An Evaluation of Video-to-Video Face Verification

Norman Poh, *Member, IEEE*, Chi Ho Chan, Josef Kittler, Sébastien Marcel, Christopher Mc Cool, Enrique Argones Rúa, José Luis Alba Castro, Mauricio Villegas, *Student Member, IEEE*, Roberto Paredes, Vitomir Štruc, *Member, IEEE*, Nikola Pavešić, Albert Ali Salah, Hui Fang, and Nicholas Costen

***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 realize 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 an evaluation of person identity verification using facial video data, organized in conjunction with the International Conference on Biometrics (ICB 2009). It involves 18 systems submitted by seven academic institutes. These systems provide for a diverse set of assumptions, including feature representation and preprocessing variations, allowing us to assess the effect of adverse conditions, usage of quality information, query selection, and template construction for video-to-video face authentication.**

***Index Terms*—Biometric authentication, face video recognition.**

I. Introduction

With an increasing number of mobile devices with built-in web-cams, e.g., PDA, mobile phones, and laptops, the face is arguably the most widely accepted means of person verification (or authentication). 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/authentication performance is to use multiple samples.

Face verification is but one of the possible tasks of high level cognition; others include face classification, identification, and face memorization [1]. Face recognition generally refers to both face verification and face identification.

Gorodnichy [2] proposed that comparing two photographic facial data (still images) is not the same as comparing two video sequences containing face images. Two justifications are given. First, arguably, photographic facial data is considered a hard biometric trait whereas face recognition in video is a soft one, i.e., having behavioral traits (e.g., one's facial expression and talking dynamics). Second, due to its bandwidth, real-time nature, and environmental constraints, faces in a video are often of significantly lower resolution compared to photographic facial images. Furthermore, the quality of faces in video is likely to be uncontrolled, e.g., located too far from the camera, or at an angle which makes recognition difficult.

Despite the above nature of faces in video, humans have excellent ability in recognizing these faces with high efficiency and accuracy. We outline below some conclusions from studies in neurobiology [3], also summarized in [4] and [5]: Humans recognize faces in grayscale images with the same ease at which he/she recognizes faces in color images. Although the color cue does not seem to be used for high-resolution face images, this is possibly not the case for low-resolution images. Motion and color are used to focus the attention of interest. Human vision is guided by salient features, and seeing is performed in a saccadic motion (from one salient feature to another). Eyes are the most salient features in a face. Rather than 3-D models, it is believed that humans use several representative face images in order to memorize a face. Humans seek and accumulate evidence (over time).

Insights drawn from human vision can be used to efficiently solve problems in computer vision, including face recognition. Some example applications of recognizing faces in video are face detection in a crowd [6], principle casts/characters detection recognition in a movie [7], video surveillance with associative memory [8], to cite a few.

In our case study, we focus on *remote* face verification. A typical scenario of this consists of a user requesting access to an online service which requires identity verification by using his face images. Examples of services are credit card authentication, access to public services (e-banking), and financial transactions. This application scenario presents a significant challenge because the users employ their own cameras and the acquisition

Manuscript received September 30, 2009; revised August 18, 2010; accepted August 23, 2010. Date of publication September 20, 2010; date of current version November 17, 2010. The work of N. Poh was supported by the Advanced Researcher Fellowship PA0022 121477 of the Swiss NSF. The work of N. Poh, C. H. Chan, and J. Kittler was supported by the EU-funded Mobio project grant IST-214324. The work of N. Costen and H. Fang was supported by the EPSRC Grant EP/D056942 and Grant EP/D054818. The work of N. Pavešić and V. Štruc was supported by the Slovenian National Research Program P2-0250(C) Metrology and Biometric System, the COST Action 2101 and FP7-217762 HIDE. The work of E. A. Rúa was supported by the Spanish Project TEC2008-05894. The work of M. Villegas and R. Paredes was supported by the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018). The work of A. A. Salah was supported by the Dutch BRICKS/BSIK project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick J. Flynn.

N. Poh, C. H. Chan, and J. Kittler are with the Centre for Vision, Speech and Signal Processing (CVSSP), School of Electronics and Physical Sciences, University of Surrey, Guildford, GU2 7XH, Surrey, U.K. (e-mail: n.poh@surrey.ac.uk; normanpoh@ieee.com).

S. Marcel and C. Mc Cool are with the Idiap Research Institute, 1920 Martigny, Switzerland.

E. A. Rúa and J. L. Alba Castro are with the Signal Technologies Group, Signal Theory and Communications Department, University of Vigo, 36310 Vigo (Pontevedra), Spain.

M. Villegas and R. Paredes are with the Universidad Politécnica de Valencia, Instituto Tecnológico de Informática, 46022 Valencia, Spain.

V. Štruc and N. Pavešić are with the Faculty of Electrical Engineering, University of Ljubljana, SI-1000 Ljubljana, Slovenia.

A. A. Salah is with the University of Amsterdam, 1098 XH Amsterdam, The Netherlands.

H. Fang was with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, M1 5GD, U.K. He is now with the Computer Science Department, Swansea University, Wales, SA2 8PP, U.K.

N. Costen is with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, M1 5GD, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2010.2077627

TABLE I
CATEGORIZATION OF SUBMITTED ALGORITHMS.

Matching approach	Parts-based approach	Holistic approach
Frame-based	11 systems	6 systems
Video-based	None	1 system (MMU)

process is not supervised. The consequence is that the quality of acquired images can vary significantly from one user to another.

While humans have excellent vision ability, recent progress shows that computers can surpass the human ability in face recognition of still images [9]. However, as far as unconstrained face recognition is concerned, the human performance easily surpasses any computer algorithm, as evidenced by the recent Multibiometric Grand Challenge (MBGC).[1]

A. Previous Face Evaluation Efforts

Previous attempts at assessing the performance of face verification algorithms have been restricted to matching still images, e.g., the three FERET evaluations[2] (1994, 1995, and 1996), the face recognition vendor tests (FRVTs 2000, 2002, and 2006)[3], and assessment on XM2VTS and BANCA databases [10], [11]. The well-known Face Recognition Grand Challenge [12] 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 aims at assessing *video-to-video* matching, i.e., in both enrollment and authentication 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 Multiple Biometric Grand Challenge (MBGC) organized by the National Institute of Standards and Technology (NIST), USA. Note that MBGC aims at "portal application" where the task is to verify the identity of a 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 consumer type devices, e.g., web-cams and camera-embedded mobile phones, where a sequence of unconstrained (talking) face images can be easily acquired.

Last but not least, it is also worth mentioning that the CLEAR evaluation [13] also contains a subtask of face video recognition from multiple cameras but in a meeting scenario. Since the evaluation was not aimed specifically at face video recognition, the submitted face systems were not thoroughly evaluated, which is the main focus of this paper.

B. About the Submitted Systems

The submitted systems can be conveniently grouped into four categories, depending on the dichotomies: parts-based versus holistic approach and frame-based versus image-set (video-to-video)-based comparison, as depicted in Table I.

The holistic approach generally refers to methods that use the *entire* face image for face recognition, e.g., the principal component analysis (PCA) method, or Eigenface, and the linear discriminant analysis (LDA) method, or Fisherface [14]. Recent advances in face recognition are dominated by the *parts-based* approach, where a facial image is divided into several regions and for each region, features are extracted and compared, independently of the other regions. The resultant comparison scores are often combined via a fusion rule (e.g., sum, min, max) or by another trained fusion classifier. This process is illustrated in Fig. 1.

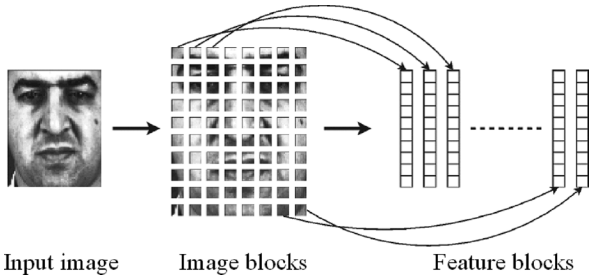


Fig. 1. Parts-based approach to face recognition.

The *frame-based* approach processes a video of facial images frame by frame, whereas the *video-based* approach treats the entire video as a single observation, or as a facial manifold defined by the set of facial images. Because of this fundamental difference, the frame-based approach often requires a separate fusion stage in order to combine the matching scores due to comparisons with several query frames (with the reference/enrollment data). This categorization is not meant to be exhaustive; a more detailed categorization of different approaches for face video recognition is surveyed in Section II.

C. Objectives, Contributions, and Paper Organization

The objectives of this paper are two-fold: 1) to assess the performance of different facial recognition techniques (parts-based versus holistic; frame-based versus image-set-based) in processing video sequences under controlled and adverse conditions, and 2) to validate the effectiveness of facial/image-related quality measures for face video matching. Our findings, carried out on the BANCA face video database, suggest that image quality can effectively be used to select video frames, as supported by the existing literature, e.g., [15]. Indeed, not only that algorithms that selectively process the video sequences using the quality information perform better in adverse conditions, but they also make a *significant savings* in terms of computational resources. This highlights the potential use of facial quality measures in video-based face recognition.

This paper is organized as follows: Section II gives a brief survey of video-based face recognition. Section III describes the submitted systems. This is followed by a description about the database in Section IV. Section V presents our evaluation methodology. Section VI presents the experimental findings. Finally, Section VII concludes the paper.

II. VIDEO-BASED FACE RECOGNITION

Face recognition should be greatly assisted by video information for two reasons:

1) the enhanced observational information about the subject conveyed by multiple frames;

[1]Available: http://face.nist.gov/mbgc

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

[3]Available: http://www.frvt.org

2) the widened range of approaches to face matching that can be adopted.

The latter benefit, in particular, makes it possible to extract additional cues about the subject, i.e., to use spatio-temporal representations of faces, rather than single frame models. The availability of video also simplifies the process of face detection and geometric normalization by exploiting the continuity of face information in videos through tracking, as well as the identity maintenance through auxiliary information such as clothing. The continuous sampling of a dynamic face also offers the use of behavioral biometric information to inform recognition.

The literature dealing with face recognition from video can be conveniently divided into two categories:

1) single-face-to-video matching;
2) video (image set)-to-video matching.

The techniques in the former category are a straightforward extension of single face-to-single face matching methods to the multiple image frames available in a video footage. The techniques in the second category formulate the problem of face recognition from video from a more pertinent stand point which is based on the assumption that video footage of an individual is available not only for recognition but also for learning the person's model. This opens the range of possibilities for model type selection and model construction. We shall structure the review according to this basic categorization.

A. Still-Face-to-Video Matching

1) Still-Frontal-Face Matching Extension: Any still-to-still face image matching method can naturally be extended to a still-to-video matching by matching each frame of the query video against the class templates [4]. The classical techniques of eigenfaces [16], probabilistic eigenfaces [14], the elastic bunch graph matching [17], [18], and the PDBNN [19] are typical examples of these approaches.

The more advanced techniques have also been extended in this way. For instance, the advanced correlation filter method of [20] has been applied to still-to-video matching in [21]. The proposed technique uses intramodal weighted averaging of the outputs of several correlation filters (MACE [20], optimal trade-off filter [22], optimal trade-off circular harmonic function [23], and polynomial filters [24]) for each frame and accumulates the combined score over time by equally weighting all the samples.

In [25], the still-to-still multiscale local binary pattern face recognition system is applied to all the frames in the query video and the sequence of scores is combined to reach the final decision. Both in [25] as well as in [26] and [27], the aim is to extract features that are invariant to various performance degrading phenomena, such as deviations from the frontal pose and image blur. While the methods [28] and [29] use statistical class models built from multiple face images or an image sequence, the actual matching against query video is conducted frame by frame. The approach of Xu *et al.* [30] also offers pose and illumination invariance.

2) Still-Frontal-Face to Pose-Corrected-Video-Frames Matching: Under the assumption that face images in a video sequence have varying poses, the full exploitation of the multiple observations in the context of matching against a single frontal face image will be possible only when all the face images in all the video frames are pose corrected. This can be achieved using face models. The basic idea is to fit such a model to each frame of the video. The by-product of the fitting process is an estimate of the 3-D pose of the subject. Based on the estimated pose, a virtual view of the subject's face can then be generated and used for recognition. The most powerful is a 3-D morphable face model which captures the 3-D shape of a face and its surface texture [31]–[33]. The most recent work using such a model for face recognition in video is that of Park and Jain [34]. As an alternative to the morphable model, one can adopt the deformable model of DeCarlo and Metaxa [35], [36] which is computationally much simpler to fit. Once such model is fitted to a 2-D image, as its tracking for the same subject involves estimating only the pose parameters and adaptation for a changing expression, it could be applied to an image sequence in real time.

