Idiap Research Report

PARTS-BASED FACE VERIFICATION USING LOCAL FREQUENCY BANDS

Chris McCool Sébastien Marcel

Idiap-RR-06-2011

MARCH 2011

Parts-Based Face Verification using Local Frequency Bands

Christopher McCool and Sébastien Marcel

Idiap Research Institute, PO Box 592, CH-1920 Martigny, Switzerland
christopher.mccool@idiap.ch and sebastien.marcel@idiap.ch

Abstract. In this paper we extend the Parts-Based approach of face verification by performing a frequency-based decomposition. The Parts-Based approach divides the face into a set of blocks which are then considered to be separate observations, this is a spatial decomposition of the face. This paper extends the Parts-Based approach by also dividing the face in the frequency domain and treating each frequency response from an observation separately. This can be expressed as forming a set of sub-images where each sub-image represents the response to a different frequency of, for instance, the Discrete Cosine Transform. Each of these sub-images is treated separately by a Gaussian Mixture Model (GMM) based classifier. The classifiers from each sub-image are then combined using weighted summation with the weights being derived using linear logistic regression. It is shown on the BANCA database that this method improves the performance of the system from an Average Half Total Error Rate of 26.59% for a baseline GMM Parts-Based system to 14.85% for a column-based approach on the frequency sub-images, for Protocol P.

1 Introduction

The face is an object that we as humans know can be recognised. It is used to verify the identity of people on a daily basis through its inclusion in passports, drivers licences and other identity cards. However, performing automatic face verification has proved to be a very challenging task. This is shown by the fact that face recognition has been an active area of research for over 25 years [1], in fact the earliest research into face recognition was conducted by Bledsoe [2] in 1966.

Many techniques have been proposed to perform face verification ranging from dimensionality reduction techniques like Principal Component Analysis (PCA) [3] and Linear Discriminant Analysis (LDA) [4] through to feature distribution modelling techniques such as Hidden Markov Models (HMMs) [5] and Gaussian Mixture Models (GMMs) [6, 7]. Several other approaches to dimensionality reduction have been proposed some of which, such as [8, 9], are improvements on pioneering techniques such as PCA and LDA. Other recent approach explore different techniques such as Local Binary Pattern (LBPs) [10]. Examples of recent LBP-based techniques include the Local Gabor Pattern Histogram Sequence [11], descriptor-based method [12] and it has even been used as a generic pre-processing technique [13]. It is not the aim of this paper to provide an extensive review of face verification techniques, rather we take a deeper look at an existing and widely used feature distribution modelling and propose an alternative method to form the features.

A recent advance in face verification has been the effective use of feature distribution modelling techniques. Two effective methods for performing feature distribution were both published in 2002, these being the work of Sanderson and Paliwal [6] and Martinez [7]; despite the earlier work of Samaria et al. [5, 14] and Nefian and Hayes [15] who used HMMs. Martinez divided the face into k pre-defined regions, then trained a PCA representation for each k-region and the variation of each region's PCA vector was then modelled using a simplified GMM, thus there were k-GMMs; we note the GMM of Martinez is simplified as it does not include a weight for each mixture component which implies a pre-set equal weight for each mixture component. By contrast, Sanderson and Paliwal proposed that the face could be divided into blocks and all of these blocks could be used to derive a *single* GMM, this method implies that at matching time each block is a assigned a probability of coming from the components of the GMM; thus there are no pre-defined regions, rather, each block is aligned probabilistically. This method of Sanderson and Paliwal, from here on referred to as the GMM Parts-Based approach, has since been used and extended by several researchers.

The approach of Sanderson and Paliwal has been employed and extended to include background model adaptation [16] and the use of LBPs as a pre-processing technique [13]. By using background model adaptation, the hope is to form a general description of the face (a background model) which can be used as a starting point to derive a more reliable client model as well as providing a description for faces that are not of the client. By using the LBP as a pre-processing technique Heusch et al. [13] showed that extra illumination robustness could be achieved leading to improved results.

In [17] we proposed a method to perform both a Spatial and Frequency based decomposition for the GMM Parts-Based approach. The frequency decomposition was achieved by collating the responses from each DCT coefficient from each block (observation) and forming a separate sub-image for each frequency. Each of these sub-images was treated separately and a GMM based classifier was generated for each sub-image. The classifiers from each sub-image were then combined using weighted summation with the weights being derived using linear logistic regression. Tests conducted on the BANCA database showed that this extension provided a significant improvement with the Average Half Total Error Rate being reduced from of 24.38% to 15.17% when compared to a baseline Parts-Based approach. It is worth noting that the baseline system and our sub-image system used exactly the same feature vectors but re-assembled them differently and so modelled in different way.

We extend upon the previous work [17] by: examining the methods robustness to its hyper parameters such as block and image size, performing a comparison of the local frequency band approach against other related state-of-the-art techniques, and finally investigating explanations for why the column-based approach outperforms all other approaches. We first begin by providing an overview of the GMM Parts-Based approach and we then explain clearly the differences between this technique and our proposed approach.

2 Related Work on GMM Parts-Based Face Verification

The Parts-Based approach divides the face into blocks, or parts, and treats each block as a separate observation of the same underlying signal (the face). In this method a feature vector is obtained from each block by applying the Discrete Cosine Transform and the distribution of these feature vectors is then modelled using GMMs. Several advances have been made upon this technique, for instance, Cardinaux et al. [16] proposed the use of background model adaptation while Lucey and Chen [18] examined a method to retain part of the structure of the face utilising the Parts-Based framework as well as proposing a relevance based adaptation.

This parts-based framework is quite different to other parts-based techniques such as that of Heisele et al. [19]. For instance Heisele et al. pre-define a set of regions (similar to Martinez [7]) and then derive an expert classifier for each region. By contrast the GMM Parts-Based approach derives a single classifier for the whole face and blocks are not associated to an expert classifier but rather are probabilistically aligned to the components of the GMM. By doing this probabilistic alignment problems associated with poorly aligned images are more easily overcome and it suggests one of the advantages of using a GMM over holistic techniques such as PCA and LDA.

2.1 Feature Extraction

The feature extraction algorithm is described by the following steps. The face is normalised, registered and cropped. This cropped and normalised face is divided into blocks (parts) and from each block (part) a feature vector is obtained. Each feature vector is treated as a separate observation of the same underlying signal (in this case the face) and the distribution of the feature vectors is modelled using GMMs. This process is illustrated in Figure 1.

The feature vectors from each block are obtained by applying the DCT. It would be possible to apply advanced feature extraction methods such as the DCTmod2 [6], however, this advanced feature extraction method uses the DCT as its basis feature vector and by using only the DCT we can treat each coefficient as a separate frequency response from the image, or block. This is because each DCT coefficient is orthogonal whereas some of the DCTmod2 coefficients are not orthogonal since they incorporate spatial information by using the deltas from neighbouring blocks. More details about the DCT can be found in [20].

