Face Video Competition

Norman Poh[1], Chi Ho Chan[1], Josef Kittler[1], Sébastien Marcel[2], Christopher Mc Cool[2], Enrique Argones Rúa[3], José Luis Alba Castro[3], Mauricio Villegas[4], Roberto Paredes[4], Vitomir Štruc[5], Nikola Pavešić[5], Albert Ali Salah[6], Hui Fang[7], and Nicholas Costen[7]

[1] CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK
normanpoh@ieee.org, c.chan@surrey.ac.uk, j.kittler@surrey.ac.uk
[2] Idiap research institute, Marconi 19, Martigny, CH
[3] Signal Technologies Group, Signal Theory and Communications Dept., University of Vigo, 36310, Spain
[4] Universidad Politécnica de Valencia, Instituto Tecnológico de Informática, Camino de Vera s/n, 46022 Valencia (Spain)
[5] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia
[6] CWI, Science Park 123, 1098 XG Amsterdam
[7] Department of Computing and Mathematics, Manchester Metropolitan University, UK M1 5GD

Abstract. Person recognition using facial features, e.g., mug-shot images, has long been used in identity documents. However, due to the widespread use of web-cams and mobile devices embedded with a camera, it is now possible to realise facial video recognition, rather than resorting to just still images. In fact, facial video recognition offers many advantages over still image recognition; these include the potential of boosting the system accuracy and deterring spoof attacks. This paper presents the first known benchmarking effort of person identity verification using facial video data. The evaluation involves 18 systems submitted by seven academic institutes.

1 Introduction

With an increasing number of mobile devices with built-in web-cams, e.g., PDA, mobile phones and laptops, face is arguably the most widely accepted means of person verification. However, the biometric authentication task based on face images acquired by a mobile device in an uncontrolled environment is very challenging. One way to boost the face verification performance is to use multiple samples.

Previous attempts at assessing the performance of face verification algorithms have been restricted to matching still images, e.g., the three FERET evaluations[1] (1994, 1995 and 1996), the face recognition vendor tests (FRVTs 2000, 2002 and 2006)[2], and assessment on XM2VTS and BANCA databases [1, 2]. The well known Face Recognition

---

[1] http://www.itl.nist.gov/iad/humanid/feret/feret_master.html
[2] http://www.frvt.org

Grand Challenge [3] includes queries with multiple still images but this is far from the vast amount of data available in video matching.

The evaluation exercise presented here is the first known effort in assessing *video-to-video* matching, i.e., in both enrolment and verification phases, the data captured is in the form of video sequence. This is different from still-image-to-video matching, one of the evaluation scenarios currently examined by the NIST Multiple Biometric Grand Challenge[3] (MBGC). Note that NIST MBGC aims at "portal application" where the task is to verify the identity of person as he/she walks through an access control check point. The video-to-video matching adopted here has a slightly different application, with a focus on mobile devices, where a sequence of unconstrained (talking) face images can be expected.

The video-to-video face verification assessment has several objectives, among which are:

– to promote the development of algorithms for analysing video sequences (e.g., exploring the talking face dynamics)
– to assess the merit of multi-template face representation
– to compare whether early integration (e.g., feature-level fusion) is better than late integration (e.g., decision-level fusion) in dealing with sequences of query images.

2   Database, Protocols, Facial Video Annotations

Towards the above goal, we have opted to use the publicly available BANCA database [4][4]. It has a collection of face and voice biometric traits of up to 260 persons in 5 different languages, but only the English subset is used here. It contains a total of 52 persons; 26 females and 26 males. The 52 persons are further divided into two sets of users, which are called g1 and g2, respectively. Each set (g1 or g2) is designed to be balanced in gender, i.e., having 13 males and 13 females. According to the experimental protocols reported in [4], when g1 is used as a development set (to build the user's template/model), g2 is used as an evaluation set. Their roles are then switched. This corresponds to a two-fold cross-validation procedure.

The BANCA database was designed to examine matching under the same recording conditions (as the enrolment session) and two different challenging conditions: recording under a noisy (adverse) environment and with a degraded device. In each of the three conditions, four recordings were performed. The clean conditions apply to sessions 1–4; adversed conditions to sessions 5–8; and degraded conditions to sessions 9–12. There are altogether seven experimental protocols specifying the sessions to be used for enrolment and for testing in an exhaustive manner. In this face video recognition evaluation, we focused on two protocols, namely the match controlled (Mc) and unmatched adversed (Ua) protocols. The first protocol was intended as a vehicle to design and tune their face verification systems. The second protocol aims at testing the systems under more realistic and challenging conditions.

---

[3] http://face.nist.gov/mbgc
[4] http://www.ee.surrey.ac.uk/CVSSP/banca

In the Mc protocol, session 1 data is used for enrolment whereas the data from sessions 2–4 are reserved for testing. In the Ua protocol, the session 1 data again is used for enrolment but the test data is taken from session 5–8 (recorded under adversed conditions). The ICB2009 face video competition was thus naturally carried out in two rounds, with the first round focusing on the Mc protocol and the second round on the Ua protocol.

In order to be consistent with the previous BANCA evaluations [1, 2], we also divided a query video sequence into 5 chunks, each containing 50 frames for convenience; the remaining frames were simply not used.

In order to standardise the evaluation, we provided a pair of eye coordinates, based on the face detector provided by the OmniPerception's SDK[5]. However, the participants could use their own face detectors. For each image in a video sequence, the SDK also annotated the following quality measurements. Note that the entire processes from detection to annotation were done automatically. No effort was made to fine tune the system parameters, and in consequence, some imperfectly cropped images were observed. The image quality measures assessed.