Another possibility is to exploit the structure from motion algorithms in computer vision so as to estimate the instantaneous 3-D positions of facial features. The pose can then be deduced from this 3-D information [37]. A 3-D model can also be built from multiple-view 2-D images annotated by pose and the position of facial landmarks [38]. The 3-D model consists of an affine geometrical model and shape and pose free texture model.

The alternative to 3-D models is to opt for 2-D active appearance models, proposed by Cootes and Taylor [39]. An appearance-model-based method for video tracking and enhancing identification was proposed in [40]. Advanced versions of Active Appearance models are computationally efficient and can be fitted to face video in real time [41], [42]. There is also a merit in using user-specific, rather than generic, Active Appearance Models, as demonstrated in [43]. An improved deformable model fitting technique has been proposed in [44], which avoids the problem of locking to local minima by replacing discrete locations in a point distribution model by their smooth versions.

In the view synthesis methods, the desired view can also be synthesized by learning the mapping function between two views [45]. In [36], the matching of frames in a video is carried out using a face manifold model. The face manifold is constructed for each subject by analyzing a sequence of frames exhibiting variation in pose.

Although 3-D models are deemed effective, fitting a 3-D morphable model to 2-D video frames is challenging computationally and practically (as landmarks are required to initiate the fitting process). Motivated by this, Heo and Savides in [46] proposed to construct a 3-D model of a face from a sequence of 2-D active appearance fitted image frames. As the reconstructed 3-D model is sparse (determined by the number of points used by the active appearance model), the vertex density is enhanced by loop subdivision (new points inserted into each triplet of points). The resulting face models look subjectively pleasing, but their effectiveness for recognition has not been demonstrated.

Rather than using a face model for pose correcting the query image, it is possible to match video frames against synthesized views exhibiting the same pose and illumination using a 3-D face model [47]. However, this approach supports only similarity based matching as detailed statistical models of face images parametrized by pose are not available.

3) Frame-by-Frame Multiview Matching: If multiview training data is available, it is possible to construct multiview models and perform pose invariant face matching by maximizing the match score over multiple pose hypotheses, or using a model selection based on pose estimation. The PCA-based pose estimation methods of Pentland *et al.* [48] can be used for this purpose. Once the pose of a face image is determined, the corresponding face model is selected and a match score for each hypothesis computed. A more sophisticated solution was proposed in [38] and [49], where the concept of an identity surface that captures joint spatial and temporal information was introduced. An identity surface is defined by projecting all the images of an individual onto a discriminating feature space parametrized by head pose. The subject identity can then be disambiguated by comparing the motion of the samples of a query video with that of the model trajectories.

A more recent approach proposed by Chai *et al.* [50] advocates the use of local linear regression for pose invariant face recognition in video.

B. Multicue Matching

The techniques discussed in the previous two subsections extend the still-to-still image matching to the still-to-video matching scenario in an uninspiring way which is not cognizant of the rich information content video provides. This can be used to extract, for instance, behavioral biometrics from a talking face, or characteristic motion patterns for each individual. The former constitutes an additional biometric modality which can significantly enhance recognition performance. The latter contributes an additional cue that can be integrated with other observational information. However, the simplest way to exploit video is to make use of motion cues to reduce the computation complexity of face detection, especially in videos containing more than one person in each frame. In this subsection, we focus on this latter category of approaches. An early example is the system proposed in [51] which combines facial appearance with motion cues.

The application of structure from motion algorithms can generate 3-D cues about the analyzed face. This information can be used as an additional modality and fused at a decision level. Alternatively, 3-D shape features can augment the 2-D appearance feature set to enhance recognition.

A better founded approach to face recognition from video, which exploits motion cues, has been advocated in [52]. The main idea is to exploit both spatial information and temporal information (the trajectories of facial features). The motion of facial feature points is approximated by a global 2-D affine transformation (accounting for head motion) plus a local deformation. The tracking is accomplished using a particle filter in a Bayesian inference setting [53], [54]. The assumption behind the method is that the motion trajectories of the same individual are more coherent than those of a different person. Using motion trajectory characteristics as features, the overlap of the *a posteriori* probability distributions of competing hypotheses can be reduced to promote a better separation of identities.

The idea was developed further in [55] which models the joint distribution of identity and motion using a video sequence as input. A marginalization of the distribution with respect to the motion variable yields the *a posteriori* probability distribution over the identity variable. In [56], a similar idea is used to endow a video-based face recognition system with the ability to cope with illumination changes. However, the authors use a much simpler model based on a first-order expansion of the image luminance function [57]. The basic method deals with illumination changes. Face dynamics is modeled by a first-order Markov model.

Lee *et al.* in [36] model spatio-temporal evolution by constructing a manifold model. The inherent nonlinearity of face manifolds is handled by approximation in terms of linear manifolds which are linked by transition probabilities in order to model face dynamics. The problem of this generative model is that it has a limited discriminatory capacity. The work in [58] extends [36] to allow on-line learning of probabilistic appearance manifolds.

The work of Matta and Dugelay [59] exploits behavioral information as well as physiological information extracted from video to realize a video-based face recognition system.

C. Video-to-Video Matching

Face recognition from video can be cast as a learning problem over image sets. A set of images may represent a variation in a face's appearance. The objective of the image set approach is to classify an unknown set of vectors to one of the training classes, each also represented by several image sets. Whereas most of the work on matching image sets exploits temporal coherence between consecutive images [36], [55], [60], it is not strictly necessary to make such a restrictive assumption. In [60], the temporal coherence of face images in a video footage is modeled by a linear dynamical system whose appearance changes with pose. Recognition is performed using the concept of subspace angles to compute distances between probe and gallery video sequences.

Relevant previous approaches for set matching can broadly be partitioned into parametric model-based [61], [62] and nonparametric sample-based methods [63]. In the model-based approaches, each set is represented by a parametric distribution function, typically Gaussian. The closeness of the two distributions is then measured by some measure of similarity.

Relatively recently, the concept of canonical correlations has attracted increasing attention for image set matching in [64]and [65]. Each set is represented by a linear subspace and the angles between two high-dimensional planes are exploited as a similarity measure of two sets. A nonlinear extension of canonical correlation has been proposed in [66], [67] and a feature selection scheme for the method in [67]. The constrained mutual subspace method (CMSM) [65], [68] is the most well known. A related method of object recognition using image sets, which is based on canonical correlations, has been proposed in [69]. The method, known as discriminative analysis canonical correlation (DACC), uses a linear discriminant function that maximizes the canonical correlations of within-class sets and minimizes the canonical correlations of between-class sets, is devised, by analogy to the optimization concept of LDA.

The problem of face-video to face-video matching can also be formulated as one of semisupervised learning, as suggested in

TABLE II
OVERVIEW OF THE SUBMITTED FACE VERIFICATION SYSTEMS

	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	No
	cwi-Eq			DCT	Max(NC)	Yes	No
	cwi-Cr			DCT	Max(NC)	No	No
	cwi-Er			DCT	Max(NC)	No	No
	upv	Local-HEQ	LF	PCA	Avg(KNN)	Yes	No
	uni-lj	ZMUV + HEQ	Gb2	KDA+PCA	WNC	Yes	No
	uvigo	Ani	Gb1		GMM	No	Yes

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 = Normalized correlation; WNC = Sum of whitened NC; LBP = local binary pattern.
Note: OmniPerception's face detector was used by all systems. The three systems which specifically considered the provided quality measures (i.e., UPV, UniLJ, and CWI systems) are described in Sections III-D–III-F, respectively.

[70]. The method exploits the properties of a face data manifold using a computationally inexpensive graph-based algorithm.

All the above approaches draw on holistic face representation. An alternative approach has been suggested by Mian in [71] which is inspired by video retrieval methods. The advantage of this method is that one does not have to worry about face registration. The approach is based on the use of image descriptors, such as scale-invariant feature transform (SIFT) features [72], which are computed at interest points detected in the image. The training to recognize a particular identity is based on a sequence of frames for which a pairwise similarity matrix is computed. The similarity of two face images is defined by the minimum and average similarity of SIFT feature vectors describing each face. The faces in the sequence are then clustered by a hierarchical clustering method and cluster representatives selected. The matching of unknown videos is then based on measuring the similarity of the SIFT features computed for the video frames to the cluster representatives for each hypothesized identity.

The continuity in time of the information in face video can be exploited more directly than what is offered by a statistical analysis of a set of frames [55]. In particular, it can be invoked to resolve the uncertainty in face localization and identification. Zhou *et al.* [55] tackle the inherent uncertainty in tracking a moving face and its identification by performing simultaneous tracking and recognition of faces in video. A detailed modeling of face dynamics using HMM models in Liu *et al.* [73] is potentially more powerful. However, learning temporal dynamics during recognition of unknown video query footage is currently computationally too demanding that renders this approach practically infeasible.

III. SYSTEM DESCRIPTIONS

Sections III-A–III-F describe the submitted systems. Section III-G then compares these systems by their attributes (see Table II).

A. University of Vigo (UVigo)

The video-based face verification system submitted by the University of Vigo for the *preregistered 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 operations. Then a two-step illumination normalization is performed on the geometrically normalized faces. The first step is the anisotropic illumination normalization described in [74]. The 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 [75] $\mathcal{J}_k^{\mathcal{V},i} = \left\{a_{k,0}^{\mathcal{V},i}, \ldots, a_{k,39}^{\mathcal{V},i}\right\}$ are extracted from the ith frame (magnitude of the responses of Gabor filters with five scales and eight orientations, encoded in the second subindex) at fixed points, k, along a rectangular grid of dimensions $D = 10 \times 10$ superimposed on each normalized face image. Frame $\mathcal{I}^{\mathcal{V},i}$ is characterized 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: $\left\{a_{k,l}^{\mathcal{V},i} \mid \mathrm{mod}(l, N_S) = m\right\}$, 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. Sixty-four mixture 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. The number of subjets N_S was fixed as a trade-off between discrimination capability and dimensionality. The first subset includes the coefficients from filters with an even filter index ($l \mid \mathrm{mod}(l, N_S) = 0$), and the second subset includes the coefficients with an odd filter index ($l \mid \mathrm{mod}(l, N_S) = 1$). 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

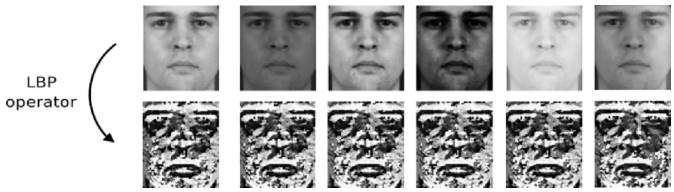


Fig. 2. Invariance of LBP to different illumination.

$f_{\mathrm{UBM},n}(\cdot)$, where $n \in \{0,\ldots,199\}$ is estimated using LBG [76] initialization and the EM algorithm. Gaussian mixtures are constrained to have diagonal covariance matrices. Input vectors for this training process are $\mathbf{x}^{\mathcal{V},i}_{\lfloor n/2\rfloor,\mathrm{mod}(n,2)}$, where $\mathcal{V} \in \mathcal{WM}$, i.e., the world model set videos. Grid node is indexed by $\lfloor n/2\rfloor$, which is the integer part of $n/2$. Subjet set is indexed by $\mathrm{mod}\,(n,2)$.

$f_{\mathrm{UBM},n}(\cdot)$ is then adapted to the corresponding vectors obtained from the user u enrollment video by means of the MAP technique [77], 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 [78]:

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

B. 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 centers and frontalness) provided by the OmniPerception SDK (to be described in Section IV). More precisely, the eye-center coordinates are used to extract the ten-best faces from each video according to the frontalness measure.

1) Geometric and Photometric Normalization: For all systems, the face is first geometrically normalized as described in [79] 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 normalized using two methods: 1) standard histogram equalization (HEQ) as in [79] or 2) a preprocessing based on local binary patterns (LBPs) as proposed in [80] (see Fig. 2).

