as a dichroic mirror than germanium on its own but at a considerable cost. The optical system produced by the Equinox Corporation is highly complex, consisting of multiple, custom-made superachromatic lenses which are responsible for focusing the light within each sub-spectrum (VIS and LWIR) onto the focal plane of the dichroic mirror. Even when this is achieved, the signal strength of the VIS component must be increased for low-level light conditions by using an intensifier.

The system does produce extremely high quality results (60fps with the images spatially registered to within 1/5 of a pixel). However, the hardware is extremely expensive in terms of research and development, as well as manufacture and is not commercially available. With few other attempts being made to produce a commercially viable system using an identical optical path for multiple sensors, the general solution to the problem of image registration has been the design and development of proprietary hardware [23, 72, 73]. Even now such devices are very limited in their availability and are mostly designed for military use e.g. target acquisition and identification, which means they are too expensive for general applications such as airport security, general access control or enrollment systems.

Only in the past decade have the continual advances in processor technology (Moore's law) and the resulting proliferation of multi-core processors made real-world image fusion systems using commercial off-the-shelf (COTS) hardware a possibility. While advances in real-time image fusion have been made [33], the computational cost of spatially registering image pairs in real-time is so high that the latest developments in fusion algorithm research are still not viable in a real-time system.

In [14, 15] a hyperspectral camera system is described which captures images in small sub-bands of the overall VIS spectrum (400-750nm) using a set of narrowband filters and a liquid crystal tunable filter (LCTF) which can be induced to transmit only the desired wavelength of light during capture. Experiments across a database of 82 subjects for various indoor and out-door lighting modes show that fusion of the hyperspectral images in match-score space produced an increase of 78% in recognition performance under outdoor lighting.

However, the single sensor design of the camera system requires the subject to remain still and compliant during capture as the filter system iterates through the wavelengths. For applications where the subject is not necessarily compliant, the system is at a large disadvantage.

A multispectral camera system is designed and built by Toet et al [83]. It applies beam-splitting technology to co-register multiple camera sensors. The system contains VIS, NIR and LWIR cameras and a series of filters which portion off the incoming light into its constituent wavebands appropriate to the response of each camera. The system is designed for defence applications where the visibility of a building in a combat area can be low if the illumination is insufficient or there are occulsions. No results for face recognition are reported. The system does, however, produce a fused video output of multiple spectral bands in real-time.

In [41] an array of cameras working in VIS, NIR, LWIR and a 3D TOF camera are used to capture face images under varying pose and illumination. It is noted that the different spectral images are not fused. The 3D face data is used for pose normalisation of registered 2D face images prior to recognition. The constituent cameras do not share a common optical path. Instead the captured images are registered by using the 3D TOF camera to estimate the translation and rotation of the TOF 2D image plane with respect to the world coordinate system. The separate spectral images (VIS, LWIR, NIR) are texture mapped to the 3D point data before being normalised using Iterative Closest Point (ICP) to correct for any variation in pose. The recognition experiments were then conducted using these 2.5D images. The combination of the normalised 3D face images with different 2D spectral images was found to outperform the use of VIS or 3D images alone.

In the few commercially available image fusion systems on the market, over half the available processing power, even for the optimised systems described in [33] is used in image registration prior to fusion. This remains an area of on-going development to produce reliable results in real-time environments. The advantages of a common optical path system are a greatly reduced demand on system resources due to the removal of an image registration step and, therefore, an increased amount of system resources available for advanced fusion algorithms and image processing. An automatically registered multispectral camera system also reduces errors induced

during image registration between sensors. In particular, facial location features can be detected in the VIS image and mapped directly to the other modality images. It is not surprising, therefore, that the most recent review literature notes an increasing interest in multispectral + 3D/2.5D camera systems [97] and predicts that such camera systems are the future of face recognition [3].

2.6 Summary of the Literature Review

From our review of the literature we have seen that face recognition can be considerably improved by using multispectral and multimodal images. As noted in [97] the comparative lack of reserach on multispectral and 2.5D/3D fusion for face recognition is primarily due to the lack of camera systems capable of capturing co-registered images across these modalities. We have therefore proposed a design for a novel mutlispectral with 2.5D/3D camera system (Section 3.1) which is capable of capturing spatially registered VIS, NIR, LWIR video and depth data through a common optical path.

Much of the multispectral image fusion research reports results using publically available databases, with the main focus on the development of fusion algorithms that can achieve higher recognition rates than previous attempts. We suggest that this risks the pursuit of an algorithm finely tuned to a specific database, rather than a more general one that is robust to changes in lighing or pose. From our literature review we would suggest that research in this area should have a more holistic approach with respect to the performance of both the fusion algorithm and camera system and that performance of the whole system under previously "unseen" lighting conditions should be more of a concern. This is also covered in our literature review where we found the problem of adaptive automatic fusion of multispectral and multimodal images for face recognition has received little attention.

With consideration to these areas of the literature, this thesis describes a contribution to the field in the development of a novel mutlispectral and 2.5D/3D camera system (Section 3.1 and Section 3.2) as well as the databases of face images we have generated using the camera system (Section 3.3). Our experimental method for investigating the fusion of these images across different levels of image fusion (Sections 4) is designed to demonstrate the over-fitting of

certain algorithms to specific image sets as well as develop an adaptive fusion model that results in accurate face recognition under new "unseen" conditions in lighting.

3 Analysis, Design and Implementation

In this section we will report on the design and development of a novel multispectral camera system and the generation of the face image databases used in our experiments. The methods of image fusion in data, feature and match-score space are also described.

3.1 Camera System

The camera system described here was first designed in 2010 as a VIS + LWIR multispectral system using an oscillating gold mirror to produce identical optical paths with near-simultaneous image capture between cameras as shown in Figure 1. A gold mirror was used because commercially available dichroic mirrors at that time could only separate out the infrared to 1200nm which is not sufficient for the thermal range (8-14μm). Gold is commonly used in IR optics because of its consistently high reflectance from the NIR into the LWIR (see Figure 2) and it retains a relatively high reflectivity (~80-90%) within part of the VIS range (400-700nm) , making it well suited as a mirror for capturing both VIS and LWIR images. The gold mirror has a protective layer of silicon dioxide which prevents tarnishing of the gold.

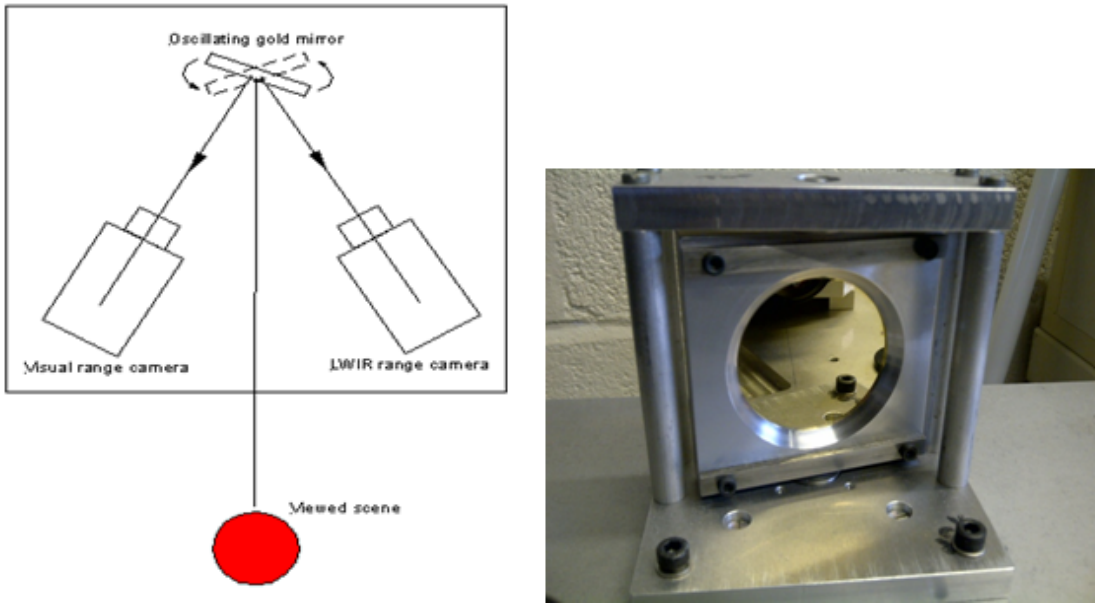

Figure 1: LEFT: Schematic of the original oscillating gold mirror camera system, RIGHT: Photograph of the oscillating gold mirror and mount.

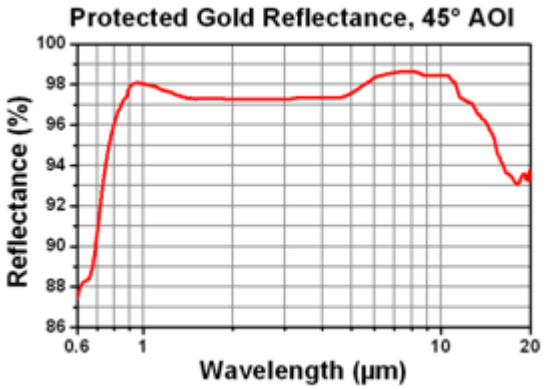

Figure 2: Reflectance spectrum of protected gold from 600nm to 20µm at 45° angle of incidence (AOI)

Recognition experiments using face images captured via this system and fused in the transform domain showed the effectiveness of the camera system and the advantage of the spatial registration of the images. However, the system was limited by the relatively low frequency oscillation of the gold mirror. This problem could not be overcome.

In early 2012, ISP-Optics announced a dichroic beam-splitter with a >70% average transmittance across the 400-700nm visible range and an >95% average reflectance across the LWIR 8-12µm range as shown in Figure 3. Replacing the gold mirror with this dichroic beam splitter not only improved the signal level for the VIS camera, but also increased the rate of image acquisition to allow full video rate image capture of the spatially registered images.

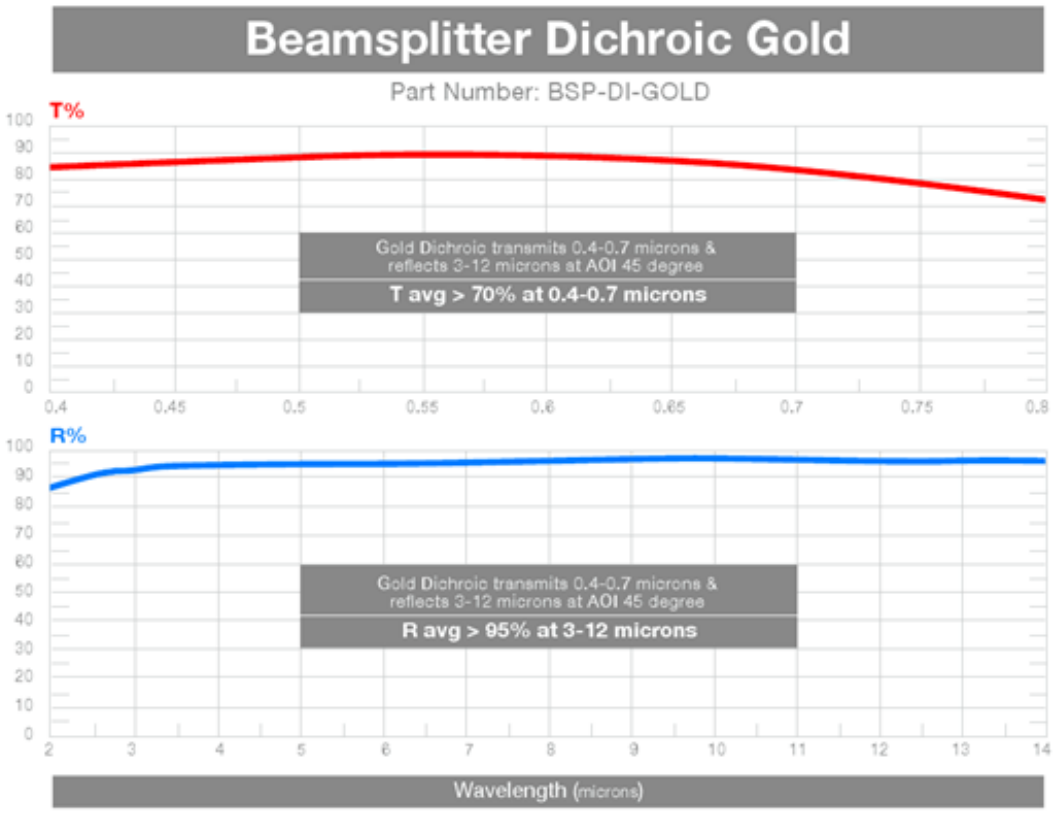

Figure 3: The VIS range transmittance (T%) and LWIR range reflectance (R%) of the ISP Optics dichroic mirror.

Finally, in July 2013, the camera system was expanded to include a Microsoft 'Kinect' depth sensor device as well as an additional dichroic mirror which separated the VIS and NIR portions of the spectrum. The 'Kinect' depth sensor was selected for several reasons: a wide range of driver and software development tools exist for the platform, it is comparatively cheap and the structured light format lends itself well to our system since the projector can be fired in parallel with the camera system. Thus a novel multi-spectral + 3D/Depth camera system that is capable of capturing spatially registered VIS, LWIR and Depth video and images has been designed and constructed. The system, as shown in Figure 4, comprises the following sensors:

- An EYE R25 thermal sensor using a 12mm (F/0.85) manual focus lens for the LWIR component.(3000-12000nm)
- A MTV-63WW100P CCD camera using a 4.5-10mm (1/2") varifocal auto iris lens VIS

camera (400-700nm)

- A Microsoft Kinect^{TM} camera and encoded light laser projector unit which operates in the NIR region of the spectrum (800-900nm).

The EYE R25 presented the best compromise between image quality and cost. For the VIS cameras an acceptable level of image quality could be achieved for a low cost. The LWIR camera has a limited availability of sensor sizes and germanium lenses. For our relatively close-range application we selected a 12mm lens on the EYE rank 25 to maximise speed and image quality. The VIS camera has a 1/3 crop sensor which when used with a 4.5-10mm zoom lens produces an image with the same focal length and field of view as the LWIR camera. However, the Kinect^{TM} camera has a fixed focal length and cannot therefore be adjusted to coincide with the LWIR and VIS camera images. The one-off registration process carried out to adjust for this is described in Section 3.2.

Due to the limitations of the Kinect hardware, the NIR laser projector cannot be turned off without also turning off the NIR camera, thus separate NIR images and video cannot be captured simultaneously with the other cameras because the NIR laser pattern is projected onto the subject's face. However, it is possible to obtain VIS, LWIR and NIR images and video by physically shuttering the NIR laser projector and illuminating the scene with a 850nm LED array (shown in Figure 5). As the NIR sensor on the Kinect does not have a broad-band response we require this additional illumination in the 850nm wavelength in order to obtain an image, however, this has obvious implications for the fairness of any comparison between lighting modes since all the NIR images are effectively fully illuminated from the front. As such the NIR images are considered separately in our image fusion experiments discussed in Section 5 on page 70. Our system handles the shuttering of the NIR laser projector and NIR LED illumination automatically when capturing an image set, with a slight temporal offset of ~0.5 seconds between the NIR image and the corresponding VIS and LWIR images.

The video streams from the LWIR and VIS cameras are captured using an IDS Falcon Quattro framegrabber running on a Microsoft Windows^{TM} PC with an Intel i3 530 processor and 4GB RAM. The software to capture images from the camera streams is written in C++ and uses the OpenCV library while the Microsoft Kinect ^{TM} is interfaced using the OpenNI library

and the Primesense^{TM} driver. Software for the control of the broadband lighting, LED lighting, NIR laser shutter and data storage is written in National Instruments Labview^{TM} and allows for quick, automatic control of the timed lighting, the NIR shutter and the capture and storage of every image stream. The user is required to enter a subject ID number, select the pose the subject is in and then select one of four lighting modes (front, left, right or low). The software then turns on the appropriate lights and the images are automatically captured, labeled and stored for that subject ID.

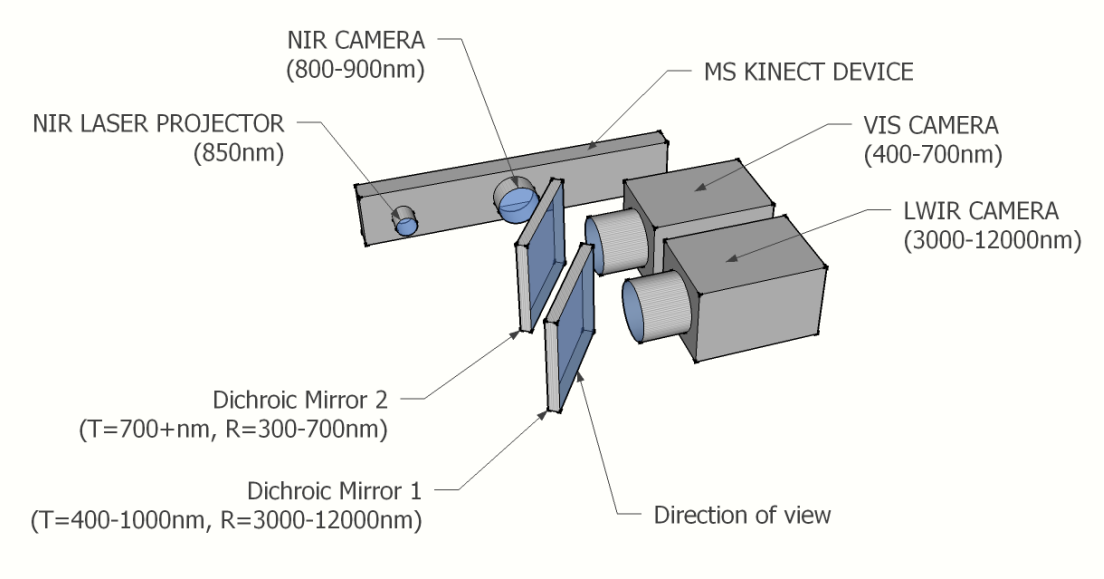

Figure 4: A schematic showing the configuration of the multispectral and 3D/Depth camera system components. Two dichroic mirrors are positioned at 45 degrees to the direction of view and reflect (R) into the VIS and LWIR cameras the wavebands of light to which they respond whilst transmitting (T) the NIR wavelengths through to the NIR camera. Each dichroic mirror transmits the remaining wavelengths through to the next dichroic mirror or camera.

The cameras are independently fixed to a machined aluminum base plate with precision engineered mounts which allow for fine adjustment of the pitch and yaw of the VIS and LWIR cameras. The dichroic mirrors are similarly mounted in precision engineered and fixed mounts such that the filters are held vertical and at 45^{o} to their respective cameras. The base plate itself is mounted on a frame to achieve a general head height from the bench top. Fine vertical adjustment of the entire frame and camera system is achieved via an optics-quality lab jack which can raise and lower the assembly. Wide views of the camera system and lighting assembly are shown in Figure 5 and Figure 7.

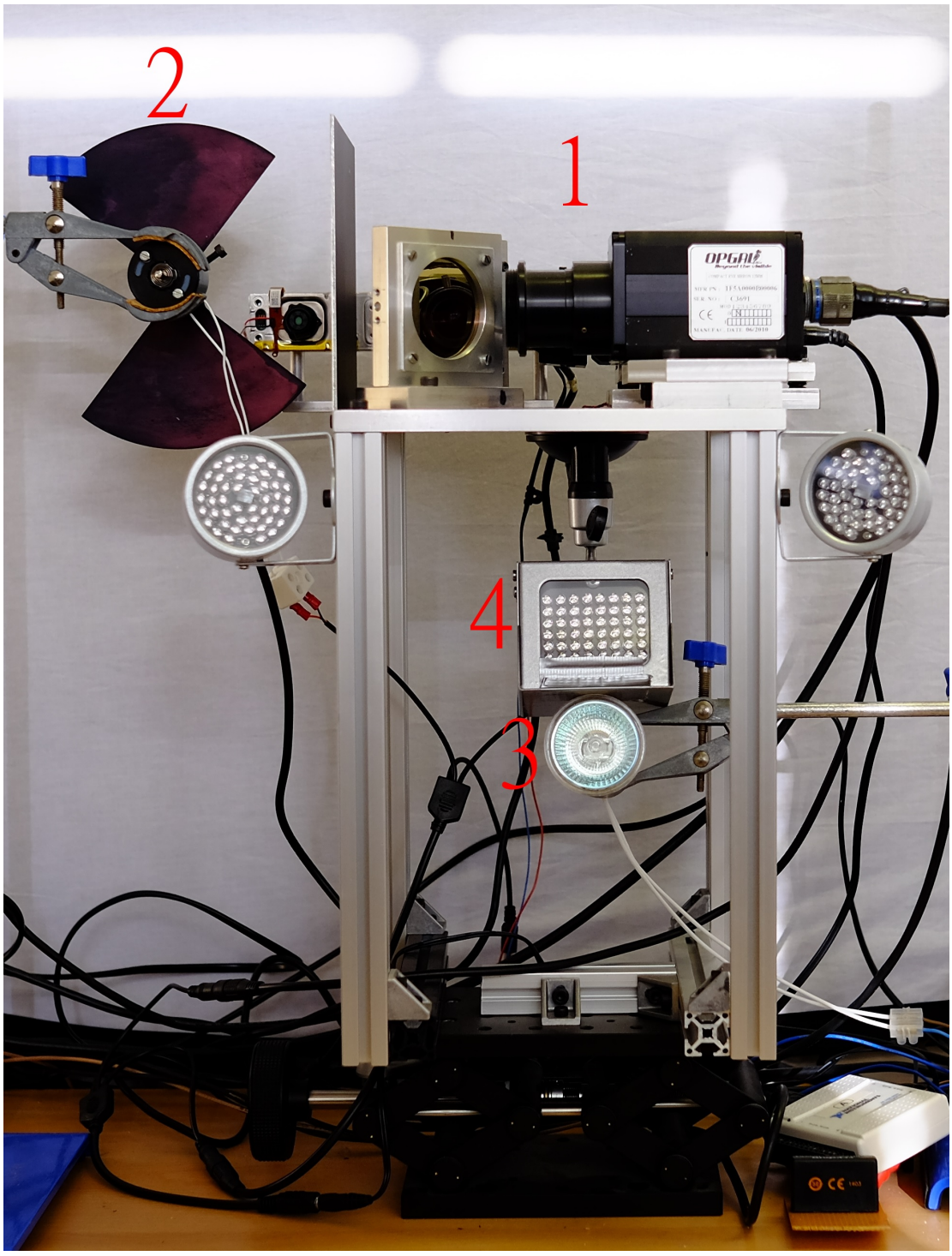

Figure 5: A subject's view of the camera system and setup: 1) The camera system 2) 'Kinect' projector shutter 3) Front lighting mode lamp 4) NIR illumination LED array

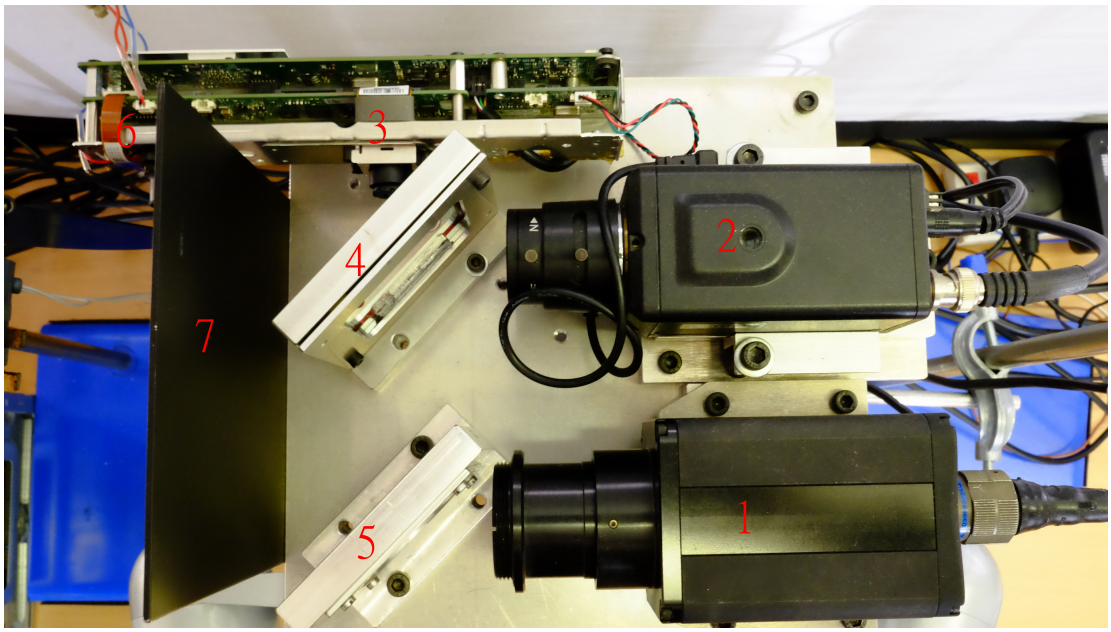

Figure 6: A plan view of the camera system: 1) LWIR camera 2) VIS camera 3) 'Kinect' NIR depth camera 4) Second dichroic 5) First dichroic 6) 'Kinect' NIR laser projector 7) NIR laser baffle

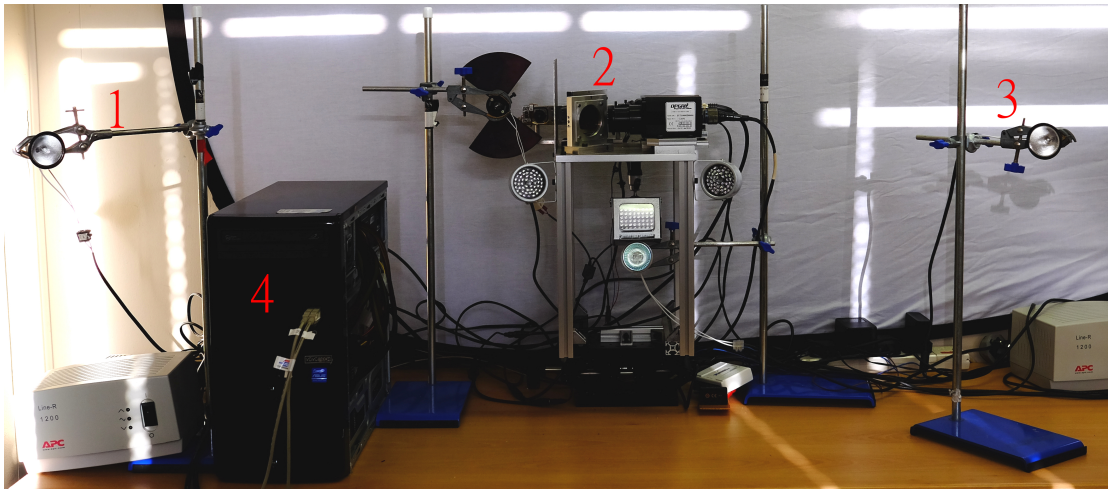

Figure 7: A wide view of the camera system and setup: 1) Left side illumination lamp 2) Camera system 3) Right side illumination lamp 4) Control and acquisition computer

3.2 Camera registration

As discussed in Section 4 the variable focal length of the VIS camera lens allows it to be adjusted such that the VIS and LWIR images coincide. However, the fixed focal length of the Kinect^{TM} prevents such an adjustment being made for the depth and NIR images. In order to calculate the affine transformation required to scale and register the VIS/LWIR images with the NIR/Depth images a registration tool was constructed. This consisted of a set of four lights which have an output covering the response of all the camera sensors, mounted within a sheet of Tufnol^{R} to thermally insulate the lights from their mounting. The registration tool was then imaged using the camera system.