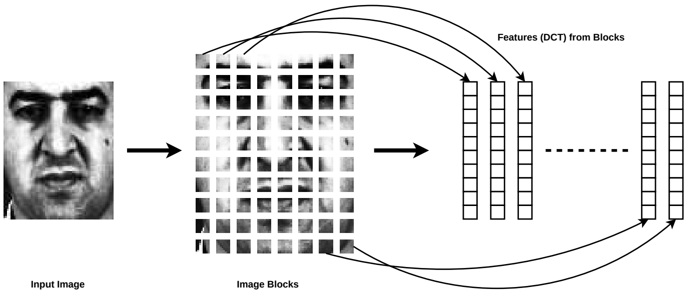


Fig. 1. A flow chart describing the extraction of feature vectors from the face image for Parts-Based approaches.

2.2 Feature Distribution Modelling

Feature distribution modelling is achieved by performing background model adaptation of GMMs [16, 18]. The use of background model adaptation is not new to the field of biometric authentication in fact it is commonly used in the field of speaker verification [21]. Background model adaptation first trains a world (background) model Ω_{world} from a set of faces and then derives the client model for the i^{th} client Ω^{i}_{client} by adapting the world model to match the observations of the client.

Two common methods of performing adaptation are mean only adaptation [22] and full adaptation [23]. Mean only adaptation is often used when there are few observations available because adapting the means of each mixture component requires fewer observations to derive a useful approximation. Full adaptation is used when there are sufficient observations to adapt all the parameters of each mode. Mean only adaptation is the method chosen for this work as it requires fewer observations to perform adaptation, this is the same adaptation method employed by Cardinaux et al. [16].

2.3 Verification

A description of the Parts-Based approach is not complete without defining how an observation is verified. To verify an observation, x, it is scored against both the client (Ω^{i}_{client}) and world (Ω_{world}) model, this is true even for methods that do not perform background model adaptation [6]. The two models, Ω^{i}_{client} and Ω_{world}, produce a log-likelihood score which is then combined using the log-likelihood ratio (LLR),

$$h(x) = \ln(p(x \mid \Omega^{i}_{client})) - \ln(p(x \mid \Omega_{world})), \tag{1}$$

to produce a single score. This score is used to assign the observation to the world class of faces (not the client) or the client class of faces (it is the client) and consequently a threshold τ has to be applied to the score $h(x)$ to declare (verify) that x matches to the i^{th} client model Ω^{i}_{client} when $h(x) \geq \tau$.

3 Local Frequency Band Approach

The proposed method is to divide the face into separate blocks and to then decompose these blocks in the frequency domain. This can be achieved by treating the frequency response from each block separately to form frequency sub-images, this process is applied to the DCT feature vectors obtained by applying the Parts-Based approach. An important property of the DCT is that each coefficient is orthogonal and thus each frequency can be considered independently.

The technique was introduced in [17] and is summarised as follows:

1. the face is cropped and normalised to a 68 × 68 image,
2. the face is divided into square blocks with an overlap of 4 pixels in the horizontal and vertical axes,
3. the D DCT coefficients from each block are separated and used to form their own frequency sub-image, and
4. a feature vector is formed by taking a block from the frequency sub-image and vectorising it.

The way in which the frequency sub-images are formed is demonstrated in Figure 2. It is important to note that the number of sub-images formed is determined by the number of D DCT coefficients retained.

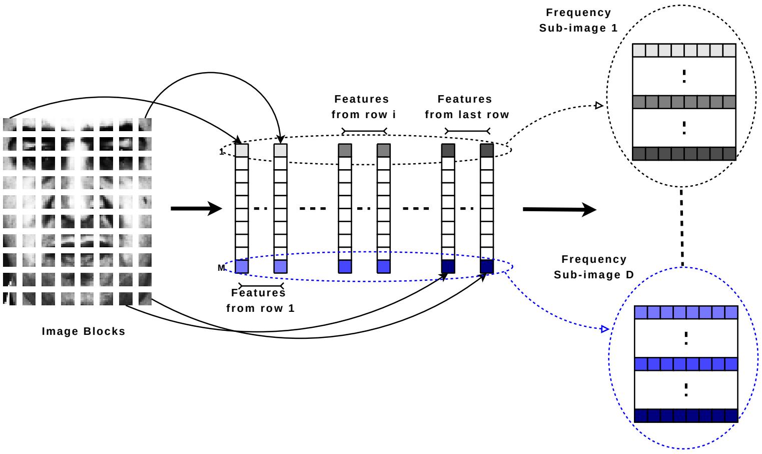


Fig. 2. The figure above describes how the face can be decomposed into separate frequency sub-bands (sub-images). The value D refers to the total number of DCT coefficients extracted from each block.

3.1 Motivation

To illustrate the differences between the frequency decomposition approach and the classic Parts-Based approach the following statements are made. For the

Parts-Based approach it is often stated that the face is broken into blocks and the distribution of each block is then modelled [6, 18], however, another stricter statement would be that the frequency information from each block is simultaneously modelled since each dimension of the feature vector represents a different sampling frequency of the DCT. By contrast the frequency decomposition approach separates the frequency information from each local block and forms many feature vectors from the resulting frequency sub-images. Thus the image is decomposed in both the spatial domain and the frequency domain.

A side effect of working on the frequency sub-images is that the feature vectors formed from these sub-images will retain extra spatial information. This is because the frequency decomposition approach gets the response from each block and then extracts a feature vector using responses from several blocks. This means that the feature vectors extracted from the frequency sub-images will actually span several blocks when compared to the Parts-Based approach, for instance the feature vector could be formed from a frequency sub-image by spanning an entire row or column of the image.

3.2 Feature Extraction

Three methods of forming a feature vector from the frequency sub-images are examined, these are to form a feature vector:

1. across a row of the frequency sub-image (row-based approach),
2. across a column of the frequency sub-image (column-based approach), or
3. from a square block of the frequency sub-image which is vectorised (block-based approach).

These three methods are described in more detail below and also visually in Figure 3.

Row-based approach (SB_{Row}): For this approach feature vectors are formed from the sub-images by performing a horizontal scan. This means that they get formed across the face capturing the spatial relationship between the left eye, right eye, the asymmetry of the nose and how the mouth varies.

Column-based approach (SB_{Col}): The feature vectors are formed from by scanning the face in a vertical manner. This means that the feature vectors are formed in stripes down the face and could capture the spatial relationship of features such as eyes to mouth, eyebrows to eyes to mouth and nose to mouth.

Block-based approach (SB_{Blk}): In this approach the feature vectors are formed in a block-based manner. Square blocks are formed, so as to not bias the technique in either the vertical or horizontal direction, and these blocks are vectorised to produce the feature vector. Conceptually this collects information from adjacent regions of the face meaning that blocks which capture the spatial relationship between the eye and the cheek or the eye and the bridge of the nose could exist.

An issue with applying these three feature extraction techniques is that each method should result in the same number of observations and the same number of dimensions for the feature vector, so as to not bias any one method. This is a particularly restrictive requirement when considering the block-based approach as there is a further requirement that the blocks are square blocks so that equal emphasis is placed on the horizontal or vertical responses.

The above requirements led to the initial face image size being 68×68 pixels. This means that when using square blocks of 8×8 pixels (with four pixels of overlap between blocks in the vertical and horizontal direction) this results in local frequency sub-images of size 16×16. Using the row- and column-based approaches this would lead to feature vectors of dimensionality 16 with 16 observations (by using all the data from a row or column). To retain the same dimensionality for the block-based method would require us to sample (from the frequency sub-image) with a non-overlapping block of size 4×4 which leads to a dimensionality of 16 with 16 observations.