1. Overall reliability
2. Brightness
3. Contrast
4. Focus
5. Bit per pixel
6. Spatial resolution (between eyes)
7. Illumination
8. Background uniformity
9. Background brightness
10. Reflection
11. Presence of glasses
12. In-plane rotation
13. In-depth rotation
14. Frontalness

In the above list, "frontalness" quantifies the degree of similarity of a query image to a typical frontal (mug-shot) face image. The overall reliability is a compounded quality measure obtained by combining the remaining quality measures. Two categories of quality measures can be distinguished: face-specific or generic. The face-specific ones strongly depend on the result of face detection, i.e., frontalness, rotation, reflection, between-eyes spatial resolution in pixels, and the degree of background uniformity (calculated from the remaining area of a cropped face image). The generic ones are defined by the MPEG standards. All the annotation data (including eye coordinates and quality measures) has been published on the website "http://face.ee.surrey.ac".

A preliminary analysis shows that when the frontalness measure is 100%, the detected face is always frontal. On the other hand, any value less than 100% does indeed suggest an imperfect face detection, or else a non-ideal (non-frontal) pose.

3 System Descriptions

3.1 University of Vigo (UVigo)

The video-based face verification system submitted by the University of Vigo for the *pre-registered test* uses the annotated eyes coordinates in order to set the eyes position in the same coordinates for all the faces, using simple rotation and scaling op-

---

[5] http://www.omniperception.com

erations. Then a two-step illumination normalisation is performed on the geometrically normalised faces. First step is the anisotropic illumination normalisation described in [5]. Second step is a local mean subtraction. We denote the video frame sequence as $\mathcal{V} = \{\mathcal{I}^{\mathcal{V},1}, \ldots, \mathcal{I}^{\mathcal{V},N_\mathcal{V}}\}$, where $\mathcal{I}^{\mathcal{V},i}$ represents the ith frame of video $\mathcal{V}$, and $N_\mathcal{V}$ is the number of frames in the video. Gabor jets [6], $\mathcal{J}_k^{\mathcal{V},i} = \left\{ a_{k,0}^{\mathcal{V},i}, \ldots, a_{k,39}^{\mathcal{V},i} \right\}$ are extracted from the ith frame (magnitude of the responses of Gabor filters with 5 scales and 8 orientations, encoded in the second subindex) at fixed points, k, along a rectangular grid of dimensions $D = 10 \times 10$ superimposed on each normalised face image. Frame $\mathcal{I}^{\mathcal{V},i}$ is characterised by all the extracted Gabor jets $\{\mathcal{J}_1^{\mathcal{V},i}, \ldots, \mathcal{J}_D^{\mathcal{V},i}\}$.

GMM-UBM verification paradigm is adapted to video-based verification. Gabor jets extracted from each grid location are divided in $N_S = 2$ separate vectors $\mathbf{x}_{k,m}^i$ constituted by sets of subjets: $\left\{ a_{k,l}^{\mathcal{V},i} \mid \mod (l, N_S) = m \right\}$, where i is the frame index, k is the grid point index, $l \in \{0, \ldots, 39\}$ is the filter index and $m \in \{0, 1\}$ is the subset index. 64 mixtures UBMs are trained for both vectors $\mathbf{x}_{k,0}^{\mathcal{V},i}$ and $\mathbf{x}_{k,1}^{\mathcal{V},i}$ at each grid location. Number of subjets N_S was fixed as a trade-off between discrimination capability and dimensionality. First subset includes the coefficients from filters with an even filter index $(l \mid \mod (l, N_S) = 0)$, and second subset includes the coefficients with an odd filter index $(l \mid \mod (l, N_S) = 1)$. Independence between the subjets from each node is assumed in order to avoid the curse of dimensionality in the UBM training. This assumption leads us to independent training for each subjet at each grid location. The nth UBM probability density function $f_{UBM,n}(\cdot)$, where $n \in \{0, \ldots, 199\}$, is estimated using LBG [7] initialisation and the EM algorithm. Gaussian mixtures are constrained to have diagonal covariance matrices. Input vectors for this training process are $\mathbf{x}_{\lfloor \frac{n}{2} \rfloor, \mod (n,2)}^{\mathcal{V},i}$, where $\mathcal{V} \in \mathcal{WM}$, i.e., the world model set videos. Grid node is indexed by $\lfloor \frac{n}{2} \rfloor$, which is the integer part of $\frac{n}{2}$. Subjet set is indexed by $\mod (n, 2)$.

$f_{UBM,n}(\cdot)$ is then adapted to the corresponding vectors obtained from the user u enrolment video by means of the MAP technique [8], obtaining user model *pdf* $f_{u,n}(\cdot)$. The verification score for the video $\mathcal{V}$ and claimed identity u is computed as the following log-likelihood ratio [9]:

$$s_{\mathcal{V},u} = \log \left(\prod_{i=1}^{N_\mathcal{V}} \prod_{n=0}^{2D-1} \frac{f_{u,n}\left(\mathbf{x}_{\lfloor \frac{n}{2} \rfloor, \mod (n,2)}^{\mathcal{V},i} \right)}{f_{UBM,n}\left(\mathbf{x}_{\lfloor \frac{n}{2} \rfloor, \mod (n,2)}^{\mathcal{V},i} \right)} \right) \qquad (1)$$

3.2 IDIAP

Two types of systems were submitted by Idiap, these being holistic (PCA and PCAxLDA) and parts-based (GMM and HMM). In all cases the world model (for PCA, LDA, GMM and HMM world) are computed on the world model data defined by the provided protocol. This results in one specific world model for each group of clients g1 and g2.

All of the face verification systems use the automatic annotations (eye centres and frontalness) provided by the OmniPerception SDK. More precisely, the eye-centre coordinates are used to extract the 10-best faces from each video according to the frontalness measure.