The images of the registration tool were then loaded into the Matlab^{TM} control point selection tool (available in the image processing toolbox). An example of this process is shown in Figure 8. The centers of the light points in the VIS and corresponding NIR image are manually marked and the affine transform required to register the two sets of points is then calculated. This registration process and affine transform calculation is only required once. The transform is then automatically applied to the NIR/Depth images at point of capture and storage.

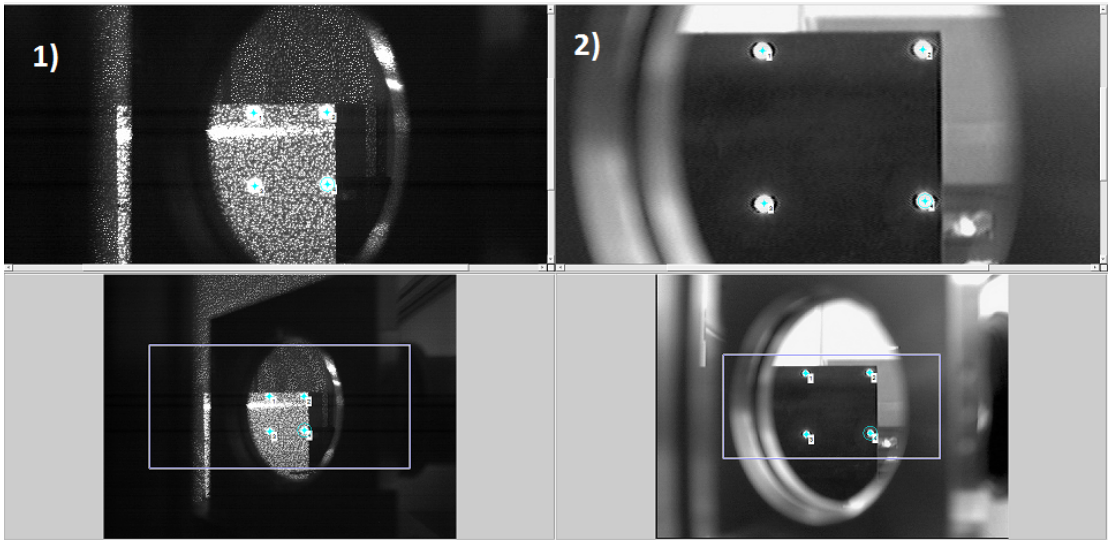

Figure 8: The registration tool as viewed by the NIR/Depth (1) and VIS (2) cameras. The registration points selected via the Matlab^{TM} registration function can be seen.

3.3 Face Databases

Using the novel, multispectral and depth camera systems described in 3.1 on page 31, a database of face images has been generated.

The database consists of 42 subjects imaged during two separate sessions using the above camera system under varying pose and illumination conditions. The subject and camera system were situated in a blacked out room to remove any ambient light effects during image capture. Illumination is applied via two 12W halogen lights placed behind and to either side of the camera system and directed at the subject. Each light has a computer controlled power supply and the lighting variations and image capture are computer controlled. All images are cropped to 128x128 pixels and manually normalised for eye position and head rotation such that the line of the eyes is parallel to the x-axis of the image plane. Each session was conducted over a month long period with approximately one year between the sessions. There was no control of the ambient room temperature during either capture session.

The images from session one are obtained as follows:

- 30 subjects
- Lighting modes consist of front, side and low lighting. The side lights (shown in Figure 7) are positioned at 30^{o} to the subject's face. The pose and lighting variations are intended to simulate an environment with uncontrolled, non-uniform illumination in which the subject changes position. This is referred to as 'Lighting mode 1' or LM1
- Each subject is imaged under front, left side and right side poses, with each lighting mode for each pose
- Each subject is imaged wearing eyeglasses and in the front pose, for each lighting mode. Where the subject did not have their own eyeglasses a pair was provided
- The capture process under each lighting and pose variation is repeated for each subject
- All images are cropped to 128x128 pixels and normalised for eye position and head rotation with the line of the eyes parallel to the x-axis of the image plane

Therefore for session one each subject has 24 images for each spectral modality and depth image modes. There are three spectral modality images (VIS, NIR, LWIR) and a 2.5D/depth image mode at each capture, thus for each subject there are $4 \times 24 = 96$ images and a total of $96 \times 30 = 2880$ images for session one. An sample showing the lighting variations for LM1 is shown in Figure 9.

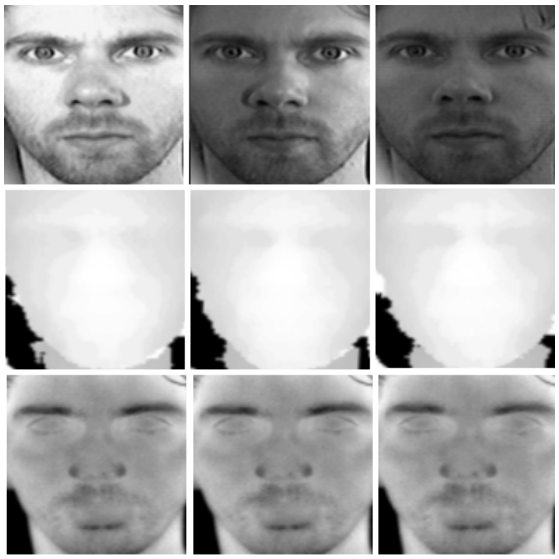

Figure 9: An example of the LM1 lighting variations for the front pose VIS (first row) depth (second row) LWIR (third row) images. From left to right: Front-lit, side-lit and low-light.

The images from session two are obtained as follows:

- 30 subjects, 18 of which were imaged in session one approximately one year previously
- Lighting modes consist of front, side and low lighting. The side lights (shown in Figure 7) are positioned at 70° to the subject's face and the power of the lamps was increased. The pose and extreme lighting variations are intended to simulate an environment with "harsh" highly directional, non-uniform, uncontrolled lighting in which the subject changes position. This is referred to as 'Lighting mode 2'
- Each subject is imaged under front, left side and right side poses, with each lighting mode for each pose

- Each subject is imaged wearing eyeglasses and in the front pose, for each lighting mode. Where the subject did not have their own eyeglasses a pair was provided.
- The capture process under each lighting and pose variation is repeated for each subject
- All images are cropped to 128x128 pixels and normalised for eye position and head rotation with the line of the eyes parallel to the x-axis of the image plane

Therefore for session two each subject has 24 images for each spectral modality and depth image modes. There are three spectral modality images (VIS, NIR, LWIR) and a 2.5D/depth image mode at each capture, thus for each subject there are $4 \times 24 = 96$ images and a total of $96 \times 30 = 2880$ images for session two. An sample showing the lighting variations for LM2 is shown in Figure 10.

Figure 10: An example of the LM2 lighting variations for the left side pose VIS (first row) depth (second row) LWIR (third row) images. From left to right: Front-lit, side-lit and low-light

With a total of 42 subjects our databse can be considered small scale in comparison with other databases used in similar research. The size is primarily due to the practicalities of obtaining subjects whilst completing the research in a reasonable timescale, as well as the self-funded nature of the research. As noted in [41] there is no database of registered multispectral VIS, NIR, LWIR and 2.5D/3D face images so making a direct comparison with our database

is not possible. The authors of [41] generated their own database of 100 subjects using their camera system although it is worth noting that the research was funded by grants from the National Research Foundation of Korea and the Ministry of Knowledge Economy (Korea) and the database is not available to the public.

There are, however, multispectral face databases available which are commonly used in multispectral face recognition research. For comparison to our own database in terms of subject numbers and image type, these are:

- Equinox Face Database [89] contains 300 subjects VIS, SWIR, MWIR and LWIR. It was funded under a US DARPA program but is no longer available to the public.
- Near Infrared Visible Light Database (ND-NIVL) [5] contains 574 subjects imaged in VIS and NIR in front pose with normal indoor lighting.
- ASU Database [96] contains 96 subjects imaged in VIS and LWIR modalities under varying indoor and outdoor lighting and with eyeglasses where the subject required them.
- IRIS Database [51] contains images of 30 subjects in VIS and LWIR taken under varying expression pose and illumination.
- IRIS-M3 Database [13] contains images of 80 subjects in VIS and LWIR along with hyperspectral divisions within the VIS band. The subjects are imaged under indoor and outdoor lighting conditions.
- WVUM Database contains images of 50 subjects in VIS and SWIR. The VIS images are take in front profile with the SWIR images under varying pose. The SWIR images are taken in 10nm sub-bands of the SWIR band.

3.4 Eyeglasses Detection and Compensation

A LWIR image of a subject wearing eyeglasses is unsuitable for face recognition. The opacity of the glass to the LWIR light results in large black areas which occlude information that is important for face recognition.

There have been attempts to develop a suitable method of compensation for the presence of eyeglasses in LWIR images prior to fusion with other spectral images. For example, in [44] an "average eye" is constructed by averaging all of the LWIR eye images in a database. The average eye is then rotated and transformed into position over the eyeglasses within an LWIR image. Alternative methods for eyeglasses compensation have also been proposed in which the LWIR features are modeled to allow the reconstruction of the LWIR image that would have been obtained without eyeglasses [93]. However the process is computationally demanding, which is undesirable for a real-time system. Also, the presented results, based on the correlation between the reconstruction and the average image for a subject, are not extensive or particularly convincing.

A more desirable solution is to use the VIS image information to estimate the occluded LWIR eye image. There has been some work to identify a transform between image modalities, specifically the VIS and IR spectral modes. In [17] a patch-based transformation of VIS and NIR image pairs is used to synthesise a VIS face image from a given NIR face image with excellent results. Whilst an explicit global manifold for the transform is not learned, a local linear mapping between VIS and NIR face images is shown to exist. The mapping between VIS and NIR modalities using Local Linear Embedding (LLE) is also shown to preserve the local geometry between two VIS and NIR images. This is in contrast to [22] where LWIR images are converted to VIS images for face recognition using Sophisticated Local Linear Embedding (SLLE) which does not assume that the local geometry is preserved between two VIS and LWIR images. This is discussed further in Section 3.4.2.

A method of synthesising the occluded LWIR eye image from the VIS image using a local linear mapping similar to those described in [17, 22, 62] is employed in this paper. A description of this method is given in Section 3.4.1.

3.4.1 Mapping of VIS to LWIR Eye Images

Using the N VIS and LWIR pairs of normalised front pose images without eyeglasses from the database described in Section 3.3, a set of VIS and LWIR eye image pairs (φ_i^1,φ_i^2) ($i = 1, 2...N$) of 64×64 pixels are extracted, where φ_i^1 is a VIS eye image and φ_i^2 is the LWIR image of the same eye. Each eye image is divided into M overlapping patches P that are16×16 pixels in size and have an overlapping region of 12×12 pixels. A set of patches for an eye image can therefore be given as $P_{j,k}$ where ($j = 0, 1...m-1$) is the number of patches sampled in the row direction and ($k = 0, 1...n-1$) is the number of patches sampled in the column direction giving a total of $M = m \times n = 13 \times 13 = 169$ patches per eye image.

For each patch in a set $P_{j,k}^1$ extracted from a VIS eye image φ_i^1we calculate the LBP histograms at 7 different resolutions using the Multiresolution Local Binary Pattern analysis (MLBP) as described in [17]. The MLBP is an efficient, rotation invariant method of texture analysis using local binary patterns [57]. In MLBP the standard LBP operator which calculates the value of a central pixel by thresholding, weighting and summing pixels within a surrounding neighbourhood is adapted to work at different resolutions. The central pixel neighbourhood for MLBP is defined by a variable radius R and sample point number P. The different LBP resolutions are obtained by varying the pixel radius R and the number of sample points P to obtain LBP histograms at various resolutions. An example how these point and radius values are configured is shown in Figure 11.

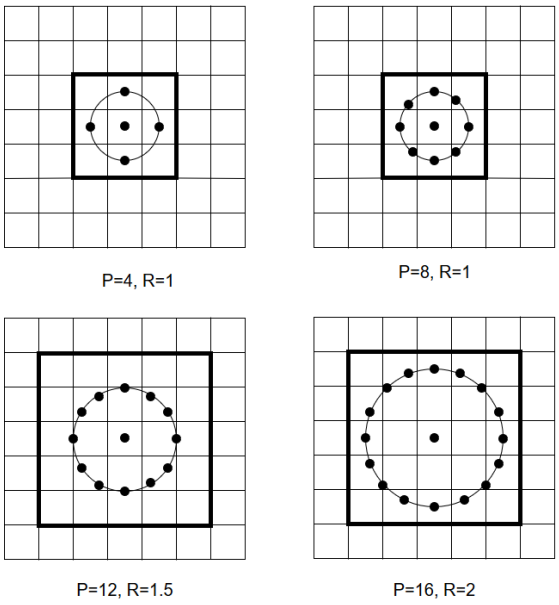

Figure 11: Examples of varying point number and radius value configurations used in MLBP.

The LBP histogram computed at each resolution is normalised and stored in a cell array such that we can reference for a given patch $P^1_{j,k}$ from eye image φ^1_i the LBP histogram $\{H_{j,k}(LBP_{P,R})\}$ where the sample point number and radius values used are (P,R)=(4,1),(8,1),(12,1.5),(8,2),(16,2),(1
By repeating this for all eye images extracted from a database we build a cell array which can be used as a dictionary D^1 consisting of an entry for each VIS eye image and the LBP histograms extracted at each resolution for each patch therein. Each row of D^1 indexes the (j,k) patch positions and each column indexes the VIS eye image extracted from the database. In parallel to this we have the dictionary D^2 consisting of the corresponding LWIR eye images φ^2_i and the LWIR eye image patches $P^2_{j,k}$.

For a new eye image pair $(\varphi^3_i, \varphi^4_i)$ where we wish to synthesise the LWIR image φ^4_i, we first compute the LBP histograms $\{H_{j,k}(LBP_{P,R})\}$ for each patch $P^3_{j,k}$ of the VIS eye image φ^3_i in the same manner as above. For each patch of φ^3_i we compute the similarity at each LBP resolution for each column in the corresponding row of dictionary D^1 . The LBP similarity is calculated using the histogram intersection such that for for two histograms H_1 and H_2 at the same resolution we have:

$$\Psi(H_1, H_2) = \sum_{t=1}^{L} min(H_{1,t}, H_{2,t}) \quad (1)$$

where L is the total number of bins in the histogram (in this case 256) and H_1 and H_2 are the histograms from the input image patch $P^3_{j,k}$ and dictionary reference patch $P^1_{j,k}$ respectively. Thus for a given LBP resolution we compare the input and dictionary LBP histograms at that resolution, select the minimum LBP histogram value at each histogram bin and sum these values together to give a similarity measure. This is repeated at every LBP resolution and the similarities combined using the product rule as described in [17, 43] to give a single similarity score S for a pair of patches in set $P^1_{j,k}$ and $P^3_{j,k}$ such that:

$$S = \prod_{i=1}^{T} s_i$$

Where s_i is the similarity score at the ith MLBP resolution and T is the total number of MLBP resolutions used. In this case $T = 7$ as we use 7 LBP resolutions.

To synthesise the LWIR patch $P^4_{j,k}$, this similarity score is used to select the K-nearest neighbours (KNN)($K = 15$) of similar VIS patches. The corresponding LWIR patches of the K-nearest neighbour VIS patches from dictionary D^2 are then combined, using their normalised similarity scores as weights, to produce the synthesised LWIR image patch. Once all 169 patches have been generated they are copied into a blank image at their corresponding (x, y) coordinates. Where the combined patches overlap, the average of pixels from each contributory patch at each pixel in the 12×12 region is used.

It is noted that under the assumption of a local geometry preservation between the VIS and LWIR manifolds as described in Section 3.4.2 below, only patches from the same image position as the input patch are considered when synthesising the LWIR patch. That is to say the similarities for a given input patch are only calculated along the corresponding row of D^1. Conversely, when no geometry preservation is assumed, all patches from all image positions are considered for synthesis. That is, the similarities for a given input patch are calculated along all rows of D^1.

3.4.2 Local Geometry Preservation

In [17] it is proposed that the local geometry is preserved when mapping between VIS and NIR images. That is, if VIS patches from the same image position of two subjects are K-nearest neighbours, their corresponding NIR patches should also be K-nearest neighbours to one another. Statistical evidence to support the proposed local geometry preservation is obtained by computing the ratio of the number of these spatially matched patch-pairs to the total number of matched pairs. In [17] the ratio is 92.5% for their NIR-VIS database.

When we consider the differences between VIS, NIR and LWIR imaging modes, i.e. reflected light (VIS, NIR) vs emitted (LWIR), the assumption of local geometry preservation between the VIS and LWIR manifolds does not seem as obvious. Any changes in lighting will affect the VIS and NIR images but not the LWIR image.

In [22] the mapping of LWIR images to VIS images is assumed to preserve no local geometry between the LWIR and VIS manifolds. During the synthesis of a VIS image from an input LWIR image, all image patches in the dictionary are measured for similarity, not just the ones from the same image location, as described in Section 3.4.1. In [22] the authors' application of this Local Linear Embedding (LLE) method without assuming local geometry preservation is referred to as 'Sophisticated LLE'. The ratio of the number of spatially matched patch-pairs to the total number of matched pairs is calculated as in [17]. Their experiments show that in the case of mapping from LWIR->VIS images, this ratio is <10%. Indeed, experiments on our own image set using the algorithm described in Section 3.4.1 have shown that for a given VIS patch <2% of the KNN patches matched assuming local geometry preservation are also KNN matches when assuming no local geometry preservation. Therefore, in the experiments described in this paper we have worked under the assumption that no local geometry preservation is assumed and all patches in all rows of D^1 are considered in the similarity measure.

3.4.3 Eyeglass Detection and Segmentation

Because of the prominent nature of eyeglasses in thermal images, it is possible to accurately segment the eyeglass lenses from the rest of the image by thresholding to produce a binary image with very reproducible results across different types of eyeglasses. By optimising the

threshold it is possible to reduce the face image to just a few simple contours which include the lenses of the eyeglasses. We have heuristically set a threshold grayscale value of 40. Once this thresholding has been achieved we can make several a priori assumptions that will allow us to automatically detect eyeglasses. Firstly we assume that the two lenses are the same size and shape and that therefore any matching pair of contours of appropriate size within the face image can be assumed to be eyeglasses. We reduce the search area over which we look for matching contours by using only 80% of the height of the face area under the assumption that the eye areas will always be contained within this upper section of the face.

We then take each contour in this area and measure the similarity between it and every other contour using the Hu moments of the two contours as given in[10]. Thus to compare two contours A and B we have:

$$m_i^A = \text{sign}(h_i^A) \cdot \log |h_i^A|$$

$$m_i^B = \text{sign}(h_i^B) \cdot \log |h_i^B|$$

$$I_1(A,B) = \sum_{i=1}^{7} |\frac{1}{m_i^A} - \frac{1}{m_i^B}|$$

and $h_i^A$and $h_i^B$are the Hu moments of A and B respectively. The Hu moments are given in [10, 35] and are expressed as:

$$hu[0] = \eta_{20} + \eta_{02}$$

$$hu[1] = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

$$hu[2] = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

$$hu[3] = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$

$$hu[4] = (\eta_{30}-3\eta_{12})(\eta_{30}+\eta_{12})\,[\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2] + (3\eta_{21}-\eta_{03})\,(\eta_{21}+\eta_{03})\,[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]$$

$$hu[5] = (\eta_{20} - \eta_{02})\,[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$hu[6] = (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\,[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 - (\eta_{21} + \eta_{03})^2]$$

$$-(\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})\,[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

and the normalised central moments η_{ij} are given by:

$$\eta_{ij} = \frac{mu_{ij}}{m_{00}^{(1+\frac{i+j}{2})}}$$

where:

$$i + j \geq 2$$

from the central moments mu_{ij} given by:

$$mu_{ij} = \sum_{x,y}((x, y)\,.\,(x - \bar{x})^j\,.\,(y - \bar{y})^j)$$

where $(\bar{x}, \bar{y})$ is the mass center:

$$\bar{x} = \frac{m_{10}}{m_{00}},\ \bar{y} = \frac{m_{01}}{m_{00}}$$

which is given by the spatial moments:

$$m_{ij} = \sum_{x,y}((x, y)\,.\,x^i\,.\,y^j)$$

If the contour comparison as measured by $I_1(A, B)$ is near to 0 then the two contours are likely to match. We found that a threshold of <0.2 gave a high rate of correct matches with an acceptably low number of spurious matches. To ensure that any matching contours are eyeglass lenses, we find the approximate total area enclosed by both contours by fitting a bounding box to each of them and calculating the combined area of both boxes. If the combined area of the matched contour pair covers $<30\%$ and $>10\%$ of the upper face area then it is assumed that eyeglass lenses are detected. The two contours of the detected lenses are then drawn in a new

Figure 13: Samples of synthesised eyeglasses taken from our database.

blank image to produce a binary mask of the eyeglass lenses.

When eyeglasses are detected in the input LWIR image as in Figure 12(a), the eye area images are synthesised as described in Section 3.4.1. A copy of the input LWIR image is taken and the synthesised eye patches are placed on top at their respective (x, y) positions as shown in Figure 12 (b). By applying the binary mask to this image we can extract the synthesised LWIR images encapsulated by the eyeglasses as shown in 12(c). The algorithm accurately and consistently segments the eyeglasses in LWIR and will work for any size or shape of eyeglasses. Samples of synthesised eyeglasses are shown in Figure 13.

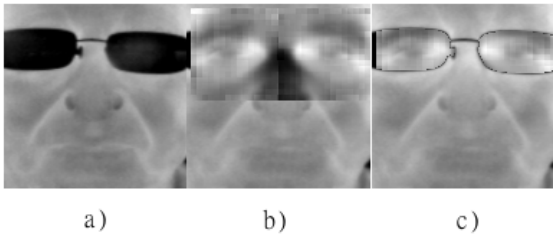

Figure 12: Stages of the eyeglass replacement in LWIR image (a) Input LWIR image (b) LWIR image with synthesised LWIR eye patches (c) Copy of LWIR image in image (a) with synthesised LWIR image copied over the eyeglasses.