The initial sampling parameter to form the frequency sub-images will change the final frequency sub-image sizes. For instance changing just the block size 12×12 would lead to local frequency sub-images of size 15×15 if we use the same overlap margin of 4 pixels. In this particular case forming a square block from the frequency sub-images would not lead to the same number of samples (and thereby an unfair comparison). Therefore for initial testing we limited ourself to sampling using an 8×8 block with a four pixel overlap. We then vary the block size and produce results where possible (in Section 5.3).

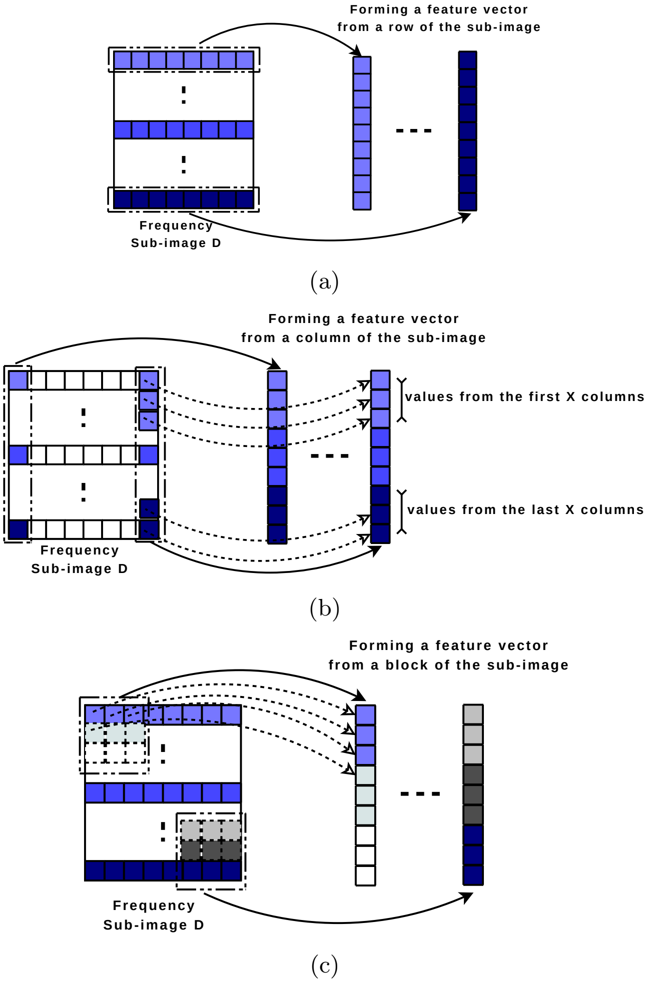
3.3 Classifier Combination

Once a set of feature vectors is obtained a classifier is then trained for each frequency sub-image. The approach is it to perform the same background model adaptation that was used for the Parts-Based approach [16]. Each classifier (C_k) is combined using weighted linear score fusion,

$$C_{weight\_sum} = \sum_{k=1}^{K} \beta_k C_k, \tag{2}$$

where β_k is the weight given to the k^{th} classifier and K is the number of classifiers (frequency sub-images) that are combined. This method is used as Kittler et al. [24] showed that the sum rule (which is what linear classifier score fusion abstracts to be) is robust to estimation errors. The weights, β_k, for the classifiers are derived using an implementation of linear logistic regression [25]. Prior to performing linear logistic regression the output of each classifier is z-score normalised.

By performing z-score normalisation the classifiers have a common frame of reference which means that the weights provide an insight into the relevance of each classifier. This common frame of reference is achieved by normalising the scores such that the impostor scores can be assumed to have zero mean and unit


Forming a feature vector
from a row of the sub-image
Frequency
Sub-image D
Forming a feature vector
from a column of the sub-image
values from the first X columns
values from the last X columns
Frequency
Sub-image D
Forming a feature vector
from a block of the sub-image
Frequency
Sub-image D

(a)

(b)

(c)

Fig. 3. Forming the feature vectors (a) along a row of the frequency sub-image, (b) along a column of the frequency sub-image and (c) from a block of the frequency sub-image. In total there are D frequency sub-images.

variance. These parameters are derived using the development set defined by each experimental protocol.

3.4 Relationship to Other Work

Some work has already proposed the use of frequency decomposotion. For instance, in the work of Zhang et al. the initial input image is transformed into as many of 40 Gabor filtered images [11]. From each of these Gabor filtered images sub-regions are defined and a histograms is obtained from each sub-region.

Finally, all of these histograms are concatenated to form one vector which then represents the image. This is quite different from the sub-bands approach as the histograms from the Gabor filtered images of Zhang et al. are forced to come from pre-defined regions. By contrast we divide the frequency sub-images into feature vectors and these feature vectors are probabilistically aligned to their associated GMM.

4 Experimental Protocol

The experiments were conducted on the BANCA English database [26]. This database has challenging conditions in terms of pose and illumination and has several well defined protocols and it has been used to evaluate face recognition techniques in two international competitions [27, 28].

4.1 BANCA Database

The 52 subjects in this database are split into two gender-balanced groups, *g1* and *g2*. Each subject participated in 12 recording sessions, grouped into three different scenarios:

- **Controlled (Sessions 1-4)**: Captured in a controlled environment using a high quality camera.
- **Degraded (Sessions 5-8)**: Captured in a less-controlled environment using a low quality camera (a web-cam).
- **Adverse (Sessions 9-12)**: Captured in an uncontrolled environment using a high quality camera.

An example of the three scenarios is provided in Figure 4.

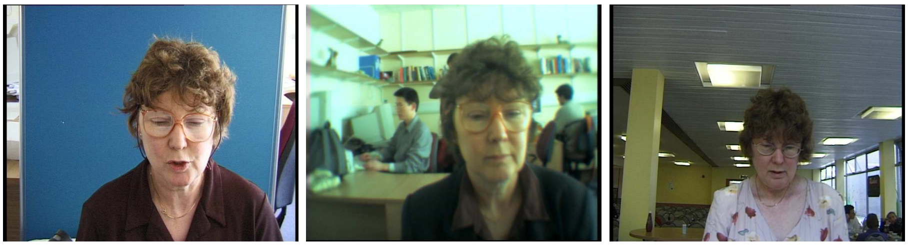


Fig. 4. Examples from the BANCA dataset, representing (from left to right) controlled, degraded and adverse capture conditions.

In each recording session two recordings were captured. One is a client access where the user matches their claimed identity and the other is an impostor attack where the user did not match their claimed identity; a different identity is claimed each time such that each client is attacked once by all the other clients. There is an additional *world model* group consisting of 30 subjects (15 male and 15 female).

Within the BANCA protocol there are three defined data sets: *world model*, *development* and *evaluation*. The *world model* data set conists of users that are external to both the *development* and *evaluation* data sets, this set is used to train background models such as the background GMM. The *development* set is used to calibrate the system for instance by deriving weights for fusion or by deriving the decision threshold τ which is used on the *evaluation* set. The *evaluation* set is used to evaluate the final system and as such it includes enrollment and testing data for all of the clients. The *development* and *evaluation* sets come from *g1* and *g2* used in a cross-validated manner.