Geometric and Photometric Normalisation: For all systems the face is first geometrically normalised as described in [10] rotated to align the eye coordinates, then cropped and scaled to a size of 64×80 (width $\times$ height pixels). The face image is then photometrically normalised using two methods: (1) standard Histogram Equalisation (HEQ) as in [10] or (2) a pre-processing based on Local Binary Patterns (LBP) as proposed in [11].

Feature Extraction: The two holistic systems are based on well-known dimensionality reduction methods, namely PCA and PCAxLDA. For PCA dimensionality reduction was achieved by retaining 96% of the variance of the vector space. This resulted in 181 and 180 dimensions being retained for groups g1 and g2 respectively, instead of the 5120 dimensions (64×80 pixels). Face images projected in the PCA subspace are then further projected into an LDA subspace (PCAxLDA), where only 55 dimensions are retained for both group.

The parts-based approaches decompose the face image into blocks and then use statistical models such as GMMs or HMMs. For each block the DCT (2D DCT) or its DCTmod2 variant is computed, as described in [10], resulting in one feature vector per block. An extension to these methods is provided where the 2D coordinate (xy) of each block is appended to its corresponding feature vector, this was done to incorporate spatial information.

Classification: Classification for the holistic methods, PCA and PCAxLDA, is examined using three different similarity measures, these being: Pearson, Normalised Correlation and Standard Correlation. Classification for the DCT and DCTmod2 features is performed using GMMs and HMMs as described in [12].

3.3 Manchester Metropolitan University (MMU)

The General Group-wise Registration (GGR) algorithm is used to find correspondences across the set of images. This shares similar ideas with others [13, 14] which seek to model sets efficiently, representing the image set and iteratively fitting this model to each image. The implementation of GGR [15] proceeds through a number of stages. Firstly, one image is selected as a reference template and all other images are registered using a traditional template match. Next, a statistical shape and texture model is build to represent the image set. Each image is represented in the model and the correspondences are refined by minimising a cost function. Finally the statistical models are updated and the fitting repeated until convergence.

The model used here is a simple mean shape and texture built by warping all the faces to the mean shape using a triangular Delauney mesh. A coarse-to-fine deformation scheme is applied to increase the number of control points and optimise their position. In the final iterations, the points are moved individually to minimise the cost. The cost function includes both shape and texture parts,

$$E = \lambda \sum_i \left(c - \frac{0.5 \| d_i - (\Delta d_i + d_{neig}) \|}{\sigma_s^2} \right) - \frac{|r|}{\sigma_r} \tag{2}$$

where r is the residue between the model and the current image after deformation, σ_r and σ_s are the standard deviations of the residue and shape, c is a constant, d_i is the

position of the ith control point, d_{neig} is the average of the positions of the neighbourhood around point i and Δd_i represents the offset of the point from the average mean shape.

A set of 68 sparse correspondent feature points are initialised manually on the mean image of the image set. When GGR has found the dense correspondences across the images, all the sparse feature points are warped to each image using the triangulation mesh. Once the correspondences have been found for the ensemble images, a combined Appearance Model [16] is built for each individual and the points are encoded on it. Pixels defined by the GGR points as part of the face are warped to a standard shape, ensuring that the image-wise and face-wise coordinates of images are equivalent. Because of the size of the database, representative frames are selected for each ensemble subject using k-means clustering of their encoding on their individual model to give approximately 10 groups (one for each 50 frames). The frame most representative of each group is then selected and used to build both an Appearance Model of the full ensemble. This provides a single 48-dimensional vector which encodes both the shape and grey-level aspects of the face for a given frame. It models the whole of the inner tile face, using 5000 grey scale samples (and the 68 feature points), describing 98% of the ensemble variation, but without any photometric normalization.

In the same sequence, regardless of parameter change due to different poses, lighting and expressions, the identity can be expected to be constant. However, in this case, the model will encode (even after averaging) both identity and non-identity variation. To remove the latter, a Linear Discriminate Analysis subspace [17] is used. This provides a subspace which maximises variation between individuals and minimises that within them. Each frame in a gallery or probe sequence is projected onto this subspace, before taking the mean of the identity parameters and assessing similarity with another sequence,

$$S_c = \frac{\bar{\mathbf{d}}_1}{|\bar{\mathbf{d}}_1|} \cdot \frac{\bar{\mathbf{d}}_2}{|\bar{\mathbf{d}}_2|}. \tag{3}$$

where S_c is the correlation-based similarity and $\bar{\mathbf{d}}$ represents the mean LDA parameters of a sequence. Behavioural consistency is a possible addition which improves discrimination performance within this framework when longer probe sequences can be exploited [18]. However, it is not useful in this short-sequence situation.

3.4 Universidad Politécnica de Valencia (UPV)

The approach we adopted for the verification of a sequence of face images was as follows. The first NA frames from the input video are analysed using the quality measures and the best NQ frames are selected. Afterwards a verification score is obtained for each of the selected frames using the local feature algorithm [19–21]. The final verification score is the average of the scores for each of the selected frames.

The parameters NA and NQ were kept fixed for all of the videos of the same scenario. For each scenario NA and NQ were varied and their value was chosen making a compromise between the performance of the algorithm on the development set and the computational cost. For the matched controlled scenario (Mc) the chosen parameters were NA=10 and NQ=5, and for the unmatched adverse scenario (Ua) the parameters

were NA=20 and NQ=6. The number of frames used to build the user models was NT=5 for both scenarios.

For each video frame several quality measures were supplied. Therefore in order to choose the best frames the quality measures were fused into a single quality value, and the frames with highest quality were selected. To fuse the quality measures we trained a classifier of good and bad frames and used the posterior probability of being a good frame as a quality measure. The classifier used was the nearest neighbour in a discriminative subspace trained using the LDPP algorithm [22]. To train this classifier the quality values of the frames of the background model videos were used, and each frame was labelled as being good or bad based on the result of face identification using the local feature algorithm [21].