2) 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 groups.

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 (2-D DCT) or its DCTmod2 variant is computed, as described in [79], resulting in one feature vector per block. An extension to these methods is provided where the 2-D coordinate (xy) of each block is appended to its corresponding feature vector, this was done to incorporate spatial information.

3) Classification: Classification for the holistic methods, PCA and PCAxLDA, is examined using three different similarity measures: Pearson, Normalized Correlation, and Standard Correlation. Classification for the DCT and DCTmod2 features is performed using GMMs and HMMs as described in [81].

It should be mentioned that a development database of images is often needed in order to obtain the PCA and PCAxLDA transformation matrices as well as the background model, or the "world model" for the GMM and HMM classifiers. This development database is also made available to the participants (see Section IV).

C. 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 [82], [83] which seek to model sets efficiently, representing the image set and iteratively fitting this model to each image. The implementation of GGR [84] proceeds through a number of stages. First, 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 minimizing 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 optimize their position. In the final iterations, the points are moved individually to minimize the cost. The cost function includes both shape and texture parts

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

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 neighborhood 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 initialized 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 [85] 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

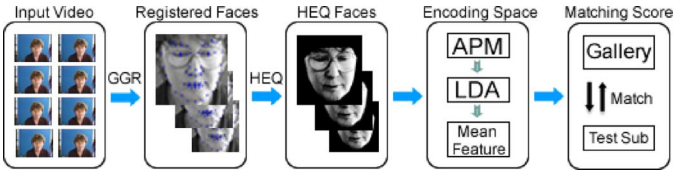


Fig. 3. Architecture of the MMU system, representing an appearance-model based approach.

subject using k-means clustering of their encoding on their individual model to give approximately ten 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 gray-level aspects of the face for a given frame. It models the whole of the inner tile face, using 5000 grayscale 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 nonidentity variation. To remove the latter, a Linear Discriminate Analysis subspace [86] is used. This provides a subspace which maximizes variation between individuals and minimizes the same within them. Each frame in a gallery or probe sequence is projected onto this subspace (see Fig. 3 for a schematic diagram). A set of images derived from a video sequence is then represented by the mean of the LDA coefficients.

When two video sequences are processed as described above, one obtains two mean vectors of LDA coefficients. Let these two vectors are $\bar{\mathbf{d}}_1$ and $\bar{\mathbf{d}}_2$, respectively. The final matching score of these two vectors are assessed by the normalized cross-correlation metric, defined as

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

More details can be found in [87].

D. 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 analyzed using the quality measures and the best NQ frames are selected. After this process, a verification score is obtained for each of the selected frames using the local feature algorithm. 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 neighbor in a discriminative subspace trained using the LDPP algorithm [88]. To train this classifier, the quality values of the frames of the background model videos were used, and each frame was labeled as being good or bad based on the result of face identification using the local feature algorithm [89].

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 neighbors 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 neighbors from the user model divided by the number of extracted local features. For further details refer to [90] and [91]. The parameters of the algorithm were chosen based on previous research and were not adjusted to minimize the error rates of the scenarios. In the algorithm grayscale 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.

E. 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 nonlinear (kernel) subspace projection techniques. The training, enrollment, and test stages of the employed approach can be summarized as follows.

1) 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 preprocessing procedure which first extracted the facial regions from the images based on manually marked eye-center locations, then geometrically aligned and ultimately photometrically normalized the facial regions by means of zero-mean-and-unit-variance normalization and a subsequent histogram equalization step. The normalized facial images cropped to a standard size of 100×100 pixels were then filtered with a family of Gabor kernels with five scales and eight orientations. From the complex filter responses features encoding Gabor-magnitude as well as Gabor-phase information [92] 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 nonlinear subspace was computed based on the multiclass kernel Fisher analysis (KFA) [93]. 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

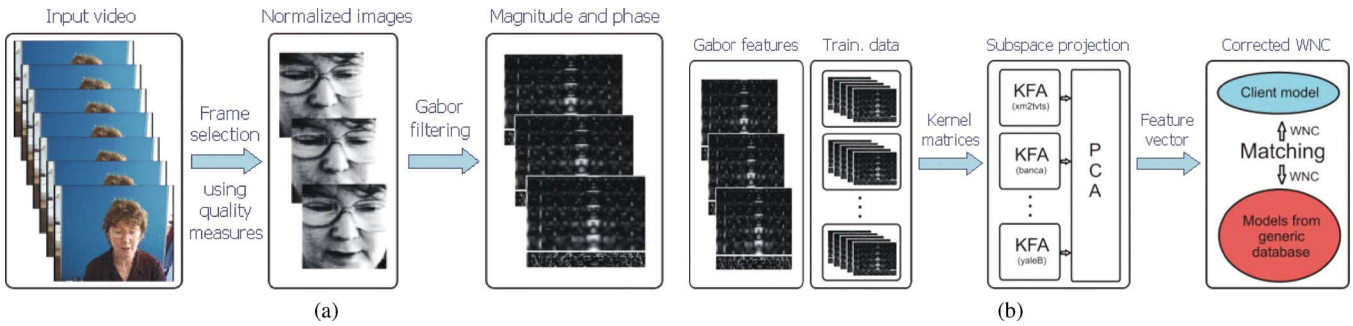


Fig. 4. (a) Feature extraction and (b) classification of UniLJ.

(PCA)[16] to further reduce their dimensionality. Fig. 4(a) illustrates the feature extraction stage.

2) The Enrollment Stage: Using the provided quality measures associated with the video sequences of the BANCA database, a small number of images was chosen from each enrollment video of a given subject.[4] 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.

3) 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 enrollment stage) was processed to extract the facial features. The resulting feature vectors were then matched with the template corresponding to the claimed identity using the nearest neighbor classifier and the whitened cosine similarity detailed in a recently proposed correction scheme [94] (to be briefly explained in the next paragraph). Depending on the cumulative value of the matching score, a decision regarding the validity of the identity claim was made in the end. A schematic diagram of the test stage is shown in Fig. 4(b).

The frame selection procedure is based solely on the quality measure describing the overall reliability of the face detector. In the training stage, a threshold is determined for this quality measure in such a way that at least 5% of all frames from each video sequence of the world set exhibit an overall reliability higher then the threshold value. During testing a random frame selection procedure is used. Here, for each selected frame, the overall reliability of the face detector is compared with the threshold learned during the training stage. Once three (or five, depending on the protocol) frames are successfully selected from the test video sequence, the procedure is terminated. If less than three (or five) frames from the test sequence pass the quality check, the remaining ones are selected randomly.

F. Centrum Voor Wiskunde en Informatica (CWI)

In the CWI approach, the automatically annotated eye locations are used to crop and rectify the face area at each frame. Each cropped frame is then normalized to 64×64, and split into 8×8 windows, from which 2-D-DCT coefficients are extracted [95]. 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 localized 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 [96]. 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 authentication of a new query video is dynamic in that the number of query images used from the video is not fixed. The top NQ ranked images (NQ being from 1 to 8) are matched to the templates in succession, and a preselected distance threshold is checked for authentication. If the similarity score is above this threshold (0.75), it is reported. Else, the next best ranked frame is evaluated, up to eight frames per video 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 preset similarity threshold is a second parameter (the first being the final score threshold for acceptance) that controls the system output.

The CWI submission has four variations to inspect the dichotomies of system complexity [Cheap (C) versus Expensive (E)] and the strategy for choosing the query samples [random (r) versus quality-based (q)]. For the so-called cheap (respectively expensive) version, five (respectively 15) templates are used for each client and only up to four (respectively up to eight) images are used for query. Increasing the number of templates for each gallery subject leads to diminishing returns. Since the DCT feature dimensionality is higher than the number of available frames, an automatic model selection approach usually justifies only a few clusters. During our simulations, we contrasted a random selection of frames versus a quality-based selection of frames. We observed that higher quality faces produced both

[4]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.

higher genuine similarity scores, and higher impostor scores, leading to greater false accept rates.

G. A Summary of the Submitted Systems

Table II summarizes the systems by their attributes. namely, the choice of illumination preprocessing, facial feature representation and extraction methods, the back-end classifier, whether or not quality measures are used, and whether or not all images are processed.

In our summary, feature extraction is distinguished from feature representation by their *purpose*. While feature representation aims to describe the facial information (e.g., using the grayscale of the cropped face image, or other intermediate representation such as appearance model and features designed to extract local information), feature extraction generally aims at making the features more compact and sometimes more discriminative with respect to the identity space.

The systems are grouped into holistic or local (i.e., parts-based) in Table II. Many of the holistic systems were submitted by IDIAP as baseline systems. These systems are only tested on the Mc protocol but not the Ua protocol due to higher illumination and pose variation of the latter data set. As can be observed, the majority of the submitted systems are parts-based. This is generally consistent with the current research trend in face recognition.

IV. Database, Protocols, Facial Video Annotations

We opted to use the publicly available BANCA database [97].[5] It is a collection of face and voice biometric traits for 260 persons in five different languages, but only the English subset is used here. The latter 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 [97], when g1 is used as a development set (to build the user's reference 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 biometric comparisons under the same recording conditions (as the enrollment 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, adverse conditions to sessions 5–8, and degraded conditions to sessions 9–12.

Apart from the g1 and g2 data sets, there is also an additional data set, called the "world model data set," that is used as a development data set. It contains a single session of video recordings of 30 subjects in controlled, adverse, and degraded conditions. This additional data set can be used to calculate the transformation matrix needed for the holistic approach (e.g., the Eigenface and Fisherface methods), as well as the background or world model [81] for the parts-based approach. When one of g1 and g2 sets is used as the test data set, the other set can be used in conjunction with the world model as the training data set.

There are altogether seven experimental protocols specifying the sessions to be used for enrollment 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 adverse (Ua) protocols. The first protocol was intended as a vehicle to design and tune a face verification system. The second protocol aims at testing the system under more realistic and challenging conditions.

In the Mc protocol, session 1 data are used for enrollment whereas the data from sessions 2–4 are reserved for testing. In the Ua protocol, the session 1 data again are used for enrollment but the test data are taken from session 5–8 (recorded under adverse conditions). The ICB2009 face video competition was thus naturally carried out in two rounds, with the first round defined by the Mc protocol and the second round by the Ua protocol [98].

In order to be consistent with the previous BANCA evaluations [10], [11], we also divided a query video sequence into five chunks, each containing 50 frames for convenience; the remaining frames were simply not used.

In order to standardize the evaluation, we provided a pair of eye coordinates, based on the face detector provided by the OmniPerception SDK.[6] However, the participants could use their own face detectors. For each image in a video sequence, the SDK also annotated the following quality measurements:

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.

Note that the entire process from detection to annotation was automatic. No effort was made to fine tune the system parameters, and in consequence, some imperfectly cropped images were observed (see Fig. 8, for instance). 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

[5]Available: http://www.ee.surrey.ac.uk/CVSSP/banca

[6]Available: http://www.omniperception.com

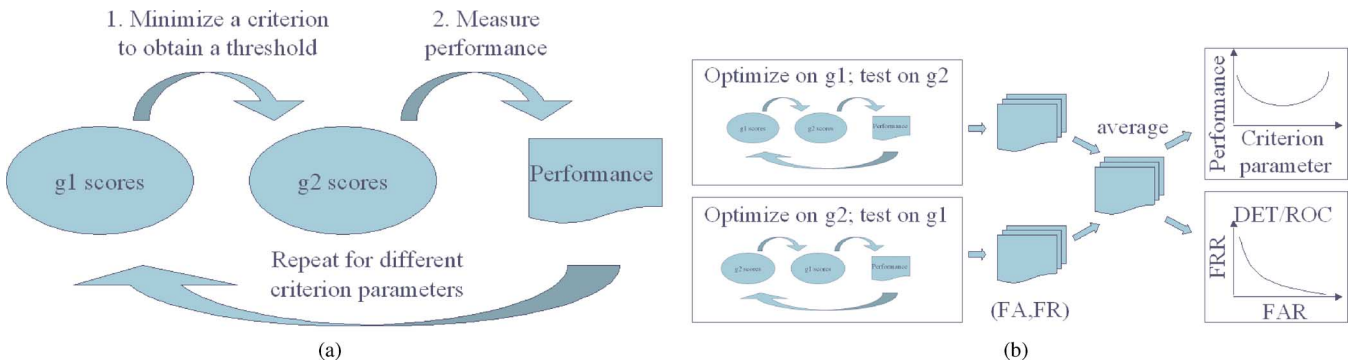


Fig. 5. Assessment methodology. (a) Threshold determination. (b) Two-fold cross-validation.

the MPEG standards.[7] All the annotation data (including eye coordinates and quality measures) have 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 nonideal (nonfrontal) pose.

