Evaluation methodology for face recognition technology in video surveillance applications

Prepared by:
Eric Granger
Université Québec
1100, rue Notre-Dame Ouest
Montréal (Québec) H3C 1K3
and
Dmitry O. Gorodnichy
Canada Border Services Agency
Ottawa ON Canada K1A 0L8

Scientific Authority: Pierre Meunier, DRDC Centre for Security Science, 613-992-0753

Contract Report
DRDC-RDDC-2014-C249
September 2014

IMPORTANT INFORMATIVE STATEMENTS

PROVE-IT (FRiV) Pilot and Research on Operational Video-based Evaluation of Infrastructure and Technology: Face Recognition in Video project, PSTP 03-401BIOM, was supported by the Canadian Safety and Security Program (CSSP) which is led by Defence Research and Development Canada's Centre for Security Science, in partnership with Public Safety Canada. Led by Canada Border Services Agency partners included: Royal Canadian Mounted Police, Defence Research Development Canada, Canadian Air Transport Security Authority, Transport Canada, Privy Council Office; US Federal Bureau of Investigation, National Institute of Standards and Technology, UK Home Office; University of Ottawa, Université Québec (ÉTS).

The CSSP is a federally-funded program to strengthen Canada's ability to anticipate, prevent/mitigate, prepare for, respond to, and recover from natural disasters, serious accidents, crime and terrorism through the convergence of science and technology with policy, operations and intelligence.

Canada Border Services Agency
Agence des services frontaliers du Canada

Science and Engineering Directorate

Border Technology Division

Division Report 2014-29 (TR)
July 2014

Evaluation methodology for face recognition technology in video surveillance applications

E. Granger, D. Gorodnichy

This page left intentionally blank

Abstract

This report reviews metrics, methodologies and data-sets used for evaluation of face recognition in video (FRiV) and establishes a multi-level evaluation methodology that is suitable for video surveillance applications. The developed methodology is particularly tailored for such video surveillance applications as screening of faces against the wanted list (still-to-video application) and matching faces across several video feeds, also known as search-and-retrieval or face re-identification problem (video-to-video application). According to the developed methodology, FRiV technologies are evaluated at several levels of analysis, each level dealing with a particular source of potential system failure. Level 1 (transaction-based analysis) deals with unbalanced target vs. non-target distributions. Level 2 (subject-based analysis) deals with robustness of the system to different types of target faces. Level 3 (time-based analysis) allows to examine the quality of the final decision while tracking a person over time. The methodology is applied to conduct an evaluation of state-of-art Commercial Off-The-Shelf (COTS) face recognition systems, the results of which are presented.

Keywords: video-surveillance, face recognition in video, instant face recognition, watch list screening, biometrics, reliability, performance evaluation

Community of Practice: Biometrics and Identity Management

Canada Safety and Security (CSSP) investment priorities:

1. Capability area: P1.6 – Border and critical infrastructure perimeter screening technologies/ protocols for rapidly detecting and identifying threats.
2. Specific Objectives: O1 – Enhance efficient and comprehensive screening of people and cargo (identify threats as early as possible) so as to improve the free flow of legitimate goods and travellers across borders, and to align/coordinate security systems for goods, cargo and baggage;
3. Cross-Cutting Objectives CO1 – Engage in rapid assessment, transition and deployment of innovative technologies for public safety and security practitioners to achieve specific objectives;
4. Threats/Hazards F – Major trans-border criminal activity – e.g. smuggling people/ material

Acknowledgements

This work is done within the PROVE-IT(FRiV) project (PSTP-03-401BIOM) funded by the Defence Research and Development Canada (DRDC) Centre for Security Science (CSS) Public Security Technical Program (PSTP) and in-kind contributions from École de technologie supérieure by the following contributors:

1. **Dr. Dmitry Gorodnichy**, Research Scientist with the Science & Engineering Directorate, Canada Border Services Agency; Adjunct Professor at École de technologie supérieure, Université du Québec.

2. **Dr. Eric Granger**, Professor of Systems Engineering at École de technologie supérieure, Université du Québec; Research Scientist with the Science & Engineering Directorate of the Canada Border Services Agency (2011-2012, on sabbatical leave from the university).

The feedback from project partners: University of Ottawa (R. Laganiere, S. Matwin), RCMP, TC, CATSA, DRDC, UK HomeOffice, FBI is gratefully acknowledged. Assistance from W. Khreich and M. De le Torre from École de technologie supérieure and E. Choy, J.-P. Bergeron, and David Bissessar from Canada Border Services Agency in conducting the experiments with commercial face recognition products is also gratefully acknowledged.

Disclaimer

In no way do the results presented in this paper imply recommendation or endorsement by the Canada Border Services Agency (CBSA), nor do they imply that the products and equipment identified are necessarily the best available for the purpose. The information presented in this report contains only the information available in the public domain. This work was conducted from September 2011 till March 2013 and may not reflect the technology development after that date.

The results presented in this report were produced in experiments conducted by the CBSA, and should therefore not be construed as vendor's maximum-effort full-capability result. Additionally, it is noted that the Cognitec technology was operated under the conditions that are outside of its specifications.