In the local feature face verification algorithm, from a face image several feature vectors are extracted. Each feature is obtained using only a small region of the image, and the features are extracted all over the image at equal overlapping intervals. Given a test image, the nearest neighbours of its local features are found among the feature vectors from the background model and the user model. The verification score is simply the number of nearest neighbours from the user model divided by the number of extracted local features. For further details refer to [19, 20]. The parameters of the algorithm were chosen based on previous research and were not adjusted to minimise the error rates of the scenarios. In the algorithm grey scale images were used, the faces were cropped to a size of 64×64 pixels, and the local features were of size 9×9 extracted every 2 pixels.

3.5   University of Ljubljana (UniLJ)

The UniLj face recognition technique is based on a feature extraction approach which exploits Gabor features and a combination of linear and non-linear (kernel) subspace projection techniques. The training, enrolment and test stages of the employed approach can be summarised as follows:

The training stage: Facial images from various sources (such as BANCAs world model, the XM2VTS, the AR, the FERET, the YaleB and the FRGC databases) were gathered to form a large image set that was employed for training. This training set was subjected to a pre-processing procedure which first extracted the facial regions from the images based on manually marked eye-centre locations, then geometrically aligned and ultimately photometrically normalised the facial regions by means of zero-mean-and-unit-variance normalisation and a subsequent histogram equalisation step. The normalised facial images cropped to a standard size of 100×100 pixels were then filtered with a family of Gabor kernels with 5 scales and 8 orientations. From the complex filter responses features encoding Gabor-magnitude as well as Gabor-phase information [23] were derived and concatenated to form the final Gabor feature vectors. Next, the constructed feature vectors were partitioned into a number of groups and for each a non-linear subspace was computed based on the multiclass kernel Fisher analysis (KFA) [24]. The Gabor feature vectors from all groups were projected into all created KFA subspaces and the resulting vectors were then subjected to a Principal Component Analysis (PCA)[25] to further reduce their dimensionality.

The enrolment stage: Using the provided quality measures associated with the video sequences of the BANCA database a small number of images was chosen from each

enrolment video of a given subject[6]. These images were processed in the same manner as the training images, i.e., feature vectors were extracted from each image by means of Gabor filtering and subsequent subspace projections. The processed images served as the foundation for computing the client templates - the mean feature vectors.

The test stage: From each test video sequence a small subset of randomly selected frames which passed our quality check (using the same quality measures as in the enrolment stage) were processed to extract the facial features. The resulting feature vectors were then matched with the template corresponding to the claimed identity using the nearest neighbour classifier and the whitened cosine similarity detailed in a recently proposed correction scheme [26]. Depending on the cumulative value of the matching score, a decision regarding the validity of the identity claim was made in the end.

3.6 Centrum voor Wiskunde en Informatica (CWI)

In CWI approach, the ground truth for eye-locations is used to crop and rectify the face area at each frame. Each cropped frame is then normalised to 64×64, and split into 8×8 windows, from which 2D-DCT coefficients are extracted [27]. Each window supplies nine coefficients in zig-zag fashion, bar the DC value, which are then concatenated into the final feature representation for the face. During testing, DCT coefficients are extracted from a face localised in a given frame and the similarity of vectors i and j is computed as:

$$S(i, j) = \frac{i.j}{|i||j|}. \tag{4}$$

During training, 15-means clustering is applied to DCT features extracted from the training images of each person, and cluster means are selected as templates. Our experimental results suggest that using a mixture model for the genuine class and one model for the generic impostor class, combined with a likelihood ratio based decision is suboptimal to the DCT-based method [28]. From each video frame, a number of relevant quality measures (i.e. bits per pixel, spatial resolution, illumination, background brightness, rotation in plane, and frontalness) are summed and a ranked list is prepared. The ranked images are evaluated in succession, and a pre-selected distance threshold is selected for authentication. If the similarity score is above this threshold (0.75), it is reported as the score. Else, the next best ranked frame is evaluated, up to eight frames per sequence. The maximum similarity score is returned as the final score. Since there is no early stopping for rejecting claims, the ROC-curves produced for this method do not fully reflect the possible operation range of the algorithm. The pre-set similarity threshold is a second parameter (the first being the final score threshold for acceptance) that controls the system output.

Cwi's submission has four variations: depending on the dichotomies: system complexity, i.e., Cheap (C) versus Exepnsive (E); and strategy for choosing the query samples, i.e, random (r) versus quality-based (q). For the so-called cheap (resp. expensive) version, 5 (resp. 15) templates are used for each client and only 4 (resp. up to 8) images are used for query. Increasing the number of templates for each gallery subject leads

---

[6] It has to be noted that only the quality measures corresponding to the overall reliability of the face detector and the spatial resolution were considered for the frame selection process.

to diminishing returns. Since the DCT feature dimensionality is high than the number of available frames, an automatic model selection approach usually justifies only a few clusters. During our simulations, we contrasted random selection of frames vs. quality-based selection of frames. We observed that higher quality faces produced both higher genuine similarity scores, and higher impostor scores, leading to greater false accept rates.

3.7 Summary

The submitted face verification systems can be categorised according to whether they are image-set-based or frame-based (comparison) approach. In the image-set based approach, a video sequence is analysed and treated as a set of images. When comparing two video sequences, this approach, in essence, compares two *sets* of images. On the other hand, the frame-based approach directly establishes similarity between two images, each obtained from their respective video sequence. If there are P and Q images in both sequences, there will be at most PQ similarity scores. The frame-based approach would select, or otherwise combine these similarity scores to obtain a final similarity score. Among the systems, only the MMU system belongs to the image-set based approach, while the remaining systems are the frame-based approach.