V. Evaluation Measures

We use two types of curves in order to compare the performance: the detection error tradeoff (DET) curve [99] and the expected performance curve (EPC) [100]. A DET curve is actually a receiver operating characteristic (ROC) curve plotted on a scale defined by the inverse of a cumulative Gaussian density function, but otherwise similar in all respects. We have opted to use EPC because it has been pointed out in [100] 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 [101]. 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 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 score sets can be conveniently used.

Fig. 5(a) shows how g1 and g2 can be used in tandem in two steps:

1) Minimize a *criterion* on g1 in order to obtain a decision threshold.
2) Apply the decision threshold in order to measure the *performance* on g2.
3) Repeat steps 1 and 2 using a different *criterion parameter* exhaustively at fine incremental steps.

[7]Available: http://www.chiariglione.org/mpeg/standards.htm

TABLE III
Performance Measures

Term	Definition
False acceptance rate	$\text{FAR} = \frac{\text{FA}}{\text{NI}}$
False rejection rate	$\text{FRR} = \frac{\text{FR}}{\text{NC}}$
Half total error rate	$\text{HTER} = \frac{1}{2}(\text{FAR} + \text{FRR})$
Weighted error rate †	$\text{WER}(\beta) = \beta\,\text{FAR} + (1-\beta)\text{FRR}$
Equal error rate	EER = FAR = FRR

†: $\beta \in [0, 1]$. FA counts the number of false acceptance cases; FR counts the number of false rejection cases; NI is the total number of impostor accesses; and, NC is the total number of genuine (client) accesses.

At this point, it is useful to distinguish the role of *system optimization criterion* (or simply referred to as criterion) and that of *performance measure*, although in practice, they may be the *same measure*. A criterion is used to determine the decision threshold, whereas performance refers to how successful the system is, which can be a verification rate (FRR for a given desired level of FAR), FAR for a given desired level of FRR, EER, HTER, and WER. These performance measures are defined in Table III. Note that among these performance measures, only EER is *not* dependent on any given decision threshold, because there is only a single point.

There are some natural pairings between a criterion and a performance measure. For instance, if the performance is FAR (respectively FRR), the criterion will necessarily be FRR (respectively FAR). If the performance measure is HTER, it is common to use EER as a criterion. In our case, the chosen performance measure is HTER and the criterion we used is WER. For the case of WER, it has a *tunable* parameter β, which penalizes between FAR and FRR. Therefore, by employing WER with different β values, we obtain an HTER curve.

Fig. 5(b) illustrates the two-fold cross validation process. By varying the β parameter in each fold, we actually obtain a set of pairs of FA and FR (as an intermediate step). The resultant two sets are collated into a single set via averaging. The collated statistics can be visualized either using a DET curve or an EPC. The generalization to k-fold cross validation, or even k data sets could be accomplished in the same manner (see, for instance, [101]). A DET curve plots FAR versus FRR and the criterion is not shown explicitly as an independent variable. In contrast, an EPC shows explicitly this relationship; it plots a chosen criterion parameter as an independent variable (in the X-axis) and the performance as the dependent variable (in the Y-axis). Such a

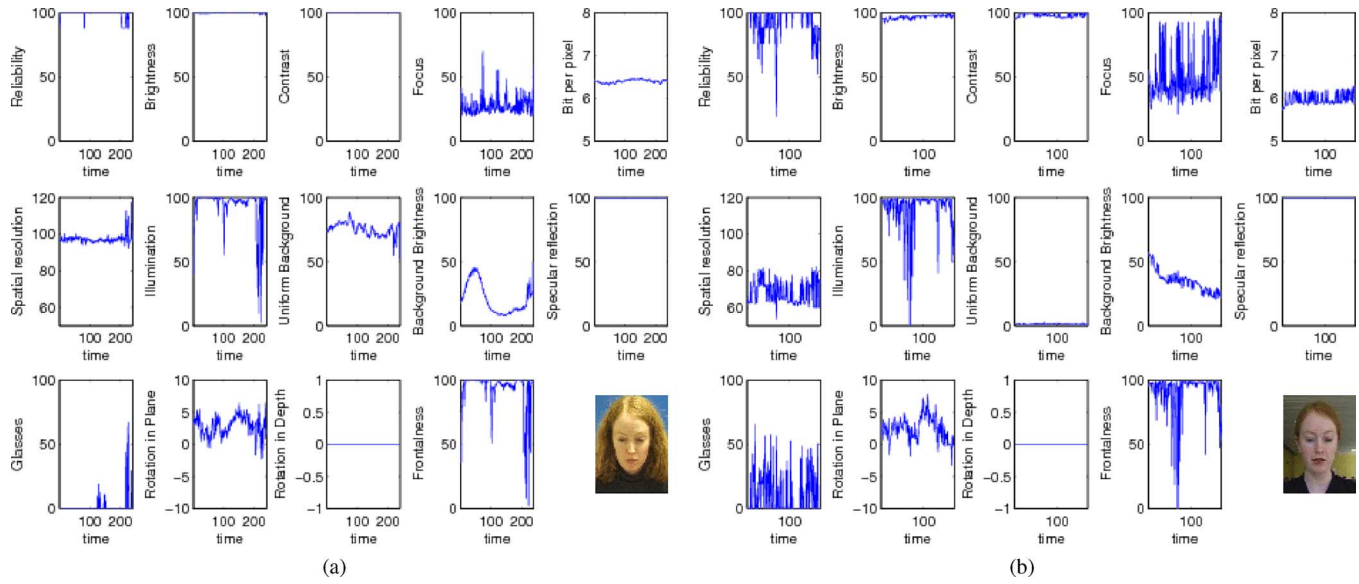


Fig. 6. Evolution of 14 quality measures over time. (a) Controlled scenario. (b) Adverse scenario.

relationship is desirable in the sense that in order to compare two systems, one only needs to pick a criterion parameter (hence choosing a relative trade off between false acceptance and false rejection that is relevant to a particular application scenario), and reads off the performance values of the two systems.

The motivation for using WER as a decision threshold criterion is that it generalizes the criterion used in the annual NIST speaker evaluation [102] as well as the three operating points used in the past face verification competitions on the BANCA database [10], [11]. The NIST speaker evaluation uses approximately $\beta = 0.9$ whereas the BANCA evaluation protocols uses the following three coefficients of β:

$$\beta = \frac{1}{1+R} \quad \text{for} \quad R = \{0.1, 1, 10\} \tag{5}$$

which yield approximately $\beta = \{0.9, 0.5, 0.1\}$, respectively. We, therefore, sampled WER with the following β values: $\{0.1, 0.2, \ldots, 0.9\}$.

In order to satisfy both camps of biometric practitioners, while retaining the advantage of EPC which makes performance comparison between systems less biased (with respect to the choice of decision threshold), we shall report the results in terms of DET curve (as a function of β values) *as well as* EPC. In this way, we actually establish a *correspondence* between EPC and DET, i.e., each point in the DET space has a corresponding point in the EPC.

Examples of DET and EPC can be found in Figs. 10 and 11. It can be observed that 1) the best system is the closest DET curve to the origin in the DET space and that 2) its corresponding EPC has the smallest HTER values.

VI. Results

The experimental results here are presented in four parts. The first part analyzes the evolution of some of the quality measures on both controlled and adverse scenarios. The second part analyzes the effect of increasing the number of enrollment samples as well as the query samples on the verification performance. The third part compares the performance of different face verification systems in both the controlled and adverse conditions. Finally, the last part investigates the performance versus time complexity.

A. Preliminary Analysis on Quality

This section aims to analyze subjectively the effectiveness of automatically derived facial quality measures. An objective analysis in terms of performance will be presented in Sections VI-C and VI-F.

In the context of video-based face recognition, two questions relevant to our scenarios are of interest here:

1) Can quality measures be used to distinguish *between* video images taken under controlled and adverse scenarios?
2) Can quality measures distinguish different quality of images *within* the same video sequence (and consequently the same application scenario)?

The ability to distinguish between controlled and adverse scenarios is important because, very often, algorithms that work well in controlled scenarios may not necessarily perform optimally under adverse scenarios (as will be attested by our evaluation results in Figs. 10 and 11). This opens the possibility of combining complementary algorithms, each of which is optimal under a particular type of conditions [103]. The second question is also of great interest because if quality measures can indeed distinguish well aligned (frontal) images from badly aligned ones, this information can be used directly for computing the final scores (e.g., selecting only the qualified ones according to some criteria).

To answer the first question, we plotted the evolution of the 14 quality measures over time on two video sequences recorded under both controlled and adverse conditions. The results are shown in Fig. 6. It should be noted that the quality measures are not designed to operate on video sequences but we applied frame by frame, ignoring the dependency between two consecutive frames over time. Over the entire video sequence, it can be observed that some quality measures can indeed be used to

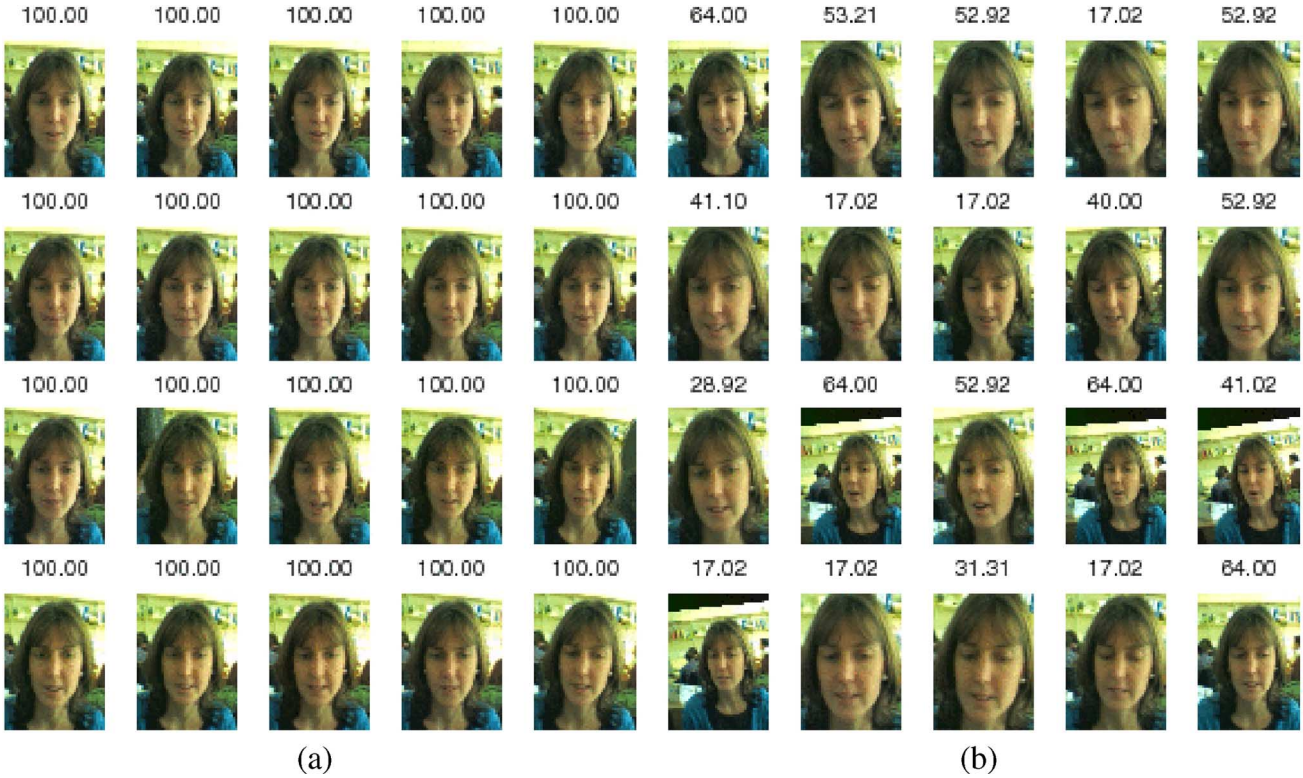


Fig. 7. Overall reliability as a quality measure, assessed on the video sequence gb_video_degraded_1003_f_g1_06_1003_en.avi. (a) High overall reliability. (b) Low overall reliability.

distinguish the two types of recording conditions. For instance, under the controlled conditions, the overall reliability is in general very high (mostly 100%, with some exceptions), whereas under the adverse conditions, it varies (this quality measure will be analyzed in a later section). Brightness and focus also turn out to be very good discriminating criteria.