Release Notes

Context: This document is part of the set of reports produced for the PROVE-IT(FRiV) project. All PROVE-IT(FRiV) project reports are listed below.

- Dmitry Gorodnichy and Eric Granger "PROVE-IT(FRiV): framework and results". Proceedings of NIST International Biometrics Performance Conference (IBPC 2014), Gaithersburg, MD, April 1-4, 2014. Online at http://www.nist.gov/itl/iad/ig/ibpc2014.cfm.
- Dmitry Gorodnichy and Eric Granger, "Evaluation of Face Recognition for Video Surveillance". Proceedings of NIST International Biometric Performance Conference (IBPC 2012), Gaithersburg, March 5-9, 2012. Online at http://www.nist.gov/itl/iad/ig/ibpc2012.cfm.
- D. Bissessar, E. Choy, D. Gorodnichy, T. Mungham, "Face Recognition and Event Detection in Video: An Overview of PROVE-IT Projects (BIOM401 and BTS402)", Border Technology Division, Division Report 2013-04 (TR).
- E. Granger, P. Radtke, and D. Gorodnichy, "Survey of academic research and prototypes for face recognition in video", Border Technology Division, Division Report 2014-25 (TR).
- D. Gorodnichy, E.Granger, and P. Radtke, "Survey of commercial technologies for face recognition in video", Border Technology Division, Division Report 2014-22 (TR).
- E. Granger and D. Gorodnichy, "Evaluation methodology for face recognition technology in video surveillance applications", Border Technology Division, Division Report 2014-27 (TR).
- E. Granger, D. Gorodnichy, E. Choy, W. Khreich, P. Radtke, J.-P. Bergeron, and D. Bissessar, "Results from evaluation of three commercial off-the-shelf face recognition systems on Chokepoint dataset", Border Technology Division, Division Report 2014-29 (TR).
- S. Matwin, D. Gorodnichy, and E. Granger, "Using smooth ROC method for evaluation and decision making in biometric systems", Border Technology Division, Division Report 2014-10 (TR).
- D. Gorodnichy, E. Granger, S. Matwin, and E. Neves, "3D face generation tool Candide for better face matching in surveillance video", Border Technology Division, Division Report 2014-11 (TR).
- E. Neves, S. Matwin, D. Gorodnichy, and E. Granger, "Evaluation of different features for face recognition in video", Border Technology Division, Division Report 2014-31 (TR).

The PROVE-IT(FRiV) project took place from August 2011 till March 2013. This document was drafted and discussed with project partners in March 2013 at the Video Technology for National Security (VT4NS) conference. The final version of it was produced in June 2014.

Supplemental material: The presentation related to this report presented at the VT4NS 2013 conference is included as Appendix. This report is also accompanied by an additional report entitled "Results from evaluation of three commercial off-the-shelf face recognition systems on Chokepoint dataset", which provides complete experimental results obtained using the methodology described in this report.

Typesetting: The tabulated content in this report was produced automatically using LaTeX for improved source control, flexibility and maintainability. The report contains automatically generated hyper-link references and table of contents for easier navigation and reading on-line.

Contact: Correspondence regarding this report should be directed to DMITRY dot GORODNICHY at CBSA dot GC dot CA.

Contents

List of Figures

List of Tables

1 Introduction

The global market for video surveillance technologies has reached revenues in the billions of $US as traditional analog technologies are being replaced by IP-based digital surveillance technologies. In this context, video surveillance based on the facial biometric modality is extremely useful. The ability to automatically recognize and track individuals of interest in crowded airports or other public places, and across a network of surveillance cameras may provide enormous benefits in terms of enhanced screening and situational analysis.

Two main types of face recognition in video (FRiV) are possible:

1. **still-to-video** recognition, which deals with matching of facial images in a video stream with still images stored in gallery.
2. **video-to-video** recognition, also referred to as **re-identification**, which deals with matching of facial images in a video stream with facial images captured in another video stream.

Still-to-video recognition is used in such applications as Watch List Screening, where faces extracted from video are matched to faces stored in a Watch List. Video-to-video recognition is used in such applications as search and retrieval, face tagging, video summarization, and face tracking and re-detection across multiple video streams.

To evaluate and compare the performance of face recognition (FR) technologies and academic systems, objective evaluation methodologies are required. Such methodologies depend on the task for which the FR technologies are used. In video surveillance applications, the task of a FR system is to detect the presence of an individual from a watch list gallery or cohort [29, 36, 34] in a variety of video surveillance environments, which can range from semi-controlled (one person, little motion) to uncontrolled environments with crowded and moving scenes.

Systems and technologies for Face Recognition in Video Surveillance (FRiVS) should be evaluated in terms of their ability to accurately and efficiently detect the presence of an individual's face under various operational conditions. In *still-to-video* recognition, a cohort typically corresponds to individuals populating a pre-established watch list, while in *video-to-video* recognition, it is typically a set of suspicious individuals to be monitored in a scene.