Face verification systems can also be further distinguished by the way a face image is treated, i.e, either holistic or local (parts-based) appearance approach. In the former, the entire (often cropped) image is considered as input to the face classifier. In the latter, the face images are divided into (sometimes overlapping) parts which are then treated separately by a classifier. Table 1 summarises the systems by this categorisation. Principal component analysis (PCA), or Eigenface, and local discriminant analysis (LDA), or Fisherface, are perhaps the most representative (and popular) examples of the holistic approach due to the pioneer work of Turk and Pentland [29]. Many of these systems were submitted by IDIAP as baseline systems, tested on the Mc protocol (and not the Ua protocol). Recent face verification research has been dominated by the local appearance approach, as exemplified by *most* of the submissions in this competition.

4   Evaluation Metrics

We use two types of curves in order to compare the performance: the Detection Error Trade-off (DET) curve [30] and the Expected Performance Curve (EPC) [31]. A DET curve is actually a Receiver Operator Curve (ROC) curve plotted on a scale defined by the inverse of a cumulative Gaussian density function, but otherwise similar in all aspects. We have opted to use EPC because it has been pointed out in [31] that two DET curves resulting from two systems are not comparable. This is because such comparison does not take into account how the decision thresholds are selected. EPC turns out to be able to make such comparison possible. Furthermore, the performance across different data sets, resulting in several EPCs, can be merged into a single EPC [32]. Although reporting performance in EPC is more meaningful than DET as far as performance comparison is concerned, it is relatively new and has not gained a widespread

	Systems	Pre-processing	Face rep.	Feature Extraction	Classifier	Quality measure used	Process all images
Holistic	idiap-pca-pearson	HEQ		PCA	Pearson	No	Yes
	idiap-pca-nc	HEQ		PCA	NC	No	Yes
	idiap-pca-cor	HEQ		PCA	StdCor	No	Yes
	idiap-lda-pearson	HEQ		PCAxLDA	Pearson	No	Yes
	idiap-lda-nc	HEQ		PCAxLDA	NC	No	Yes
	idiap-lda-cor	HEQ		PCAxLDA	StdCor	No	Yes
	mmu		AM	LDA	Avg(NC)	No	Yes
Local	idiap-dcthmmt-v1	HEQ		DCT	HMM	No	Yes
	idiap-dcthmmt-v2	HEQ		DCT	HMM	No	Yes
	idiap-dctgmm	HEQ		DCTmod2+xy	GMM	No	Yes
	idiap-LBP-dctgmm		LBP	DCTmod2+xy	GMM	No	Yes
	cwi-Cq			DCT	Max(NC)		Yes
	cwi-Eq			DCT	Max(NC)		Yes
	cwi-Cr			DCT	Max(NC)		No
	cwi-Er			DCT	Max(NC)		No
	upv	Local-HEQ	LF	PCA	Avg(KNN)	Yes	No
	uni-lj	ZMUV + HEQ	Gb2	KDA+PCA	WNC	Yes	No
	uvigo	Ani	Gb1		GMM	Yes	No

Table 1. Overview of the submitted face verification systems.

The following keys are used: AM = Appearance model, ZMUV = zero mean and unit-variance, Ani = Anisotropic+local mean subtraction, LF = Local feature Gb1 = Gabor(magnitude) Gb2 = Gabor(phase+magnitude, NC = Normalised correlation, WNC = Sum of whitened NC Note: OmniPerception's face detector was used by all systems.

acceptance in the biometric community. As such, we shall also report performance in DET curves, but using only a subset of operating points.

The EPC curve, however, is less convenient to use because it requires two sets of match scores, one used for tuning the threshold (for a given operating cost), and the other used for assessing the performance. In our context, with the two-fold cross-validation defined on the database (as determined by g1 and g2), these two match scores can be conveniently used.

According to [31], one possible, and often used criterion is the weighted error rate (WER), defined by:

$$\text{WER}(\beta, \Delta) = \beta \, \text{FAR}(\Delta) + (1 - \beta) \, \text{FRR}(\Delta), \tag{5}$$

where FAR is the false acceptance rate, FRR is the false rejection rate at a given threshold Δ and $\beta \in [0, 1]$ is a user-specified coefficient which balances FAR and FRR. The WER criterion generalises the criterion used in the annual NIST's speaker evaluation [33] as well as the three operating points used in the past face verification competitions on the BANCA database [1, 2]. In particular the following three coefficients of β are used:

$$\beta = \frac{1}{1 + R} \text{ for } R = \{0.1, 1, 10\}$$

which yields approximately $\beta = \{0.9, 0.5, 0.1\}$, respectively.

The procedure to calculate an EPC is as follows: Use g1 to generate the development match scores; and g2, the evaluation counterpart. For each chosen β, the development score set is used to minimise (5) in order to obtain an operational threshold. This threshold is then applied to the evaluation set in order to obtain the final pair of false acceptance rate (FAR) and false rejection rate (FRR). The EPC curve simply plots half total error rate (HTER) versus β, where HTER is the average of FAR and FRR. Alternatively, the generalisation performance can also be reported in WER (as done in the previous BANCA face competitions). To plot the corresponding DET curve, we use the pair of FAR and FRR of all the operating points, as determined by β. Note that this DET curve is a *subset* (in fact discrete version) of a conventional continuous DET curve because the latter is plotted from continuous empirical functions of FAR and FRR. By plotting the discrete version of the DET curve, we establish a *direct correspondence* between EPC and DET, satisfying both camps of biometric practitioners, while retaining the advantage of EPC which makes performance comparison between systems less biased.