4 Experimental Method

In this Section we will discuss the different methods of feature extraction and image fusion used in our experiments. The methods for image fusion are divided into several distinct classes, namely; transform space (Section 4.1) which is at the signal level, feature space (Section 4.2) which is at the feature level, and match-score space (Section 4.3) which can be considered as the symbolic level. The discussion on image fusion in the transform space (Section 4.1) includes subsections 4.1.3 and 4.1.4 which discuss the non-adaptive and adaptive coefficient selection methods we have applied in these experiments.

4.1 Image Fusion in Transform Space

For our research we have applied two commonly used and well known image transforms; the Discrete Wavelet Transform (DWT) and the Wavelet-Based Contourlet Transform (NSCT). These are selected because of their well known computational efficiency and effectiveness in capturing texture and edge detail in images. As discussed in Section 2.1, the NSCT is shift invariant and is more efficient at capturing texture and contour details than the DWT but has a higher computational cost. It is intuitive to assume that the increased efficiency in texture and contour capture of the NSCT would provide better fused output for face recognition compared to the DWT. However, to our knowledge there has been no experimental work using the NSCT to fuse VIS, LWIR and depth images for face recognition.

Both transforms produce coefficient sub-images which, for purposes of visualisation, can be noramalised to the 0-255 grayscale range (e.g. Figure 17). When considered in this context the edges in the coefficient sub-images represent potentially useful information to be included in the final fused image. The adaptive fusion methods described in this chapter are therefore selected for their ability to measure edge detail in the coefficient images of the VIS, LWIR and depth images. As discussed in Section 2.3, such metrics have been used to adaptively fuse images for medical or surveilance purposes, but have not been tested for multispectral/multimodal image fusion for face recognition.

In the sections below we will describe the nature of the transforms used as well as the various non-adaptive and adaptive methods we use for selecting the transform coefficients during fusion.

4.1.1 Discrete Wavelet Transform (DWT)

For these experiments the component images are fused in the transform domain using the Discrete Wavelet Transform (DWT) and Haar wavelets, which are ideal for real-time applications because of their low computational cost. The DWT and Haar wavelets are well established for image fusion with applications to face recognition [31, 44, 71].

The decomposition of a an image using the DWT and Haar wavelets is described. For an input image M which is m x n pixels in size we move pairwise along the rows of the image and for each pair of pixels we compute the 'low-pass' and 'high-pass' components L and H respectively:

$$L(x,y) = \frac{(2x,y) + (2x+1,y)}{\sqrt{2}}$$

$$H(x,y) = \frac{(2x,y) - (2x+1,y)}{\sqrt{2}}$$

$$x = 0, 1, 2...\frac{n}{2}$$

$$y = 0, 1, 2...m$$

where n =the pixel width of image M and m =the pixel height of image M.

To complete the transform at the first level the same operation is applied along the y-axis of the images L and H. Thus for the low-pass image we take the pixel pair $L(P_1(x,y))$ and $L(P_2(x,y))$ and similarly for the high-pass image we have the pixel pair $H(P_1(x,y))$ and $H(P_2(x,y))$. The four coefficient images are then given as:

$$LL(x,y) = \frac{L(x,2y) + L(x,2y+1)}{\sqrt{2}}$$

$$HH(x,y) = \frac{H(x,2y) - H(x,2y+1)}{\sqrt{2}}$$

$$HL(x,y) = \frac{H(x,2y) + H(x,2y+1)}{\sqrt{2}}$$

$$LH(x,y) = \frac{L(x,2y) - L(x,2y+1)}{\sqrt{2}}$$

$$y = 0, 1, 2...\frac{m}{2}$$

$$x = 0, 1, 2...n$$

where m and n are the pixel height and width of images L and H respectively.

The four resulting images LL, HH, HL, LH are half the width and height of the input image M and can be considered as the average, diagonal detail, vertical detail and horizontal detail coefficients of input image M respectively. In order to decompose the image at further levels, the transform is repeated using the average coefficient image LL as the input image M. The fusion of two images is performed by combining these coefficients at the desired transform level such that a single set of average and detail coefficients is produced. The final fused image is then produced by applying the inverse of the Discrete Wavelet Transform (IDWT).

4.1.2 Contourlet Transform (NSCT)

The Contourlet Transform (CT), first proposed by Do and Vetterli in [21], is a multiresolution, multidirectional transform which captures contours and textures more efficiently than the DWT. For our experiments we have used an improved version of the original CT, namely the Nonsubsampled Contourlet Transform (NSCT) as presented in [20] and has been for image fusion [26]. The NSCT is realised by combining a Nonsubsampled Pyramid (NSP) structure and Nonsubsampled Directional Filter Bank (NSDFB). The NSP and NSDFB provide multiresolution and multidirectional analysis respectively. Unlike the Laplacian pyramid used in the CT, the NSP does not upsample or downsample the image and as such the NSCT is also shift invariant.

The structure and ideal frequency response of the fan NSFB which comprises the NSDFB is shown in Figure 14a. The first stage horizontal and vertical directional analysis filters are shown

as $U_0(Z)$ and $U_1(Z)$ respectively. The complementary synthesis filters are shown as $V_0(Z)$ and $V_1(Z)$.

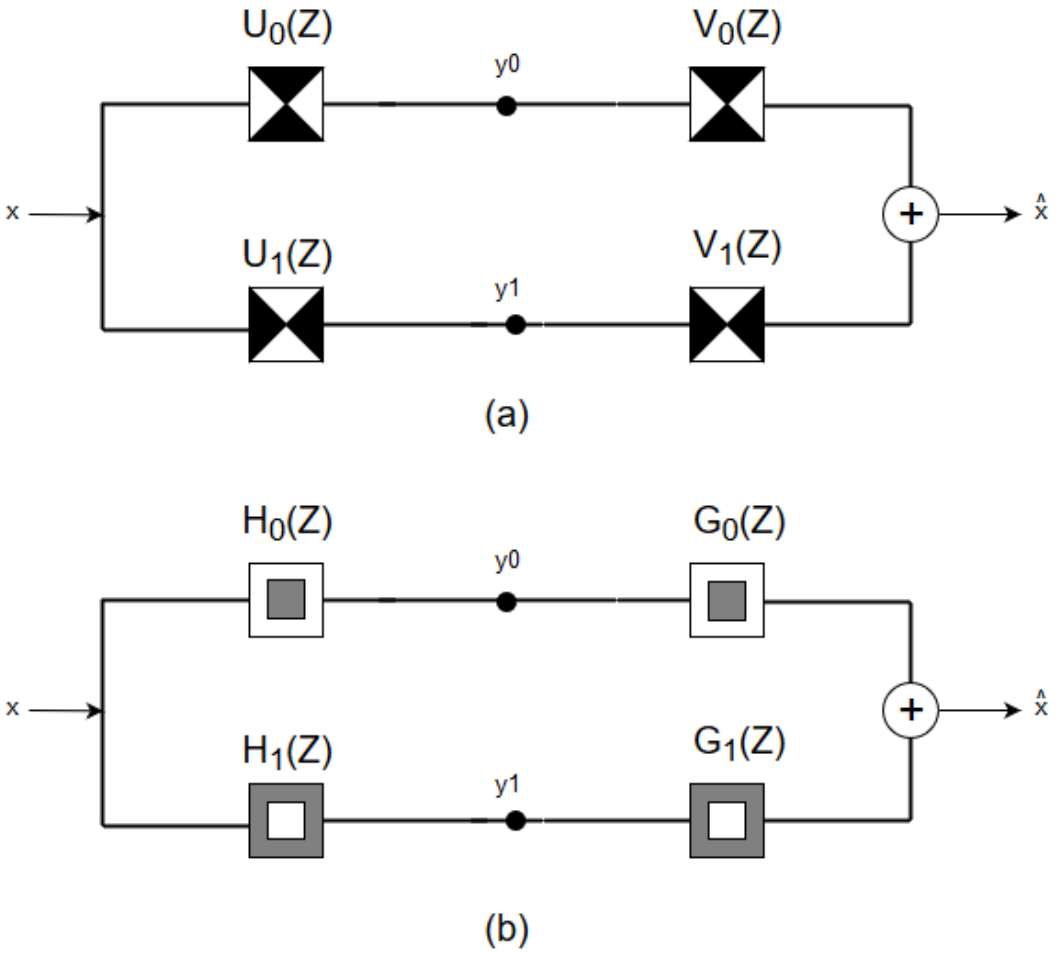

Figure 14: The structure of the two channel nonsubsampled filter banks (NSFBs) that comprise a) the NSDFB, b) the NSP

In Figure 15a the structure of a four channel NSDFB is shown. The tree-structure built from the fan NSFBs allows additional directions to be analysed by upsampling the filters at each stage of the tree. Thus for an l^{th} stage NSFB there will be 2^l directional subbands.

In Figure 15 the upsampled fan NSFBs at the second level of the NSDFB are shown as $U_i(Z^Q), i = 0, 1$ have a "checker board" frequency support and produce 4 directional subbands $y_k, k = 0, 1, 2, 3$. The directional frequency partitioning produced by the combination of the first

and second stage filters of the NSDFB is shown in Figure 15b.

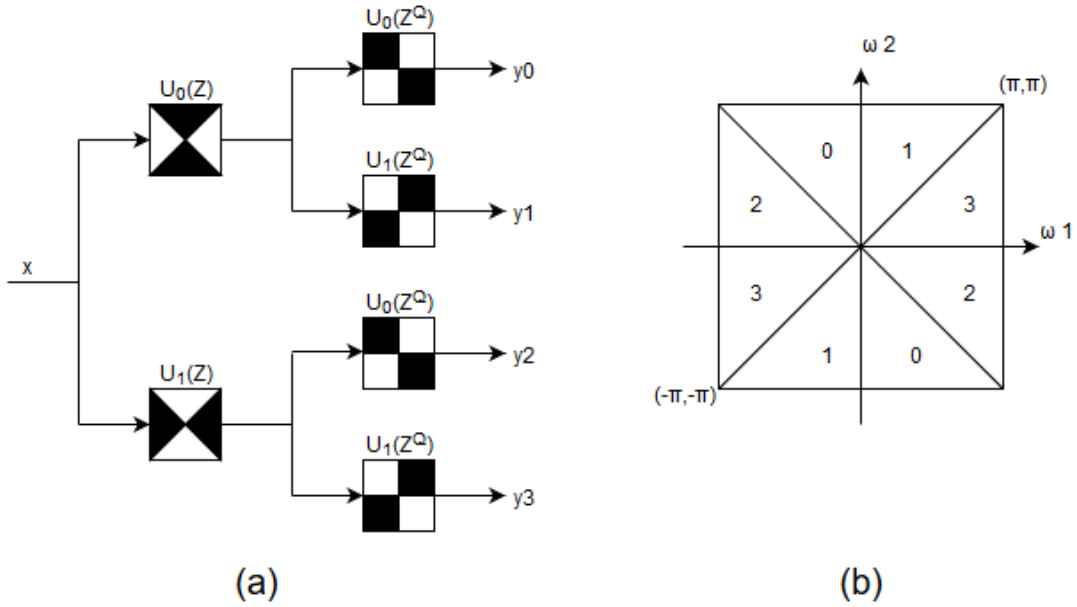

Figure 15: a) The tree structure of a four channel NSDFB comprising fan filter banks b) Partitioning of the 2D frequency plane by the four channel NSDFB

The structure and ideal frequency response of the two channel filter bank which comprises the NSP can be seen in Figure 14b. The first stage lowpass and bandpass filters of the pyramid Nonsubsampled Filter Bank (NSFB) are identified as $H_0(Z)$ and $H_1(Z)$ respectively, along with the complementary synthesis filters $G_0(Z)$ and $G_1(Z)$.

Each pass of the NSP produces a lowpass filtered image and a bandpass filtered image. Further levels of analysis can be achieved by resampling the lowpass image of the previous level in a similar manner to the Nonseparable Wavelet Frame Transform (NSWFT) (discussed in section 2.2).

The output from the NSP is then fed into the NSDFB which constitutes the NSCT. Figure 16 shows a two stage NSP combined with a NSDFB producing 8 directions of analysis at the first scale and 4 directions of analysis at the second scale. The first stage NSP filters are shown as $H_k(Z), k = 0, 1$ and the second stage filters as $H_l(Z^2), l = 0, 1$

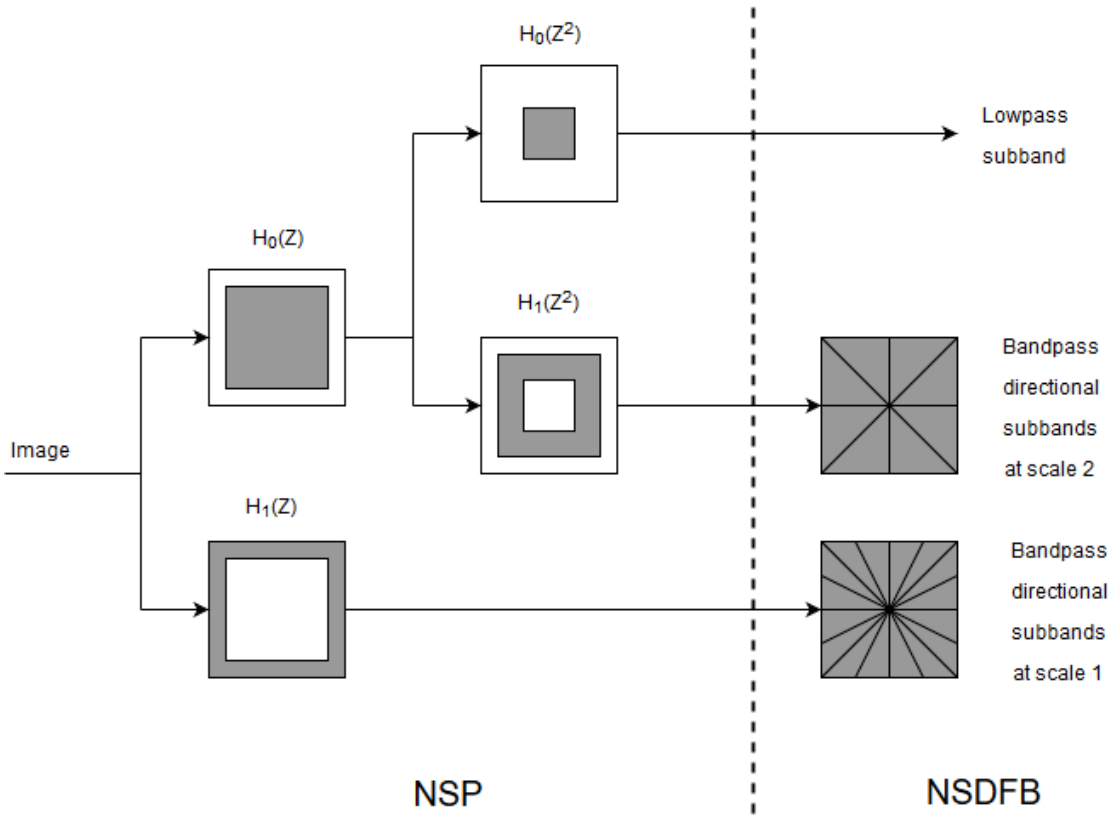

Figure 16: The combination of a NSP and NSDFB in the NSCT.

For our experiments we have used Haar filters for the NSP and the 'pkva' filters [60] for the NSDFB. We took a selection of our database and calculated the energy of the resulting NSCT coefficient images for varying NSP levels and a varying number of DSFB directions. From these results we heuristically use a 3 level NSP with 4 directions of analysis at each level in order to maximise the coefficient energy. For the fusion we consider the NSCT lowpass subband images and directional bandpass image as the DWT average and detail coefficient images respectively. For our experiments using images fused in the NSCT space we have used adaptive fusion methods to select the directional subband coefficients in the same we we select the DWT detail coefficients as described in Section 4.1.4. These adaptive selection methods are applied separately to each directional subband coefficient image i.e. there is no discrimination or weighting for directional subbands during fusion. The NSP lowpass coefficients are fused by taking a mean in the same way as the DWT average coefficients.

We have used the NSCT functions provided by the contourlet toolbox for Matlab^{TM} for our experiments. An example of the NSCT applied to a sample face image from our database is shown in Figure 17.

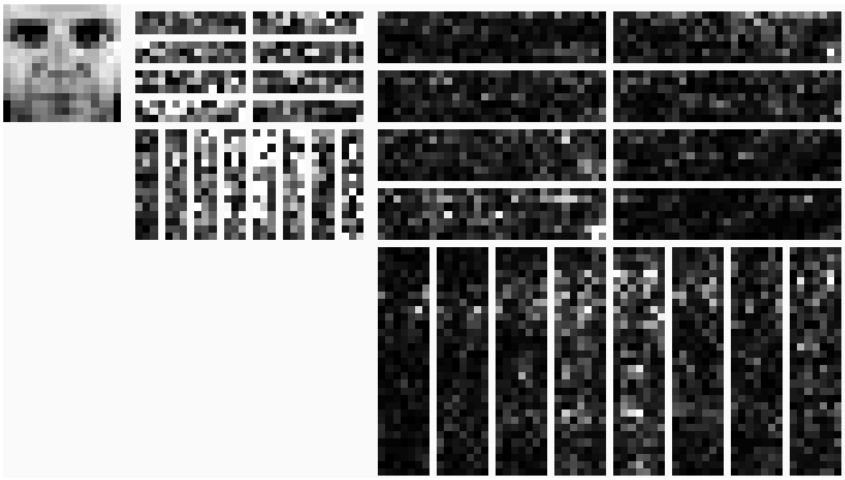

Figure 17: An example of the contourlets produced after applying the NSCT at 3 levels of DWT with 8 directions at the finest level. Coefficients have been normalised to 0-255 for better visualisation

4.1.3 Non-Adaptive Fusion In Transform Space

As discussed in Section 2.3 on page 19, weighting the transform coefficients during image fusion can compensate for variations or degredation in the component images by biasing the coefficients from a specific modality at the risk of over-fitting to a particular database [38]. For our experiments we have used a non-adaptive weighted fusion in the DWT space (Section 4.1.1).

The VIS and LWIR images from the database are fused at the 3rd level DWT transform and the linear fusion weights are applied to the average and detail coefficient values of each component image:

$$W_{\varphi}^{VIS/LWIR} = (\omega_{avg}^{VIS} . W_{\varphi}^{VIS}) + (\omega_{avg}^{LWIR} . W_{\varphi}^{LWIR})$$

$$W_{\psi}^{VIS/LWIR} = (\omega_{det}^{VIS} \,.\, W_{\psi}^{VIS}) \,+\, (\omega_{det}^{LWIR} \,.\, W_{\psi}^{LWIR})$$

Where W_{φ} and W_{ψ} are average and detail coefficients respectively and where the fusion weights ω_{avg}^{VIS} and ω_{det}^{VIS} are constrained by the inequalities:

$$0.1 \leq \omega_{avg}^{VIS} \leq 1.0$$

$$0.1 \leq \omega_{det}^{VIS} \leq 1.0$$

These were evaluated in increments of 0.1 and the values substituted in the expressions below to provide the corresponding fusion weights $\omega_{avg}^{LWIR}$and ω_{det}^{LWIR}:

$$\omega_{avg}^{LWIR} = 1.0 \,-\, \omega_{avg}^{VIS}$$

$$\omega_{det}^{LWIR} = 1.0 \,-\, \omega_{det}^{VIS}$$

The corresponding depth images are then also decomposed to the 3rd level DWT and the average and detail coefficient values are fused using the linear fusion weights:

$$W_{\varphi}^{VIS/LWIR/DEPTH} = (\omega_{avg}^{VIS/LWIR} \,.\, W_{\varphi}^{VIS/LWIR}) \,+\, (\omega_{avg}^{DEPTH} \,.\, W_{\varphi}^{DEPTH}))$$

$$W_{\psi}^{VIS/LWIR/DEPTH} = (\omega_{det}^{VIS/LWIR} \,.\, W_{\psi}^{VIS/LWIR}) \,+\, (\omega_{det}^{DEPTH} \,.\, W_{\psi}^{DEPTH}))$$

where the fusion weights are constrained by the following inequalities:

$$0.1 \leq \omega_{avg}^{VIS/LWIR} \leq 1.0$$

$$0.1 \leq \omega_{det}^{VIS/LWIR} \leq 1.0$$

These were evaluated in increments of 0.1 and the values substituted in the expressions below to provide the corresponding fusion weights ω_{avg}^{DEPTH} and ω_{det}^{DEPTH}:

$$\omega_{avg}^{DEPTH} = 1 - \omega_{avg}^{VIS/LWIR}$$

$$\omega_{det}^{DEPTH} = 1 - \omega_{det}^{VIS/LWIR}$$

Thus as we increase the the VIS average coefficient fusion weight from $0.1 - 1.0$ the LWIR average coefficient fusion weight will decrease from $1.0 - 0.1$ and the same for the detail coefficient weights. For each combination of VIS and LWIR fusion weights we can also vary the depth image coefficient weights in a similar manner i.e. the fused VIS/LWIR average coefficient weight is increased from $0.1 - 1.0$ as the depth average coefficient weight is decreased from $1.0 - 0.1$ and so on for the detail coefficient weights.

4.1.4 Adaptive Fusion in Transform Space

As discussed in Section 2.3 the ability of an image fusion method to adapt to the quality, or variations within the component images is desirable. In order to achieve this we have experimented with several methods of transform coefficient selection using the following measures.

- ***Energy Measure:*** We apply a scrolling 3x3 window to each pixel in turn. The energy is measured as in [32] as the sum of the squares of the pixel values. So taking the VIS horizontal DWT coefficient image HH^{VIS} from Section 4.1.1 for example, the energy measure at pixel position $HH_E^{VIS}p(i,j)$ is defined as:

$$HH_E^{VIS} = \sum_{i,j} HH^{VIS}p(i,j)^2$$

We calculate this for each DWT subband image in each component image to be fused. As in [26] we then define a fusion map by selecting the coefficient corresponding with the maximum energy measure. For the fusion of two images, each position in the fusion map F_E will have a value 0 or 1 representing the coefficient selection from either image A or B respectively. Taking the energy measures from two horizontal DWT coefficient images HH_E^{VIS} and HH_E^{LWIR} for example, the fusion map is decided by:

$$F_E = \begin{cases} 1 & \text{if } HH_E^{VIS} \geq HH_E^{LWIR} \\ 0 & \text{if } HH_E^{VIS} < HH_E^{LWIR} \end{cases} \tag{2}$$

A consistency verification similar to that used in [26] is applied to the fusion map where for a coefficient selection $F_E(i,j)$ the majority of the surrounding coefficients in a 3x3 window must also be from the same component image. If not, the central coefficient selection is changed to the majority vote. In the event of a 4/4 split the coefficient selection remains unchanged. However, in our experiments we are fusing three images which requires a small modification to the process above. Our method produces a fusion map where each position has a value of either 1,2 or 3 representing a selection of either VIS, LWIR or Depth respectively. Since this does not produce a binary matrix it cannot be applied directly to the coefficient images. Instead the continuity check is applied by counting the occurrences of the values in a 3x3 neighbourhood. As before, changes are then made by a majority vote with ties resulting in no change. Three separate binary masks for the VIS, LWIR and Depth coefficients are then generated and applied to the separate coefficient images.

- ***Weighted Sum-Modified Laplacian Activity Measure:*** We use the WSML activity measure as described in [26] to calculate an activity level measure of the area surrounding a coefficient. For a single coefficient $p(i,j)$ in a subband detail image, the Modified Laplacian (ML) value is given as:

$$ML_p(i,j) = |2p(i,j) - p(i-1,j) - p(i+1,j)| + |2p(i,j) - p(i,j-1) - p(i,j+1)|$$

The WSML of $p(i,j)$ can therefore be expressed as:

$$WSML(p(i,j)) = \sum_{x=-1}^{1} \sum_{y=-1}^{1} w(x+1, y+1) \, . \, ML_p(i+x, j+y)$$

where w is the city block distance weight matrix:

$$w = \frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

The WSML activity measure scores are then used to create a fusion map in the same way as described for the **energy measure** above and given in (2). Thus, using the same example as before, the WSML measures for two horizontal subband DWT coefficient images, HH^{VIS}_{WSML} and HH^{LWIR}_{WSML} are used to decide the WSML fusion map F_{WSML}

$$F_{WSML} = \begin{cases} 1 & \text{if } HH^{VIS}_{WSML} \geq HH^{LWIR}_{WSML} \\ 0 & \text{if } HH^{VIS}_{WSML} < HH^{LWIR}_{WSML} \end{cases}$$

A consistency verification as described in the **energy measure** description above is then applied.