The divergence and uncertainty of facial models w.r.t. faces collected in real-world video scenes (due to different cameras and uncontrolled and changing environments) underscores the need to assess image quality, context distortion of input Regions of Interest (ROIs) [35, 40, 43]. FR in video surveillance corresponds to an *open-set* or *open-world* FR problem [29], in which only a very small proportion of captured faces correspond to an individual of interest in a restrained cohort. FR systems are typically designed using the samples from a Universal Model (UMs) to set decision thresholds and/or design accurate matchers. Some individuals are naturally more difficult to detect than others, and the risk associated with detection errors varies from one individual to another. Therefore, performance must be also assessed with skewed samples and cost-sensitive decisions scenarios in mind. Finally, video surveillance networks are comprised of a growing number of IP-based surveillance cameras and must transmit or archive massive quantities of data. Storage and processing time of different systems is also an important consideration.

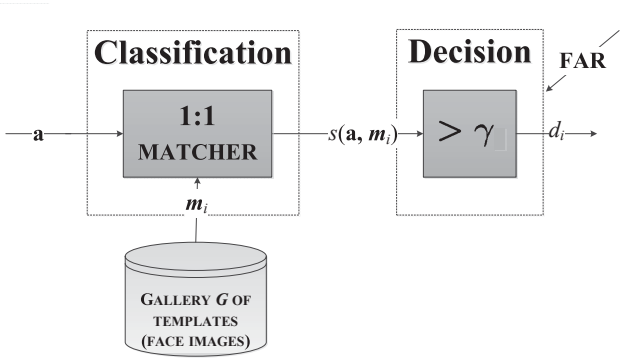


Figure 1: Illustration of a basic matcher used in a biometric system. Decision threshold is set based on the allowable False Accept Rate (FAR).

For evaluation, a FR system in video surveillance should not be viewed as an identification problem, where each input ROI is simultaneously compared against all user templates and where the system outputs a ranking of the best matching candidate [1], but rather should be viewed in terms of independent user-specific detection (1:1 classification) problems, each one implemented using cost-effective template matchers or one- or two-class pattern classifiers, and a threshold applied to its output scores $s(a,m_i)$ [2, 3, 13]. In surveillance applications, the objective is to detect the presence of an individual of interest among a universe of people, regardless of other individuals in the cohort. The system outputs a list of all possible matching individuals, that is, cases where $d_i = 1(s(a,m_i) \geq \gamma)$. As it is observed in practice however, most biometric technologies, including those developed for FR in video surveillance, implement screening tasks as a sequence of independent 1:1 classification tasks – comparisons of each input ROI (pattern **a**) against user templates, over a gallery G of templates m_i (see Figure 1).

In this report, we establish experimental methodologies for large-scale performance evaluation of state-of-the-art commercial technologies and academic systems for both still-to-video and video-to-video FRiVS. These methodologies are particularly tailored for such video surveillance applications as screening of faces against wanted lists (still-to-video application) and matching a face/person across several video feeds, also

[1] For a system that is implemented an an identification problem system, the Cumulative Match Curve (CMC) is a well established performance measure that provides ranked lists of candidates. It has been shown by Bolle *et al.* [5] that the CMC is related to the false positive/negative rates (fpr and fnr) of a 1:1 classifier used to rank the candidates by sorting their classification scores from high to low. As a consequence, when a 1:1 classifier is applied to identification (i.e., for sorting classification scores), the CMC can be derived from the fpr and fnr, and does not provide any additional information beyond ROC or DET curves. As emphasized recently by DeCann and Ross [10], the opposite however is not true: the same CMC curve can be generated by different ROC or DET curves.

Table 1: Datasets for FR in video-surveillance.

DATASET	TARGET APPLICATIONS
1) CMU-MoBo [25] Carnegie Mellon University Motion of Bodies	subjects performing different walking patterns on a treadmill
2) CMU-FIA [15] Carnegie Mellon University Faces in Action	subjects mimicking passport checkpoint at airport
3) Chokepoint [47]	video-surveillance subjects walking though portals
4) MOBIO [32] EC FP7 Mobile Biometry	m-modal unconstrained authentication on mobile device (face + voice)
5) ND-Q0-Flip [33] Notre-Dame Crowd Data	detection of questionable observers that appear often in crowd videos
6) NIST-MBGC(http://www.nist.gov/itl/iad/ig/mbgc.cfm) National Institute of Standards and Technology - Multiple Biometric Grand Challenge	m-modal verification of subjects walking through portal or access control checkpoint (still- and video-to-video)
7) NRC-FRiV [19] National Research Council - Face Recognition in Video	user identification for secured computer login
8) XM2VTS(http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/) Multi-Modal Verification for Teleservices and Security Applications	multi-modal verification for tele-service and security

known as search-and-retrieval, face re-identification problem (video-to-video application)

The report is organized as follows. First, we survey publicly-available reference data-sets (Section 2) that can be used for the purpose. Then we survey performance metrics (Section 3) that are suitable for video surveillance applications. Finally, we present a benchmarking protocol (Section 4), which is developed for testing commercial and academic FR technologies for video surveillance applications. The complete results obtained from the evaluation of commercial FR products using the presented protocol are presented in a separate report [23].

2 Reference data-sets

2.1 Publicly-available data

Table 1 presents a survey of publicly-available data-sets that are suitable for medium- to large-scale benchmarking of systems for mono-modal and multi-modal recognition of face in video-surveillance. Due to constraints of standard evaluation protocols and performance metrics, these data-sets must contain at least 10 unique individuals, and a considerable number of frames per person. In addition, it is desirable to use common data-sets that allow for comparison with results found in open literature.