4.2 Normalisation

Each face image is cropped and scaled to a size of 68×68 pixels with a distance between the eyes of 33 pixels. Illumination normalisation is applied to each image as a two stage process, the image is histogram equalised and then encoded using a Local Binary Pattern (LBP) [10]. This is the same normalisation strategy employed by Heusch et al. [13] which was also applied to a GMM parts-based system (using DCT feature vectors). The parameters used by Heusch et al. were an LBP of radius of $R = 2$ and with $P = 8$ sampling points along the circle. A visual description of the LBP encoding process is given in Figure 5 and more formally it is written as,

$$LBP(x_c, y_c) = \sum_{i=0}^{P} s(p_i - p_c) * 2^i, \tag{3}$$

where (x_c,y_c) is a given pixel position in the image, p_i is the i^{th} sampling point on the LBP circle, p_c is the center of the LBP circle (which is the point x_c,y_c), and where the function s is,

$$s(x) = \begin{cases} 1\, if\ x \geq 0 \\ 0\, if\ x < 0. \end{cases} \tag{4}$$

4.3 Protocols

The following experimental configurations are defined for the BANCA dataset: Matched Controlled (Mc), Matched Degraded (Md), Matched Adverse (Ma), Unmatched Degraded (Ud), Unmatched Adverse (Ua), Pooled test (P) and Grand test (G). The Table in 1 summarises the usage of the different sessions in each configuration. “TT” refers to the client enrollment while “T” depicts the client and imposter test sessions. For example, in configuration Mc the true client data from session 1 is used for enrollment and the true client data from sessions 2, 3 and 4 are used for testing. The imposter attack data from all sessions are used for imposter testing.

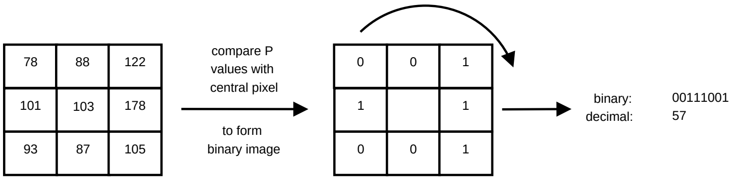


Fig. 5. A diagram illustrating how the LBP operator can be applied to a set of pixels from an image.

Session	P	G	Ud	Ua	Mc	Md	Ma
1	TT	TT	TT	TT	TT		
2	T	T			T		
3	T	T			T		
4	T	T			T		
5		TT				TT	
6	T	T	T			T	
7	T	T	T			T	
8	T	T	T			T	
9		TT					TT
10	T	T		T			T
11	T	T		T			T
12	T	T		T			T

Table 1. Usage of different sesisons in BANCA configurations.

4.4 Performance Measures

Results are presented numerically using the Average Half Total Error Rate (HTER) and graphically using Detection Error Tradeoff (DET) plots. The HTER is an average of the False Acceptance Rate (FAR) and False Rejection Rate (FRR) at a given threshold such that

$$HTER = \frac{FAR + FRR}{2}, \tag{5}$$

where the threshold is obtained from the Equal Error Rate (where the FAR equals the FRR) on the *development* data. For the BANCA experiments we present the Average HTER which is the average HTER from g_1 and g_2 (when used as the *evaluation* set),

$$Average\ HTER = \frac{g1_{HTER} + g2_{HTER}}{2}, \tag{6}$$

and this is derived in a cross-validated manner by tuning parameters, such as the threshold, on one data set (taken as the *development* set) and then calculating the HTER on the other data set (taken as the *evaluation* set). The second method of presenting results, DET plots, provide a more complete description of the system performance by plotting the percentage of FAR versus the percentage of FRR on a log scale, for more details on the use of DET plots for assessing system performance readers are referred to [29].

Parameters such as the number of mixture components M and block size B are derived in a global manner while other parameters such as the decision threshold $\tau_{decision}$ and classifier weights β_k are derived in a cross-validated manner. When deriving the parameters in a global manner the best results from Protocol P of the BANCA database are used whereas the cross-validated parameters are derived on the independent *development* set defined for each Protocol.

5 Analysis of Verification Performance

To analyse the effectiveness of the local frequency band approach a consistent basis for comparison is needed and for this work we provide two ways of comparing the system performance. First, using exactly the same images and exactly the same DCT features a baseline GMM Parts-Based system is derived: this means that 68×68 face images are used, the same image normalisation procedure is applied and the top 15 dimensions of the DCT are retained as a feature vector. More details about this system are provided below. Second, the performance of state-of-the-art systems are presented and contrasted with the current system. Where possible we present results using both manual and automatic annotations.

The effect of using manual and automatic eye positions is examined since any deployed face verification system will need to cope with errors introduced from an automatic face detection system. The manually annotated eye positions were provided with the BANCA database and the automatically annotated eye positions were obtained using a face detector based on a cascade of LBP-like features [30] [1]. There were 93 images (out of $6,540$ images) where the automatic face detector could not find the face, these images were excluded from *training*, *development* and *evaluation* of the automatic systems.

5.1 Baseline System

Two baseline systems are considered for this work one that uses DCT feature vectors and another that uses DCTmod2 feature vectors. DCTmod2 feature vectors are examined as it was previously found to be more robust than DCT feature vectors [6]. The size of the feature vectors, $D = 15$ for DCT feature vectors and $D = 18$ for DCTmod2 feature vectors, was chosen based on work conducted in [6]. Both baseline systems were developed using 68×68 face images and varying the number mixtures by powers of 2 such that $M = [16, 32..., 512]$.

Results on the Development set of the P protocol found that a system using DCTmod2 feature vectors, for both manual and automatic eye positions, provides the best performance. For the manual annotations $M = 128$ provided the

[1] This detector has been implemented with the Torch3vision computer vision library (torch3vision.idiap.ch)

best performance while for automatic annotations $M = 256$ provided the best performance. These two systems were then used to produce results for all of the BANCA protocols and are presented in Table 2. These results are different to those presented in [17] where DCT feature vectors were found to provide the optimal system.

	P	G	Ud	Ua	Mc	Md	Ma
manual	26.59%	11.68%	27.31%	30.37%	8.78%	14.71%	16.81%
automatic	27.84%	11.98%	25.51%	30.12%	7.55%	15.27%	17.33%

Table 2. This table presents the average HTER for the baseline Parts-Based verification system. This system uses DCTmod2 feature vectors and results are presented for all of the BANCA protocols. We present results when using both manual and automatic annotation of the eye positions.

For the baseline system it can be seen that the performance is consistent between manual and automatic eye annotations. The absolute performance degradation for Protocol P is 1.25%. This consistency in performance is attributed to the fact there are 93 faces which are not located and consequently are completely ignored and also that the Parts-Based approach is robust to imprecise face localisation.