5   Results

The DET curves of all submitted systems for the g1 and g2 data sets, as well as for the Mc and Ua protocols, are shown in Figure 1. By merging the results from g1 and g2, we plotted the EPCs for Mc and Ua in Figure 2 (plotting β versus HTER). To be consistent with the previous published BANCA evaluations [1, 2], we also listed the individual g1 and g2 performance, in terms of WER, in Table 2 for the Mc protocol and in Table 3 for the Ua protocol.

The following observations can be made:

– **degradation of performance under adversed conditions**: It is obvious from Figure 2 that all systems systematically degrade in performance under adversed conditions.
– **holistic vs. local appearance methods**: From Figure 1(a) and (b) as well as Figure 2(a), we observe that the performance of the holistic appearance methods (PCA and LDA) is worse than that of the local appearance methods, except for the CWI classifier (where photometric normalisation was not performed). Thus, we can expect that the performance of CWI to be similar to the performance of other local appearance methods in the raw image space, such as idiap-dctgmm, idiap-dcthmmt-v2 and upv if photometric normalisation were to be performed.
– **still vs. video comparison**: Among the submitted systems, only IDIAP's DCT-HMM system was involved in the previously reported results for the Mc protocol [1] which was based on 5 still images taken from a video sequence (as opposed to five video chunks as done here). The results for this classifier are shown in Table 2 (comparing rows 1-2 with row 3). In theory, one would expect the classifier tested on video sequence to be better than still images. Unfortunately, such conclusion cannot be made except for $R = 0.1$.
– **Pre-processing**: In dctgmm methods, the performance of applying HEQ is better than that of applying LBP as a pre-processing method for Mc protocol. However,

the case is reversed for Ua protocol because HEQ enhances shadows while LBP features are invariant to such monotonic transformation (in relation to the neighbourhood pixels cast under shadows). In other words, the selection of the pre-processing methods should be dependent on the environmental conditions.

– **Sample size**: Cwi's submission has four variations: depending on the dichotomies: system complexity, i.e., Cheap (C) versus Exepnsive (E); and strategy for choosing the query samples, i.e, random (r) versus quality-based (q) (see Section 3.6). Two observations can be noted: First, the performance of cwi-Eq and cwi-Er are better than that of cwi-Cq and cwi-Cr. Second, using *more* template and query features can improve the cwi system. A rigorous and systematic design of experiments is still needed to find out the usefulness of the provided quality measures, and more importantly, the most effective ways of using such auxilliary information. This is a challenging problem for two reasons. First, not all 14 quality measures provided are relevant to a face matching algorithm, e.g., an algorithm that is robust to illumination changes would, in principle, be invariant to some photometric measures used here (brightness, contrast, etc). This implies that a quality measure selection strategy is needed. Second, quality measures are themselves not discriminatory for distinguishing subjects but discriminatory in distinguishing environmental conditions.

– **Multi resolution Contrast Information**: The best algorithm of this competition for MC protocol is UVigo where the WER at R=1 is 0.77% for G1 and 2.31% for G2. For UA protocol, the best algorithm is uni-lj where WER at R=1 is 8.78% for G1 and 6.99% for G2. In fact, the performance of these two systems is very close but uni-lj is slightly better overall as the average of WER at different R is 3.96% for G1 and 3.98% for G2, while the result of UVigo is 3.97% for G1 and 4.34% for G2. The success of these two algorithms derives from the use of multi resolution contrast information.

6 Discussion and Future Evaluation

Because the target application scenario of this assessment is on mobile devices, computational resources are crucial. For this reason, when benchmarking a face verification algorithm, the cost of computation has to be considered. For instance, a fast and light algorithm, capable of processing all images in a sequence, may be preferred over an extremely accurate algorithm only capable of processing a few selected images in a sequence. However, the former algorithm may be able to achieve better performance since it can process a much larger number of images within the same time limit and memory requirement. The above scenario highlights that the performance of two algorithms cannot be compared on equal grounds, unless both use comparable computation costs, taking the time, memory and computational resources into consideration.

The current evaluation has not taken this cost factor into consideration, but this will be carried out in future. The idea is to request each participant to run a benchmarking program, executable in any operating system. The time registered by the program will be used as a *standard unit time* for the participant's system. Thus the time to process

Table 2. Performance of g1 and g2 based on the Mc protocol using video sequences

systems	WER (%)					
	$R = 0.1$		$R = 1$		$R = 10$	
	G1	G2	G1	G2	G1	G2
idiap-dcthmm†	7.52	4.90	5.45	0.64	2.56	0.12
idiap-dcthmm‡	7.78	3.76	5.13	2.08	1.17	2.74
idiap-dcthmmT-v2	1.34	2.03	4.20	4.29	1.92	3.93
idiap-dctgmm	0.82	5.14	1.12	5.48	0.82	1.96
idiap-LBP-dctgmm	0.75	6.26	1.63	7.37	1.22	2.77
uvigo	1.05	0.42	0.77	2.31	0.45	4.20
mmu	5.94	2.14	9.84	9.07	5.21	9.64
upv	3.01	1.81	5.06	7.50	4.00	5.86
cwi-Cq	3.80	9.84	14.20	18.14	7.28	12.76
cwi-Cr	3.66	11.72	13.14	18.69	6.49	12.40
cwi-Eq	2.84	9.51	10.90	16.83	6.32	11.49
cwi-Er	2.59	9.73	9.87	16.63	6.25	11.68
uni-lj	0.86	2.18	2.34	4.81	2.32	2.02

†: Experimental results on *still* images, taken from [1] with automatic localisation. ‡: Similar to †, except with manual localisation.

a video file for a participant, for instance, will be reported in terms of multiples (or fractions) of the participant's standard unit time.