For evaluation of video surveillance techniques and technologies, the CMU-FIA, Chokepoint and ND-Q0-Flip data-sets are are the most suitable for mono-modal recognition and tracking of faces over one or

more cameras. For applications involving multi-modal recognition and tracking, MOBIO (face and voice modalities) and NIST-MBGC[2] (face and iris modalities) data-sets are the most suitable for benchmarking studies. Specific evaluation scenarios that involve these data-sets are presented in the next sections.

In Tables 2 through 4, CMU-FIA and Chokepoint, MOBIO data-sets from Table 1 are characterized on the basis of:

- demographics: the distribution of people per session and in the entire data-set.
- complexity in scene: the systematic variation of illumination, occlusion, motion, expression and/or pose for some target application;
- capture properties: the number and type of cameras, duration of video sequences, frame rate and resolution

Properties for CMU-MoBo and NRC-FRiV are also shown for reference. Finally, some additional details are provided only for data-sets used in case studies.

[2]Note that NIST-MBGC includes several data-sets, and is most relevant here for combined face-iris recognition at primary inspection lanes.

Table 2: Demographics of reference data-sets.

Dataset	**no. subjects per sequence**	**no. unique subjects in population**	**male:female ratio**	**age (years)**
CMU-MoBo	1	25	23:2	N/A
CMU-FIA	1	session 1: 214 session 2: 180 (subset) session 3: 153 (subset)	session 1: 61.3%:38.7%	session 1: 18-57 (mean: 25.4)
Chokepoint	1 and 24	portal 1: 25 portal 2: 29	portal 1: 19:6 portal 2: 23:6	N/A
MOBIO	1	152	100:52	N/A
ND-Q0-Flip	4 to 12	90 (5 appear in several)	N/A	N/A
NRC-FRiV	1	11	10:1	N/A

Table 3: Complexity of scene.

Dataset	**illumination**	**occlusion**	**expressions, gestures, pose and other motion**	**ageing**
CMU-MoBo	indoor (controlled)	partial	y	no
CMU-FIA	indoor (controlled): fixed background, fluorescent lighting outdoor (natural): different seasons and climate	y	y	3 sessions over 10 months
Chokepoint	indoor (controlled)	y	y	2 sessions 1 month apart (portal 1 and 2)
MOBIO				6 sessions on 6 sites over 2 years
ND-Q0-Flip	2 indoor (controlled) 12 outdoor: 6 overcast; 6 sunny	partial	y	14 sessions over 7 months
NRC-FRiV	indoor (controlled) similar background	partial	y	none, 2 consecutive sessions

Table 4: Camera Capture Properties.

Dataset	**no. and type of cameras**	**resolution**	**frame rate**	**no. and duration of sequences**	**camera angle and placement**
CMU-MoBo	6 Sony DXC9000	VGA: 640x480	30	6 cameras x 4 motion sequences of 11 sec	overhead view with 6 horizontal angles
CMU-FIA	6 Dragonfly (Point Grey Research) Sony ICX424 sensor	3 at 100x100 (4mm foc.len.) 3 at 300x300 (8mm foc.len.)	30	54 sequences of 20 sec	cameras at 0.83m, mounted on carts at 3 horiz. angles 0^0 (frontal) and $\pm 72.6^0$
Chokepoint	3	SVGA: 800x600	30	x sequences of x sec	cameras placed above portals at 3 different angles
MOBIO	camera in Nokia N93i + iSight webcam in 2008 MacBook	VGA: 640x480	30	6 sessions/subject	cameras inside mobile devices (phone + laptop)
ND-Q0-Flip	1 CISCO Flip Q0 handheld camcorder	VGA: 640x480 (H.264 compression)	30	14 sequences of 25-59 sec	frontal view, camera pans and zooms over a crowd of subjects
NRC-FRiV	1 USB Webcam	160x120 (Intel video codec)	20	2 sequences of approx. 15 sec	computer monitor mounted at eye level

2.1.1 NRC-FRiV data

The NRC-FRiV [19] data-set is composed of 22 video sequences captured from eleven individuals positioned in front of a computer. For each individual, two color video sequences of about fifteen seconds are captured at a rate of 20 frames per seconds with an Intel web cam of a 160×120 resolution that was mounted on a computer monitor. Of the two video sequences, one is dedicated to training and the other to testing. They are taken under approximately the same illumination conditions, the same setup, almost the same background, and each face occupies between $1/4$ to $1/8$ of the image. This data base contains a variety of challenging operational conditions such as motion blur, out of focus factor, facial orientation, facial expression, occlusion, and low resolution. The number of ROIs detected varies from class to class, ranging from 40 to 190 for one video sequences. Figure 2 shows some frames captured for that data-set.

Figure 2: Video frames captured in the reference NRC-FRiV data-set [19].

2.1.2 CMU-MoBo data

The CMU-MoBo [25] data-set, and was collected at Carnegie Mellon University under the HumanID project. Each video sequence shows one of 25 different individuals on a tread-mill so that they move their heads naturally to four different motion types when walking: slowly, fast, on an inclined surface, and while carrying an object. Six Sony DXC 9000 cameras, with a resolution of a 640×480 pixels, are positioned at different locations around the individuals. Only the video sequences with visible faces were kept: full frontal view and both sides with an angle of about 70° with the full frontal view. Figure 3 shows some frames captured for that data-set.