- **Sobel Edge Strength Measure:** Since the coefficients in the transform spaces represent predominantly edge information, we apply a similar approach to [91] whereby a Sobel edge strength is calculated for each coefficient. In theory this should provide a better basis for coefficient selection than the maximum selection rule as the neighbouring coefficients are considered. In our experiments we expand the edge strength calculation to include four Sobel operators [65]. For a given input image A the operators are:

$$S^H_A = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{3}$$

$$S^V_A = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{4}$$

$$S_A^{DU} = \begin{bmatrix} -2 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \tag{5}$$

$$S_A^{DD} = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix} \tag{6}$$

which give impulse responses for horizontal (3),vertical (4), diagonally up from left to right (5) and diagonally down from left to right (6) edges respectfully. The Sobel edge strength $g_A(i,j)$ for a coefficient $p(i,j)$ is then calculated as:

$$g_A(i,j) = \sqrt{S_A^H(i,j)^2 + S_A^V(i,j)^2 + S_A^{DU}(i,j)^2 + S_A^{DD}(i,j)^2}$$

The Sobel edge strength measures are then used to build a fusion map as described in Equation(2) above. The map is then used to select the coefficients to be fused into the final image.

4.2 Discrete Cosine Transform Features

The Discrete Cosine Transform (DCT) is commonly used for image compression and is used in the international transform coding systems standards. The transform is from a spatial domain to a frequency domain using a separable, orthonormal basis of cosine functions to encode the gray level values of an image. The resulting components of the DCT image represent the DCT frequencies from low to high, running from the top left to the bottom right in a zig-zag formation. The first pixel in a DCT transform image B represents the average gray level value for the entire input image A, known as the $'DC'$component, with vertical texture/edge frequencies increasing from left to right and horizontal texture/edge frequencies increasing from top to bottom.

Here the DCT is applied as described in [29, 37] where for an input image A with row and column length M and N respectively, the transformed image B is calculated as:

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos \frac{\pi(2m+1)p}{2M} \cos \frac{\pi(2n+1)q}{2N}$$

for:

$$0 \leq p \leq M-1$$

$$0 \leq q \leq N-1$$

where:

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}} & , p = 0 \\ \sqrt{\frac{2}{M}} & , 1 \leq p \leq M-1 \end{cases}$$

and:

$$\alpha_q = \begin{cases} \frac{1}{\sqrt{N}} & , q = 0 \\ \sqrt{\frac{2}{N}} & , 1 \leq q \leq N-1 \end{cases}$$

In order to extract discriminating features from the images in DCT space we need to identify the DCT components for each image modality which have a large inter-subject variability (i.e.

differ between images of subjects) and low intra-subject variability (i.e. small difference between images of the same subject). For this we apply the Fisher Linear Discriminant (FLD) [76] which has been used to select discriminating features in VIS, NIR and LWIR face images [25].

Using the method described in [25], for a database with a total number of subjects P with a total of F images per subject in the training set we have:

The image f of a subject p

$$i_{p,f}(x,y) \tag{7}$$

where $p = 1 \dots P$ and $f = 1 \dots F$

The DCT of the image given in Equation (7)

$$I_{p,f}(f_1,f_2) = 2D - DCT\{i_{p,f}(x,y)\} \tag{8}$$

The average of each DCT frequency component for the entire training image set

$$m(f_1,f_2) = \frac{1}{P \times F} \sum_{p=1}^{P} \sum_{f=1}^{F} I_{p,f}(f_1,f_2) \tag{9}$$

The average of each DCT frequency component for a subject in the training image set

$$m_p(f_1,f_2) = \frac{1}{F} \sum_{f=1}^{F} I_{p,f}(f_1,f_2) \;\; p = 1 \dots P \tag{10}$$

The variance of each DCT frequency component for a subject in the training image set

$$\sigma_p^2(f_1,f_2) = \frac{1}{F} \sum_{f=1}^{F} (I_{p.f}(f_1,f_2) \; - \; m_p(f_1,f_2))^2 \tag{11}$$

The variance of each DCT frequency component for the entire training image set

$$\sigma_{INTER}^2(f_1,f_2) = \frac{1}{P \times F} \sum_{p=1}^{P} \sum_{f=1}^{F} (I_{p,f}(f_1,f_2) \; - \; m(f_1,f_2))^2 \tag{12}$$

The average of the variance of each DCT frequency component for a subject in the training

image set

$$\sigma^2_{INTRA}(f_1, f_2) = \sum_{p=1}^{P} \sigma^2_{p,f}\,(f_1, f_2) \tag{13}$$

Thus the FLD ratio for a DCT frequency component of a subject in the training image set is given as:

$$FLD_{RATIO}(f_1, f_2) = \frac{|m_p(f_1, f_2) - m(f_1, f_2)|}{\sqrt{\sigma^2_{INTRA}(f_1, f_2) + \sigma^2_{INTER}(f_1, f_2)}} \tag{14}$$

The FLD ratio of the DCT frequencies measures the inter-subject and intra-subject variability of a particular DCT frequency i.e. a DCT frequency with a large FLD ratio represents a high variation between subjects but low variation between images of the same subject. To select the DCT frequencies with the largest inter-subject and lowest intra-subject variability we calculate the FLD ratio for each of the DCT frequencies for each training image set in the database. Surface plots of these ratios are shown in Figure 18 (the larger the spike the greater the discriminating power of the DCT frequency in the training image set). From these plots it is clear that the majority of the discriminating information is contained withing the low-frequency DCT components (the furthest corner of the plot), specifically in a 40x40 area for the VIS, NIR and LWIR images and a reduced 20x20 area for the depth images. Thus for our experiments using the fusion of DCT features, we extract the first 40x40 components from the VIS and LWIR images and 20x20 components from the depth images and concatenate them into a single feature vector.

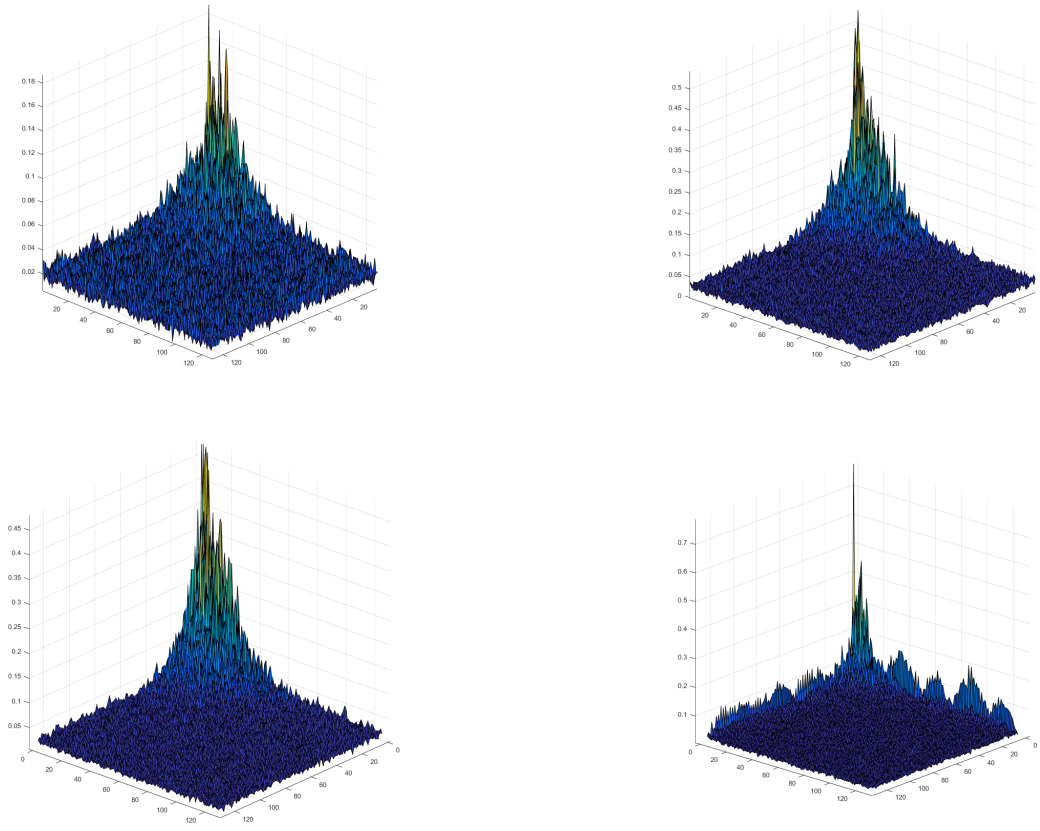

Figure 18: Averages of the FLD ratios of the DCT frequencies for each of the VIS (top left), LWIR (top right), NIR (bottom left) and depth (bottom right) images in a training set.

4.3 Adaptive Image Fusion in Match-Score Space

As discussed in Section 2.4 the fusion of multimodal face images in match-score space has been found to improve recognition performance [8, 9, 81, 85]. Under this fusion method the component images are processed separately for feature extraction and distance calculation between an unknown image (probe, or test image) and known image (gallery, or training image) sets to produce a similarity matrix. In the similarity matrix each row represents an image from the probe set and each column represents an image from the gallery set. The matrix is populated with the match scores, or similarity measures, between each probe and gallery image. An example is shown in Figure 19 with rows representing probe images labeled 1...j and subject numbers 1...n where j =the number of probe images tested for each subject and n =the number of subjects in the database. Similarly, the gallery images listed in the columns are labeled1...m for subject numbers 1...n where j =the number of gallery images used for each subject.

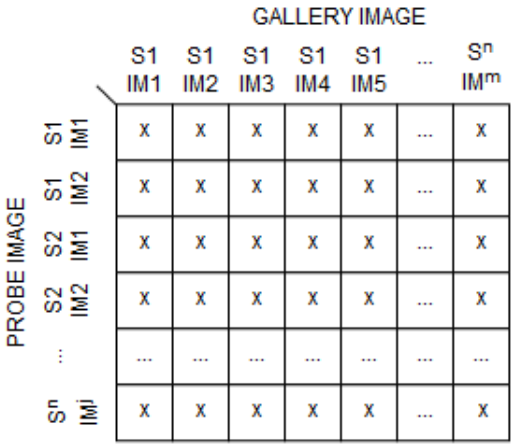

Figure 19: The construction of a similarity matrix. Each position marked x contains a similarity score.

The similarity matrices for each separate modality (i.e. VIS, LWIR and Depth) are calculated and fused by adding the matrices together, element-wise. The fused similarity matrix can then be used to calculate the most likely matches for each comparison and also derive receiver operating characteristic (ROC) and cumulative match score (CMC) curves for the recognition experiment. The fusion can also be weighted in order to enhance or reduce the contribution of a particular modality's similarity matrix to the final recognition score. A flow diagram of the fusion routine is shown in Figure 20 on the following page.

We also adapt the score fusion further in an attempt to reduce the impact any degradation of the VIS image has on the recognition accuracy. To do this we take a global luminance measure by calculating the mean of the gray level values of the VIS image for each recognition comparison made. If the average gray level of a VIS image was found to be < 20, the respective similarity scores in the similarity matrix are set to zero, thus discounting them from the final score fusion and recognition decision.

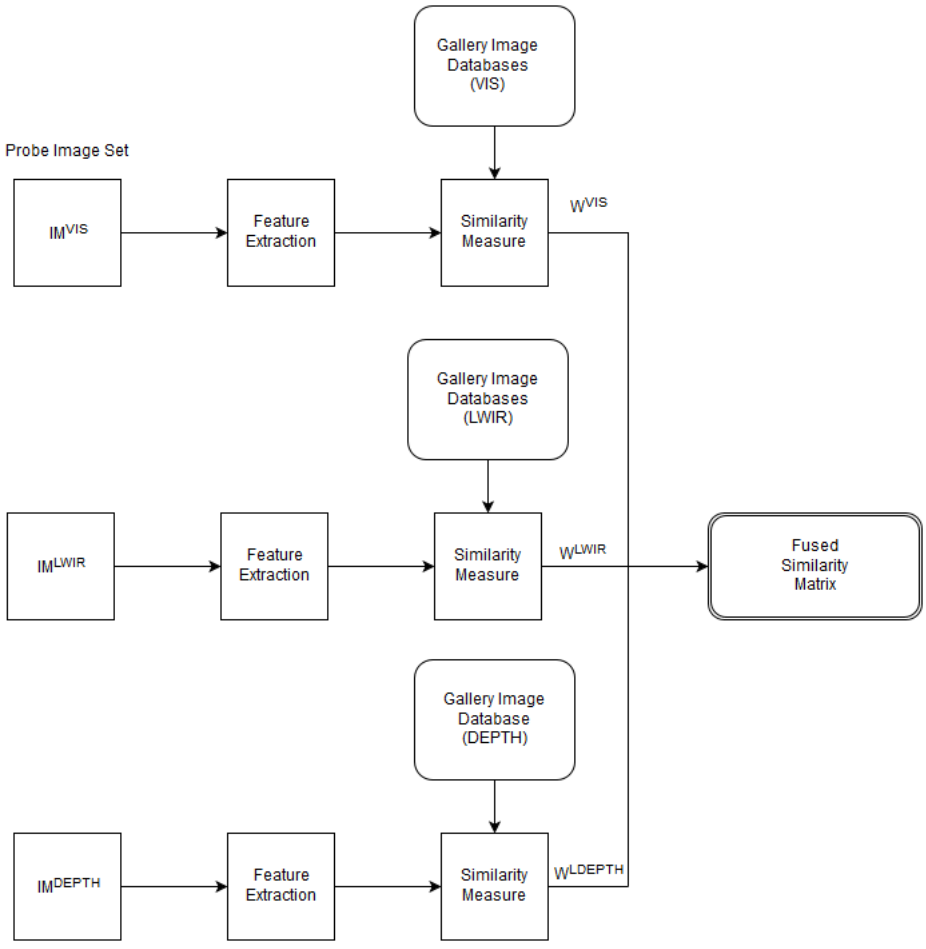

Figure 20: Flow diagram of image fusion in match-score space via fusing the single modality similarity matrices. Each similarity matrix is weighted prior to fusion.

5 Results

In this Section we will present our results and observations from recognition experiments using fused images captured with our camera system. To facilitate easier cross-referencing, the methods of fusion are presented in the same order as in Section 4.

Recognition and verification performance statistics are collected for each set of images using our own, modified version of the 'PhD Toolbox' for Matlab [77, 78]. The recognition methods used in these experiments are:

- Principal Components Analysis using the Mahalanobis cosine distance measure (PCA+MAHCOS) as described in [86]
- Kernel Principal Components Analysis using the Mahalanobis cosine distance measure (KPCA+MAHCOS) as described in [64]
- Kernel Fisher Analysis using the Euclidean distance measure (KFA+EUC) as described in [48]
- Linear Discriminant Analysis using the cosine distance measure (LDA+COS) as described in [4, 28]

For each experiment the images for each subject are divided into 3 training images and 21 test images i.e. for each modality database of 30 subjects we have 90 training images and 630 test images. The recognition and verification tests are repeated 10 times with the training images for each subject randomised. The mean recognition rate and mean verification rate are obtained.

We present results for both recognition (identifying an unknown face image by matching it to a set of known face images i.e. one-to-many) and verification (confirming or denying that a face image is that of its claimed identity i.e. one-to-one). The recognition results are given as percentage accuracy at increasing rank of recognition. A rank 1 recognition accuracy is the percentage of matches the method has achieved where the top match calculated is correct. A rank 2 recognition accuracy is the percentage of matches where the correct match is in the top 2 matches, and so on.

The verification results are given as Verification Rates (VAR) and their False Acceptance Rates (FAR). For example a VAR accuracy at FAR=0.1% represents the percentage of correct verifications with a 0.1% chance of a false acceptance.

The equal error rates (EER) for the verification performance are also given. The EER is the error rate at which the false rejection rate and false acceptance rate are equal.

5.1 Single Modality Images

In order to compare the recognition accuracy of the non-fused images with the accuracy obtained in our image fusion experiments we first present the recognition and verification results for the single modality images. It should be noted that due to the limitations of the Kinect projector and sensor, as discussed in Section 3.1 on page 31, the NIR images do not contain the same lighting biasing as the VIS, LWIR and depth images. We have therefore left the NIR images out of the image fusion experiments in the following sections in order to present a fair comparison under varying lighting conditions.

In Figure 21 and Figure 22 the cumulative match characteristic curves (CMC) for the single modality images are shown for the Lighting Mode 1 (LM1) and Lighting Mode 2 (LM2) data sets (descriptions of the lighting modes were given in Section 3.3 on page 39). The receiver operator characteristic curves (ROC) for the best performing single modality images are shown in Figure 23. The full recognition scores for these curves are given in Table 1 along with the verification and Equal Error Rates (EER) in Table 2.

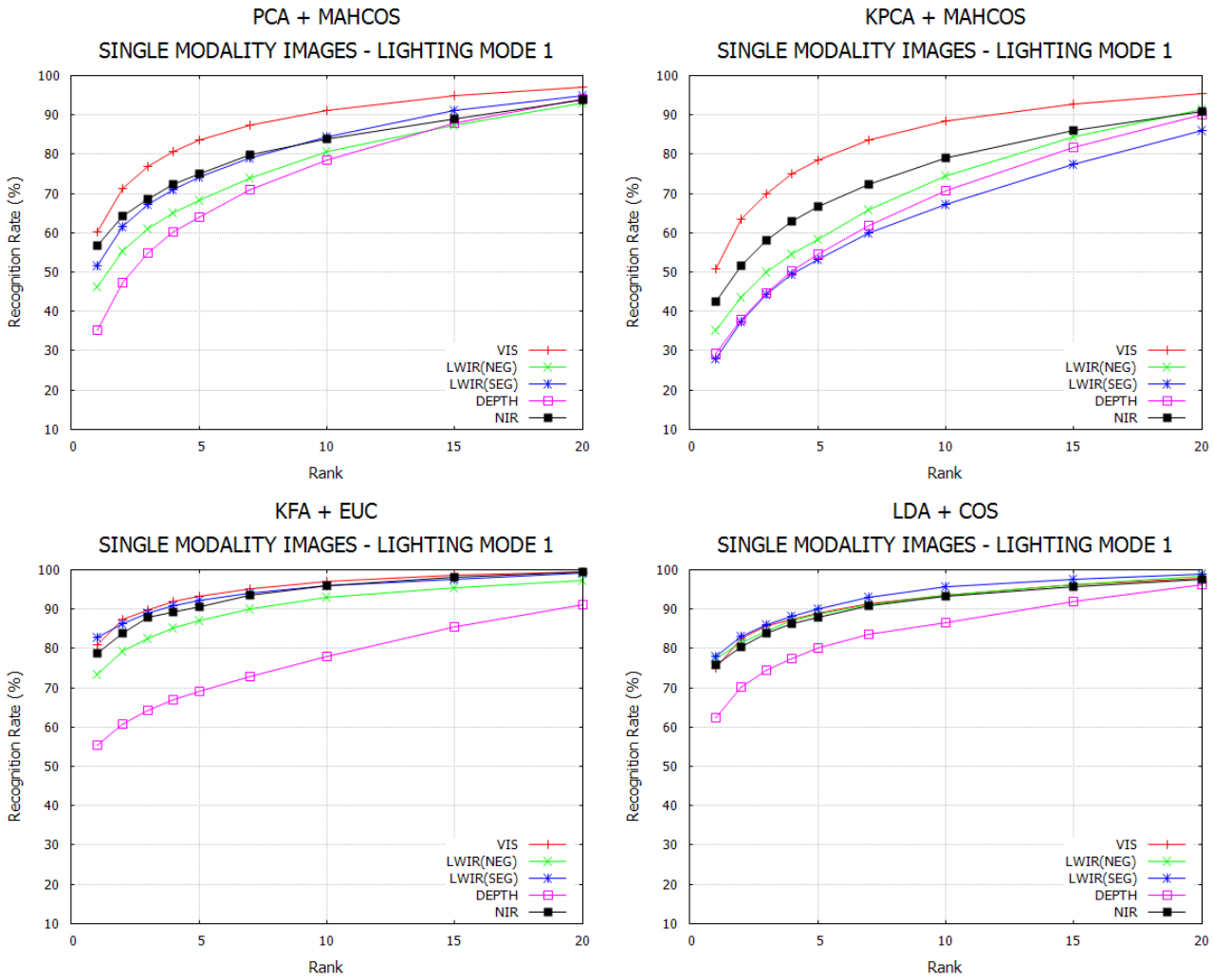

Figure 21: CMC curves and recognition rates for the single modality images captured using Lighting Mode 1

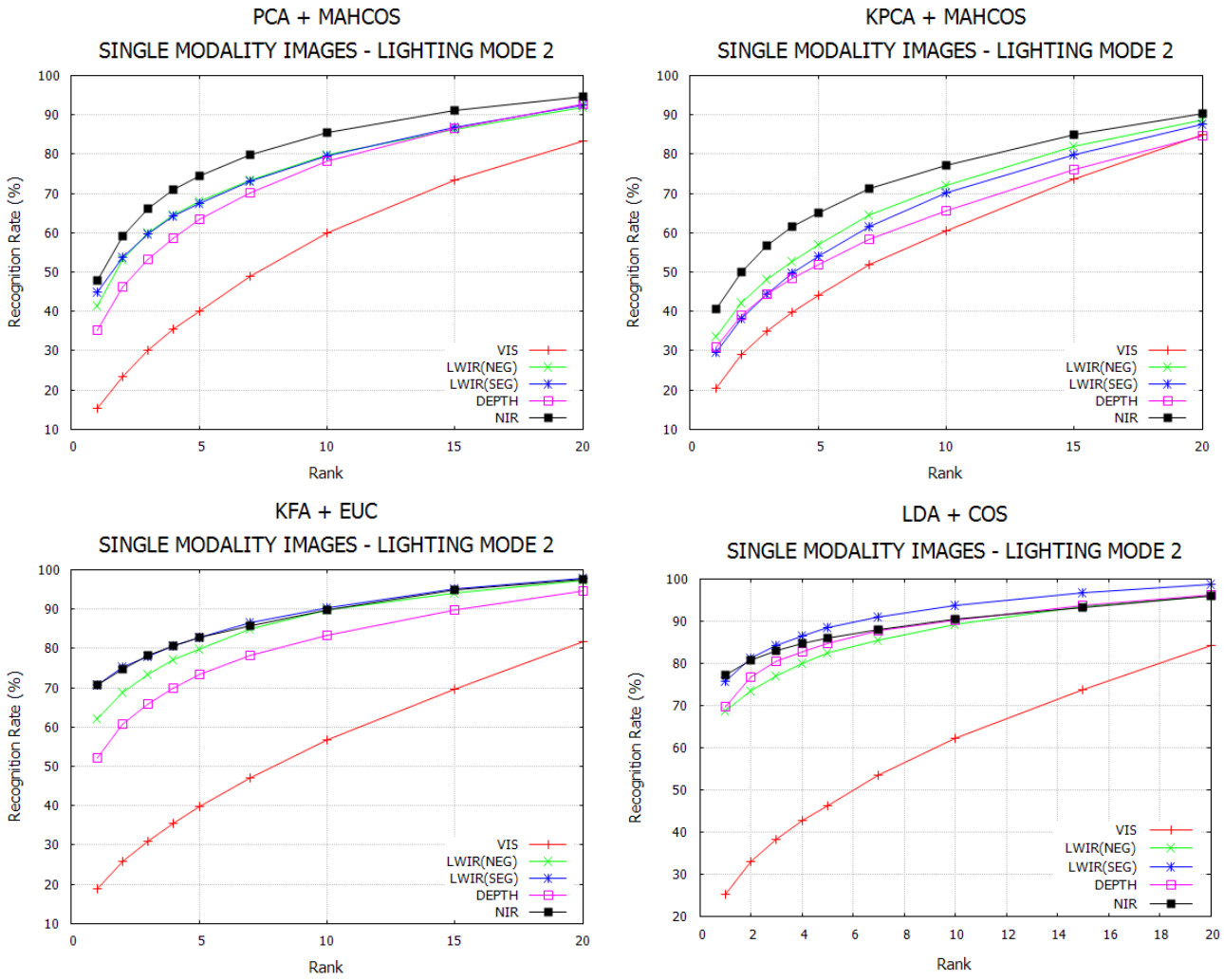

Figure 22: CMC curves and recognition rates for the single modality images captured using Lighting Mode 2