7   Conclusions

This paper presents a comparison of video face verification algorithms on BANCA database. Eighteen different video-based verification algorithms from a variety of academic institutions participated in this competition. The results show that the performance of the local appearance methods is better than that of the holistic appearance methods. Secondly, using more query and selected template features to measure similarity improve the system performance. Finally, the best algorithm in this competition clearly shows that multi resolution contrast information is important for face recognition.

8   Acknowledgement

Table 3. Performance of g1 and g2 based on the Ua protocol

systems	WER (%)					
	$R = 0.1$		$R = 1$		$R = 10$	
	G1	G2	G1	G2	G1	G2
idiap-dcthmmT-v2	8.52	8.66	18.65	17.08	6.37	12.61
idiap-dctgmm	9.10	11.03	27.31	24.49	10.54	13.31
idiap-LBP-dctgmm	8.34	10.08	23.85	24.94	10.58	11.47
uvigo	2.81	5.06	8.75	9.49	10.00	4.55
mmu	13.61	9.88	27.72	31.96	10.97	18.21
upv	4.00	6.60	9.29	13.46	3.98	11.45
cwi-Cq	9.06	14.18	28.08	34.46	16.54	11.19
cwi-Cr	9.43	11.41	26.60	31.79	14.50	11.79
cwi-Eq	8.72	14.73	24.23	27.98	16.50	8.48
cwi-Er	8.00	12.23	21.38	24.29	12.86	8.80
uni-lj	4.67	3.03	8.78	6.99	4.78	4.83

References

1. 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. Czyz, L. Vandendorpe, J. Ng, H. Cheung, and B. Tang, "Face authentication competition on the banca database," in *Intl. Conf. Biometric Authentication*, 2004, pp. 8–15.
2. K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostin, 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-L. Shen, Y. Wang, Chiang Yueh-Hsuan, H-C. Liu, Y-P. Hung, A. Heinrichs, M. Muller, A. Tewes, C. vd Malsburg, R. Wurtz, Zg. Wang, Feng Xue, Yong Ma, Qiong Yang, Chi Fang, Xq. Ding, S. Lucey, R. Goss, , and H. Schneiderman, "Face authentication test on the banca database," in *Int'l Conf. Pattern Recognition (ICPR)*, 2004, vol. 4, pp. 523–532.
3. P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 947–954.
4. E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *LNCS 2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA 2003*. 2003, Springer-Verlag.
5. Ralph Gross and Vladimir Brajovic, "An Image Preprocessing Algorithm for Illumination Invariant Face Recognition," in *Audio- and Video-Based Biometric Person Authentication*, Springer Berlin / Heidelberg, Ed. June 2003, vol. 2688/2003 of *Lecture Notes in Computer Science*, pp. 10 – 18, Springer.
6. Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775 – 779, July 1997.
7. Yoseph. Linde, Andrés. Buzo, and Robert M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transaction on Communications*, vol. 28, no. 1, pp. 84 – 94, January 1980.

8. Jean-Luc Gauvain and Chin Hui Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291 – 298, April 1994.

9. José Luis Alba Castro, Daniel González Jiménez, Enrique Argones Rúa, Elisardo González Agulla, and Enrique Otero Muras, "Pose-corrected Face Processing on Video Sequences for Webcam-based Remote Biometric Authentication," *Journal of Electronic Imaging*, vol. 1, no. 17, January 2008.

10. F. Cardinaux, C. Sanderson, and S. Marcel, "Comparison of MLP and GMM classifiers for face verification on XM2VTS," in *Intl Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA)*. 2003, Springer.

11. G. Heusch, Y. Rodriguez, and S. Marcel, "Local Binary Patterns as an Image Preprocessing for Face Authentication," in *IEEE Intl Conf. on Automatic Face and Gesture Recognition (AFGR)*, 2006, pp. 9–14.

12. F. Cardinaux, C. Sanderson, and S. Bengio, "User Authentication via Adapted Statistical Models of Face Images," *IEEE Trans. on Signal Processing*, vol. 54, no. 1, pp. 361–373, 2005.

13. S. Baker, I. Matthews, and J. Schneider, "Automatic construction of active appearance models as an image coding problem," *IEEE Transations on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1380–1384, 2004.

14. T. F Cootes, S. Marsland, C. J. Twining, Smith K., and Taylor C. J., "Groupwise diffeomorphic non-rigid registration for automatic model building," in *Proc. ECCV*, 2004, pp. 316–327.

15. T. F. Cootes, C. J. Twining, V. Petrovic, R. Schestowitz, and C. J. Taylor, "Groupwise construction of appearance model using piece-wise affine deformations," in *Proc. BMVC*, 2005, pp. 879–888.

16. T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transations on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

17. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transations on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, 1997.

18. H. Fang and N. Costen, "Behavioral consistency extraction for face verification," in *Proc. COST 2102 2nd Conference*, 2008.

19. M. Villegas and R. Paredes, "Illumination invariance for local feature face recognition," in *1st Spanish Workshop on Biometrics*, Girona (Spain), June 2007.

20. M. Villegas, R. Paredes, A. Juan, and E. Vidal, "Face verification on color images using local features," *Computer Vision and Pattern Recognition Workshops, 2008. CVPR Workshops 2008. IEEE Computer Society Conference on*, pp. 1–6, June 2008.

21. R. Paredes, J. C. Pérez, A. Juan, and E. Vidal, "Local Representations and a direct Voting Scheme for Face Recognition," in *Proc. of the Workshop on Pattern Recognition in Information Systems (PRIS 01)*, Setúbal (Portugal), July 2001.

22. M. Villegas and R. Paredes, "Simultaneous learning of a discriminative projection and prototypes for nearest-neighbor classification," *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, 2008.

23. V. Štruc, B. Vesnicer, and N. Pavešić, "The phase-based gabor fisher classifier and its application to face recognition under varying illumination conditions," in *Proceedings of the 2nd International Conference on Signal Processing and Communication Systems*, Gold Coast, Australia, 15-17 December 2008.