Figure 3: Video frames from the reference CMU-MoBo data-set [25].

2.1.3 CMU-FIA data

The FIA face database [15] is composed of 20-second videos of face data from 221 participants, mimicking a passport checking scenario, in both indoor and outdoor scenario. Videos have been captured from three different angles, with two different focal length for each, at a resolution of 640x480 pixels at 30 images per second. Data were captured in three sessions, with at least a month between each one. On the first session, 221 participants were present, 180 of whom returned for the second session, and 153 for the third.

The simulations have been performed using pictures captured by the two frontal cameras, in the first two indoor sessions – the first one for training and the one for testing. Among the all individual, 45 have been selected to populate the watch list because they are present in every session, with at least 150 ROIs for training and 300 ROIs for testing. This guarantees up to 15 samples per fold when performing 10-fold cross-validation, and thus the possibility to experiment with different amounts of training samples. Among the remaining 176 classes, 88 have been randomly chosen to build the UM for training, which guarantees the presence of another unknown 88 individuals for testing. For each of the 3 sessions, the FIA dataset have been separated into 6 subsets, according to the different cameras (left, right and frontal view, with 2 different focal length for each one). Figure 4 shows some frames captured for that data-set.

Figure 4: Video frames from the reference CMU-FIA data-set [15].

2.1.4 Chokepoint data

The Chokepoint [47] video surveillance data-set is a publicly available data-set that features video streams with either one or several subjects walking in a naturally way through several indoor portals (chokepoints of pedestrian traffic) mounted with modern array of IP network cameras. This setup is similar to the surveillance environments setup observed in airports [22] where individuals pass in a natural free-flow way in a narrow corridor. The array of three cameras is mounted just above a door, used for simultaneously recording

the entry of a person from three viewpoints. Figures 5 and 13 show examples of frames captured in the data-set.

The data consists of 25 subjects (19 male and 6 female) in portal 1, and 29 subjects (23 male and 6 female) in portal 2. Videos were recorded over two sessions 1 month apart. In total, it consists of 54 video sequences and 64,204 labeled face images. Each sequence was named according to the recording conditions, where P, S, and C stand for portal, sequence and camera, respectively. E and L indicate subjects either entering or leaving the portal. Frames were captured with the 3 cameras at 30 fps with an SVGA resolution (800X600 pixels), and faces incorporate variations of illumination, expression, pose, occlusion, sharpness and misalignment due to automatic frontal detection.

The Chokepoint data-set is suitable for medium- to large-scale benchmarking of systems for mono-modal recognition and tracking of faces over one or more cameras in watch list applications. It is provided with the ground truth (person ID, eye location and ROIs for each frame), as well as a high-resolution mug shot for each individual in the data-set. These stills images can be used as facial models of people in a watch list.

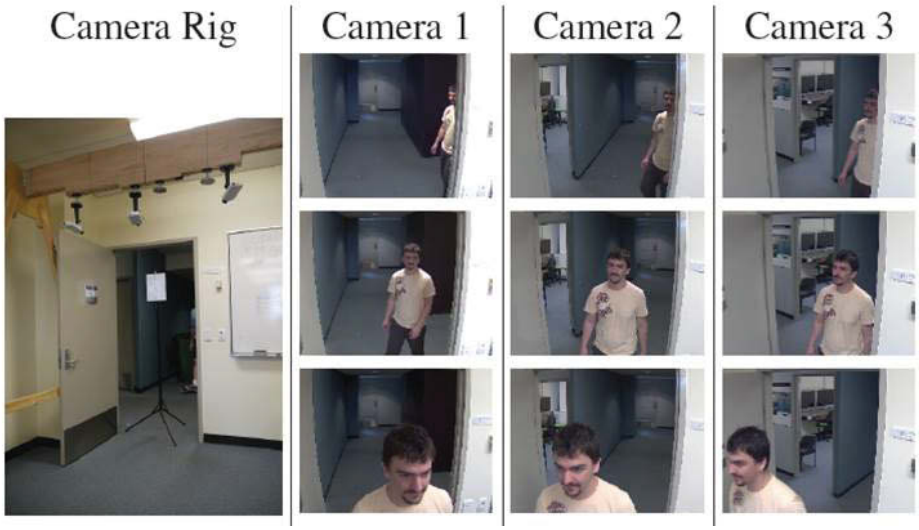


Figure 5: Video frames from the Chokepoint data-set [47].

3 Performance Metrics

This section provides a summary of metrics for transaction-based, subject-based and time-based analysis, as well as clustering and image-based quality measure and computational complexity measures proposed for performance evaluation. These metrics are used to establish a multi-level benchmarking protocol for testing face recognition systems in video surveillance applications, similar to the multi-order evaluation of

iris systems performed for border control applications described in [18]. Section 5 presents evaluation results of commercial products in the form of score cards, obtained used the defined levels of analysis.

3.1 Acquisition errors

Regardless of the accuracy of the classification systems, the performance of any biometric system is compromised by individual that cannot enroll or if they cannot present a satisfactory samples during operations.