		Recognition Rate % at Rank								
Image Set	Method	1	2	3	4	5	7	10	15	20
VIS	PCA + Mahcos	60.13	71.33	76.91	80.73	83.61	87.37	91.20	94.98	97.13
Lighting Mode 1	KPCA + Mahcos	50.73	63.37	69.83	74.86	78.55	83.50	88.33	92.66	95.36
	KFA + Euc	**80.78**	87.22	89.76	91.89	93.33	95.09	97.00	98.59	99.34
	LDA + Cos	74.97	82.42	85.63	87.45	88.89	91.35	93.60	96.27	97.80
VIS	PCA + Mahcos	15.31	23.53	30.14	35.61	39.98	48.82	60.00	73.49	83.31
Lighting Mode 2	KPCA + Mahcos	20.57	28.98	34.99	39.89	44.00	51.85	60.45	73.60	84.83
	KFA + Euc	18.93	25.87	30.82	35.55	39.71	47.01	56.80	69.60	81.68
	LDA + Cos	**25.25**	32.90	38.18	42.68	46.14	53.38	62.26	73.63	84.14
LWIR	PCA + Mahcos	46.29	55.49	61.07	64.94	68.24	73.87	80.53	87.34	93.11
Lighting Mode 1	KPCA + Mahcos	35.25	43.67	50.00	54.62	58.29	65.79	74.57	84.50	91.38
No synthesised eyeglasses	KFA + Euc	73.42	79.24	82.59	85.11	87.21	89.92	93.02	95.42	97.32
(NEG)	LDA + Cos	**77.33**	81.36	84.39	87.03	88.59	91.17	93.39	96.31	98.39
LWIR	PCA + Mahcos	41.48	53.25	59.83	64.50	67.96	73.40	79.75	86.33	91.94
Lighting Mode 2	KPCA + Mahcos	33.60	42.25	48.20	52.75	56.90	64.38	72.08	82.04	88.81
No synthesised eyeglasses	KFA + Euc	62.08	68.84	73.46	77.18	79.80	84.98	89.75	94.14	97.18
(NEG)	LDA + Cos	**68.82**	73.37	76.92	80.00	82.38	85.45	89.22	93.38	95.89
LWIR	PCA + Mahcos	51.67	61.42	67.08	70.83	74.31	78.92	84.26	90.97	94.81
Lighting Mode 1	KPCA + Mahcos	36.57	44.77	51.18	56.27	60.15	66.33	73.63	83.00	89.66
Synthesised eyeglasses	KFA + Euc	76.60	82.34	85.70	87.65	88.93	91.90	94.29	96.92	98.84
(SEG)	LDA + Cos	**77.89**	83.03	85.96	88.20	90.03	93.06	95.54	97.43	98.92
LWIR	PCA + Mahcos	44.78	53.65	59.55	64.34	67.54	73.11	79.61	86.89	92.43
Lighting Mode 2	KPCA + Mahcos	29.66	38.08	44.37	49.78	54.10	61.44	70.18	79.93	87.71
Synthesised eyeglasses	KFA + Euc	70.81	75.32	78.04	80.61	82.65	86.49	90.35	95.24	97.76
(SEG)	LDA + Cos	**75.60**	81.15	84.12	86.43	88.46	91.06	93.69	96.64	98.70
depth	PCA + Mahcos	35.35	47.22	54.74	60.19	63.90	70.83	78.47	87.97	93.98
Lighting Mode 1	KPCA + Mahcos	29.35	37.88	44.64	50.30	54.69	61.80	70.62	81.81	90.12
	KFA + Euc	55.24	60.83	64.24	66.89	69.01	72.82	77.91	85.45	91.00
	LDA + Cos	**62.36**	70.17	74.56	77.45	80.06	83.50	86.51	92.04	96.18
depth	PCA + Mahcos	35.12	46.19	53.22	58.65	63.39	70.16	78.27	86.54	92.83
Lighting Mode 2	KPCA + Mahcos	30.84	38.88	44.24	48.51	51.95	58.25	65.61	76.17	84.70
	KFA + Euc	52.10	60.79	65.91	69.80	73.31	78.33	83.34	89.77	94.51
	LDA + Cos	**69.73**	76.79	80.43	82.61	84.78	87.75	90.25	93.76	96.18
NIR	PCA + Mahcos	56.65	64.34	68.60	72.27	74.92	79.71	83.83	88.95	93.71
Lighting Mode 1	KPCA + Mahcos	42.40	51.70	57.93	62.98	66.63	72.38	79.03	86.02	90.83
	KFA + Euc	78.63	83.75	87.85	89.28	90.69	93.42	95.90	98.06	99.39
	LDA + Cos	**75.83**	80.40	83.87	86.35	87.85	90.74	93.15	95.67	97.50
NIR	PCA + Mahcos	47.98	59.07	66.04	71.03	74.43	79.72	85.44	91.09	94.71
Lighting Mode 2	KPCA + Mahcos	40.53	50.04	56.70	61.58	65.02	71.27	77.13	84.95	90.29
	KFA + Euc	70.58	74.63	78.16	80.69	82.69	85.64	89.65	94.73	97.62
	LDA + Cos	**77.27**	80.66	82.89	84.62	86.03	88.07	90.33	93.27	95.84

Table 1: Full recognition results for the single modality images

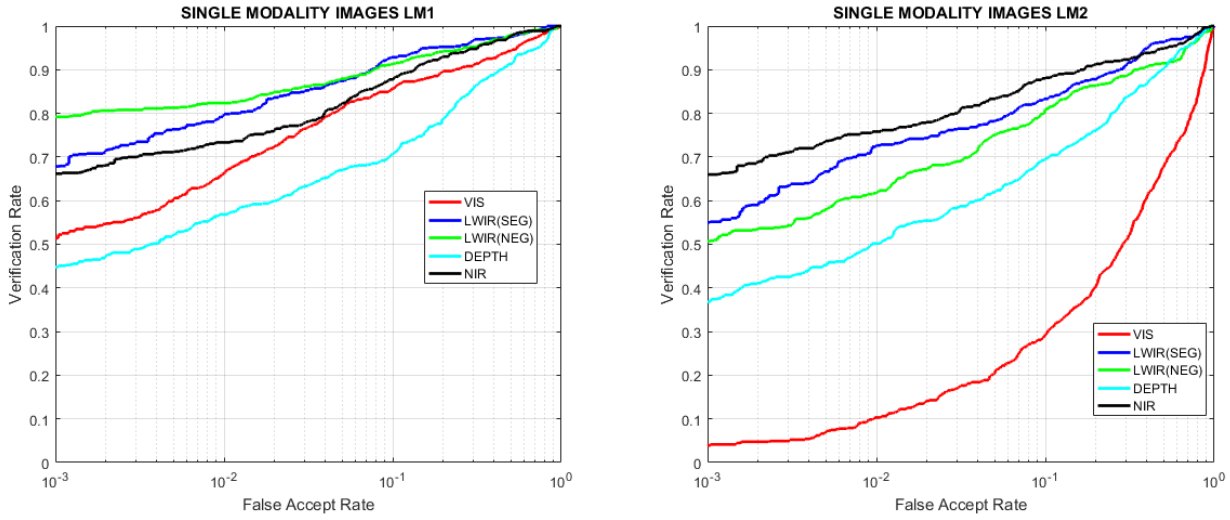

Figure 23: ROC curves for the best single modality image results.

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Method	EER	FAR = 1	FAR = 0.1	FAR = 0.01
VIS	PCA + Mahcos	8.59	71.23	57.52	47.12
Lighting Mode 1	KPCA + Mahcos	10.50	63.77	41.46	22.24
	KFA + Euc	17.11	57.28	40.55	31.14
	LDA + Cos	**10.37**	74.03	61.54	51.98
VIS	PCA + Mahcos	35.92	6.97	2.59	1.34
Lighting Mode 2	KPCA + Mahcos	39.30	12.04	4.62	1.78
	KFA + Euc	46.44	5.52	2.77	2.07
	LDA + Cos	**35.33**	13.13	6.98	4.95
LWIR	PCA + Mahcos	12.88	67.56	59.26	54.63
Lighting Mode 1	KPCA + Mahcos	13.12	65.87	50.18	40.75
No synthesised eyeglasses	KFA + Euc	23.40	60.11	55.73	52.64
(NEG)	LDA + Cos	**4.58**	90.61	84.50	80.21
LWIR	PCA + Mahcos	**12.87**	65.44	59.25	55.50
Lighting Mode 2	KPCA + Mahcos	16.40	62.87	54.33	49.32
No synthesised eyeglasses	KFA + Euc	29.69	51.10	46.96	44.88
(NEG)	LDA + Cos	13.14	69.09	61.85	59.54
LWIR	PCA + Mahcos	**5.42**	90.59	82.56	74.97
Lighting Mode 1	KPCA + Mahcos	19.35	59.36	51.29	43.31
Synthesised eyeglasses	KFA + Euc	14.89	72.53	64.80	60.77
(SEG)	LDA + Cos	10.82	74.56	66.55	62.22
LWIR	PCA + Mahcos	13.17	68.56	59.78	53.65
Lighting Mode 2	KPCA + Mahcos	15.32	77.26	67.57	56.48
Synthesised eyeglasses	KFA + Euc	21.82	63.26	59.41	56.91
(SEG)	LDA + Cos	**12.52**	69.00	61.83	57.95
depth	PCA + Mahcos	**12.75**	66.97	46.82	30.13
Lighting Mode 1	KPCA + Mahcos	23.34	35.47	21.94	12.19
	KFA + Euc	30.79	36.83	26.80	18.55
	LDA + Cos	15.96	62.66	45.82	31.95
depth	PCA + Mahcos	**13.17**	68.79	50.49	33.26
Lighting Mode 2	KPCA + Mahcos	24.07	51.39	38.15	26.37
	KFA + Euc	33.69	32.87	22.22	15.77
	LDA + Cos	13.53	69.42	54.36	36.83
NIR	PCA + Mahcos	**8.95**	78.61	72.07	67.56
Lighting Mode 1	KPCA + Mahcos	12.27	73.02	62.93	56.47
	KFA + Euc	16.10	69.82	66.63	64.45
	LDA + Cos	10.20	76.10	68.75	65.46
NIR	PCA + Mahcos	**7.83**	78.44	69.50	58.30
Lighting Mode 2	KPCA + Mahcos	14.00	66.97	57.47	50.44
	KFA + Euc	20.20	62.06	57.16	52.71
	LDA + Cos	11.66	77.53	72.20	64.14

Table 2: Full verification results for the single modality images

The degradation of the recognition accuracy between LM1 and LM2 is clearly shown for the VIS results, as expected. The rank one (rank 1) recognition accuracy drops to 76% for the KFA+EUC recognition method under LM2. Conversely, and rather interestingly, under LM1 conditions the VIS images gave a rank 1 recognition of 80.78%, the highest recorded for the single modality image sets showing that, while VIS is extremely useful for face recognition, it is indeed very sensitive to the lighting variations between LM1 and LM2. The results also show that the depth images do not provide a high recognition accuracy using these recognition methods but appear comparatively consistent under lighting variations.

The results show that of the recognition methods used here, the LDA+COS method is the most consistent between the LM1 and LM2 images. This is to be expected as in [4] it is shown that the application of LDA to the PCA subspace improves recognition accuracy for images with varying lighting. The combined use of LDA with PCA also deals with the dimensionality problem inherent in the PCA recognition method [28].

While LWIR sensors have been shown to have a small sensitivity to lighting variations [89], the slight variations in the LWIR results between between the LM1 and LM2 image sets are likely due to a combination of the unconstrained temperatures of the subjects and the room during capture. The two sets also contain different subjects. Variances in recognition rates of this size have been previously reported in the literature [25] and are therefore expected. There is, however, a change in recognition rate across the two image sets for the depth image results from 62% rank 1 for LM1 to 69% rank1 for LM2. This is counter intuitive considering the points raised in Section 2.4 where highly directional lighting was found to cause "blooms" in the depth images which can reduce recognition accuracy. As the depth sensor is not sensitive to changes in temperature, the variation in recognition between the LM1 and LM2 image sets must therefore be due to the variation in subjects.

Comparisons between the LWIR images containing non synthesised eyeglasses (NEG) and synthesised eyeglasses (SEG) show a small but measurable improvement in recognition accuracy using the PCA+MAHCOS and KFA+EUC methods for LM1 images. However, a substantial reduction in verification accuracy for the same images was also measured for the PCA+MAHCOS method (an increase in EER from 4.58% to 10.82%). For the LDA+COS recognition method which proved the most accurate across both databases the improvement was minimal for LM1 images (an increase from 77.33% to 77.89%) but greater for the LM2 images (an increase from 68.82% to 75.60%) which meant the LWIR SEG images were more accurate under LM2 conditions.

5.2 Non-Adaptive Image Fusion in Transform Space

For our experiments using non-adaptive image fusion we have used fixed fusion weights as described in Section 4.1.3.

We have first conducted an extensive exploration of the possible fusion weight combinations that can be applied to the VIS, LWIR and depth image coefficients in the DWT space. For three component images (VIS, LWIR and depth), each with average and detail coefficient subbands receiving a fusion weight between 0.1-1.0 there are a possible 10^4different combinations to test (10x10 weight combinations between VIS and LWIR images and 10x10 weight combinations between the fused VIS and LWIR coefficients and the depth coefficients). With each experiment consisting of 10 recognition experiments with randomised training images, the results presented in this section are therefore taken from a total of 10^5recognition experiments we have conducted. As such we have only used the LDA+COS recognition method, as it was shown in Section 5.1 to be the most accurate.

In Figure 24 we see the CMC curves for the top four weight configurations found for the images in the LM1 set. The same weight configurations were then applied to the images in the LM2 set. The resulting CMC curves are also plotted for comparison in Figure 24. The full recognition and verification scores for both image sets are given in Table 3 and ROC curves for the top LM1 fusion weight configuration versus the LM2 images fused using the same weights are shown in Figure 25.

These results demonstrate that an optimised set of fusion weights derived under certain lighting conditions do not necessarily give similar results under different lighting conditions. The fused LM1 images all produce a >90% rank 1 recognition rate which rises to >96% for a rank 5 match with a verification EER between 3.4-4.0%. In comparison, the same weights applied to the LM2 images show a 12% drop in recognition accuracy with the best rank 1 recognition rate being 80% and only achieving >90% by rank 7. The verification EER for the LM2 images is similarly increased to 9.4%. We assume this is because the LM2 VIS image coefficients add a considerable amount of noise to final fused image as even a relatively low weighting (0.3) causes the recognition and verification accuracy of the fused images to be greatly reduced.

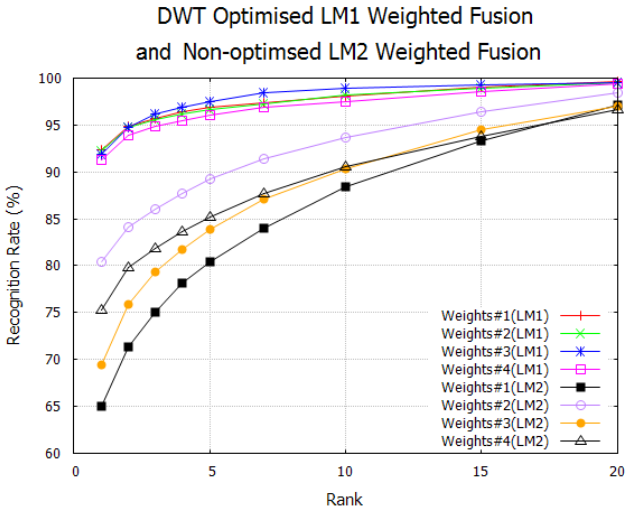

Figure 24: CMC curves for the top 4 fusion weights for LM1 and the same weights applied to LM2

		Recognition Rate % at Rank									Fusion Weight			
Image Set	Result Rank	1	2	3	4	5	7	10	15	20	ω_{avg}^{VIS}	ω_{det}^{VIS}	$\omega_{avg}^{VIS/LWIR}$	$\omega_{det}^{VIS/LWIR}$
DWT LM1	1	**92.29**	**94.86**	**95.68**	**96.41**	**96.92**	**97.34**	**98.09**	**99.03**	**99.66**	**0.3**	**0.8**	**0.7**	**1.0**
LDA+COS	2	92.22	94.71	95.60	96.21	96.61	97.29	98.17	98.89	99.46	0.1	0.7	0.7	0.6
(SEG)	3	91.83	94.68	96.13	96.92	97.47	98.41	98.88	99.32	99.54	0.3	1.0	0.8	1.0
	4	91.22	93.92	94.81	95.50	96.00	96.91	97.54	98.59	99.37	0.2	0.7	0.8	0.7
DWT LM2	-	**64.98**	**71.33**	**75.05**	**78.10**	**80.42**	**84.04**	**88.41**	**93.36**	**97.15**	**0.3**	**0.8**	**0.7**	**1.0**
LDA+COS	-	80.39	84.16	86.02	87.69	89.21	91.40	93.63	96.38	98.44	0.1	0.7	0.7	0.6
(SEG)	-	69.44	75.90	79.38	81.76	83.91	87.05	90.33	94.46	97.02	0.3	1.0	0.8	1.0
L1 weights	-	75.27	79.75	81.81	83.59	85.17	87.67	90.57	93.82	96.59	0.2	0.7	0.8	0.7

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Result Rank	EER	FAR = 1	FAR = 0.1	FAR = 0.01
DWT non-adaptive weights - LM1	1	**3.47**	**93.70**	**88.24**	**83.25**
LDA+COS	2	3.82	93.28	88.06	82.08
(SEG)	3	3.02	93.52	87.65	81.38
	4	4.04	92.91	87.86	82.44
DWT non-adaptive weights- LM2	-	15.75	65.05	52.45	42.50
LDA+COS	-	**9.40**	**80.74**	**73.82**	**67.84**
(SEG)	-	13.41	69.54	56.00	41.64
Weights selected by LM1 results	-	11.87	75.43	68.06	59.43

Table 3: The recognition and verification scores for the top 4 LM1 fusion weights and the same weights applied to LM2

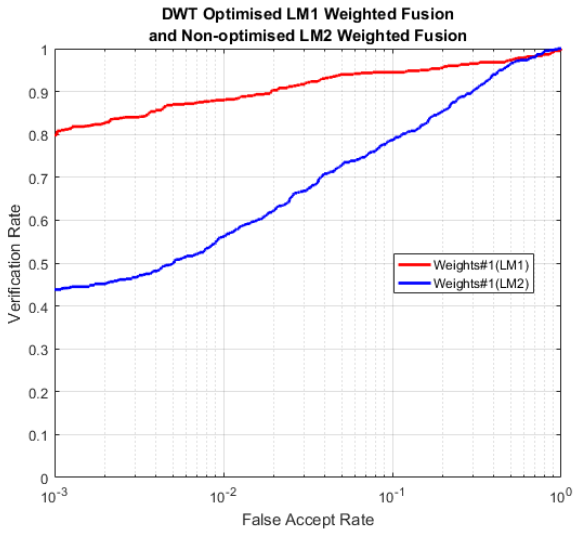

Figure 25: ROC curves for the fused LM1 and LM2 images using weights optimised using LM1 images.

5.3 Adaptive Image Fusion in Transform Space

On considering the results presented in Section 5.2 above, we have conducted experiments using the DWT and NSCT (Section 4.1.2) spaces as well as a series of automatic, adaptive coefficient selection methods described in Section 4.1.4 in order to identify an adaptive method for image fusion that produces images with a consistent recognition accuracy across both LM1 and LM2 lighting modes.

5.3.1 Fusion in DWT Space - Energy Measure

For these experiments the images are decomposed to the 3rd level of the DWT transform, as described in Section 4.1.1. The approximation coefficients are fused by taking the mean and the detail coefficients are fused using an energy measure as described in Section 4.1.4. The CMC curves for the best fused and single modality image results are shown in Figure 26. The full recognition and verification scores for the fused images in these experiments are given in Table 4.

The CMC curves for the fused images show that for the LM1 image set the VIS images outperform the fused image set until the rank 20 (rank 20) match score. The fused images show less sensitivity to the lighting variations in the LM2 image set compared to the VIS single

modality, however their overall recognition accuracy is comparable to the LWIR images for LM1 and are even slightly worse in terms of verification accuracy and EER.

The energy measure used in these experiments is well known in the application of image fusion for surveillance and remote sensing (i.e. satellite imagery) where the fused image is intended for a human operator. While it seems intuitive that the energy measure would be efficient at selecting coefficients for face recognition, the fused images are noisy and are outperformed by the single modality LWIR images.

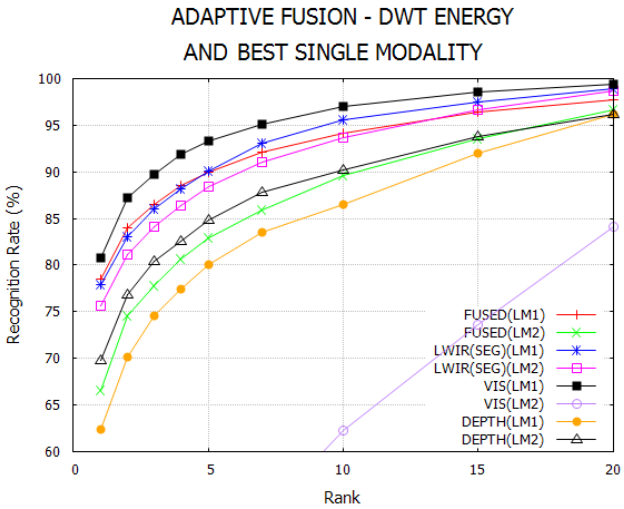

Figure 26: CMC curves for the images adaptively fused in DWT space shown as FUSED(LM1) and FUSED(LM2). Average coefficient selection by mean, detail coefficient selection by energy. Best single modality results shown for comparison.

		Recognition Rate % at Rank								
Image Set	Method	1	2	3	4	5	7	10	15	20
Adaptive Fusion - LM1	PCA + Mahcos	58.14	67.43	72.98	76.35	79.36	83.81	88.41	93.45	96.52
Average - MEAN	KPCA + Mahcos	54.37	65.13	70.70	75.05	78.26	82.86	87.85	92.54	95.47
Detail - ENERGY	KFA + Euc	73.61	79.41	82.28	84.91	86.81	89.48	92.12	95.45	97.35
(SEG)	LDA + Cos	**78.44**	**83.97**	**86.47**	**88.57**	**89.97**	**92.09**	**94.15**	**96.37**	**97.69**
Adaptive Fusion - LM2	PCA + Mahcos	43.30	53.97	60.28	65.14	68.89	75.07	81.42	89.36	94.55
Average - MEAN	KPCA + Mahcos	38.54	49.11	55.67	60.69	64.76	71.33	78.39	87.45	93.64
Detail - ENERGY	KFA + Euc	60.09	67.47	72.21	75.23	77.76	82.22	86.64	92.52	96.34
(SEG)	LDA + Cos	**66.61**	**74.60**	**77.76**	**80.68**	**82.87**	**85.95**	**89.57**	**93.57**	**96.64**

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Method	EER	FAR = 1	FAR = 0.1	FAR = 0.01
Adaptive Fusion DWT - LM1	PCA + Mahcos	9.06	77.54	64.59	55.15
Average - MEAN	KPCA + Mahcos	**7.78**	**83.43**	**69.13**	**57.76**
Detail - WSML	KFA + Euc	18.48	64.96	55.76	46.15
(SEG)	LDA + Cos	9.09	79.30	70.39	61.02
Adaptive Fusion DWT - LM2	PCA + Mahcos	**13.59**	**66.56**	**49.20**	**36.32**
Average - MEAN	KPCA + Mahcos	15.77	59.06	39.90	27.94
Detail - WSML	KFA + Euc	21.74	48.31	36.41	28.79
(SEG)	LDA + Cos	14.70	65.54	50.15	39.22

Table 4: Recognition and verification results for the adaptively fused images using DWT and energy measure

5.3.2 Fusion in DWT Space - WSML Measure

Here the images are decomposed to the 3rd level of the DWT transform, as discussed in Section 4.1.1. The approximation coefficients are fused by taking the mean and the detail coefficients are fused using a WSML selection as described in Section 4.1.4. The full recognition and verification results for these fused images are given in Table 5 and the CMC curves for the best recognition results achieved for each set of fused images and the single modality images are presented in Figure 27. It is interesting to note that the LDA+COS recognition method gave the highest recognition and verification accuracy for all image sets with the exception of the VIS images under LM1 which, as shown in Section 5.1 and the results here, gave good results with the KFA+EUC recognition method under LM1 conditions but worse results under LM2.

Looking at Figure 27 we can see that the adaptively fused images have higher rank 1 recognition rates of 86% and 84% for the LM1 and LM2 images respectively when compared to the best single modality results. More specifically, this equates to a 10.5% and 20.8% increase in rank 1 recognition accuracy over the LWIR images for the LM1 and LM2 lighting modes respectively. The fused images show a decrease of 2.3% in rank 1 recognition accuracy under extreme lighting changes which is slightly less than the 2.94% decrease in the LWIR images and stands in stark contrast to the 68% decrease in the VIS image accuracy.

The adaptively fused images also show a much improved verification accuracy, see Table 5. The fused images reduce the EER to approximately 5-6% across both LM1 and LM2 image sets which represents a reduction in EER by 48% and 68% compared to the best LWIR and VIS results respectively. Similarly the verification rates at a FAR of 1% are 18% and 24% higher compared to the LWIR images for LM1 and LM2 respectively and as we decrease FAR to 0.01% the verification rates for LM1 and LM2 are still 21% and 12% higher than for the LWIR images.

These results are considerably better than those reported for the energy measure fusion routine (Section 5.3.1) where the fused images were outperformed by the single modality LWIR images. In [26] a direct comparison of fused MRI images using the energy and WSML measures shows the WSML method captures more edge and texture detail in the final fused image, although the performance was only marginally better. In our experiments using the WSML measure for fusion for face recognition we can see a much greater difference in performance. This suggests that the WSML is better at removing noise from the VIS component image.