The robustness to imprecise face localisation can be seen by examining the baseline systems using the two Parts-Based systems, one which uses DCT feature vectors and one which uses DCTmod2 feature vectors. Tables 2 and 3 present the optimal performance for the DCTmod2 and DCT feature vectors respectively. It can be seen that the performance using manual and automatic annotations is very similar, in fact both sets of features have very similar performance except for the case of Mc, Md and Ma. For these three matched protocols it can be seen that the DCTmod2 feature vectors perform significantly better than their DCT counterparts with the performance degrading on average by 4.8% when using the DCT feature vectors; this average is formed across the Mc, Md, and Ma protocols across both the manual and automatic annotations. This implies that one major advantage of using the DCTmod2 feature, over just DCT features, occurs when the conditions between enrollment and testing are matched.

	P	G	Ud	Ua	Mc	Md	Ma
manual	26.58%	15.83%	27.07%	30.77%	11.54%	19.15%	21.14%
automatic	27.53%	19.54%	28.76%	30.14%	13.91%	20.20%	23.33%

Table 3. This table presents the average HTER for the Parts-Based verification system using DCT features for all of the BANCA protocols with an optimal value of $M = 64$ for both manual and automatic annotations.

5.2 Subband Approaches

The initial experiments indicated that all of the local frequency sub-band approaches, across all BANCA protocols, provided significantly improved performance when compared to the baseline system. When using manual annotations, see Table 4 and Figure 6 for full results, the optimal local frequency sub-band approach is the column-based approach followed by the block-based approach and finally the row-based approach. The column-based approach provides an absolute improvement over the baseline system (using DCTmod2 features) of 11.74% for Protocol P with the average HTER reducing from 26.59% to 14.85%; the frequency sub-band systems are optimised in a similar manner to the baseline systems, however, because there are fewer observations ($o = 16$ observations for each frequency sub-image whereas $o = 256$ for the Parts-Based approach) the number of mixtures were constrained to $M = [2, 4, 8..., 32]$.

	P	G	Ud	Ua	Mc	Md	Ma
baseline	26.59%	11.68%	27.31%	30.37%	8.78%	14.71%	16.81%
row-based	19.73%	8.13%	18.91%	23.17%	6.6%	11.35%	12.71%
block-based	18.05%	7.64%	16.41%	24.58%	8.65%	9.86%	**12.55%**
column-based	**14.85%**	**5.04%**	**12.90%**	**21.71%**	**5.77%**	**8.45%**	12.87%

Table 4. This table presents the average HTER for the local frequency sub-band approaches and the optimal baseline system (see Table 2) when using **manually annotated** eye locations for all of the BANCA protocols. Each system has $M_{row} = 8$, $M_{blk} = 4$ and $M_{col} = 8$ mixture components respectively. Highlighted are the best performing systems.

Examining the performance of the sub-band approach for automatic annotations it was found that the column-based frequency sub-band approach was again superior to that of the baseline system. This result is true across all test conditions, see Table 5 for full results, except Mc where the baseline system outperforms the column-based approach by absolute difference of 0.33% which is not a significant amount. Further analysis shows that the column-based approach is also more robust than either the row-based or block-based approaches. For instance when comparing the performance of manual and automatic eye locations on Protocol P the column-based approach has an absolute performance degradation of 1.77% whereas the block-based and row-based approaches have a degradation of 3.52% and 5.97% respectively. A similar result happens for the other protocols with the average performance degradation, between manual and automatic eye locations, for all of the BANCA protocols being 1.65%, 5.05% and 4.59% for the column-, block- and row-based techniques respectively.

The experiments presented thus far have demonstrated that choosing the correct method for forming a feature vector has a significant impact on the local frequency sub-band approach. It has been shown empirically that the column-based approach is more robust to localisation errors than either the row-based

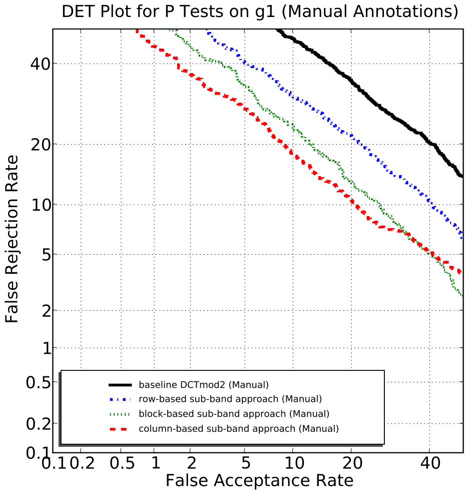


Fig. 6. In this figure we present the DET plots of the four systems (baseline, row-based, block-based and column-based) for protocol P for g1 of the BANCA database. It can be seen that the column-based approach outperforms every other approach.

	P	G	Ud	Ua	Mc	Md	Ma
baseline	27.84%	11.98%	25.51%	30.12%	**7.55%**	15.27%	17.33%
row-based	26.58%	9.55%	27.28%	28.00%	9.32%	17.03%	15.83%
block-based	21.57%	12.29%	24.43%	25.17%	14.7%	16.53%	18.34%
column-based	**16.62%**	**7.40%**	**17.97%**	**19.86%**	7.88%	**10.41%**	**13.02%**

Table 5. This table presents the average HTER for the local frequency sub-band approaches and optimal baseline system (see Table 2) on **automatically annotated** eye locations for all of the BANCA protocols. Each system has $M_{row} = 16$, $M_{blk} = 4$ and $M_{col} = 8$ mixture components respectively. Highlighted are the best performing systems.

or block-based approaches. Also, the column-based performs better than either the row-based or block-based approaches for all of the test conditions. This fact could be explained by suggesting that the features of the face are more stable when scanned in a vertical manner, particularly when there is misalignment of the face image. Another argument is that there is more variance in the features when they are scanned in a vertical rather than horizontal manner. This second line of reasoning, regarding the variance of the features, forms the basis for the experiments in the later sections. Before proceeding to these experiments the issue of block and image size is examined and then a comparison with state-of-the-art feature distribution modelling techniques is provided.

5.3 Block Size and Image Resolution

The initial experiments above (and in [17]) analysed the system performance using a block size of $B = 8$, these experiments were extended in two ways. The first is to include three other blocks sizes $B = 4$, $B = 12$ and $B = 16$ for the 68×68 images and the second is to apply a similar procedure to higher resolution images.

The results for varying the block sizes can be found in Table 6. It can be seen in this Table that when we use larger blocks (than $B = 8$) there is a minor degradation in performance, while if we decrease the block size (to $B = 4$) there is a significant degradation in performance. Given these results we have retained the optimal block-size of $B = 8$ for the 68×68 images.

	$B = 4$	$B = 8$	$B = 12$	$B = 16$
row-based	24.91%	19.73%	20.68%	22.98%
block-based	N/A	18.05%	N/A	N/A
column-based	16.21%	**14.85**%	15.04%	15.96%

Table 6. This table presents the average HTER (%) on Protocol P of the BANCA database using manual annotations while varying the block size. The value N/A is presented when it is not possible to fairly apply this technique.

To better understand the effect of block size a higher resolution image with several block sizes was examined. The cropped image size was changed to be approximately one and a half times bigger than the original 68×68 pixels. The final image size used was 104×104 pixels so that a constant 4 pixel shift between each block could be used. A range of block sizes were considered ($B = 8, 12, 16, 20, 24, 28$) and it was found that block size does have an impact on the performance of the column-based systems; we did not consider a block-size of $B = 4$ as it already shown to degrade performance in smaller images. However, the performance of the system is fairly stable over the range of block sizes $B = 12, 16, 20$ because for these block sizes the difference in performance is less than 1% and for all the block sizes the difference in performance is at most 2.11%, see Table 7. It can also be seen that there is a minimal performance increase of 0.12% when using the higher resolution 104×104 images (best average HTER of 14.73%) compared to the 68×68 images (best average HTER of 14.85%).