24. C.Liu, "Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 725–737, 2006.

25. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neurosicence*, vol. 3, no. 1, pp. 71–86, 1991.

26. V. Štruc and N. Pavešić, "The corrected normalized correlation coefficient: A novel way of matching score calculation for lda-based face verification," in *Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery*, Jinan, China, 18-20 October 2008, pp. 110–115.

27. H.K. Ekenel and R. Stiefelhagen, "Local Appearance based Face Recognition Using Discrete Cosine Transform," in *Proceedings of the 13th European Signal Processing Conference (EUSIPCO), Antalya, Turkey*, 2005.

28. K. Nandakumar, Y. Chen, S.C. Dass, and A.K. Jain, "Likelihood Ratio-Based Biometric Score Fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 342–347, 2008.

29. M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscienc*, vol. 3, no. 1, pp. 71–86, 1991.

30. A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proc. Eurospeech'97*, Rhodes, 1997, pp. 1895–1898.

31. S. Bengio and J. Marithoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.

32. N. Poh and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," *Pattern Recognition*, vol. 39, no. 2, pp. 223–233, February 2005.

33. A. Martin, M. Przybocki, and J. P. Campbell, *The NIST Speaker Recognition Evaluation Program*, chapter 8, Springer, 2005.
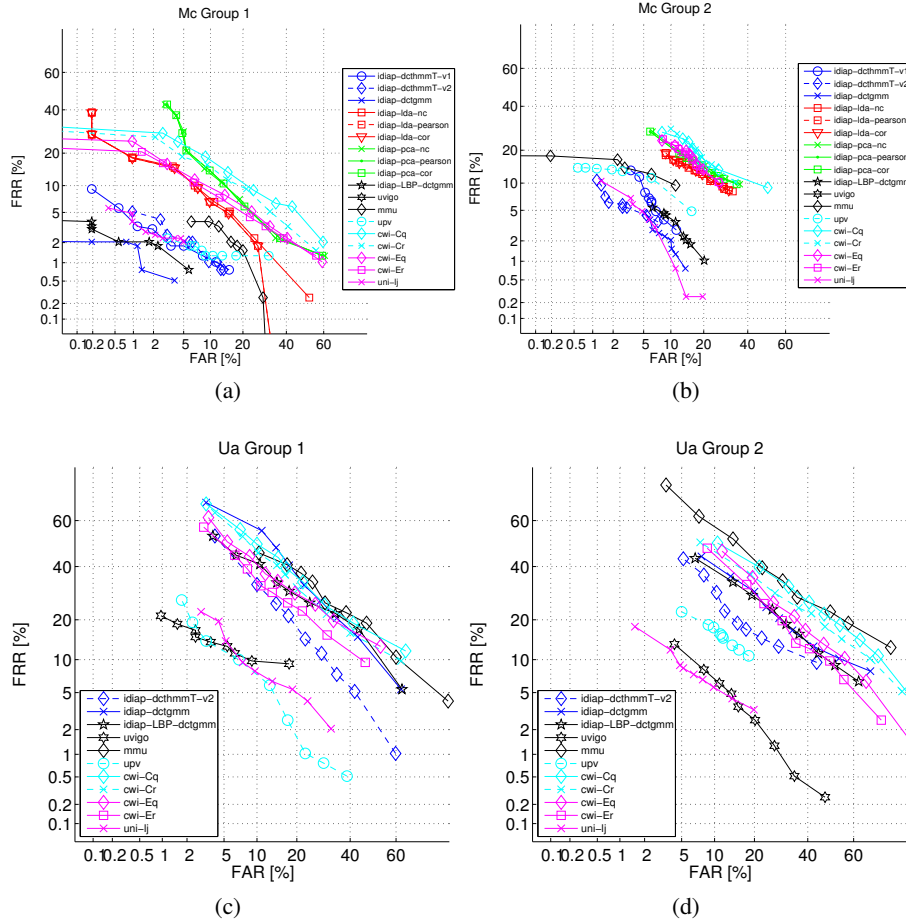
Fig. 1. DET curves of the submitted systems evaluated on the g2 (evaluation set) of the BANCA video based on the Mc protocol. Note that the uvigo system achieved zero EER on the Mc g2 datasets. As a result, its DET curve reduces to a single point at the origin $((\infty, \infty))$ in the above normal inverse scales.
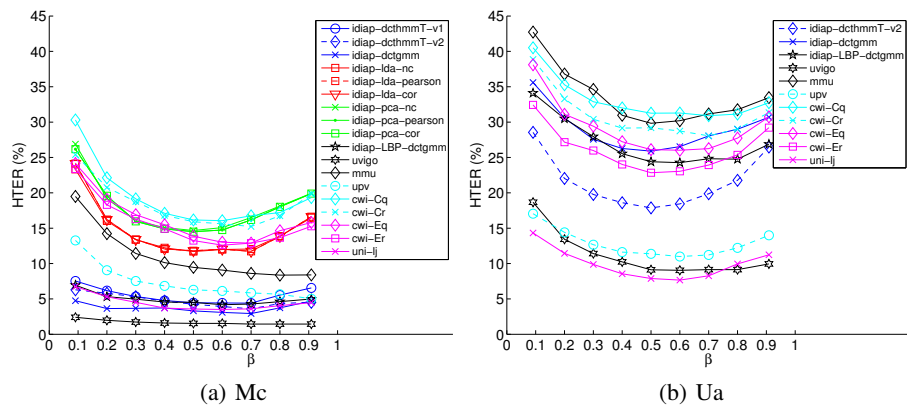
(a) Mc   (b) Ua

Fig. 2. EPC curves of the submitted systems evaluated on the g2 (evaluation set) of the BANCA video based on the Mc protocol.