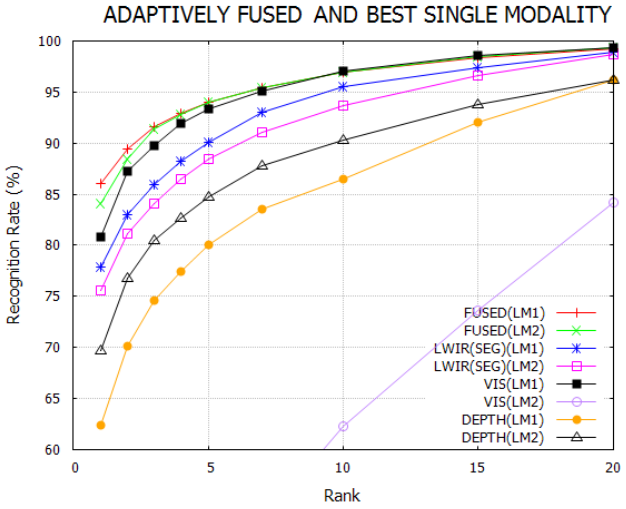

Figure 27: CMC curves for the images adaptively fused in DWT space shown as FUSED(LM1) and FUSED(LM2). Average coefficient selection by mean, detail coefficient selection by WSML. Best single modality results shown as comparison.

		Recognition Rate % at Rank								
Image Set	Method	1	2	3	4	5	7	10	15	20
Adaptive Fusion - LM1	PCA + Mahcos	60.41	69.57	75.10	78.41	80.65	84.56	88.95	94.06	97.44
Average - MEAN	KPCA + Mahcos	50.02	60.11	65.58	69.34	72.42	77.40	83.55	89.65	93.67
Detail - WSML	KFA + Euc	73.99	79.98	83.13	85.52	86.89	89.68	92.47	95.37	98.20
(SEG)	LDA + Cos	**86.07**	**89.37**	**91.58**	**92.89**	**93.98**	**95.37**	**96.92**	**98.31**	**99.22**
Adaptive Fusion - LM2	PCA + Mahcos	41.55	50.55	55.89	60.60	64.47	70.48	77.61	85.84	92.50
Average - MEAN	KPCA + Mahcos	33.09	44.11	51.45	56.18	60.09	66.72	74.37	84.48	91.30
Detail - WSML	KFA + Euc	55.09	62.98	68.45	72.16	75.12	80.43	86.38	92.63	96.68
(SEG)	LDA + Cos	**84.07**	**88.45**	**91.33**	**92.78**	**93.95**	**95.41**	**96.96**	**98.52**	**99.39**

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Method	EER	FAR = 1	FAR = 0.1	FAR = 0.01
Adaptive Fusion DWT - LM1	PCA + Mahcos	8.73	80.78	65.80	55.28
Average - MEAN	KPCA + Mahcos	12.80	70.38	52.15	39.68
Detail - WSML	KFA + Euc	16.61	66.82	59.25	52.67
(SEG)	LDA + Cos	**5.43**	**87.48**	**80.45**	**75.70**
Adaptive Fusion DWT - LM2	PCA + Mahcos	15.44	64.69	53.56	45.03
Average - MEAN	KPCA + Mahcos	18.91	51.87	35.53	24.80
Detail - WSML	KFA + Euc	21.30	43.92	31.28	23.59
(SEG)	LDA + Cos	**6.24**	**86.20**	**75.74**	**65.18**

Table 5: Recognition and verification results for the adaptively fused images using DWT and WSML measure

5.3.3 Fusion in DWT Space - Sobel Measure

For these experiments the images are decomposed to the 3rd level of the DWT transform, as described in Section 4.1.1. The approximation coefficients are fused by taking the mean and the detail coefficients are fused using the Sobel selection as described in Section 4.1.4. The CMC curves for the best fused image results and the best single modality image results are shown in Figure 28 and the full recognition and verification results for the other recognition methods are given in Table 6.

It is clear from the CMC curves that the Sobel selection for image fusion does not produce images robust to changes in lighting. There is a 13% decrease in rank 1 recognition accuracy for the fused images between LM1 and LM2 lighting scenarios. This is better than the variation in the VIS single modality for recognition accuracy but shows only a small improvement over the single modality LWIR images. There is, however a slight improvement in EER and verification accuracy for the fused images compared to the LWIR images.

Comparing these results to those for the WSML selection we suggest that, as the VIS coefficients contain much more edge and texture information when the source image is well lit (LM1), the choice of the coefficients to be fused under LM1 conditions is biased towards the VIS images which perform comparatively well in LM1 as shown in Figure 22. When the VIS images are

degraded under LM2 conditions it appears the Sobel selection, even with a continuity check for the fusion map, tends to inject noise from the VIS coefficients into the final fused image. Thus the fused images under this coefficient selection method are more susceptible to variations in lighting.

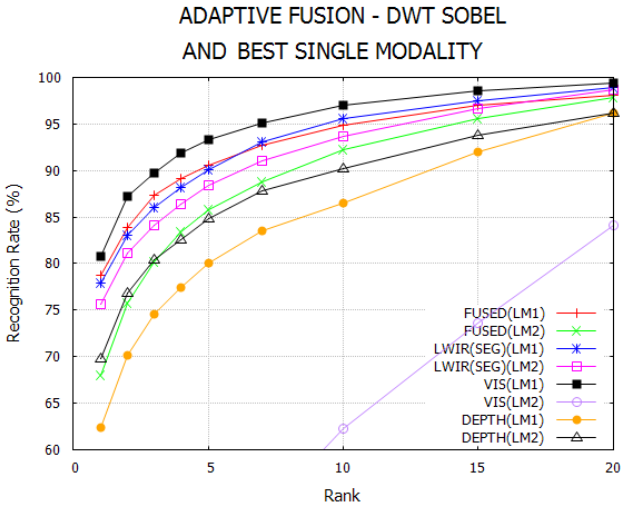

Figure 28: CMC curves for the images adaptively fused in DWT space shown as FUSED(LM1) and FUSED(LM2). Average coefficient selection by mean, detail coefficient selection by Sobel. Best single modality results shown as comparison.

		Recognition Rate % at Rank								
Image Set	Method	1	2	3	4	5	7	10	15	20
Adaptive Fusion - LM1	PCA + Mahcos	58.36	67.59	72.62	76.36	79.20	83.71	88.00	93.22	96.45
Average - MEAN	KPCA + Mahcos	50.32	60.59	65.88	69.52	72.41	77.13	82.38	88.97	93.15
Detail - WSML	KFA + Euc	72.00	77.79	81.09	83.46	85.15	88.15	91.74	95.70	98.42
(SEG)	LDA + Cos	**78.71**	**83.92**	**87.35**	**89.11**	**90.51**	**92.72**	**94.88**	**96.95**	**98.08**
Adaptive Fusion - LM2	PCA + Mahcos	44.80	54.46	60.54	65.06	68.75	74.15	80.95	88.75	93.88
Average - MEAN	KPCA + Mahcos	38.45	48.14	54.83	60.13	64.00	70.49	77.67	86.37	92.58
Detail - WSML	KFA + Euc	60.27	67.58	71.84	75.24	77.72	81.82	86.26	92.37	96.61
(SEG)	LDA + Cos	**68.03**	**75.75**	**80.13**	**83.37**	**85.76**	**88.78**	**92.17**	**95.62**	**97.88**

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Method	EER	FAR = 1	FAR = 0.1	FAR = 0.01
Adaptive Fusion DWT - LM1	PCA + Mahcos	**8.39**	**80.96**	**70.62**	**61.33**
Average - MEAN	KPCA + Mahcos	13.27	70.66	50.44	41.30
Detail - WSML	KFA + Euc	17.82	61.25	51.33	42.03
(SEG)	LDA + Cos	8.50	78.66	66.29	57.55
Adaptive Fusion DWT - LM2	PCA + Mahcos	13.03	**68.62**	**53.64**	**40.99**
Average - MEAN	KPCA + Mahcos	16.45	57.99	40.92	28.32
Detail - WSML	KFA + Euc	20.69	49.87	35.16	26.46
(SEG)	LDA + Cos	**12.27**	68.54	52.10	40.12

Table 6: Recognition and verification results for the adaptively fused images using DWT and SOBEL measure

5.3.4 Fusion in NSCT Space - Energy Measure

For our experiments using the NSCT the images are decomposed to 3 levels of the NSP with 4 directions ([4 4]) of analysis taken at each level, as described in Section 4.1.2. The CMC curves for the best recognition results using the fused images along with the best single modality recognition results for LM1 and LM2 are shown in Figure 29. The full recognition and verification results for fused images using the other recognition methods are given in Table 7.

The recognition results for the energy measure fusion using NSCT are slightly worse than for the DWT, which is surprising as the NSCT is more efficient at capturing edge and texture detail from the component images. As shown in Section 5.3.1 the energy measure failed to reduce the noise injected from the VIS image. It may be that the NSCT is more sensitive to noise than the DWT and when applied to a low quality image such as the LM2 VIS images results in much more noise in the fused coefficients. This would also explain why the fused LM1 images are also outperformed by the equivalent DWT fused images.

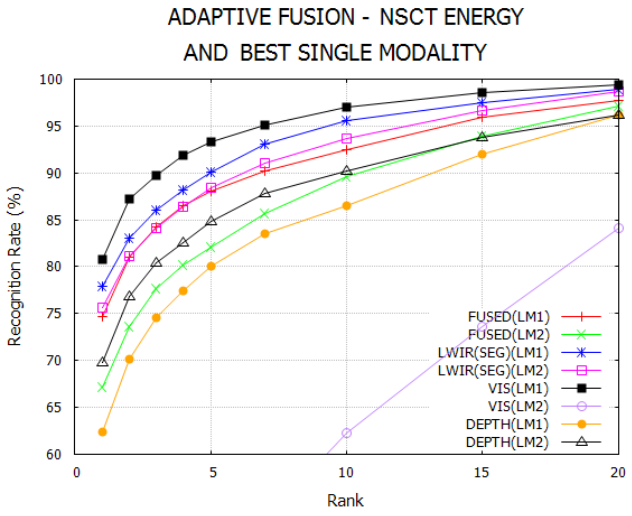

Figure 29: CMC curves for the images adaptively fused in NSCT space shown as FUSED(LM1) and FUSED(LM2). Average coefficient selection by mean, detail coefficient selection by energy. Best single modality results shown for comparison.

		Recognition Rate % at Rank								
Image Set	Method	1	2	3	4	5	7	10	15	20
Adaptive Fusion - LM1	PCA + Mahcos	59.53	67.81	72.85	76.88	80.14	84.25	88.40	92.97	96.13
Average - MEAN	KPCA + Mahcos	47.99	56.88	62.73	66.85	70.25	75.16	80.82	87.66	92.17
Detail - WSML	KFA + Euc	73.96	80.60	83.93	86.42	87.97	90.47	93.20	96.18	98.63
(SEG)	LDA + Cos	**74.71**	**80.98**	**84.17**	**86.54**	**88.01**	**90.16**	**92.51**	**95.87**	**97.77**
Adaptive Fusion - LM2	PCA + Mahcos	45.84	55.80	62.02	66.23	70.03	76.02	82.23	89.22	94.18
Average - MEAN	KPCA + Mahcos	35.41	48.49	56.87	62.32	66.37	72.86	79.60	87.50	93.27
Detail - WSML	KFA + Euc	**68.30**	**74.47**	**78.63**	**81.17**	**83.31**	**86.33**	**89.81**	**94.10**	**96.61**
(SEG)	LDA + Cos	67.19	73.61	77.70	80.18	82.07	85.64	89.58	93.93	97.10

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Method	EER	FAR = 1	FAR = 0.1	FAR = 0.01
Adaptive Fusion DWT - LM1	PCA + Mahcos	**8.22**	**80.67**	**70.22**	**60.82**
Average - MEAN	KPCA + Mahcos	15.08	63.69	48.51	37.08
Detail - WSML	KFA + Euc	19.43	64.52	56.38	47.92
(SEG)	LDA + Cos	10.98	74.92	59.95	48.97
Adaptive Fusion DWT - LM2	PCA + Mahcos	**13.70**	**68.55**	**56.06**	**47.22**
Average - MEAN	KPCA + Mahcos	14.71	57.69	39.95	28.27
Detail - WSML	KFA + Euc	21.56	54.08	40.64	31.40
(SEG)	LDA + Cos	14.85	66.31	53.28	42.32

Table 7: Recognition and verification results for the adaptively fused images using NSCT and ENERGY measure

5.3.5 Fusion in NSCT Space - WSML Measure

The images in this set of experiments are also fused using the NSP with 4 ([4 4]) directions of analysis taken at each level as described in Section 4.1.2 on page 54. The approximation coefficients are fused by taking the mean and the bandpass coefficients for each direction are fused using a WSML selection as described in Section 4.1.4 on page 60 and also applied to the DWT coefficients in the experiments discussed in Section 5.3.2. The CMC curves for the best fused image results and the best single mode images are given in Figure 30 and the full recognition and verification scores for the other recognition methods tested are shown in Table 8.

Interestingly the results for the WSML measure fusion in the NSCT transform space show the selection method doesn't work as well or as consistently as in DWT space. While the rank 1 match rate for the LM1 fused images outperforms the single modality images, the lighting variations in LM2 cause a 33% drop in recognition performance and a large increase in the EER rate to 19%. As shown from the single modality image results, the VIS images become almost unusable for recognition under LM2 and these results suggest the deterioration in image quality is effecting the fused image. It appears that the WSML selection method in NSCT space is including noise from the VIS images in LM2 which is consistent with what was observed in Section 5.3.4 where the NSCT fused images using the energy measure failed to improve on the

DWT fused images.

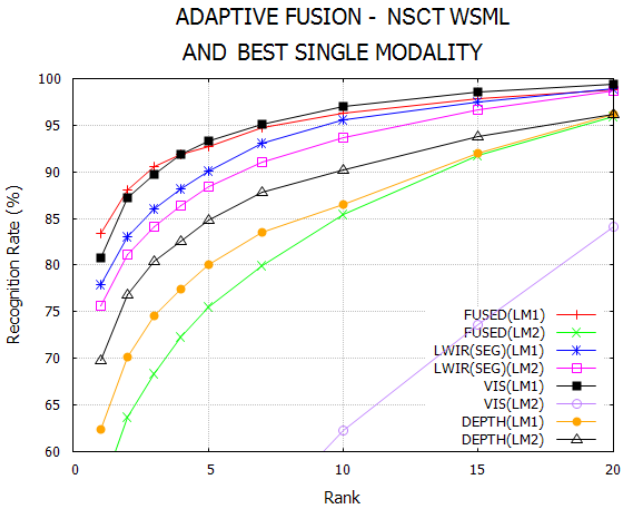

Figure 30: CMC curves for the images adaptively fused in NSCT space using WSML are shown as FUSED(LM1) and FUSED(LM2). Average coefficient selection by mean, detail coefficient selection by WSML. Best single modality results shown as comparison.

		Recognition Rate % at Rank								
Image Set	Method	1	2	3	4	5	7	10	15	20
Adaptive Fusion - LM1	PCA + Mahcos	58.04	66.31	71.49	75.31	78.00	82.82	87.74	93.51	96.86
Average - MEAN	KPCA + Mahcos	52.06	62.12	67.79	71.88	75.65	80.64	85.38	90.95	95.05
Detail - WSML	KFA + Euc	82.63	87.72	89.87	91.26	92.31	93.87	95.77	97.70	99.02
(SEG)	LDA + Cos	**83.38**	**88.07**	**90.55**	**91.81**	**92.74**	**94.70**	**96.23**	**97.78**	**98.81**
Adaptive Fusion - LM2	PCA + Mahcos	42.67	53.20	59.87	64.74	68.93	74.64	80.83	88.75	93.90
Average - MEAN	KPCA + Mahcos	35.81	46.57	53.68	59.26	63.75	70.30	78.14	86.45	92.04
Detail - WSML	KFA + Euc	53.29	63.30	68.31	72.45	75.41	80.53	85.97	92.17	96.14
(SEG)	LDA + Cos	**55.19**	**63.65**	**68.41**	**72.30**	**75.49**	**79.93**	**85.44**	**91.75**	**95.98**

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Method	EER	FAR = 1	FAR = 0.1	FAR = 0.01
Adaptive Fusion DWT - LM1	PCA + Mahcos	9.26	77.70	69.65	64.01
Average - MEAN	KPCA + Mahcos	10.77	77.41	66.36	56.12
Detail - WSML	KFA + Euc	10.97	75.54	64.54	53.48
(SEG)	LDA + Cos	**7.01**	**85.13**	**75.41**	**65.96**
Adaptive Fusion DWT - LM2	PCA + Mahcos	**16.01**	**58.25**	**42.19**	**32.84**
Average - MEAN	KPCA + Mahcos	16.24	56.13	39.82	29.08
Detail - WSML	KFA + Euc	21.84	40.65	26.52	19.37
(SEG)	LDA + Cos	19.13	52.65	36.62	26.06

Table 8: Recognition and verification results for the adaptively fused images using NSCT and WSML measure

5.3.6 Fusion in NSCT Space - Sobel

For these experiments the images are decomposed using the NSCT transform with 4 directions ([4 4]) of analysis taken at each level via as described in Section 4.1.2. The lowpass coefficients are fused by taking the mean and the directional analysis coefficients are fused using the Sobel measure selection as described in Section 4.1.4 on page 60. The CMC curves for the best fused

image results and the best single modality image results are shown in Figure 31 and the full recognition and verification results for the other recognition methods are given in Table 9. Here we can see that the Sobel measure selection method produces poor recognition results that show a 11% drop in recognition rate under the LM2 lighting conditions. Interestingly we see a similar performance for the DWT images fused with the Sobel selection, suggesting that even with the improved texture and edge capture of the NSCT, the Sobel method is still ineffective in terms of noise reduction in the final fused image.

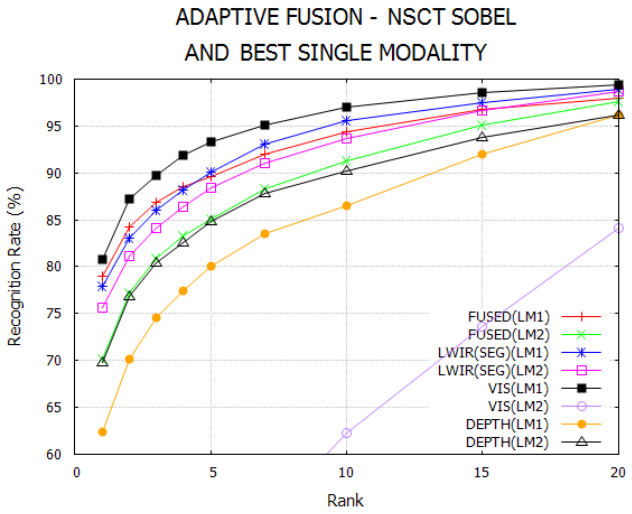

Figure 31: CMC curves for the images adaptively fused in NSCT space using Sobel selection shown as FUSED(LM1) and FUSED(LM2).Average coefficient selection by mean, detail coefficient selection by SOBEL. Best single modality results shown as comparison.

		Recognition Rate % at Rank								
Image Set	Method	1	2	3	4	5	7	10	15	20
Adaptive Fusion - LM1	PCA + Mahcos	55.29	64.01	69.22	72.80	76.08	80.85	85.94	91.82	95.33
Average - MEAN	KPCA + Mahcos	53.07	62.10	67.91	71.91	74.97	79.79	84.66	90.67	94.40
Detail - WSML	KFA + Euc	78.26	83.83	86.79	88.84	90.33	92.76	95.14	97.22	98.92
(SEG)	LDA + Cos	**79.03**	**84.18**	**86.80**	**88.57**	**89.60**	**92.02**	**94.37**	**96.74**	**97.94**
Adaptive Fusion - LM2	PCA + Mahcos	47.44	56.48	61.46	65.39	68.68	74.26	80.48	88.13	93.52
Average - MEAN	KPCA + Mahcos	35.99	46.27	53.47	58.31	62.46	68.91	76.83	86.16	92.49
Detail - WSML	KFA + Euc	63.08	69.90	73.57	76.19	78.70	82.69	86.95	93.60	96.79
(SEG)	LDA + Cos	**70.17**	**77.18**	**80.87**	**83.32**	**85.07**	**88.24**	**91.22**	**95.08**	**97.57**

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Method	EER	FAR = 1	FAR = 0.1	FAR = 0.01
Adaptive Fusion DWT - LM1	PCA + Mahcos	11.29	74.01	57.27	51.43
Average - MEAN	KPCA + Mahcos	9.95	76.83	62.16	51.53
Detail - WSML	KFA + Euc	15.17	68.27	59.53	50.43
(SEG)	LDA + Cos	**9.07**	**79.36**	**67.70**	**57.49**
Adaptive Fusion DWT - LM2	PCA + Mahcos	15.75	65.05	52.45	42.50
Average - MEAN	KPCA + Mahcos	**9.40**	**80.74**	**73.82**	**67.84**
Detail - WSML	KFA + Euc	13.41	69.54	56.00	41.64
(SEG)	LDA + Cos	11.87	75.43	68.06	59.43

Table 9: Recognition and verification results for the adaptively fused images using NSCT and WSML measure

5.4 Image Fusion in Match-Score Space

5.4.1 Semi-Adaptive Match-Score Fusion

For these experiments we fuse the recognition scores from each single modality in match-score space using the method outlined in Section 4.3 on page 68. We use a semi-adaptive fusion method in that the weights applied are fixed but we use a luminance measure of the VIS image to threshold its inclusion in the match-score fusion. Weights are applied to the similarity matrix of each image mode as shown in the flow diagram 20.

In order to demonstrate a considerable increase in the recognition accuracy using this method we have conducted a full set of recognition experiments for every weight combination for the VIS, LWIR and depth images. For three component images each with a similarity matrix weight in the range 0.1-1.0 there are 10^3different combinations to test. With each experiment consisting of 10 recognition experiments with randomised training images, the results presented in this section are taken from a total of 10^4 recognition experiments. The optimum weight combination for both LM1 and LM2 images are identified.

The CMC curves for the best recognition results for the LM1 and LM2 images fused in match-score space are shown in Figure 32 along with the best recognition results for the single modality images for comparison. The ROC curves for the fused images are shown in Figure 33. The top four fusion weight combinations for both the LM1 and LM2 images are shown in table 10 as well as recognition and verification results for the other recognition methods used.

The fusion in match-score space shows a significant improvement over the single modalities in terms of both recognition and verification accuracy. For the LM1 and LM2 fusion the EER is reduced to 2.5% and 3% respectively which, for LM1, is an improvement even on the optimised DWT fusion weights shown in Section 5.2. Observing the fusion weights for the top LM1 and LM2 results we can see that the LM2 score fusion reduces the VIS contribution by half in order to achieve the same recognition accuracy as the LM1 images because of the increase in VIS image degradation. This in addition to the semi-adaptive method of removing any measured "low luminance" VIS images from the final score fusion. Thus we can see that when the noise from the VIS images is sufficiently reduced the fused recognition performance can be considerably

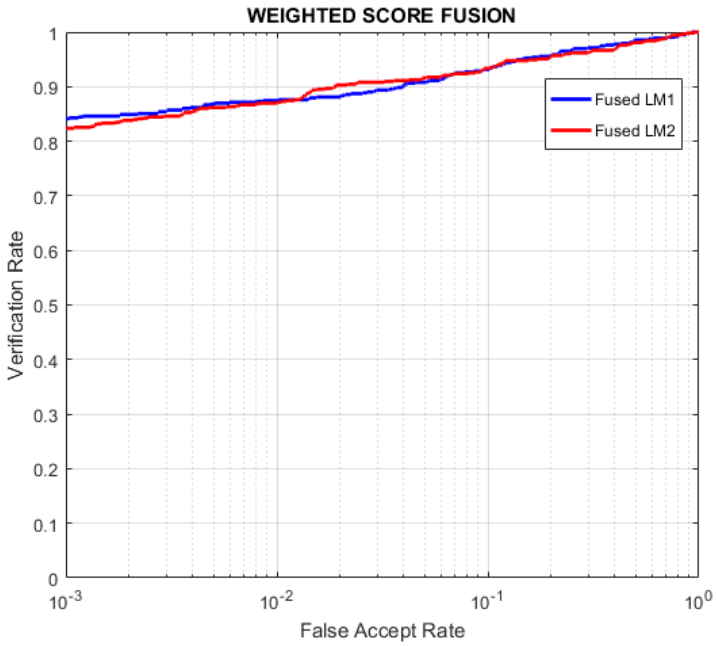

Figure 33: ROC curve for the FUSED(LM1) and FUSED(LM2) verification results.

improved, even under challenging lighting conditions.