The *failure to enroll* (FTE) rate is estimated as the proportion of individuals in the population that cannot be enrolled under the pre-determined enrollment policy. Individuals for whom the system is unable to generate repeatable templates include those unable to present the required face, those unable to produce an image of sufficient quality at enrolment, and those who cannot reliably classify their template in attempts to confirm a usable enrollment.

The *failure to acquire* (FTA) rate is estimated as the proportion of recorded transactions (both genuine and impostor) for which the system is unable to capture a face of sufficient quality. This involves failures at capture, feature extraction, or quality control phases. The FTA rate may depend on adjustable thresholds for face quality. Low-quality samples trigger a FTA, and may prompt the user to provide more training samples.

In performance evaluations, analysis is often based on a previously-collected database and there will be no problem in obtaining a sample image. However, there may be enrollment or acquisition failures, for example, when the ROI sample is of too low a quality for features to be extracted.

3.2 Transaction-based analysis

Transaction-based performance analysis allows to evaluate quality at the decision level. A crisp detector outputs a binary decision $Y \in \{0,1\}$ in response to each input sample (vector $\mathbf{x}$ extracted from an ROI), while a soft detector assigns scores to the input samples (feature vector $\mathbf{x}$) by the means of a scoring function $s : \mathbf{x} \rightarrow \Re$. As illustrated in Figure 6, the higher the score value $s(\mathbf{x})$, the more likely the prediction of the positive event ($Y = 1$). A soft detector outputs binary decisions by applying a threshold to scores. That is, sample $\mathbf{x}$ leads to a positive decision ($Y = 1$) if $s(\mathbf{x}) \geq \tau$, and negative decision ($Y = 0$) otherwise. These two mutually exclusive cases are denoted by p (for positive or target) and n (for negative or non-target).

By presenting a sample (input ROI) to a matcher or classifier applied to detection, and comparing it against facial models, the four possible outcomes may be tabulated in the confusion matrix shown in Figure 7. When a positive test sample (p) is presented to the detector and predicted as positive ($\hat{p}$) then it is counted as a true positive (TP); if it is however predicted as negative ($\hat{n}$) then it is counted as a false negative (FN). On the other hand, a negative test sample (n) that is predicted as negative ($\hat{n}$) is a true negative (TN), while it is a false positive (FP) if predicted as positive ($\hat{p}$). Given the responses of a detector over a test set of test samples, the true positive rate (tpr) is therefore the proportion of positives correctly detected (as positives) over the total number of positive samples in the test. The false positive rate (fpr) is the proportion of negatives incorrectly detected (as positives) over the total number of negative samples

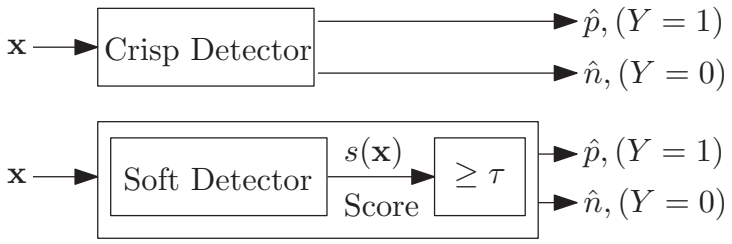


Figure 6: Illustration of a crisp versus soft detector.

in the test. Similarly, the true negative rate ($tnr = 1 - fpr$) and false negative rate ($fnr = 1 - tpr$) can be defined over the negative cases.

Performance evaluation of 1:1 classification (detection) systems is achieved by estimating tpr, tnr, fpr and fnr over a test set, and using these estimates to construct a curve or compute a scalar metric that expresses different performance tradeoffs. Receiver Operating Characteristic (ROC) and Detection Error Trade-off (DET) curves are well-accepted graphical representations to express the performance of 1:1 classification, although there are others to evaluate detection quality at the transaction-level, such as ROC isometrics, Precision-Recall curves, Cost Curves, Lift Charts, etc. This section will focus on the more relevant graphical representations and scalar metrics for FR in video surveillance

3.2.1 ROC analysis

Assume that each detector is implemented using a template matchers or using a 1- or 2-class classifiers, the performance of a user-specific detector for a set of test samples may be characterize in the ROC space [14]. As shown in Figure 8(a), a ROC curve is a parametric two-dimensional curve in which the tpr is plotted against the fpr over all threshold values. In practice, an empirical ROC curve is obtained by connecting the observed (tpr, fpr) pairs of a soft detector at each threshold. By sorting the output scores (decision thresholds) from the most likely to the least likely positive, a soft detector can efficiently produce an empirical ROC curve. Positive and negative classes are often assumed to have equal prior probabilities, and the optimal operational point is the closest point on the ROC graph to the upper-left point (1, 0) of the plane (point with maximum difference between tpr and fpr).