	$B = 8$	$B = 12$	$B = 16$	$B = 20$	$B = 24$	$B = 28$
column-based	16.84%	15.79%	**14.73**%	15.44%	16.22%	16.19%

Table 7. This table presents the average HTER (%) on Protocol P of the BANCA database using manual annotations while varying the block size for images of size 104×104 pixels.

Given the comparable performance between these two image resolutions it can be seen that this method is particularly suited to lower resolution images. When using higher resolution images there is not a significant improvement in performance and so it is difficult to justify using the higher resolution image. It can also be seen that for both the 68 images and 104 × 104 images there is an effect in varying the block size, however, this effect on performance is relatively minor within a range of appropriate block sizes. Therefore, for the remainder of this article only the results for the 68 × 68 images with a block size of $B = 8$ is considered.

5.4 Sampling Rate

The formation of the frequency sub-images is influenced by the sampling rate from the initial image. From the above experiments our initial image size was 68 and the sampling rate was every 4 pixels, using a set of 8 × 8 blocks to sample this image led to frequency sub-images of size 16 × 16. Therefore, we increased the sampling rate by factors of 2 to yield two new sampling rates: every 2 pixels and every 1 pixel. This provides us with frequency sub-images of size 31 × 31 (when sampling every 2 pixels) and 61 × 61 (when sampling every 1 pixel). This has the dual effect of increasing our dimensionality for the SB_{col} system but also increasing the number of observations.

In Table 8 we present the performance for sampling every 1 pixel for manual and automatic annotations on the BANCA protocols. We only present the full performance for sampling every 1 pixel as this performed better than sampling every 2 pixels. It can be seen that for manual annotations we have a reasonable gain in performance when compared to the column-based technique in Table 4. However, when we compare the performance when using automatic annotations we can see there is minimal difference between the default sampling rate of every 4 pixels (the column-based entry in Table 5) and sampling every 1 pixel (Table 8). Given these results and given that sampling every 1 pixel provides us with almost 4 times the data and feature vectors that are almost 4 times larger we chose to produce the final results on the default system which samples every 4 pixels.

	P	G	Ud	Ua	Mc	Md	Ma
manual	12.64%	3.85%	10.00%	17.04%	4.87%	7.15%	10.06%
automatic	16.57%	6.36%	17.60%	19.84%	6.83%	11.36%	12.18%

Table 8. This table presents the average HTER (%) for all of the BANCA protocols. Results are presented for manual and automatic annotations using a sampling rate of every 1 pixel.

5.5 Performance Comparison

To provide an overview of the column-based sub-band approach its performance is compared to state-of-the-art techniques for the BANCA database. Previous state-of-the-art face verification systems, tested on the Mc protocol of BANCA, are taken from the work of Heusch and Marcel [31] and Cardinaux et al. [32]. These two sets of results provide a comparison to a Bayesian Network classifier (BN), a Partial Shape Collapse GMM classifier (PSC-GMM), state-of-the-art GMM Parts-Based classifier and two state-of-the-art HMM-based classifiers; the two HMM classifiers are the 1D HMM (which models vertical transitions) and P2D HMM (which models vertical and horizontal transitions). When quoting the results from Cardinaux et al. [32] only the results on g2 are available. All of these results are reproduced in Table 9 where it can be seen that the column-based sub-band (SB_{Col}) approach outperforms the three other systems, including the more complex BN and HMM approaches. Furthermore, the column-based approach has fewer parameters as each sub-band consists of an $M = 8$ component GMM. However, this only compares the case of the matched condition.

	HTER on g1 (%)	HTER on g2 (%)	Total Parameters
SB_{Col}	**8.1**	**3.5**	**4,095**
P2D HMM [32]	N/A	4.6	73,728
Bayesian Network [31]	9.0	5.4	5,225
1D HMM [32]	N/A	6.9	4,032
GMM [32]	N/A	8.9	9,216
PSC-GMM [31]	11.3	11.3	$6 \times 33,280$

Table 9. This table presents the average HTER (%) on the BANCA database using Manual annotations on the Mc protocol for the column-based frequency sub-band approach, a Bayesian Network approach, PSC-GMM method and a state-of-the-art GMM approach. Because some results are only available to the first decimal place then all results are rounded to one decimal place.

In [32] further results for unmatched conditions (Ud and Ua) and for the P protocol are also available and so we compare the SB_{Col} approach to these results for both Manual and Automatic annotations, see Table 10. It can be seen that the SB_{Col} approach clearly outperforms both the 1D HMM and GMM systems particularly for Automatic eye locations, however, the SB_{Col} system performs worse than the P2D HMM system. It has to be noted that the P2D HMM system is a much more complex system which requires an order of magnitude more parameters to be estimated and used when compared to the SB_{Col} approach. Therefore, the SB_{Col} approach can be viewed as a tradeoff between accuracy and computational complexity when compared to the P2D HMM system.

Manual	SB_{Col}	P2D HMM [32]	1D HMM [32]	GMM [32]
Ud	**13.6**	15.3	16.3	17.3
Ua	18.9	**13.1**	17.0	20.9
P	14.6	**13.5**	14.7	17.0
Automatic	SB_{Col}	P2D HMM [32]	1D HMM [32]	GMM [32]
Ud	18.9	**15.9**	25.9	21.0
Ua	19.7	**14.7**	23.4	24.8
P	15.5	**14.7**	21.7	19.5

Table 10. This table presents the HTER (%) on g2 of the BANCA database using Manual and Automatic annotations for the Ud, Ua and P protocols. The results are presented for SB_{Col}, P2D HMM, 1D HMM and GMM system. Because some results are only available to the first decimal place then all results are rounded to one decimal place.

6 Analysis of Information Content

In the following experiments we analyse potential explanations as to why the column-based sub-band approach performs significantly better than the row-based or block-based methods. The first possible explanation for this difference in performance is that the variance in the column-based features is higher than that of either the row-based or block-based features.

To analyse the variance within the row- and column-based feature vectors Principal Component Analysis (PCA) was applied. PCA was used to represent the features extracted using the row- and column-based methods in their most compact form, this is based on variance. All the row-based (or column-based) observations, feature vectors, from a particular sub-band j were used to derive a vector space V_j^{row} (or V_j^{col}) using PCA. The eigenvalues of the resulting vector spaces represent the variance of each dimension (the ability to compress information). It was envisaged that the column-based method would have a more even spread for the resulting eigenvalues whereas for the row-based method there would be a greater variance in the first few eigenvalues or dimensions of the resulting vector spaces.

The results, in Figure 7, show that both techniques encode a similar quantity of information with the exception of sub-bands $j = [1, 2, 6, 7]$. These sub-bands correspond to the 0^{th} coefficient (mean or DC value of the blocks) and the first three vertical-only responses of the DCT; because this is a 2D-DCT there is both a horizontal and vertical cosine function, therefore the vertical-only response corresponds to the case where only the vertical cosine function varies. To analyse sub-bands $j = [1, 2, 6, 7]$ in more detail two approaches were taken: i) a visual inspection and ii) a performance-based inspection.