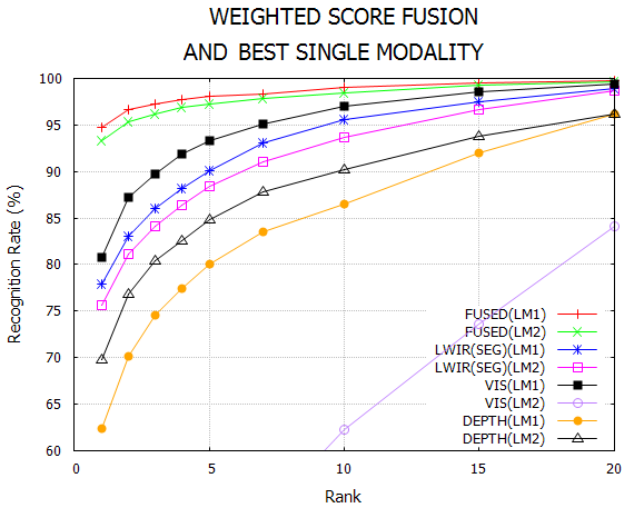

Figure 32: CMC curves showing the best weighted score fusion results shown as FUSED(LM1) and FUSED(LM2). Best single modality results shown as comparison.

		Recognition Rate % at Rank									Fusion Weight		
Image Set	Result Rank	1	2	3	4	5	7	10	15	20	VIS	LWIR	depth
Match-Score Fusion - LM1	1	**94.78**	**96.60**	**97.27**	**97.71**	**98.03**	**98.36**	**98.98**	**99.51**	**99.78**	**0.6**	**0.9**	**1**
LDA+COS	2	94.63	96.53	97.11	97.52	97.92	98.42	98.97	99.46	99.76	0.8	0.9	0.7
(SEG)	3	94.61	96.06	97.01	97.50	98.00	98.32	98.71	99.22	99.59	0.2	0.3	0.2
	4	94.38	95.89	96.87	97.52	98.04	98.49	98.92	99.46	99.79	0.6	0.9	0.5
Match-Score Fusion - LM2	1	**93.28**	**95.34**	**96.19**	**96.84**	**97.24**	**97.87**	**98.47**	**99.24**	**99.63**	**0.3**	**0.8**	**0.9**
LDA+COS	2	90.49	92.87	93.78	94.69	95.30	96.40	97.69	99.13	99.54	0.2	0.9	0.8
(SEG)	3	89.62	91.76	92.82	94.02	94.48	95.56	96.55	97.62	98.60	0.2	0.5	0.5
	4	89.56	92.68	94.13	95.42	96.22	97.19	98.11	98.85	99.34	0.4	1	0.8

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Result Rank	EER	FAR = 1	FAR = 0.1	FAR = 0.01
Match-Score Fusion - LM1	1	**2.53**	**95.47**	**92.03**	**85.74**
LDA+COS	2	2.53	95.17	92.40	89.45
(SEG)	3	2.57	95.17	91.27	86.41
	4	2.67	94.96	91.23	88.04
Match-Score Fusion - LM2	1	**3.70**	**93.11**	**88.03**	**82.29**
LDA+COS	2	4.84	89.94	85.91	79.07
(SEG)	3	5.65	89.81	85.85	81.76
	4	4.66	89.50	83.63	78.71

Table 10: The top 4 recognition and verification results using non-adaptive weighted score fusion

5.4.2 Match-Score Fusion of Fused DCT Features and DWT Fused Images

The results presented in Section 5.4.1 show the effectiveness of a match-score fusion method coupled with a multispectral camera system. However, the method can only be considered semi-adaptive in terms of thresholding the VIS image contribution, as the main fusion weights were manually derived from thousands of possible combinations and were varied with the lighting. In order to improve on this we propose a fully adaptive fusion method using the input from feature fusion and transform fusion with the resulting similarity matrices being fused in match-score space as described in this section.

The results in Section 5.3.2 show that of the adaptive selection methods used, the fusion of the DWT coefficients using the WSML selection method produced the best results in terms of consistency across the two lighting modes LM1 and LM2. We therefore use these fused images as inputs for our match-score fusion.

We simultaneously take the VIS, LWIR and depth component images and extract their LBP images [58]. We then apply the DCT transform to the LBP component images and extract the discriminant features from each modality image and concatenate them into a single feature vector as described in Section 4.2 on page 64.

The recognition method is then applied separately to the the fused DWT-WSML images and

the fused DCT features, producing two similarity matrices. The similarity matrices are then fused by taking the mean between the two and the final classification for each image is made using the fused similarity matrix. A flow diagram of this method is shown in Figure 34.

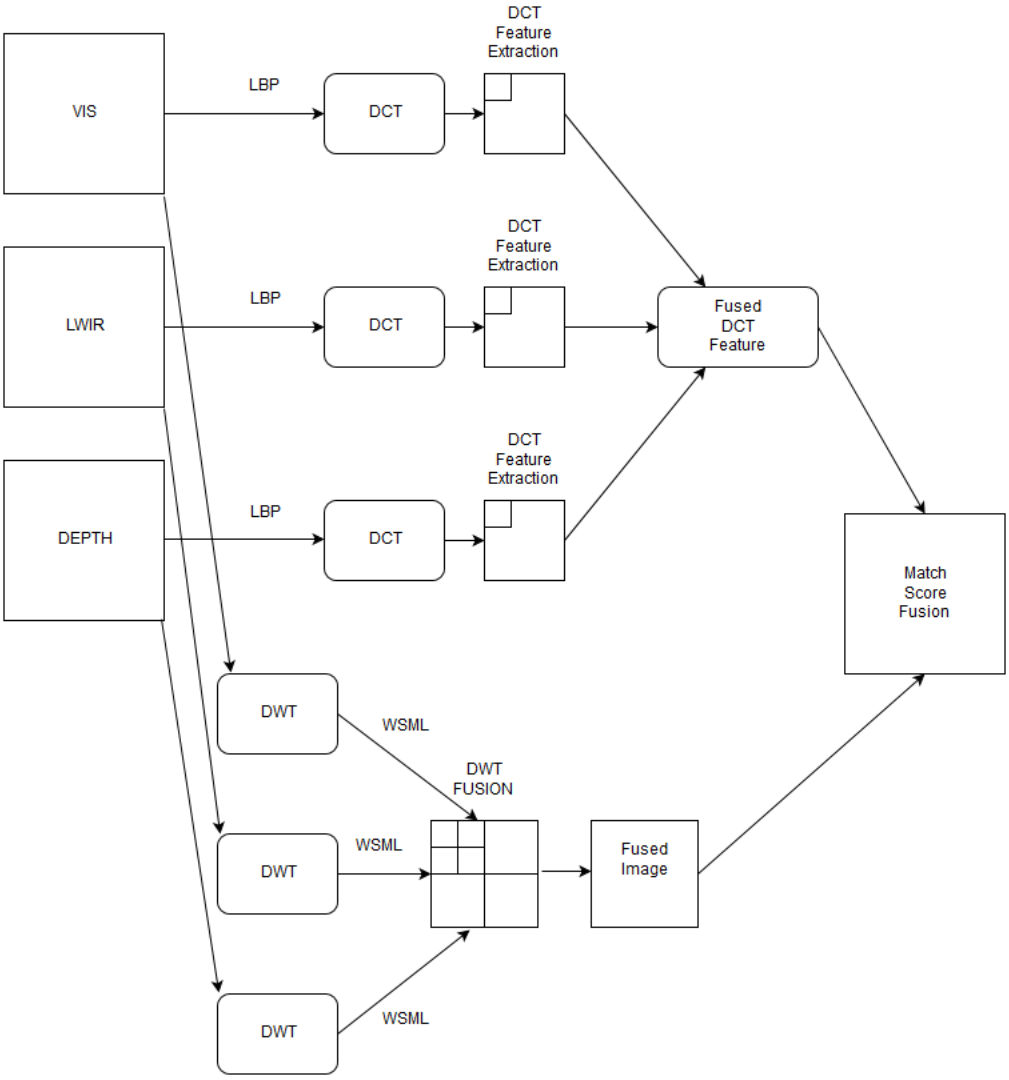

Figure 34: Flow diagram for the DCT+DWT-WSML features and match score fusion

The CMC curves for the match-score fused DCT fused features and DWT-WSML fused images, as well as the best single modality results are shown in Figure 35. The ROC curves for the fused images are shown in Figure 36 The full recognition and verification scores for the match-score fusion are given in Table 11. It is clearly shown that this method of fusion outperforms all of the single modality images and maintains a >90% rank 1 recognition rate

even under the LM2 lighting modes. The rank 1 recognition rate for the LM1 images using this method is 93.5%, which is 1.2% less than the optimised, manually weighted score fusion. The EER is reduced to 2.6% for the LM1 images and rises to 4% under LM2 conditions which is better than the EER reported using the highly optimised, weighted score fusion in Section 5.4.1.

As shown in Section 5.3.2, the WSML selection method was effective in selecting coefficients from the component images for face recognition while reducing the noise from the VIS image during fusion. This gave a more consistent recognition accuracy between then two lighting modes. By further fusing the match-scores of the fused images with the DCT features we have improved the recognition accuracy further still while maintaining this consistency across lighting modes.

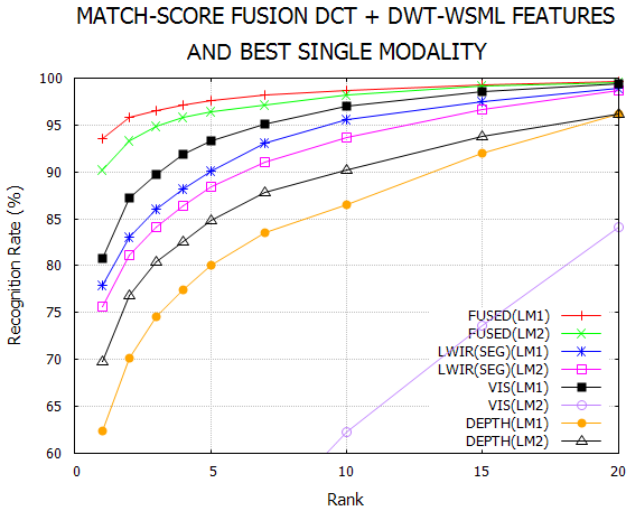

Figure 35: CMC curves for the adaptive match-score and feature fusion results shown as FUSED(LM1) and FUSED(LM2). Best single modality results shown as comparison.

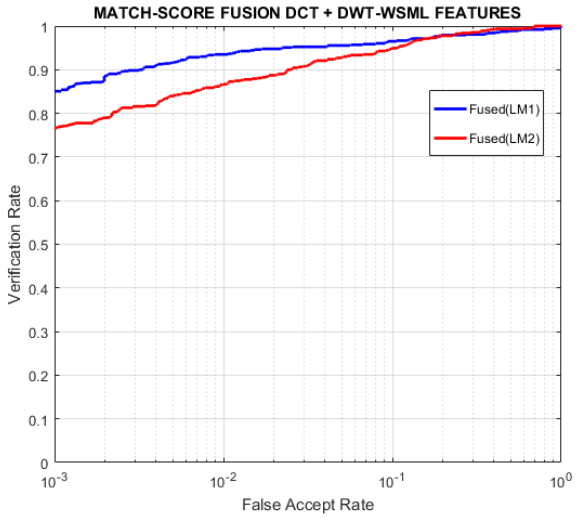

Figure 36: ROC curve for the FUSED(LM1) and FUSED(LM2) verification results.

		Recognition Rate % at Rank								
Image Set	Method	1	2	3	4	5	7	10	15	20
Adaptive Fusion - LM1	PCA + Mahcos	78.92	84.67	87.74	89.77	91.19	93.09	95.37	97.51	98.94
Average - MEAN	KPCA + Mahcos	52.08	62.31	69.75	75.43	79.21	83.72	87.83	92.88	96.44
Detail - WSML	KFA + Euc	84.43	88.00	90.29	91.50	92.62	94.28	95.89	97.92	98.85
(SEG)	LDA + Cos	**93.57**	**95.81**	**96.54**	**97.12**	**97.64**	**98.15**	**98.71**	**99.27**	**99.59**
Adaptive Fusion - LM2	PCA + Mahcos	64.76	72.08	76.49	79.26	81.51	84.99	89.00	93.22	96.50
Average - MEAN	KPCA + Mahcos	43.08	52.98	58.40	62.41	65.45	70.31	75.90	84.59	91.86
Detail - WSML	KFA + Euc	76.14	81.03	83.83	86.11	88.04	90.77	93.33	95.98	97.71
(SEG)	LDA + Cos	**90.17**	**93.33**	**94.79**	**95.82**	**96.42**	**97.12**	**98.17**	**99.10**	**99.49**

		Equal Error Rate %	Verification Rate at False Acceptance Rate % (FAR)		
Image Set	Method	EER	FAR = 1	FAR = 0.1	FAR = 0.01
Adaptive Fusion DWT - LM1	PCA + Mahcos	3.67	93.46	87.75	82.65
Average - MEAN	KPCA + Mahcos	9.51	77.19	70.00	62.78
Detail - WSML	KFA + Euc	6.50	84.60	77.88	72.49
(SEG)	LDA + Cos	**2.61**	**95.11**	**90.00**	**85.25**
Adaptive Fusion DWT - LM2	PCA + Mahcos	7.77	83.47	74.15	66.36
Average - MEAN	KPCA + Mahcos	17.18	63.50	51.56	42.35
Detail - WSML	KFA + Euc	10.36	74.63	63.96	51.21
(SEG)	LDA + Cos	**4.05**	**91.81**	**85.44**	**79.41**

Table 11: Recognition and verification results for match score fusion of fused DCT features and fused DWT-WSML images

5.5 Image Fusion In Feature Space

In order to demonstrate the effectiveness of the DCT feature extraction from the LBP component images fused with the DWT-WSML images used in Section 5.4.2 we have conducted experiments to fuse these two images in feature space. The fused features are then used to train a multi-class linear SVM classifier which is then used to classify the remaining test images. The results of these experiments are reported in this Section.

Our method of feature extraction and fusion is similar to the flow diagram shown in Figure 34 on page 94 with the exception that we extract the LBP image from the fused DWT-WSML image. The LBP image is then vectorised and concatenated with the fused DCT feature before being normalised using z-score normalisation. The z-score normalisation value z for an element x is given by:

$$z = \frac{(x - \mu)}{\sigma}$$

where μ and σ are the mean and standard deviation of the elements of the feature vector.

The modified flow diagram for this process is shown in Figure 37.

For our experiments here we have trained a "one versus all", multi-class linear SVM classifier for each set of images tested. Each SVM was trained on three images for that image set: front light front pose, front light left side pose and front light right side pose and uses a $k - fold$ cross validation check where $k = 5$. Thus for every classification made by the SVM the probe image has variations in lighting in the training images. The confusion matrices for the SVM classifiers are shown in Figure 38 and a table comparing the recognition rates for the fused DCT+DWT-WSML features as well as the single mode DCT features is given in Table 12.

These results show that the fused DCT+DWT-WSML feature vector has a better recognition accuracy compared to the LBP-DCT feature vectors of the single modality images, achieving a maximum recognition accuracy of 98% compared to the next highest score of 95% for the LM1 VIS images. The fused feature vector is also consistent across both LM1 and LM2 lighting

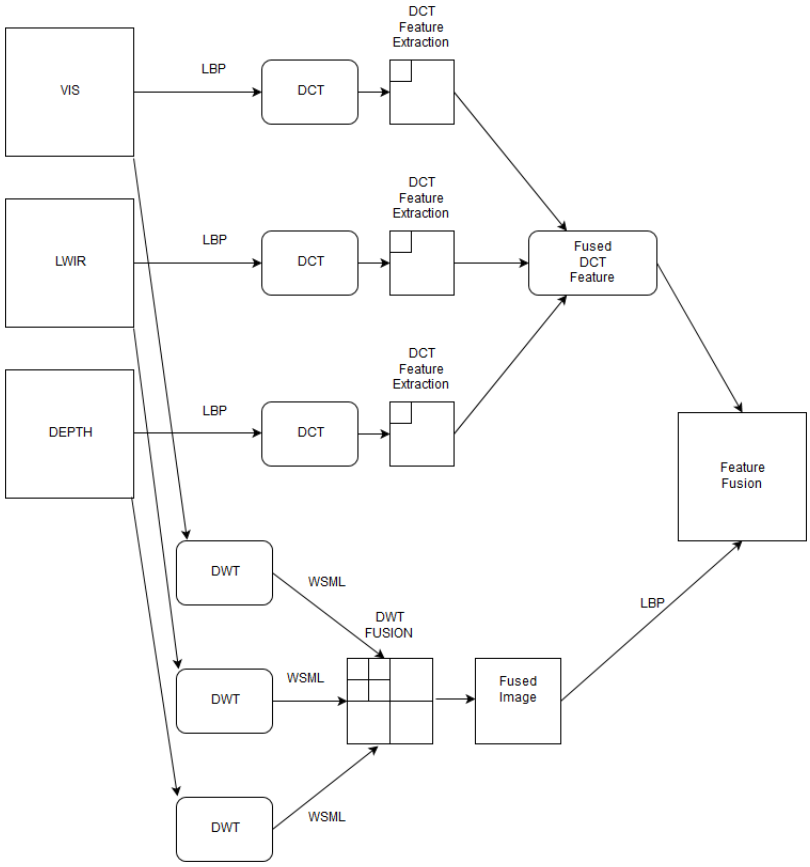

Figure 37: Flow diagram for the DCT+DWT-WSML feature fusion

modes while the VIS LM2 results drop to a 45% recognition accuracy, showing the same sensitivity to lighting variations as in the experiments above. The LWIR images show less variation across the lighting variations with a 6% drop in recognition accuracy between LM1 and LM2. Interestingly the LBP-DCT features extracted from the depth images produce their best single-modality recognition accuracy compared to the previous experiments, demonstrating that the discriminatory features identified in Section 4.2 are an effective contribution to the final, fused feature vector.

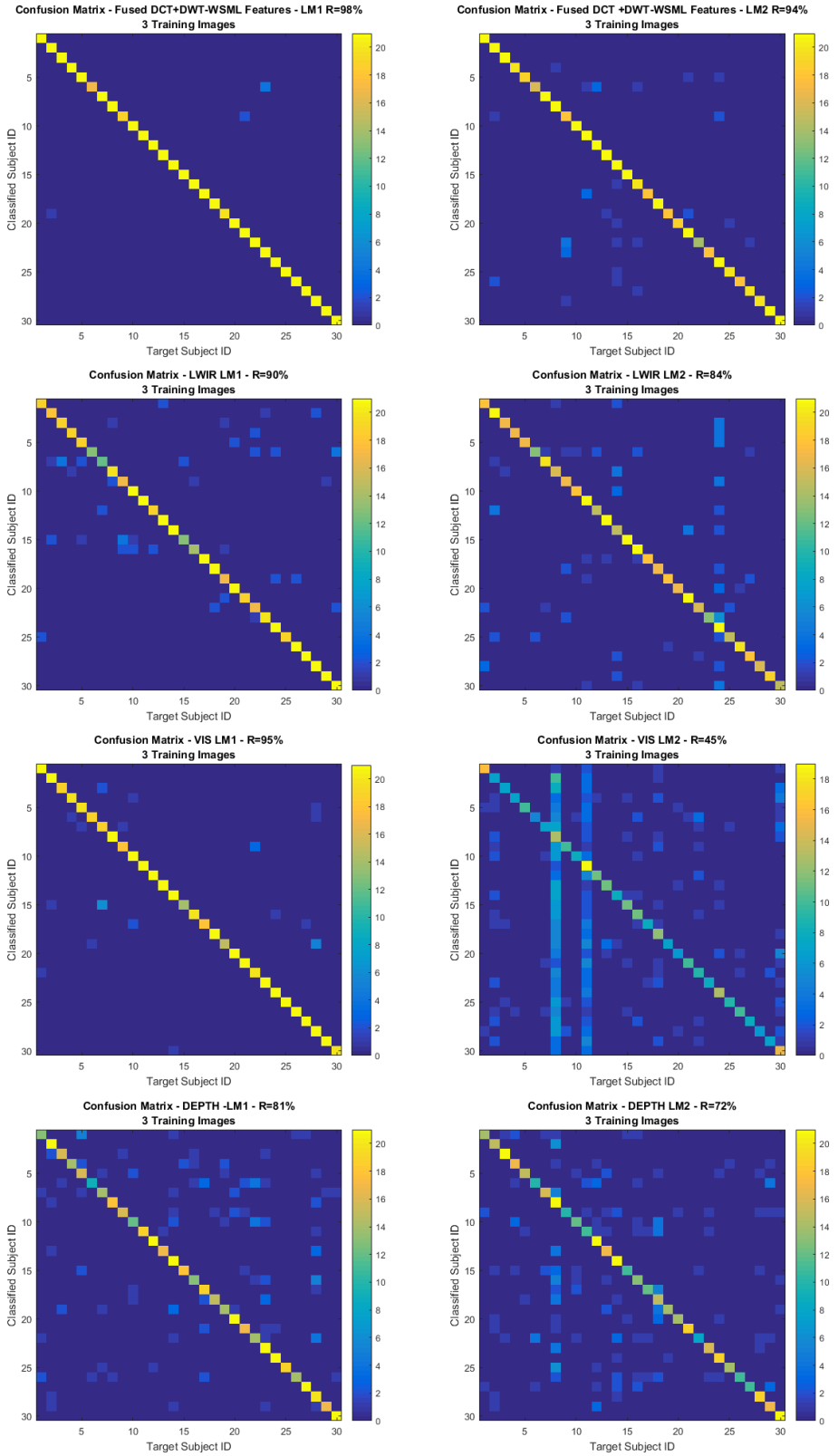

Figure 38: Colour confusion matrices for the fused DCT+DWT-WSML feature recognition experiments and the single mode LBP-DCT features

Image Set	SVM Recognition Accuracy (%)
FUSED DCT+DWT-WSML Feature (LM1)	98
FUSED DCT+DWT-WSML Feature (LM2)	94
VIS DCT Feature (LM1)	95
VIS DCT Feature (LM2)	45
LWIR DCT Feature (LM1)	90
LWIR DCT Feature (LM2)	84
depth DCT Feature (LM1)	81
depth DCT Feature (LM2)	72

Table 12: Recognition results for the fused DCT+DWT-WSML features using SVM classifier

6 Conclusions

In the review of the research literature in Section 2 it was observed that multispectral image fusion can produce images that have improved face recognition accuracy under varying lighting conditions. The most common fusion method has been the weighted fusion of transform coefficients between the component images to produce a single, fused image. The initial set of recognition experiments using the single modality images alone (Section 5.1) demonstrated how sensitive the VIS image modality is to changes in lighting and how comparatively insensitive the LWIR and depth images are. The LWIR images with synthesised eyeglasses show an improvement in recognition accuracy in the majority of our experiments. For the LDA+COS recognition method the results show an increase of 11% and 9.8% in recognition accuracy for the LM1 and LM2 images respectively. However, for the LM1 images a substantial reduction in verification performance was seen as the EER increased from 4.58% to 10.82% and the verification rate was similiarly reduced. The synthesised eyeglasses method appears to work but is unstable in terms of recognition and verification performance, which could be due the VIS image sensitivity to changes in lighting or reflections off of the lenses. During testing our method of detection and segmentation of the eyeglasses (discussed in Section 3.4) proved reliable and worked in real time.

Our experiments on non-adaptive fusion in Section 5.2 demonstrated that a fixed-weight approach to image fusion leads to an optimisation of fusion weights for that particular set of images rather than a universally optimised set of weights. Indeed, the optimised set of weights for the LM1 image set produced a 92% rank 1 recognition rate which was reduced to a 64% rank 1 recognition rate when the same weights were applied to the LM2 image set which features harsher, more directional lighting. When we consider the deterioration of the VIS image recognition accuracy measured in Section 5.1 we can see that in fused images using poorly lit VIS images, any inclusion of the VIS coefficients effectively injects noise into the image fusion, especially for the low light and side light images. Thus the fixed weights derived for the LM1 image set were too heavily weighted for the VIS coefficients when applied to the LM2 images, resulting in the measured drop in recognition accuracy.

From the coefficient selection methods tested in Section 5.3 it is clear that the WSML method produces the most consistent results across the two lighting modes, specifically when applied to

the DWT coefficients. It is interesting to note that the NSCT did not perform as well using the same method or indeed overall, compared to the DWT. Indeed we have seen the NSCT used for human and medical image fusion purposes [26, 80] as well as feature extraction for face recognition [19, 90] but we are unaware of it having been used for multispectral or multimodal image fusion for face recognition.