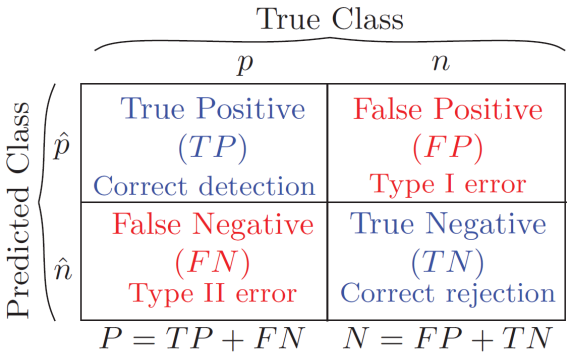


Figure 7: Confusion matrix and common performance measures. Rows accumulate the number of detector predictions as either positive or negative, and columns indicate the true classification of samples.

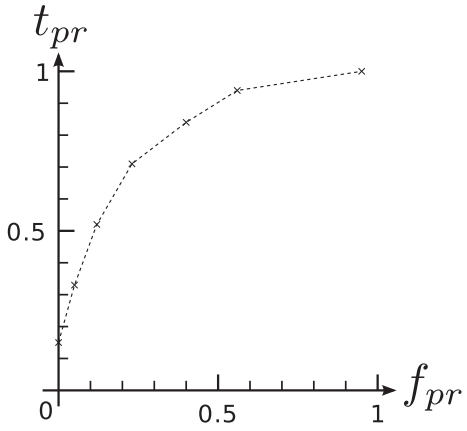


Figure 8: Example of a ROC curve.

ROC curves are commonly used for evaluating the performance of detectors at different operating points, without committing to a single decision threshold [27]. ROC analysis is robust to imprecise class distributions and misclassification costs [39]. A crisp detector outputs a binary decision and produces a single operational data point in the ROC space, while a soft detector assigns scores to the input samples, which can be converted to a crisp detector by thresholding the scores. A ROC curve is obtained by varying the threshold that discriminates between genuine and impostor classification scores. These scores are converted into a compact set of operational points, which indirectly convey information about score distributions.

Each operation point on the ROC curve corresponds to a particular threshold applied to scores. When the optimal operation points are obtained on a ROC, the thresholds of scores are also obtained. The operation points are tunable, and can be optimized with respect to accuracy. Given two operation points, say a and b, in the ROC space, a is defined as *superior* to b if $fpr_a \leq fpr_b$ and $tpr_a \geq tpr_b$. If a ROC curve has

$tpr_x > fpr_x$ for all its operation points x then, it is a *proper* ROC curve. In practice, an ROC plot is a step-like function which approaches a true curve as the number of samples approaches infinity. Therefore, it is not necessarily convex and proper.

For scalar performance measurement, it is possible to measure the detection rate tpr for a fixed point of operation (threshold of fpr) of particular interest for an application. However, summarizing performance into a single number, by fixing the operational threshold, leads to some information loss in terms of errors and costs trade-offs. Accuracy, or its complement error-rate, is defined as:

$$error = \frac{FN+FP}{P+N} = P_p fnr + P_n fpr, \tag{1}$$

an estimates the overall probability of correctly predicting a test sample, but combines results for both classes in proportion to the class priors, P_p and P_n.

The area under the ROC curve (AUC) or the partial AUC (over a limited range of fpr values) is largely known as a robust scalar measure of detection accuracy [44] over the entire range of tpr and fpr. The AUC is equivalent to the probability that the classifier will rank a randomly chosen positive sample higher than a randomly chosen negative sample. The AUC assesses ranking in terms of class separation – the fraction of positive–negative pairs that are ranked correctly. For instance, with an $AUC = 1$, all positives are ranked higher than negatives indicating a perfect discrimination between classes. A random classifier has an $AUC = 0.5$, and both classes are ranked at random. For an empirical ROC curve, the AUC may be obtained through trapezoidal interpolation or by means of the Wilcoxon-Mann-Whitney statistic. When the ROC curves cross, It is possible for a high-AUC classifier to perform worse in a specific region of ROC space than a low-AUC classifier. In such case, the partial area under the ROC curve (pAUC) could be useful for comparing the specific regions of interest.

3.2.2 Detection Error Tradeoff analysis

The Detection Error Trade-off (DET) space [31] resembles the ROC space, but it plots the fpr versus the fnr, where $fnr = 1 - tpr$ over all possible thresholds. The ROC curve is designed to depict the ranking performance, and treat positives differently than negatives. In contrast, DET curves [30] focus on the miss-classification errors made by the detector, giving uniform treatment to both error types. DET curves were introduced to evaluate detection techniques for speaker and language recognition with the advantage over ROC curve of presenting performance results where a tradeoff between two error types (fpr and fnr) is involved. In the DET curve gives a more uniform treatment to both types of error, and use a scale for both axes which spreads the plot and better distinguishes different well performing systems.