Upon visual inspection it can be seen that these sub-bands correspond to the most face-like sub-bands. In Figure 8 the average response over the entire BANCA database for each sub-band is presented. In this figure it can seen that $j = [1, 2, 6, 7]$ represent those sub-bands which are most face-like, however, sub-band $j = 15$ also appears to be face-like. These sub-bands ($j = [1, 2, 6, 7, 15]$)

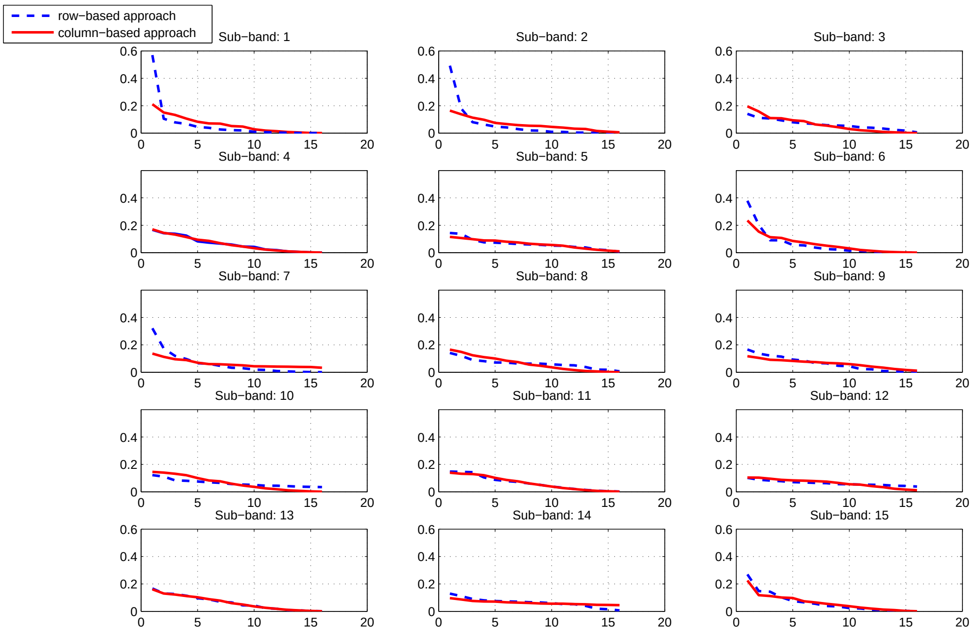
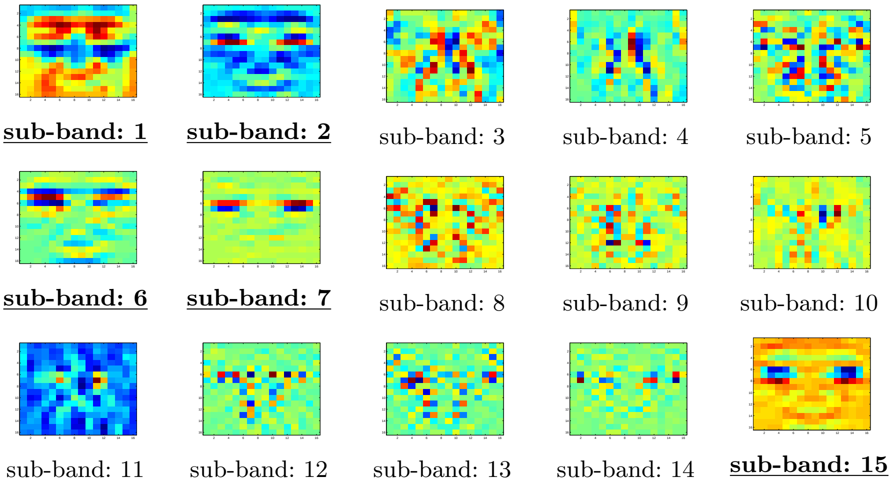


Fig. 7. In this figure the percentage of variance (represented by the eigenvalues) for the vector space of each sub-band is presented for both the row-based and column-based feature extraction approaches. The percentage of variance (y-axis) is calculated as a percentage of the total eigenvalues (x-axis) for the PCA vector space of the particular sub-band j.

correspond to the mean or DC coefficient ($j = 1$) and the vertical-only responses of the DCT ($j = [2, 6, 7, 15]$) with a frequency of $\theta = [\pi/16, \pi/8, 3\pi/16, \pi/4]$ respectively. They are also the sub-bands for which there was more redundant information for the row-based feature extraction approach (high values in the first few eigen-values) from the PCA analysis.



sub-band: 1 **sub-band: 2** sub-band: 3 sub-band: 4 sub-band: 5

sub-band: 6 **sub-band: 7** sub-band: 8 sub-band: 9 sub-band: 10

sub-band: 11 sub-band: 12 sub-band: 13 sub-band: 14 **sub-band: 15**

Fig. 8. In this figure the average response of the 15 sub-bands for the entire BANCA database is provided. It can be seen that there are sub-bands (or DCT coefficients) which appear significantly more face-like than others, these being sub-bands $j = [1, 2, 6, 7, 15]$. Note that each sub-band has been normalised to have the range [0...1] and so that each colourmap is comparable.

To further examine the importance of sub-bands $j = [1, 2, 6, 7]$ the performance of the column-based approach was analysed with and without these sub-bands. Using these sub-bands the Average HTER on Protocol P was found to be 16.44%, which is only 1.69% worse than the full sub-bands system. This result suggests that the majority of the discriminatory information is held in these particular sub-bands. This seems to be a very useful result if we want to implement an efficient face verification system as it implies only four sub-bands need to be calculated. To confirm this result the performance of these sub-bands was plotted against the full system and a system consisting of the remaining sub-bands ($j = [3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15]$) and is presented in Figure 9.

The results in Figure 9 suggest that the Average HTER is a misleading error rate for performance evaluation, when used in isolation. This is because the performance of the column-based approach using just four sub-bands ($j = [1, 2, 6, 7]$) is similar to the full column-based approach for parts of the DET plot around the point where the error rates are equal. However, the performance of the system using just the four sub-bands degrades significantly in the top-left hand corner of the DET plot, this part of the DET plot represents the area of most interest for authentication purposes as it represents a highly secure system with a low False Acceptance Rate. From these results we can see that a significant amount of information is retained in just the four sub-bands $j = [1, 2, 6, 7]$, however, in spite of this the other eleven sub-bands $j = [3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15]$ also contain important information when used in combination with these four sub-bands.

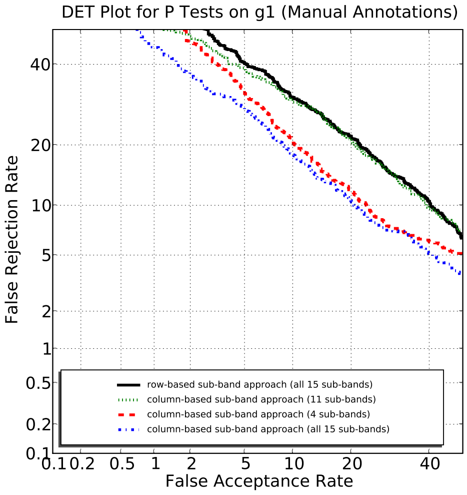


Fig. 9. In this figure the DET plot for the full row-based approach, the full column-based approach, the column-based approach using $j = [1, 2, 6, 7]$ and the column-based approach using $j = [3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15]$ are all presented.