As discussed in Section 2.2, match-score fusion of separate modality images has been suggested as an alternative or enhancement to a single image fusion for face recognition[12, 25, 69]. From the results of our match-score experiments in Section 5.4 we conclude that for our VIS, LWIR and depth images, match-score fusion produces a considerable increase in recognition performance. The match score fusion of the single modalities using fixed, optimised weights shows an 18% increase in recognition accuracy over the best VIS image score for LM1 and achieving a rank 1 recognition rate of 93% under LM2 conditions with a verification rate of 93% at a FAR of 1% and an EER of 3.7%. However it is worth noting that while the VIS component matches were thresholded, the fusion weights were manually optimised from thousands of combinations and, as mentioned above, fixed weight fusion routines are not necessarily universal as is shown by the difference in weights for LM1 and LM2 image sets.

The match-score fusion of the DWT-WSML fused images and the fused LBP-DCT features shows the effectiveness of a applying match-score fusion to two separate image features. Without using fusion weights this fusion method produces a rank 1 recognition rate of 93% under LM1 and 90% under LM2, a drop of 3.2% which shows a similar consistency between lighting modes as the DWT-WSML fused images which only varied by 2.3%. The EER is slightly increased to 2.6% and 4% compared to the fixed weight fusion experiments, although this is a better EER than the manually optimised DWT fusion experiments. The verification rate of 90% and 85% at FAR=0.1% for the LM1 and LM2 image sets respectively shows that the adaptive fusion method is capable of producing accurate recognition and verification even under extreme changes in lighting. We conclude that for the recognition methods tested in these experiments, the match-score fusion of a transform fused image and fused feature vector presents the best solution to multispectral+2.5D image fusion for face recognition. The results in Section 5.5 using a multi-class SVM classifier further show that the combination of fusion of the LBP-DCT

features with the LBP-DWT-WSML image in feature space produces an increase in recognition accuracy and a decreased sensitivity to lighting, pose and the presence of eyeglasses. Our feature fusion method results in a recognition accuracy of 98% and 94% across both LM1 and LM2 image sets compared to 95%-45% for the VIS only images and 90%-84% for the LWIR images.

In the literature review in Section 2 it was shown that the majority of image fusion for face recognition research has used VIS and a complementary image in either the NIR or LWIR wavebands. One consideration that has not been explored within these results is whether a fusion of the VIS and LWIR images alone would perform as well as our VIS, LWIR and depth fusion algorithm. In [97] a review of the current state of single vs multimodal face recognition shows that the trend in the existing research is the greater number of modalities used in the recognition, the higher the recognition accuracy. This is also suggested in [39] which is also reported in [97] where fusion of VIS, IR and 3D data outperforms any combination of the three modes.

From the results using our novel camera system we can conclude that our design for a multispectral +2.5D/3D camera system does indeed work and is capable of capturing co-registered images in several spectral modes and a 2.5D depth image in still as well as video format. We are unaware of any other similar camera system for face recognition. Our results show the advantages of the multispectral and multimodal image formats for match-score fusion between the separate modalities and separate feature streams in order to improve recognition and decrease sensitivity to pose and eyeglasses.

In conclusion the research presented here is in agreement with the predictions made by the current research literature: the future of face recognition is very much tied to multispectral and multimodal camera systems [3, 97].

7 Further Work

As discussed in Section 3.1 the inability to turn off the Kinect laser projector without also turning off the NIR camera as well as the narrow-band response of the NIR camera prevents us from collecting the NIR images concurrently with the other image streams. An obvious improvement that could be made is to use a more modern depth sensor such as the 'Kinect One' sensor and use a separate, broad-band NIR camera for the NIR image stream. This would provide higher quality NIR images that could be captured fully in sync with the VIS, LWIR and depth images. This would be particularly interesting as it is suggested that the inclusion of additional image modalities for fusion would improve the recognition performance [38, 67, 97]. This suggestion seems intuitive but it should be tested experimentally to quantify what kind of improvements, if any, can be made.

Further work on the expansion of the face image database captured with the camera system is also important in order to establish the effect of an increased number of subjects on the recognition results. It has been found, particularly with 3D face recognition where initial databases consisted of 6-10 subjects, that an increase in the size of the dataset produces a decrease in recognition accuracy. While our databases are of comparable size to those reported in the research literature, it is important to establish if our results are reproducible for an even larger number of subjects. Capturing face images under outside and therefore under less constrained conditions would also be of great interest, as would capturing face images of moving subjects. This would test the limits of the camera system, including the range of the depth sensor and the ability of the LWIR sensor, which has the slowest refresh rate, to image a moving subject.

While computationally expensive, the use of 3D point cloud data for correcting facial pose for face recognition is an increasingly popular research area. Future experiments are planned in which the multispectral camera system will be used to capture co-registered 3D point cloud data and texture map the 2D images before correcting the pose alignment.

8 Summary

We have presented our design and development of a novel multispectral+2.5D/3D camera system that can capture three separate spectral images along with depth data via a common optical path in video or as still frames. To our knowledge no similar camera system has been developed for face recognition. Using our camera system we have collected two sets of face images with each set containing 30 subjects under varied lighting, pose and with and without eyeglasses. Two lighting modes were designed. The first produces a varied illumination of the face while the second is an extremely challenging, highly directional lighting mode with the low-light images being taken in near darkness.

Our experiments have looked at image fusion at varying levels of abstraction; namely the transform, feature and match-score space, with adaptive techniques designed to automatically optimise the fusion applied at each level. A method of synthesising the occluded eyeglass patches in the LWIR images was also developed and was found to provide a small improvement in recognition accuracy using the LWIR images. The fused images have been tested using an array of well established recognition algorithms as well as multi-class SVM classifier techniques.

A method of automatic detection, segmentation and synthesis of eyeglass patches in LWIR images was designed and implemented. Experimental results showed that our method of detection and segmentation of the eyeglass lenses was reliable and worked in real time. However, improvements to recognition accuracy were measured for both lighting scenarios, but proved unstable across all of the recognition and verification results.

Our research has demonstrated that attempts to optimise the fusion of a set of multispectral images into a single image, which are widely reported in the research literature, can lead to over-fitting to a particular image set. The same recognition accuracy is not reproducible under new lighting conditions. This is due to the inability of a single fused image to fully exclude noise in the fused coefficients when the component images become degraded due to poor lighting, pose variation or noise. This is particularly relevant for the VIS image which is highly sensitive to lighting variations. The experiments presented here have shown that one effective approach for accurate and consistent face recognition with a multimodal camera system is to use an

semi-adaptive match-score fusion of the single mode images whereby the VIS image can be effectively removed from the match score when poor lighting is detected. The LWIR and depth image streams are much less sensitive to changes in lighting. The results showed that the semi-adaptive approach with optimised, fixed weights was capable of producing a high rate of recognition accuracy across both lighting modes.

In order to develop a fully adaptive image fusion method we then looked at the match-score fusion of images fused in transform space with the extracted DCT features of the single modality images. Our results suggest that for the recognition methods in our experiments, a fused image using the DWT-WSML selection method which is then fused in match-score space with the DCT features extracted from the LBP images of the single modalities produces the best results for recognition and for verification across the variations in lighting used in both of our our image sets.

Finally we demonstrated the effectiveness of fusing the LBP-DCT features with the DWT-WSML LBP features by applying a multi-class SVM classifier. The results showed that the fused feature vector outperforms the single modality images across all lighting modes used in our experiments, achieving a 98% recognition accuracy under the most challenging lighting condition used.

Publication

The research and results presented in this thesis were submitted as a research manuscript entitled 'A Novel Multispectral and 2.5D Image Fusion Camera System for Enhanced Face Recognition' to the journal 'Information Fusion' on 16/04/17. The article was under review at the time of the final submission of this thesis.

References

[1] Naseer Al-Jawad and Sabah Jassim. Wavelet Based Image Quality Self Measurements. *Proc. SPIE*, 7708:77080J–77080J–12, 2010.

[2] Boulbaba Ben Amor, Karima Ouji, Mohsen Ardabilian, and Liming Chen. 3d Face Recognition by ICP-based Shape Matching. *LIRIS Lab, Lyon Research Center for Images and Intelligent Information Systems, UMR*, 5205, 2005.

[3] Shwetank Arya, Neeraj Pratap, and Karamjit Bhatia. The Future of Face Recognition: A Review. *Procedia Computer Science*, 58:578–585, 2015.

[4] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[5] J. Bernhard, J. Barr, K. W. Bowyer, and P. Flynn. Near-ir to visible light face matching: Effectiveness of pre-processing options for commercial matchers. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, Sept 2015.

[6] Mrinal Kanti Bhowmik, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu, and Mahantapas Kundu. Image Pixel Fusion For Human Face Recognition. *arXiv preprint arXiv:1007.0628*, 2010.

[7] Elhocine Boutellaa, Messaoud Bengherabi, Samy Ait-Aoudia, and Abdenour Hadid. How Much Information Kinect Facial Depth Data Can Reveal About Identity, Gender and Ethnicity? In *European Conference on Computer Vision*, pages 725–736. Springer, 2014.

[8] Kevin W Bowyer, Kyong Chang, and Patrick Flynn. A Survey of Approaches and Challenges in 3D and Multi-modal 3D+ 2D Face Recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.

[9] Kevin W Bowyer, Kyong I Chang, Patrick J Flynn, and Xin Chen. Face Recognition using 2-d, 3-d, and Infrared: Is Multimodal Better Than Multisample? *Proceedings of the IEEE*, 94(11):2000–2012, 2006.

[10] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Inc., 2008.

[11] Peter J Burt and Edward H Adelson. Merging Images Through Pattern Decomposition. In *29th Annual Technical Symposium*, pages 173–181. International Society for Optics and Photonics, 1985.

[12] Pierre Buyssens and Marinette Revenu. Fusion Levels of Visible and Infrared Modalities for Face Recognition. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–6. IEEE, 2010.

[13] Hong Chang, H Harishwaran, Mingzhong Yi, A Koschan, B Abidi, and M Abidi. An indoor and outdoor, multimodal, multispectral and multi-illuminant database for face recognition. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 54–54. IEEE, 2006.

[14] Hong Chang, Andreas Koschan, B Abidi, and M Abidi. Fusing Continuous Spectral Images for Face Recognition Under Indoor and Outdoor Illuminants. *Machine Vision and Applications*, 21(2):201–215, 2010.

[15] Hong Chang, Andreas Koschan, Besma Abidi, and Mongi Abidi. Physics-based Fusion of Multispectral Data for Improved Face Recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1083–1086. IEEE, 2006.

[16] Kyong I Chang, Kevin W Bowyer, and Patrick J Flynn. An Evaluation of Multimodal 2D+3D Face Biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):619–624, 2005.

[17] Jie Chen, Dong Yi, Jimei Yang, Guoying Zhao, Stan Z Li, and Matti Pietikainen. Learning Mappings for Face Synthesis from Near Infrared to Visual Light Images. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 156–163. IEEE, 2009.

[18] Xuerong Chen, Zhongliang Jing, and Zhenhua Li. Image Fusion for Face Recognition. In *2005 7th International Conference on Information Fusion*, volume 2, pages 5–pp. IEEE, 2005.

[19] Yong Cheng, Yingkun Hou, Chunxia Zhao, Zuoyong Li, Yong Hu, and Cailing Wang. Robust Face Recognition Based On Illumination Invariant in Nonsubsampled Contourlet Transform Domain. *Neurocomputing*, 73(10):2217–2224, 2010.

[20] A. L. Da Cunha, J. Zhou, and M. N. Do. The nonsubsampled contourlet transform: Theory, design, and applications. *IEEE Transactions on Image Processing*, 15(10):3089–3101, Oct 2006.

[21] Minh N Do and Martin Vetterli. Contourlets: A Directional Multiresolution Image Representation. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–357. IEEE, 2002.

[22] Mingsong Dou, Chao Zhang, Pengwei Hao, and Jun Li. Converting Thermal Infrared Face Images Into Normal Gray-level Images. In *Asian Conference on Computer Vision*, pages 722–732. Springer, 2007.

[23] David J Dwyer, Moira I Smith, Jason L Dale, and Jamie P Heather. Real-time Implementation of Image Alignment and Fusion. In *European Symposium on Optics and Photonics for Defence and Security*, pages 85–93. International Society for Optics and Photonics, 2004.

[24] Ahmet M Eskicioglu and Paul S Fisher. Image Quality Measures and their Performance. *IEEE Transactions on communications*, 43(12):2959–2965, 1995.

[25] Virginia Espinosa-Duró, Marcos Faundez-Zanuy, and Jiří Mekyska. A New Face Database Simultaneously Acquired in Visible, Near-infrared and Thermal Spectrums. *Cognitive Computation*, 5(1):119–135, 2013.

[26] Padma Ganasala and Vinod Kumar. CT and MR Image Fusion Scheme in Nonsubsampled Contourlet Transform Domain. *Journal of Digital Imaging*, 27(3):407–418, 2014.

[27] Reza Shoja Ghiass, Ognjen Arandjelović, Hakim Bendada, and Xavier Maldague. Infrared Face Recognition: A Literature Review. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–10. IEEE, 2013.

[28] Shaogang Gong, Stephen J McKenna, and Alexandra Psarrou. *Dynamic Vision: From Images to Face Recognition*. Imperial College Press, 2000.

[29] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Pearson Education, 2009.

[30] A Ardeshir Goshtasby and Stavri Nikolov. Image Fusion: Advances in the State of the Art. *Information Fusion*, 8(2):114–118, 2007.

[31] Mohammad Hanif and Usman Ali. Optimized Visual and Thermal Image Fusion for Efficient Face Recognition. In *2006 9th International Conference on Information Fusion*, pages 1–6. IEEE, 2006.

[32] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, Nov 1973.

[33] JP Heather, MI Smith, J Sadler, and D Hickman. Issues and Challenges in the Development of a Commercialised Image Fusion System. In *SPIE Defense, Security, and Sensing*, pages 77010A–77010A. International Society for Optics and Photonics, 2010.

[34] Gabriel Hermosilla, Javier Ruiz-del Solar, Rodrigo Verschae, and Mauricio Correa. A Comparative Study of Thermal Face Recognition Methods in Unconstrained Environments. *Pattern Recognition*, 45(7):2445–2459, 2012.

[35] Ming-Kuei Hu. Visual Pattern Recognition by Moment Invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

[36] Tri Huynh, Rui Min, and Jean-Luc Dugelay. An Efficient LBP-based Descriptor for Facial Depth Images Applied to Gender Recognition using RGB-D Face Data. In *Asian Conference on Computer Vision*, pages 133–145. Springer, 2012.

[37] Anil K Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc., 1989.

[38] Alex Pappachen James, Dhiya Al-Jumeily, Sabu M. Thampi, Shwetank Arya, Neeraj Pratap, and Karamjit Bhatia. The Future of Face Recognition: A Review. *Procedia Computer Science*, 58:578 – 585, 2015.

[39] Ioannis A Kakadiaris, Georgios Passalis, Theoharis Theoharis, George Toderici, Ioannis Konstantinidis, and N Murtuza. Multimodal Face Recognition: Combination of Geometry with Physiological Information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1022–1029. IEEE, 2005.

[40] Xin Chen Patrick J Flynn Kevin and W Bowyer. Visible-light and Infrared Face Recognition. In *Workshop on Multimodal User Authentication*, page 48. Citeseer, 2003.

[41] Joongrock Kim, Sunjin Yu, Ig-Jae Kim, and Sangyoun Lee. 3D Multi-Spectrum Sensor System with Face Recognition. *Sensors*, 13(10):12804–12829, 2013.

[42] Nick Kingsbury. Complex Wavelets for Shift Invariant Analysis and Filtering of Signals. *Applied and computational harmonic analysis*, 10(3):234–253, 2001.

[43] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, mar 1998.

[44] Seong G Kong, Jingu Heo, Faysal Boughorbel, Yue Zheng, Besma R Abidi, Andreas Koschan, Mingzhong Yi, and Mongi A Abidi. Multiscale Fusion of Visible and Thermal IR Images for Illumination-invariant Face Recognition. *International Journal of Computer Vision*, 71(2):215–233, 2007.

[45] John J Lewis, Robert J O'Callaghan, Stavri G Nikolov, David R Bull, and Nishan Canagarajah. Pixel and Region-based Image Fusion with Complex Wavelets. *Information Fusion*, 8(2):119–130, 2007.

[46] Billy YL Li, Ajmal S Mian, Wanquan Liu, and Aneesh Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 186–192. IEEE, 2013.

[47] Hui Li, BS Manjunath, and Sanjit K Mitra. Multisensor Image Fusion Using the Wavelet Transform. *Graphical models and image processing*, 57(3):235–245, 1995.

[48] Chengjun Liu. Capitalize on Dimensionality Increasing Techniques for Improving Face Recognition Grand Challenge Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):725–737, 2006.

[49] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017.

[50] Jing Luo, Shuze Geng, Zhaoxia Xiao, and Chunbo Xiu. A Review of Recent Advances in 3D Face Recognition. In *Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*, pages 944303–944303. International Society for Optics and Photonics, 2015.

[51] Roland Miezianko. Ieee otcbvs ws series bench. *Terravic research infrared database*, 2, 2006.

[52] S Naveen and RS Moni. A Robust Novel Method for Face Recognition from 2D Depth Images Using DWT and DFT Score Fusion. In *Computational Systems and Communications (ICCSC), 2014 First International Conference on*, pages 1–6. IEEE, 2014.

[53] S Naveen and RS Moni. Contourlet and Fourier Transform Features Based 3D Face Recognition System. In *Intelligent Systems Technologies and Applications*, pages 411–425. Springer, 2016.

[54] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[55] Chuong T Nguyen and Joseph P Havlicek. Linear Adaptive Infrared Image Fusion. In *Image Analysis and Interpretation (SSIAI), 2014 IEEE Southwest Symposium on*, pages 117–120. IEEE, 2014.

[56] Stavri Nikolov, Paul Hill, David Bull, and Nishan Canagarajah. Wavelets for Image Fusion. In *Wavelets in signal and image analysis*, pages 213–241. Springer, 2001.

[57] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, jul 2002.

[58] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[59] Gang Pan, Shi Han, Zhaohui Wu, and Yueming Wang. 3D Face Recognition Using Mapped Depth Images. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 175–175. IEEE, 2005.

[60] See-May Phoong, Chai W Kim, PP Vaidyanathan, and Rashid Ansari. A New Class of Two-Channel Biorthogonal Filter Banks and Wavelet Bases. *IEEE Transactions on Signal Processing*, 43(3):649–665, 1995.

[61] Gemma Piella. A General Framework for Multiresolution Image Fusion: From Pixels to Regions. *Information fusion*, 4(4):259–280, 2003.

[62] Sam T Roweis and Lawrence K Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[63] Khalid Sayood et al. Statistical Evaluation of Image Quality Measures. *Journal of Electronic imaging*, 11(2):206–223, 2002.

[64] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural computation*, 10(5):1299–1319, 1998.

[65] Olle Seger. Generalized and Separable Sobel Operators. *Machine Vision for Three-Dimensional Scenes*, page 347, 2012.

[66] Ivan W Selesnick, Richard G Baraniuk, and Nick C Kingsbury. The Dual-tree Complex Wavelet Transform. *IEEE Signal Processing Magazine*, 22(6):123–151, 2005.

[67] Harin Sellahewa and Sabah A Jassim. Illumination and Expression Invariant Face Recognition: Toward Sample Quality-based Adaptive Fusion. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–6. IEEE, 2008.

[68] Harin Sellahewa and Sabah A Jassim. Image-quality-based Adaptive Face Recognition. *IEEE Transactions on Instrumentation and measurement*, 59(4):805–813, 2010.

[69] Richa Singh, Mayank Vatsa, and Afzel Noore. Hierarchical fusion of multi-spectral face images for improved recognition performance. *Information Fusion*, 9(2):200–210, 2008.

[70] Richa Singh, Mayank Vatsa, and Afzel Noore. Integrated Multilevel Image Fusion and Match Score Fusion of Visible and Infrared Face Images for Robust Face Recognition. *Pattern Recognition*, 41(3):880–893, 2008.

[71] Saurabh Singh, Aglika Gyaourova, George Bebis, and Ioannis Pavlidis. Infrared and Visible Image Fusion for Face Recognition. In *Proceedings of SPIE*, volume 5404, pages 585–596, 2004.

[72] Moira I Smith, Adrian N Ball, and David Hooper. Real-time Image Fusion: A Vision Aid for Helicopter Pilotage. In *AeroSense 2002*, pages 30–41. International Society for Optics and Photonics, 2002.

[73] Moira I Smith and Jamie P Heather. A Review of Image Fusion Technology in 2005. In *Defense and Security*, pages 29–45. International Society for Optics and Photonics, 2005.

[74] Diego A Socolinsky and Andrea Selinger. A Comparative Analysis of Face Recognition Performance ith Visible and Thermal Infrared Imagery. Technical report, DTIC Document, 2002.

[75] Diego A Socolinsky, Lawrence B Wolff, Joshua D Neuheisel, and Christopher K Eveland. Illumination Invariant Face Recognition Using Thermal Infrared Imagery. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–527. IEEE, 2001.

[76] Chris Solomon and Toby Breckon. *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.

[77] Vitomir Štruc and Nikola Pavešić. Gabor-based Kernel Partial-Least-Squares Discrimination Features for Face Recognition. *Informatica*, 20(1):115–138, 2009.

[78] Vitomir Štruc and Nikola Pavešić. The Complete Gabor-Fisher Classifier for Robust Face Recognition. *EURASIP Journal on Advances in Signal Processing*, 2010(1):1–26, 2010.

[79] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[80] Lei Tang and Zong-gui Zhao. The Wavelet-based Contourlet Transform for Image Fusion. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on*, volume 2, pages 59–64. IEEE, 2007.

[81] M. Tistarelli, M. Cadoni, A. Lagorio, and E. Grosso. *Blending 2D and 3D Face Recognition*, pages 305–331. Springer International Publishing, 2016.

[82] Alexander Toet, Maarten A Hogervorst, Stavri G Nikolov, John J Lewis, Timothy D Dixon, David R Bull, and Cedric Nishan Canagarajah. Towards Cognitive Image Fusion. *Information Fusion*, 11(2):95–113, 2010.

[83] Alexander Toet, Maarten A Hogervorst, Rob van Son, and Judith Dijk. Augmenting Full Colour-fused Multi-band Night Vision Imagery with Synthetic Imagery in Real-time. *International Journal of Image and Data Fusion*, 2(4):287–308, 2011.

[84] Alexander Toet, Lodewik J Van Ruyven, and J Mathee Valeton. Merging Thermal and Visual Images by a Contrast Pyramid. *Optical Engineering*, 28(7):287789–287789, 1989.

[85] Filareti Tsalakanidou, Sotiris Malassiotis, and Michael G Strintzis. Integration of 2D and 3D Images for Enhanced Face Authentication. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 266–271. IEEE, 2004.

[86] Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[87] Zhou Wang and Alan C Bovik. A Universal Image Quality Index. *IEEE signal processing letters*, 9(3):81–84, 2002.

[88] Joseph Wilder, P Jonathon Phillips, Cunhong Jiang, and Stephen Wiener. Comparison of Visible and Infra-red Imagery for Face Recognition. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 182–187. IEEE, 1996.

[89] Lawrence B Wolff, Diego A Socolinsky, and Christopher K Eveland. Quantitative Measurement of Illumination Invariance for Face Recognition Using Thermal Infrared Imagery. In *International Symposium on Optical Science and Technology*, pages 140–151. International Society for Optics and Photonics, 2003.

[90] Xiaohua Xie, Jianhuang Lai, and Wei-Shi Zheng. Extraction of Illumination Invariant Facial Features from a Single Image Using Nonsubsampled Contourlet Transform. *Pattern Recognition*, 43(12):4177–4189, 2010.

[91] CS Xydeas and V Petrovic. Objective Image Fusion Performance Measure. *Electronics letters*, 36(4):308–309, 2000.

[92] Stefanos Zafeiriou, Gary A Atkinson, Mark F Hansen, William AP Smith, Vasileios Argyriou, Maria Petrou, Melvyn L Smith, and Lyndon N Smith. Face Recognition and Verification Using Photometric Stereo: The Photoface Database and a Comprehensive Evaluation. *IEEE Transactions on Information Forensics and Security*, 8(1):121–135, 2013.

[93] Haitao Zhao, Shaoyuan Sun, and Zhongliang Jing. Visible-information-aided eyeglasses removing for thermal image reconstruction. In *Information Fusion, 2007 10th International Conference on*, pages 1–7. IEEE, 2007.

[94] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face Recognition: A Literature Survey. *ACM Computing Surveys (CSUR)*, 35(4):399–458, 2003.

[95] Wenzhi Zhao and Shihong Du. Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4544–4554, 2016.

[96] Yufeng Zheng. Orientation-based face recognition using multispectral imagery and score fusion. *Optical Engineering*, 50(11):117202–117202–9, 2011.

[97] Hailing Zhou, Ajmal Mian, Lei Wei, Doug Creighton, Mo Hossny, and Saeid Nahavandi. Recent Advances on Singlemodal and Multimodal Face Recognition: A Survey. *IEEE Transactions on Human-Machine Systems*, 44(6):701–716, 2014.