In general, both bit per pixel, focus, spatial resolution (between eyes), and illumination has a higher variation for the adverse conditions (than the controlled ones). These are examples where not only the absolute quality measures are important, but their variance over the entire video sequence is also a meaningful indicator of the signal quality.

We selected some cropped face images and show them in Fig. 7. As can be observed, when the overall quality measure is 100%, the face images are usually better registered (aligned), whereas the images with low values are usually not well registered. Hence, the overall quality measure can be used as a tool for selecting good query images in a video sequence. Plotted in Fig. 8(a) is the evolution of the overall quality measure over the entire video sequence and its histogram over the entire video sequence is shown in (b). As can be observed, well-aligned face images (with 100% value) constitute a small fraction of the entire video sequence. For some matching algorithms (e.g., holistic-based ones), it may be useful to select the detected face images based on the overall quality measure.

B. Number of Templates and Query Images Versus Performance

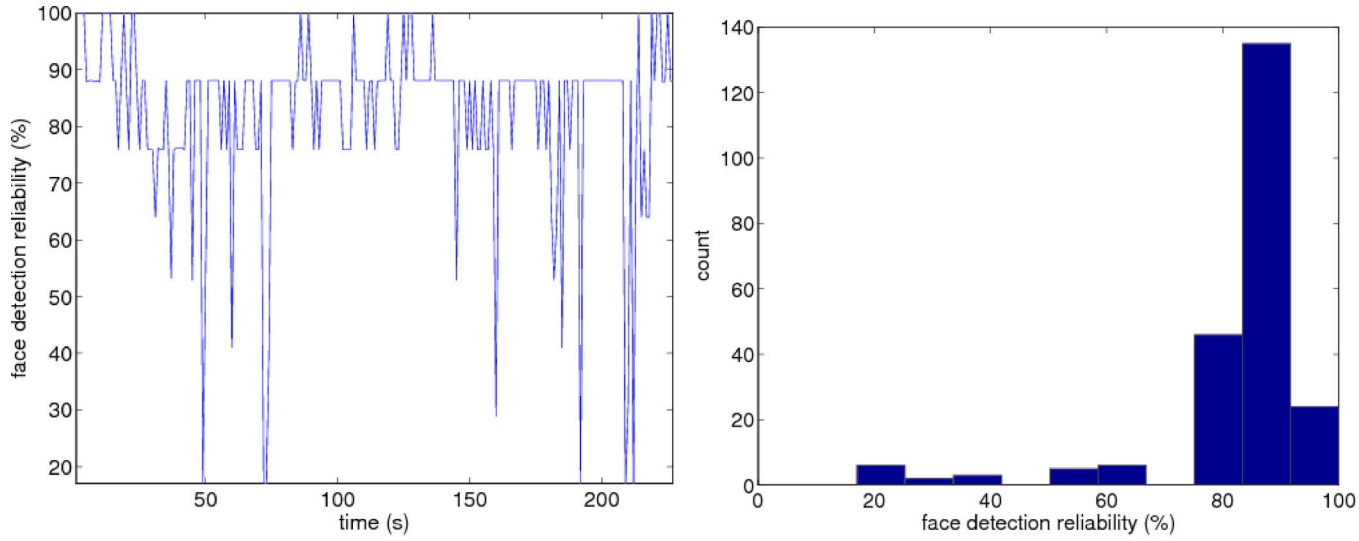
In this section, we shall examine the effect of varying the number of templates and query images. This study was performed with the system supplied by CWI on the Mc protocol (set g2). We varied the number of templates (either one or five) and the number of query images (from 1 to 4). Each time, the maximum similarity score is used as the final score.

The query images are selected according to their ranked quality as explained in Section III-D. We observe that using more queries improves performance, which means quality-based ranking is not detrimental to the diversity of the query images.

The template images are obtained by offline clustering of the training video frames to ensure diversity. By using P templates and Q query images, the number of comparisons is $P \times Q$, which is directly related with the method complexity. As can be observed in Fig. 9, a more complex system (using more comparisons) actually generalizes better. Since each hypothesis provides additional evidence for being a genuine access versus the impostor one, combining a set of scores in supporting a particular hypothesis can improve the confidence in the selected hypothesis via variance reduction [104].

C. Competition Results

The DET and EPC curves of all submitted systems for the g1 and g2 data sets, as well as for the Mc and Ua protocols, are shown in Figs. 10 and 11, respectively. These results are obtained by merging the results from g1 and g2. The EPCs here plot HTER versus β, a parameter of WER. To be consistent with the previous published BANCA evaluations [10], [11], we also listed the individual g1 and g2 performance figures, in terms of WER, in Table IV for the Mc protocol and in Table V for the Ua protocol. We also report results as a function of R, as defined



(a) (b)

Fig. 8. (a) Evolution of overall reliability as a quality measure of the video sequence gb_video_degraded_1003_f_g1_06_1003_en.avi over time and (b) its corresponding histogram over the entire video.

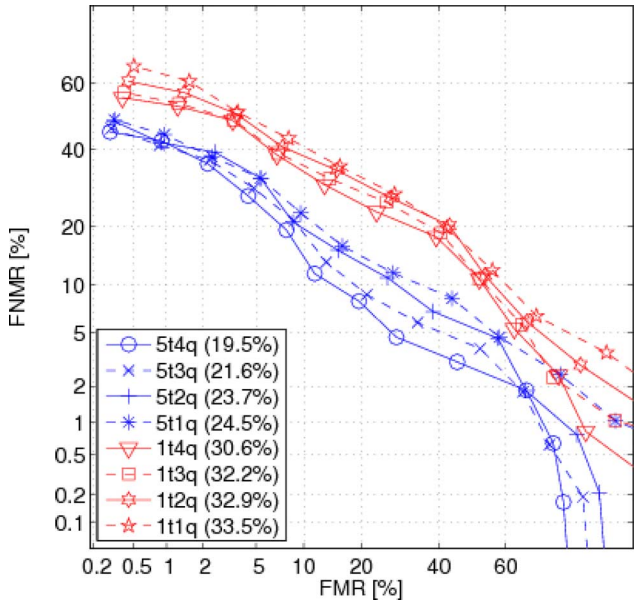
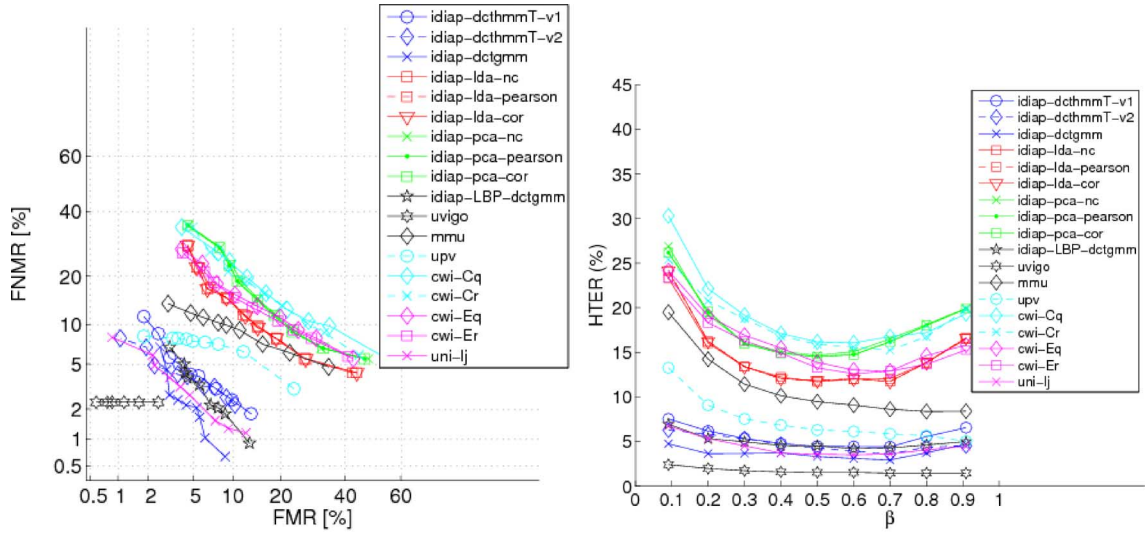


Fig. 9. Effects of the number of templates and query images on the system performance. In the legend, $ntmq$ means a particular experimental configuration using n templates and m query images. The numbers quoted in the legend are EER in percentage.

in (5). Recall that for each value of $R = \{0.1, 1, 10\}$, there is a corresponding approximate value of $\beta = \{0.9, 0.5, 0.1\}$, respectively.

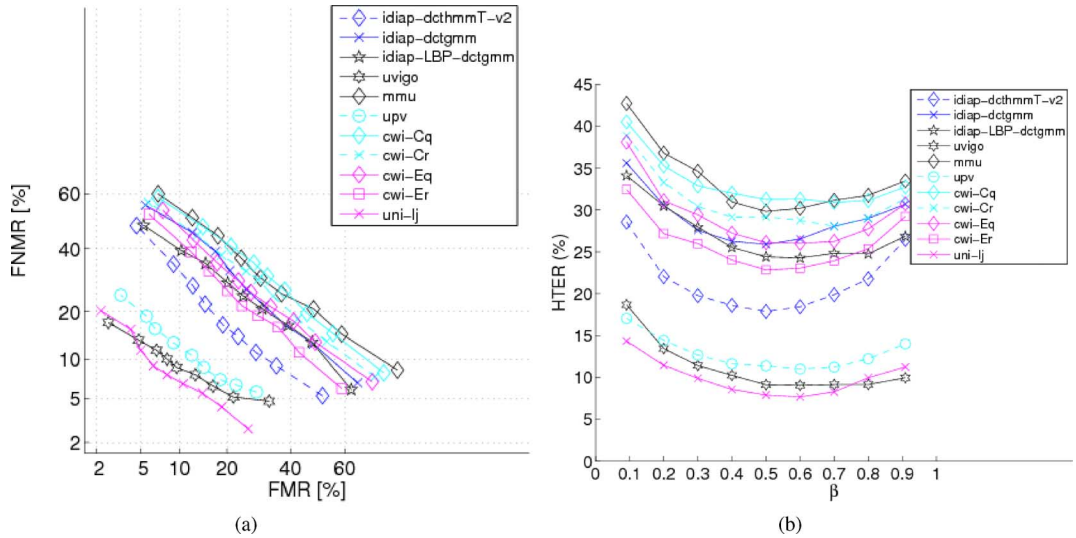
The following observations can be made:

1) **Degradation of performance under adverse conditions**: It is obvious from Figs. 10 and 11 that all systems systematically degrade in performance under adverse conditions. In order to have a better picture of the degradation, the HTER of all systems from Mc to Ua are shown in Fig. 12.
2) **Holistic versus local appearance methods**: Comparing Fig. 10 with Fig. 11, 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 normalization 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 the images are photometrically normalized.
3) **Preprocessing**: In dctgmm methods, the performance of applying HEQ is better than that of applying LBP as a preprocessing method for the Mc protocol. However, the case is reversed for Ua protocol because HEQ enhances shadows while LBP features are invariant to such monotonic transformation. Selection of the preprocessing method should be dependent on the environmental conditions. Advancement in image processing now shows that a semiautomatic procedure to achieve this is realizable [105].
4) **Sample size**: CWI's submission has four variations: depending on the dichotomies: system complexity, i.e., Cheap (C) versus Expensive (E); and strategy for choosing the query samples, i.e., random (r) versus quality-based (q) (see Section III-F). Two observations can be noted: First, the performance of cwi-Eq and cwi-Er is better than that of cwi-Cq and cwi-Cr. Second, using *more* templates and query images improves the system performance. A rigorous and systematic design of experiments is still needed to assess the usefulness of the provided quality measures, and more importantly, the most effective ways of using such auxiliary 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 strategy for quality measure selection is needed. Second, quality measures themselves are not discriminatory in relation



(a) (b)

Fig. 10. DET and EPC of the submitted systems evaluated using the Mc BANCA defined cross-validation protocol (on both g1 and g2 data sets). (a) DET. (b) EPC.