7 Conclusions and Future Work

In this paper the local frequency band approach to Parts-Based approach has been analysed and reasons for the improved performance of the column-based approach have been examined. The local frequency band approach, using the column-based method for forming feature vectors, was found to provide an absolute improvement in the HTER of 11.74% when compared to a similar baseline system (using DCTmod2 features). It was also shown that the this method is only slight worse than a state-of-the-art P2D HMM technique but with significantly fewer parameters (and complexity).

To ascertain the reason for the improved performance of the column-based approach an analysis of the variance and the performance each the frequency sub-band was performed. It was found that the most important sub-bands were $j = [1, 2, 6, 7]$ which corresponded to the vertical-only responses of the DCT (as well as the DC value of the DCT). These four sub-bands account for a significant amount of the performance difference for the column-based approach when compared to the row-based approach, however, it must also be stated that the other eleven sub-bands also contain important information when used in conjunction with these four sub-bands.

Future work will examine the applicability of this technique to a Parts-Based HMM or Bayesian Network framework. Also under consideration is how this system performs when the image normalisation technique is biased to the vertical direction since the face is normally cropped to have more vertical pixels than horizontal pixels as the face is usually considered to be longer than it is wider.

8 Acknowledgements

This work has been performed by the MOBIO project 7th Framework Research Programme of the European Union (EU), grant agreement number: 214324. The authors would like to thank the EU for the financial support and the partners within the consortium for a fruitful collaboration. For more information about the MOBIO consortium please visit http://www.mobioproject.org. The authors would also like to thank Hayley Hung, Sileye Ba and Guillaume Heusch for their input on this work.

References

1. W. Zhao, R. Chellappa, P. Phillips, A. Rosenfeld, Face recognition: A literature survey, ACM Computing Surveys 35 (4) (2003) 399–459.
2. W. W. Bledsoe, The model method in facial recognition, Technical report for Panoramic Research Inc.
3. M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
4. P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
5. F. Samaria, F. Fallside, Face identification and feature extraction using hidden markov models, Image Processing: Theory and Applications (1993) 295–298.
6. C. Sanderson, K. K. Paliwal, Fast feature extraction method for robust face verification, Electronic Letters 38 (25) (2002) 1648–1650.
7. A. M. Martinez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class, IEEE Transactions on Pattern Analysis and Machine Intelligence (2002) 748–763.
8. O. C Hamsici, A. M. Martinez, Bayes Optimality in Linear Discriminant Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence (2008) 647–657.
9. D. Tao, X. Li, X. Wu, S. J. Maybank, Geometric Mean for Subspace Selection, IEEE Transactions on Pattern Analysis and Machine Intelligence (2009) 260–274.

10. T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987. doi:http://dx.doi.org/10.1109/TPAMI.2002.1017623.
11. W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhan, Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition, In Proceedings of Tenth IEEE International Conference on Computer Vision (2005) 786–791.
12. L. Wolf, T. Hassner, Y. Taigman, Descriptor based methods in the wild, In Proceedings of ECCV Workshop on Faces in Real-Life Images Workshop (2008).
13. Y. R. G. Heusch, S. Marcel, Local binary patterns as an image preprocessing for face authentication, International Conference on Automatic Face and Gesture Recognition (2006) 9–14.
14. F. Samaria, S. Young, Hmm-based architecture for face identification, Image and Vision Computing 12 (8) (1994) 537–543.
15. A. Nefian, M. H. H. III, Hidden markov models for face recognition, Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing 5 (1998) 2721–2724.
16. F. Cardinaux, C. Sanderson, S. Marcel, Comparison of mlp and gmm classifiers for face verification on xm2vts, International Conference on Audio- and Video-based Biometric Person Authentication (2003) 1058–1059.
17. C. McCool, S. Marcel, Parts-based face verification using local frequency bands, in: Advances in Biometrics (2009) 259–268.
18. S. Lucey, T. Chen, A gmm parts based face representation for improved verification through relevance adaptation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2004, pp. 855–861.
19. B. Heisele, P. Ho, J. Wu, T. Poggio, Face recognition: component-based versus global approaches, in: Computer Vision and Image Understanding, Vol. 91, 2003, pp. 6–21.
20. K. R. Rao, P. Yip, Discrete Cosine Transform, Academic Press, 1990.
21. G. Doddington, M. Przybocki, A. Martin, D. Reynolds, The NIST speaker recognition evaluation — overview, methodology, systems, results, perspective, Speech Communication 31 (2-3) (2000) 225–254.
22. D. Reynolds, Comparison of background normalization methods for text-independent speaker verification, Proc. European Conference on Speech Communication and Technology (Eurospeech) 2 (1997) 963–966.
23. C. Lee, J. Gauvain, Bayesian adaptive learning and MAP estimation of HMM, Kluwer Academic Publishers, Boston, Massachusetts, USA, 1996, pp. 83–107.
24. J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 226–239.
25. N. Brummer, Tools for fusion and calibration of automatic speaker detection systems, `http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm`.
26. E. Bailly-Bailliere, S. Bengio, F. Bimbo, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, J.-P. Thiran, The banca database and evaluation protocol, Lecture Notes in Computer Science (2003) 625–638.
27. K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, K. Kryszczuk, J. Ng, H. Cheung, B. Tang, Face authentication competition on the BANCA database, in: Proceedings of the International Conference on Biometric Authentication (2004) 15–17.

28. K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, F. Cardinaux, S. Marcel, S. Bengio, C. Sanderson, N. Poh, Y. Rodriguez, J. Czyz, L. Vandendorpe, C. McCool, S. Lowther, S. Sridharan, V. Chandran, R. P. Palacios, E. Vidal, L. Bai, L. Shen, Y. Wang, C. Yueh-Hsuan, L. Hsien-Chang, H. Yi-Ping, A. Heinrichs, M. Muller, A. Tewes, C. von der Malsburg, R. Wurtz, Z. Wang, F. Xue, Y. Ma, Q. Yang, C. Fang, X. Ding, S. Lucey, R. Gross, H. Schneirderman, Face authentication test on the BANCA database, in: Proceedings of the 17th International Conference on Pattern Recognition (2004) 523–532.
29. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, in: Eurospeech, Vol. 4, 1997, pp. 1895–1898.
30. B. Fröba, A. Ernst, Face detection with the modified census transform, IEEE Conference on Automatic Face and Gesture Recognition (2004) 91–96.
31. G. Heusch, S. Marcel, Face authentication with Salient Local Features and Static Bayesian network, in: IEEE / IAPR Intl. Conf. On Biometrics (ICB), 2007.
32. F. Cardinaux, C. Sanderson, S. Bengio, User authentication via adapted statistical models of face images, IEEE Trans. Signal Processing (2006) 361-373.