(a) (b)

Fig. 11. DET and EPC of the submitted systems evaluated using the Ua BANCA defined cross-validation protocol (on both g1 and g2 data sets). (a) DET. (b) EPC.

TABLE IV
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-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

TABLE V
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

to subjects, but can help in distinguishing environmental conditions. The above research issues are further tackled in our recent work [106].

5) **Multiresolution 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

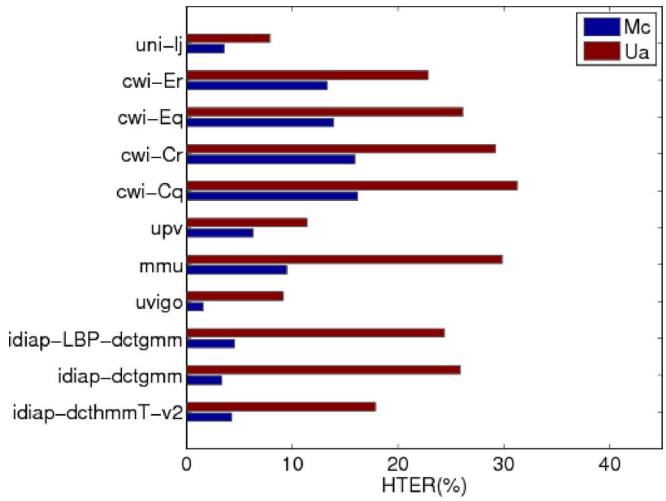


Fig. 12. Degradation from the Mc to Ua protocol when the operating threshold is tuned at the EER point ($\beta = 0.5$).

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.

D. Single-Image Versus Multi-Image Matching

One aspect that is lacking about the competition is that it was not possible to compare the result between still face versus video face matching. A particular characteristic of video-based matching is the availability of multiple images. We therefore compare single-image versus multi-image matching here.

For this purpose, the upv system was rerun to compare the results of these two approaches. The performance of the still face matching is based on a single image chosen at random. In comparison, the multi-image approach processes five images (based on the supplied quality measures) for the Mc protocol and six images for the Ua protocol (the DET curves of both configurations are taken directly from the competition submission).

The results are shown in Fig. 13. As can be observed, using multiple face images in a video sequence consistently outperforms the strategy of choosing a single face image.

E. Image-to-Image Versus Manifold-to-Manifold Matching

Another characteristic about video-based matching is the possibility of deriving a manifold from a set of images. Therefore, this section compares image-to-image versus manifold-to-manifold matching, while keeping the underlying matching classifier the same for both cases.

For the above purpose, we rerun the mmu system, which is a manifold-to-manifold matching technique. This system can easily be converted to image-to-image matching. Starting from the initial distance metric, which is calculated using (3), The new scores are calculated as

$$S_c = \max_{t,q} \frac{\mathbf{d}_t}{|\mathbf{d}_t|} \cdot \frac{\mathbf{d}_q}{|\mathbf{d}_q|} \tag{6}$$

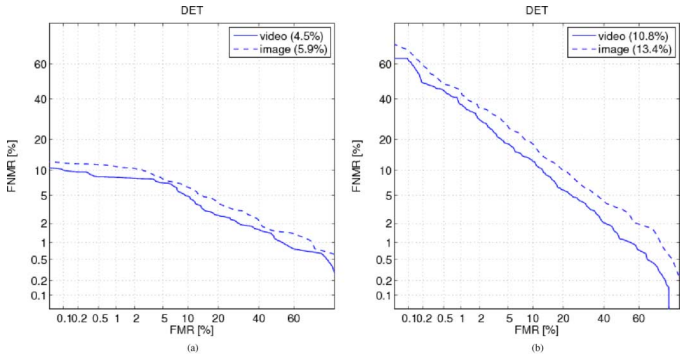


Fig. 13. Comparison of performance when using a single image (dashed line) versus multiple images (solid line) in face verification for (a) the Mc protocol, and (b) the Ua protocol. For the system with multiple images, five images were used for the Mc protocol and six for the Ua protocol. The numbers quoted in the legend are EER in percentage.

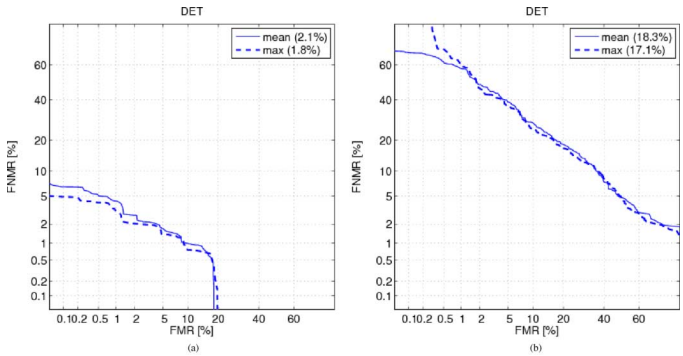


Fig. 14. Comparison of performance when using manifold-to-manifold matching (dashed line) versus image-to-image matching (solid line) in face verification for (a) the Mc protocol, and (b) the Ua protocol. The manifold-to-manifold matching is based on the mean aligned image (labeled as "mean") whereas in the image-to-image matching, maximal correlation is taken as output [hence labeled as "max," see (6)]. The numbers quoted in the legend are EER in percentage.

where $t \in [1, \ldots, T]$ and $q \in [1, \ldots, Q]$ are, respectively, the index of images in the template and the query video. Note that the above computation involves T by Q comparisons whereas the original manifold-to-manifold formation, as shown in (3), has only a single comparison.

The performance of these two systems is shown in Fig. 14. In both the Mc and Ua protocols, we observe that both strategies are not significantly different, despite the claims in the literature on manifold-to-manifold. Below is a possible explanation: The success of manifold-to-manifold matching depends on several crucial factors: an appropriate manifold representation, a distance metric between two manifolds, and the length of video frames. In our exercise here, the length of video frames is limited to 50. This hinders us from using higher order moments such as covariance matrix that has been reported to be effective in [87].

It should be borne in mind that the exhaustive image-to-image comparison involves TR computation of normalized correlation whereas the manifold-to-manifold comparison involves only a single computation of this metric. A fair comparison should clearly take into consideration of the complexity or computational cost. This subject is discussed in the next section.

TABLE VI
TIME

Systems	No. of frames processed	Time per frame (ms)		Time (s)		Standard -ized time (%)	Efficiency scale
		Feature Extraction	Classification	Total time	Benchmark time		
uvigo	50	686	76.9	38.1	290.66	13.11	1
idiap-dctgmm	50	180	74	12.7	277.37	4.58	2
idiap-LBP-dctgmm	50	186	74	13.0	277.37	4.69	2
idiap-dcthmm	50	298	322	31.0	277.37	11.18	2
upv (Mc)	5	9.7	5.68	0.077	319.43	0.112	3
upv (Ua)	6			0.092		0.134	
ulj (Mc)	3	130	2547.5	8.03	396.57	2.02	4
ulj (Ua)	5			13.39		3.38	4
cwi-Cq	max 5 2.2†	107.9	25.9	0.29	486.5	0.078	4
cwi-Eq	max 8 4.8†			0.642		0.149	4
mmu	50	80	262	0.34	372.73	0.092	4

†: Average frames processed per video

F. Complexity Versus Performance

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 ground, unless both use comparable computation costs, taking the time, memory, and computational resources into consideration.

In order to take this cost factor into consideration, we requested each participant to run a benchmarking program that is executable in any operating system. Let the time registered by the program be T_{unit}. We then asked each participant to record the time needed to process a video of 50 frames (with preprocessing, feature extraction, and classification), but excluding the time needed to load a video file. Let this time be T_{process}. The standardized time is then defined as

$$\text{standardized time} = \frac{T_{\text{process}}}{T_{\text{unit}}} \times 100.$$

Table VI lists the time taken for each system for processing a video sequence of 50 frames, T_{process} (in column five), as well as the time needed to complete the benchmark software, T_{unit} (column six). The ratio of these two time measurements are shown in column seven. T_{process} can be broken down to feature extraction (including photometric normalization, and facial feature alignment) and classification. The time of these two processes are reported in the third and fourth columns in Table VI.

It should be mentioned that in timing the systems, we assumed that all software modules are loaded in memory, including the video sequence. As a result, we excluded the time needed to load a video sequence, which is highly dependent on the implementation platform. For instance, Matlab can take as long as 4 s to load a complete video into the CPU memory before processing the video whereas a highly optimized C++ code may process a video sequence in a single pass, eliminating the need to load the entire video images into the memory. Recall that our goal here is to report the system complexity in terms of standardized time, which should be as independent as possible from a particular choice of implementation platform. This is, in practice, an aspiration that is difficult to achieve. Therefore, for the sake of completeness, we also reported the efficiency in four subjective scales, as listed below:

1) Streamlined for speed—C or C++ implementation, multi-threading, using Intel library.
2) Highly efficient—C or C++ implementation, single-threading (does not exploit multiple processors or specific processor architecture).
3) Highly portable—Java implementation, single- or multi-threading.
4) Prototype—Interpreted scripts, e.g., Matlab or Python implementation.

The performance of each participant as a function of standardized time, as calculated based on Table VI, is shown in Fig. 15(a) for the controlled conditions and Fig. 15(b) for the adverse conditions. Although the HTER for the criterion WER with $\beta = 0.5$ is used here, similar results are obtained with other β values. As can be observed, more complex systems generally perform better. However, the usefulness of quality measures are not very obvious under the controlled conditions, whereas under the adverse conditions, this benefit of quality measures is immediately apparent. Not only that the two top performing systems, i.e., upv and uni-lj, achieve high generalization performance, but they achieved so with a minimal complexity (in terms of standardized time). Both systems actually exploit the quality measures in selecting images of better quality.

VII. CONCLUSIONS

This paper has presented 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 submitted systems can be conveniently grouped into four categories, depending on the dichotomies: parts-based versus holistic approach and frame-based versus image-set (video-to-video)-based comparison. While there are a number of findings, we highlight the following significant ones: First, the parts-based approach gener-

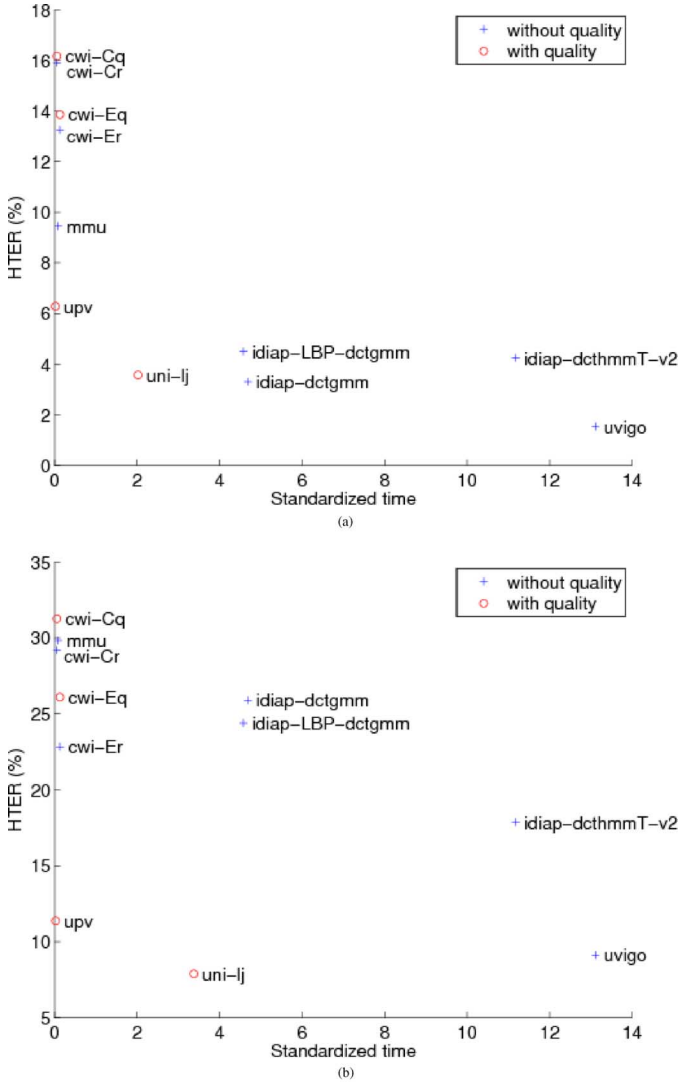


Fig. 15. Complexity versus performance of different algorithms for the (a) Mc and (b) Ua protocols.

ally outperforms the holistic one. Second, for the frame-based approach, using more query images and templates, improves the system performance. Third, the best algorithms in this evaluation clearly show the importance of selecting images based on their quality.

Two future potential research directions can be identified. First, there is a need for developing parts-based algorithms capable of comparing image-sets, hence exploiting both the robustness of parts-based algorithms and the immense potential of the temporal information. Second, in order to understand the relationship between a system output and a given set of quality measures, a systematic analysis or modeling technique is critically needed. Only with a better understanding of this relationship, one can devise algorithms capable of exploiting quality measures as a vector, rather than a scalar value.

REFERENCES

[1] D. O. Gorodnichy, "Introduction to the first IEEE workshop on face processing in video," in *Proc. Conf. Computer Vision and Pattern Recognition Workshop (CVPRW '04)*, Jun. 2004, pp. 61–61.

[2] D. O. Gorodnichy, "Video-based framework for face recognition in video," in *Proc. 2nd Canadian Conf. Computer and Robot Vision*, May 2005, pp. 330–338.

[3] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[4] D. O. Gorodnichy, "Seeing faces in video by computers. editorial for special issue on face processing in video sequences," *Image Vision Comput.*, vol. 24, no. 6, pp. 551–556, Jun. 2006.

[5] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proc. IEEE*, vol. 94, no. 11, pp. 1948–1962, Nov. 2006.

[6] H. Wechsler, *Reliable Face Recognition Methods: System Design, Implementation and Evaluation*. New York: Springer, 2007.

[7] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image Vision Comput.*, vol. 27, no. 5, pp. 545–559, 2009.

[8] D. O. Gorodnichy and O. P. Gorodnichy, "Using associative memory principles to enhance perceptual ability of vision systems," in *Proc. 2004 Conf. Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, Washington, DC, 2004, vol. 5, p. 79, IEEE Computer Society.

[9] A. J. O'Toole, P. J. Phillips, F. Jiang, J. H. Ayyad, N. Penard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1642–1646, Sep. 2007.

[10] 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.

[11] 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, C. Yueh-Hsuan, H.-C. Liu, Y.-P. Hung, A. Heinrichs, M. Muller, A. Tewes, C. vd Malsburg, R. Wurtz, Z. Wang, F. Xue, Y. Ma, Q. Yang, C. Fang, X. Ding, S. Lucey, R. Goss, and H. Schneiderman, "Face authentication test on the BANCA database," in *Int. Conf. Pattern Recognition (ICPR)*, 2004, vol. 4, pp. 523–532.

[12] 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 *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2005, pp. 947–954.

[13] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 evaluation," in *Proc. Multimodal Technologies for Perception of Humans: Int. Evaluation Workshops (CLEAR 2007 and RT 2007)*, Baltimore, MD, May 8–11, 2007, pp. 3–34, Revised Selected Papers.

[14] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, Jul. 1997.

[15] A. Fourney and R. Laganiere, "Constructing face image logs that are both complete and concise," in *Proc. 4th Canadian Conf. Computer and Robot Vision*, 2007, pp. 488–494.

[16] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, pp. 72–86, 1991.

[17] K. Okada, J. Steffans, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. Von der Malsburg, "The bochum/USC face recognition system and how it fared in the FERET phase iii test," in *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Berlin: Springer-Verlag, 1998, pp. 186–205.

[18] L. Wiskott, J.-M. Fellous, and C. Von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.

[19] S. Lin, S. Y. Kung, and L. J. Lin, "Face recognition/detection by probabilistic decision based neural network," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 114–132, Jan. 1997.

[20] B. V. K. Vijaya Kumar, M. Savides, C. Xie, K. Venkataramani, J. Thornton, and A. Mahalanobis, "Biometric verification with correlation filters," *Appl. Opt.*, vol. 43, no. 2, pp. 391–402, 2004.

[21] C. Xie, B. V. K. Vijaya Kumar, S. Palanivel, and B. Yegnanarayana, "A still-to-video face verification system using advanced correlation filters," in *Proc. Int. Conf. Biometric Authentication*, 2004, vol. 3072, Springer LNCS, pp. 102–108.

[22] P. Refregier, "Optimal trade-off filters for noise robustness, sharpness of the correlation peak, and horner efficiency," *Opt. Lett.*, vol. 16, pp. 829–831, 1991.

[23] B. V. K. Vijaya Kumar, A. Mahalanobis, and A. Takessian, "Optimal trade-off circular harmonic function correlation filter methods providing controlled in-plane rotation response," *IEEE Trans. Image Process.*, vol. 9, no. 6, pp. 1025–1034, 2000.

[24] B. V. K. Vijaya Kumar and A. Mahalanobis, "Recent advances in composite correlation filter designs," *Asian J. Phys.*, vol. 8:4, pp. 407–420, 1999.

[25] C. H. Chan, J. Kittler, N. Poh, T. Ahonen, and M. Pietikainen, "(Multiscale) local phase quantization histogram discriminant analysis with score normalisation for robust face recognition," in *Proc. 1st IEEE Workshop Video-Oriented Object and Event Classification*, Kyoto, Japan, 2009.

[26] Y.-L. Wu, L. Jiao, G. Wu, E. Y. Chang, and Y.-F. Wang, "Invariant feature extraction and biased statistical inference for video surveillance," in *Proc. IEEE Conf. Advanced Video and Signal Based Surveillance*, 2003, pp. 284–289.

[27] A. R. Dick and M. J. Brooks, "Issues in automated visual surveillance," in *Proc. 7th Digital Image Comp. Tech. and Appl.*, 2003, pp. 195–204.

[28] W. Liu, Z. F. Li, and X. Tang, "Spatio-temporal embedding for statistical face recognition from video," in *Proc. Eur. Conf. Computer Vision*, 2006, vol. 3952, Pt. II, Springer LNCS, pp. 374–388.

[29] Y. Zhang and A. M. Martinez, "A weighted probabilistic approach to face recognition from multiple images and video sequences," *Image Vision Comput.*, vol. 24, pp. 626–638, 2006.

[30] Y. Xu, A. Roy Chowdhury, and K. Patel, "Pose and illumination invariant face recognition in video," in *Proc. CVPR*, 2007, pp. 1–7.

[31] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. SIGGRAPH99*, 1999, pp. 187–194.

[32] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[33] A. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Generative models for recognition under variable pose and illumination," in *Proc. Automatic Face and Gesture Recognition*, 2000, pp. 277–284.

[34] U. Park and A. K. Jain, "3D model-based face recognition in video," in *Proc. Int. Conf. Biometrics*, 2007, pp. 1085–1094.

[35] D. DeCarlo and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *Int. J. Comput. Vis.*, vol. 38, pp. 99–127, 2000.

[36] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2003, pp. 313–320.

[37] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment – A modern synthesis," in *Proc. Int. Workshop on Vision Algorithms (ICCV '99)*, 2000, pp. 298–372, Springer-Verlag.

[38] Y. Li, S. Gong, and H. Liddell, "Modelling face dynamics across views and over time," in *Proc. Int. Conf. Computer Vision*, 2000, pp. 554–559.

[39] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[40] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Learning to identify and track faces in image sequences," in *Proc. 6th Int. Conf. Computer Vision*, Jan. 1998, pp. 317–322.

[41] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2D+3D active appearance models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 535–542.

[42] R. Gross, I. Matthews, and S. Baker, "Constructing and fitting active appearance models with occlusion," in *Proc. Conf. Computer Vision and Pattern Recognition Workshop, 2004 (CVPRW '04)*, Jun. 2004, p. 72.

[43] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image Vision Comput.*, vol. 23, no. 1, pp. 1080–1093, 2005.

[44] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. Int. Conf. Computer Vision*, 2009, pp. 1034–1041.

[45] D. Beymer and T. Poggio, "Face recognition from one example view," in *Proc. IEEE Int. Conf. Computer Vision*, 1995, pp. 500–507.

[46] J. Heo and M. Savides, "In between 3D active appearance models and 3D morphable models," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops, 2009 (CVPR Workshops 2009)*, Jun. 2009, pp. 20–26.

[47] A. Roy Chowdhury and R. Chellappa, "Face reconstruction from monocular video using uncertainty analysis and a generic model," *Comput. Vision Image Understanding*, vol. 91, no. 1-2, pp. 188–213, 2003.

[48] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 1994 (CVPR '94)* , Jun. 1994, pp. 84–91.

[49] Y. Li, S. Gong, and H. Liddell, "Constructing facial identity surfaces in a nonlinear discriminating space," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 2001 (CVPR 2001)*, 2001, vol. 2, pp. II-258–II-263.

[50] X. Chai, S. Shan, X. Chen, and W. Gao, "Local linear regression (llr) for pose invariant face recognition," in *Proc. Automatic Modeling of Face and Gesture*, 2006, pp. 631–636.

[51] S. J. McKenna and S. Gong, "Non-intrusive person authentication for access control by visual tracking and face recognition," in *Proc. Int. Conf. Audio- and Video-Based Person Authentication*, 1997, pp. 177–183.

[52] B. Li and R. Chellappa, "Face verification through tracking facial features," *J. Opt. Soc. Amer.*, vol. 18, pp. 2969–2981, 2001.

[53] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," in *Proc. Eur. Conf. Computer Vision*, 1996, pp. 343–356.

[54] J. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *J. Amer. Statist. Assoc.*, vol. 93, pp. 1031–1041, 1998.

[55] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Comput. Vision Image Understanding*, vol. 91, no. 1, pp. 214–245, 2003.

[56] J. M. Buenaposada, J. Bekios, and L. Baumela, "Apperance-based tracking and face identification in video sequences," in *Proc. AMDO*, 2008, pp. 349–358.

[57] J. M. Buenaposada, E. Munoz, and L. Baumela, "Efficiently estimating facial expression and illumination in appearance-based tracking," in *Proc. British Machine Vision Conf.*, 2006, pp. 57–66.

[58] K. C. Lee and D. Kriegman, "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," in *Proc. CVPR*, 2005, vol. I, pp. 852–859.

[59] F. Matta and J.-L. Dugelay, "Video face recognition: A physiological and behavioural multimodal approach," in *Proc. ICIP (4)*, 2007, pp. 497–500.

[60] G. Aggarwal, A. K. Roy Chowdhury, and R. Chellappa, "A system identification approach for video-based face recognition," in *Proc. ICPR (4)*, 2004, pp. 175–178.

[61] G. Shakhnarovich, J. W. Fisher, and T. Darrel, "Face recognition from long-term observations," in *Proc. Eur. Conf. Computer Vision*, 2002, Springer LNCS, pp. 851–868.

[62] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 2005 (CVPR 2005)*, Jun. 2005, vol. 1, pp. 581–588.

[63] S. Satoh, "Comparative evaluation of face sequence matching for content-based video access," in *Proc. Fourth IEEE Int. Conf. Automatic Face and Gesture Recognition, 2000*, 2000, pp. 163–168.

[64] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 10, 1998, pp. 318–323.

[65] M. Nishiyama, O. Yamaguchi, and K. Fukui, "Face recognition with the multiple constrained mutual subspace method," in *Lecture Notes in Computer Science*, 2005, vol. 3546/2005, pp. 71–80.

[66] L. Wolf and A. Shashua, "Learning over sets using kernel principal angles," *J. Mach. Learning Res.*, vol. 4, no. 10, pp. 913–931, 2003.

[67] T. K. Kim, O. Arandjelović, and R. Cipolla, "Learning over sets using boosted manifold principal angles (BoMPA)," in *Proc. IAPR British Machine Vision Conf. (BMVC)*, Oxford, U.K., Sep. 2005, pp. 779–788.

[68] K. Fukui and O. Yamaguchi, "Face recognition using multi-viewpoint patterns for robot vision," in *Robotics Research, Springer Tracts in Advanced Robotics*, 2003, vol. 15/2005, pp. 192–201.

[69] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Patern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.

[70] E. Kokiopoulou and P. Frossard, "Video face recognition with graph-based semi-supervised learning," in *Proc. ICME*, 2009, pp. 1564–1565.

[71] A. Mian, "Unsupervised learning from local features for video-based face recognition," in *Proc. 8th IEEE Int. Conf. Automatic Face Gesture Recognition, 2008 (FG '08)*, Sep. 2008, pp. 1–6.

[72] D. Lowe, "Distinctive image features from scale-invariant key points," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[73] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden markov models," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2003, pp. 340–345.
[74] R. Gross and V. Brajovic, "An image preprocessing algorithm for illumination invariant face recognition," in *Audio- and Video-Based Biometric Person Authentication*. Berlin/Heidelberg: Springer, Jun. 2003, vol. 2688/2003, Lecture Notes in Computer Science, pp. 10–18.
[75] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, Jul. 1997.
[76] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–94, Jan. 1980.
[77] J.-L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 291–298, Apr. 1994.
[78] J. L. Alba Castro, D. González Jiménez, E. A. Rúa, E. González Agulla, and E. Otero Muras, "Pose-corrected face processing on video sequences for webcam-based remote biometric authentication," *J. Electron. Imaging*, vol. 17, no. 1, p. 011004, Jan. 2008.
[79] F. Cardinaux, C. Sanderson, and S. Marcel, "Comparison of MLP and GMM classifiers for face verification on XM2VTS," in *Lecture Notes in Computer Science*, 2003, vol. 2688/2003, pp. 1058–1059, Springer.
[80] G. Heusch, Y. Rodriguez, and S. Marcel, "Local binary patterns as an image preprocessing for face authentication," in *IEEE Int. Conf. Automatic Face and Gesture Recognition (AFGR)*, 2006, pp. 9–14.
[81] F. Cardinaux, C. Sanderson, and S. Bengio, "User authentication via adapted statistical models of face images," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 361–373, Jan. 2005.
[82] S. Baker, I. Matthews, and J. Schneider, "Automatic construction of active appearance models as an image coding problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1380–1384, Oct. 2004.
[83] T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C. J. Taylor, "Groupwise diffeomorphic non-rigid registration for automatic model building," in *Proc. ECCV*, 2004, pp. 316–327.
[84] 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.
[85] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[86] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[87] H. Fang and N. Costen, "Behavioral consistency extraction for face verification," in *Revised Selected and Invited Papers, Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions: COST Action 2102 Int. Conf.*, Prague, Czech Republic, Oct. 15–18, 2008, pp. 291–305.
[88] M. Villegas and R. Paredes, "Simultaneous learning of a discriminative projection and prototypes for nearest-neighbor classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2008)* , 2008, pp. 1–8.
[89] R. Paredes, J. C. Pérez, A. Juan, and E. Vidal, "Local representations and a direct voting scheme for face recognition," in *Proc. Workshop on Pattern Recognition in Information Systems (PRIS 01)*, Setúbal, Portugal, Jul. 2001.
[90] M. Villegas and R. Paredes, "Illumination invariance for local feature face recognition," in *Proc. 1st Spanish Workshop on Biometrics*, Girona, Spain, Jun. 2007.
[91] M. Villegas, R. Paredes, A. Juan, and E. Vidal, "Face verification on color images using local features," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2008)*, Jun. 2008, pp. 1–6.
[92] V. Štruc, B. Vesnicer, and N. Pavešić, "The phase-based gabor fisher classifier and its application to face recognition under varying illumination conditions," in *Proc. 2nd Int. Conf. Signal Processing and Communication Systems*, Gold Coast, Australia, Dec. 15–17, 2008.
[93] C. Liu, "Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 725–737, May 2006.
[94] V. Štruc and N. Pavešić, "The corrected normalized correlation coefficient: A novel way of matching score calculation for lda-based face verification," in *Proc. 5th Int. Conf. Fuzzy Systems and Knowledge Discovery*, Jinan, China, Oct. 18–20, 2008, pp. 110–115.
[95] H. K. Ekenel and R. Stiefelhagen, "Local appearance based face recognition using discrete cosine transform," in *Proc. 13th Eur. Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, 2005.
[96] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio-based biometric score fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 342–347, Feb. 2008.
[97] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA database and evaluation protocol," in *Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, 2003, vol. 2688, Springer-Verlag LNCS, p. 1057.
[98] N. Poh, C. H. Chan, J. Kittler, S. Marcel, C. Mc Cool, E. A. Rúa, J. L. Alba Castro, M. Villegas, R. Paredes, V. Štruc, N. Pavešić, A. A. Salah, H. Fang, and N. Costen, "Face video competition," in *Proc. 3rd Int. Conf. Biometrics*, Sardinia, 2009, pp. 715–724.
[99] 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.
[100] S. Bengio and J. Mariéthoz, "The expected performance curve: A new assessment measure for person authentication," in *Proc. Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.
[101] N. Poh and S. Bengio, "Database, protocol and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recognit.*, vol. 39, no. 2, pp. 223–233, Feb. 2005.
[102] A. Martin, M. Przybocki, and J. P. Campbell, *The NIST Speaker Recognition Evaluation Program*. New York: Springer, 2005, ch. 8.
[103] N. Poh, G. Heusch, and J. Kittler, "On combination of face authentication experts by a mixture of quality dependent fusion classifiers," in *Multiple Classifiers System (MCS)*, Prague, 2007, vol. 4472, LNCS, pp. 344–356.
[104] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
[105] *Digital Image Processing*, R. C. Gonzalez and R. E. Woods, Eds., 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 2008.
[106] M. Villegas and R. Paredes, "Fusion of qualities for frame selection in video face verification," in *20th Int. Conf. Pattern Recognition (ICPR 2010)*, Istanbul, Turkey, Aug. 23–26, 2010, pp. 1–4.

Norman Poh (S'02–M'06) received the Ph.D. degree in computer science from the Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland, in 2006.

He is a research fellow, leading the EU-funded Mobile Biometry Project at the University of Surrey, U.K.

Dr. Poh was the recipient of five best paper awards in the areas of biometrics and human computer interaction and two fellowship research grants from the Swiss National Science Foundation.

Chi Ho Chan received his Ph.D. degree in electrical and computer engineering from the University of Surrey, U.K. in 2008.

He has served a researcher at ATR International (Japan) from 2002 to 2004. Currently, he is a research fellow working with Prof. Josef Kittler. His areas of expertise are image processing, pattern recognition, biometrics, and vision-based human–computer interaction.

Josef Kittler received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge in 1971, 1974, and 1991, respectively.

He heads the Centre for Vision, Speech and Signal Processing at the School of Electronics and Physical Sciences, University of Surrey, U.K. He teaches and conducts research in the subject area of machine intelligence, with a focus on biometrics, video and image database retrieval, automatic inspection, medical data analysis, and cognitive vision. He published a Prentice-Hall textbook on *Pattern Recognition: A Statistical Approach* and several edited volumes, as well as more than 600 scientific papers, including in excess of 170 journal papers. He serves on the Editorial Board of several scientific journals in pattern recognition and computer vision.

Sébastien Marcel received the Ph.D. degree in signal processing from "Université de Rennes I" in France (2000) at CNET, the research center of France Telecom (now Orange Laboratories).

He is senior research scientist at the Idiap Research Institute, Martigny, Switzerland. In 2010, he was appointed Visiting Professor at the University of Cagliari (IT) where he taught a series of lectures in "face recognition." He leads a research team and manages research projects on biometric person recognition. He is currently interested in multimodal biometric person recognition, man-machine and content-based multimedia indexing and retrieval.

Christopher Mc Cool received the Ph.D. degree from Queensland University of Technology, Australia, in 2007.

He is a postdoctoral researcher at the Idiap Research Institute, Martigny, Switzerland. His research interests include pattern recognition and computer vision with a particular empahsis on biometrics, 3-D face verification, 2-D face verification, and face detection.

Enrique Argones Rúa received the Ph.D. degree from the Signal Theory and Communications Department, University of Vigo, in 2008, where he is performing post-doctoral research.

His research interests include pattern recognition in dynamic signals, information fusion and their applications to biometrics, including speaker recognition, video-based face recognition, signature recognition, and multimodal fusion.

José Luis Alba Castro received the Ph.D. degree in telecommunications engineering from the University of Vigo, Spain, in 1997.

He is currently an Associate Professor at the University of Vigo, teaching image processing, statistical pattern recognition, machine learning, and biometrics. His research interests include signal and image-based biometrics, computer vision for driver assistance, and computer vision for quality control. He has been the Leader of several research projects and contracts on these topics, and is the head of the computer vision laboratory of the University of Vigo.

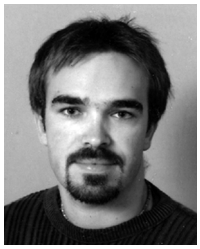
Mauricio Villegas (S'02) received the B.S. degree in telecommunications engineering from the Universidad Santo Tomas, Colombia, in 2004. He received the M.S. degree from the Departamento de Sistemas Informáticos y Computación of the Universidad Politécnica de Valencia (UPV), in 2008, and in the same university, he is currently pursuing the Ph.D. degree.

Since 2005, he has been with the Instituto Tecnológico de Informática, Velencia, Spain, working on biometrics and pattern recognition projects. His research interests include pattern recognition, dimensionality reduction, image processing, biometrics, and computer vision.

Roberto Paredes received the Ph.D. degree in computer science from the Universidad Politécnica de Valencia, Spain, in 2003.

From 1998 to 2000, he was with the Instituto Tecnológico de Informática working on computer vision and pattern recognition projects. In 2000, he joined the Departamento de Sistemas Informáticos y Computación of the Universidad Politécnica de Valencia (UPV), where he is an Assistant Professor in the Facultad de Informática. His current fields of interest include statistical pattern recognition and biometric identification. In these fields, he has published several papers in journals and conference proceedings.

Dr. Paredes is a member of the Spanish Society for Pattern Recognition and Image Analysis (AERFAI).



Vitomir Štruc (M'10) received the B.Sc. degree in electrical engineering from the University of Ljubljana, in 2005. He is currently working toward the Ph.D. degree from the University of Ljubljana, Slovenia.

He works as a software developer at Alpineon Ltd. His research interests include pattern recognition, machine learning, and biometrics.

Mr. Štruc is a member of the Slovenian Pattern Recognitionb Society.

Nikola Pavešić was born in 1946. He received the B.Sc. degree in electronics, the M.Sc. degree in automatics, and the Ph.D. degree in electrical engineering from the University of Ljubljana, Slovenia, in 1970, 1973, and 1976, respectively.

Since 1970, he has been a staff member at the Faculty of Electrical Engineering in Ljubljana, where he is currently head of the Laboratory of Artificial Perception, Systems and Cybernetics. His research interests include pattern recognition, neural networks, image processing, speech processing, and information theory. He is the author and coauthor of more than 100 papers and three books addressing several aspects of the above areas.

Albert Ali Salah received the Ph.D. degree at the Perceptual Intelligence Laboratory, Boğaziçi University, in 2007.

He is currently with the Informatics Institute at the University of Amsterdam. His research interests are biologically inspired models of learning and vision, with applications to pattern recognition, biometrics, and human behavior understanding.

With his work on facial feature localization, Dr. Salah received the inaugural EBF European Biometrics Research Award in 2006.

Hui Fang received the B.Sc. degree from Science and Technology University, Beijing, China, in 2000, and the Ph.D. degree from Bradford University, U.K., in 2006.

He worked as a Postdoctoral Research Associate in Manchester Metropolitan University. He is now a research officer in the Computer Science Department, Swansea University, Wales, from 2009. His research interests include facial recognition and analysis, image registration, and image modeling.

Nicholas Costen received the B.A. degree in experimental psychology from the University of Oxford, and the Ph.D. degree in mathematics and psychology from the University of Aberdeen.

He has undertaken research at the Advanced Telecommunications Research Laboratory, Kyoto, and at the Division of Imaging Science and Biomedical Engineering, University of Manchester. He is a Senior Lecturer at Manchester Metropolitan University, U.K., where his interests include face recognition, human motion analysis, and ultrasound interpretation.