3.2.3 Precision-Recall analysis

Accuracy is well-established and commonly used to measure the frequency of correct binary decisions, but is prone to biased performance evaluations when faced with highly imbalanced class distributions. Estimates of class priors may not reflect real operational data, and may vary over time. Moreover, traditional ROC

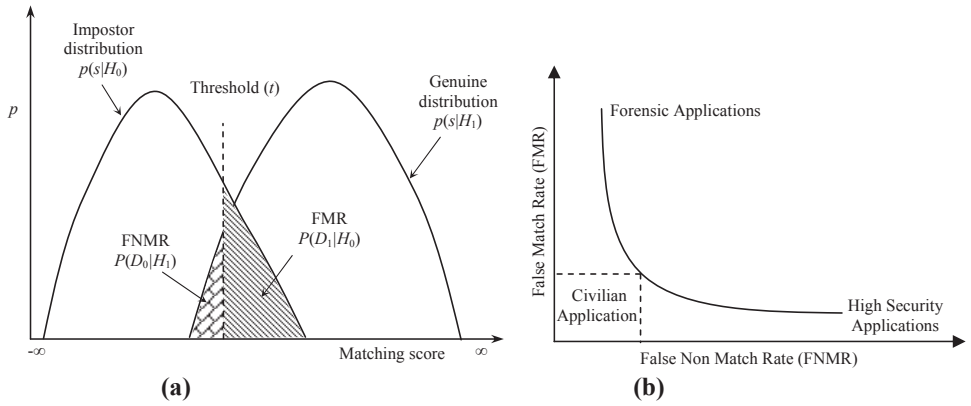


Figure 9: Error rates employed for a DET curve (a) FMR and FNMR for a given threshold τ are displayed over the genuine and impostor score distributions; (b) The DET curve plots the False Match Rate, where $fpr = FMR$, versus the False Non-Match Rate (FNMR), where $tpr = 1 - FNMR$.

analysis cannot distinguish between two classifiers for specific class miss-classification costs. Accuracy of a detector estimates the percentage of the actual cases predicted correctly for either p or n class. ROC curves and the AUC allow for a performance evaluation that is independent of costs and priors by integrating performance over a range of decision thresholds.

FR in video surveillance translates imbalanced settings, where the prior probability of the positive target class (π_p) is significantly less than that of the negative class (π_n). It is important to measure performance as the proportion of the correctly predicted positive ROIs out of the total number of ROIs predicted to belong to a given individual. Otherwise, when processing highly imbalanced data, and the minority (positive) samples are of interest, a detector may outperform others by predicting a very large number of samples as minority, resulting in an increased tpr at the expense of an increased fpr. Accuracy is inadequate as a performance measure since it becomes biased towards the majority (negative) class [45]. That is, as the skew[3] increases, accuracy tends towards majority class performance, effectively ignoring the recognition

[3] Skew, L, is defined at the ratio of the prior probability of the positive class to that of the negative class, $L = P_p/P_n$

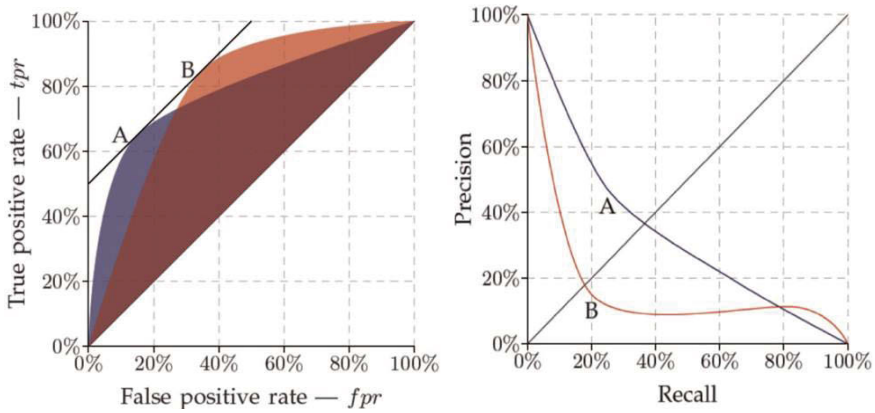


Figure 10: Illustration of (a) ROC curve and (b) Precision-Recall curves for systems with balanced (A) and imbalanced (B) class distributions.

capability with respect to the minority class [26].

In this situation, using an estimate of precision (in conjunction with recall, same as tpr) is more appropriate, as it remains sensitive to the performance on each class. In these applications recall only makes sense when combined with precision, as the prior class probabilities are unknown or highly variable. In these situations, end-users relate to precision-recall curves as they indicate how many true positives are likely to be found in a typical search. It is possible to measure what the database image retrieval community calls *precision rate*:

$$pr = \frac{TP}{TP+FP}, \tag{2}$$

the proportion of positive predictions that are actually correct. Precision effectively estimates an overall posterior probability and is therefore a meaningful performance measure when detecting rare events. The pr is relevant for problems with highly imbalanced data because both the TP and pr are zero if the detector predicts all samples as belonging to the majority class. Therefore, the precision, $pr = \frac{TP}{TP+FP}$, and recall rate, $tpr = \frac{TP}{TP+FN}$, measure the precision and accuracy for the positive instances, respectively, while $tnr = \frac{TN}{TN+FP}$ measures the accuracy for the negative instances. Figure 10 compares ROC and Precision-Recall curves for 2 systems, A and B, that have different class distributions.

It is possible to combine tpr and pr into scalar performance measures such as the geometric mean and the F-measure. The geometric mean of tpr and pr, $\sqrt{tpr \cdot pr} \in [0,1]$, combines the two conflicting measures into the square root of their product, taking the value 1 when both components are equal to 1, and the value 0 when either components are equal to 0. In general, the tpr increases with the number of minority cases in the dataset, while the pr decreases. Thus, an increase of the geometric mean indicates that the achieved increase in tpr is beneficial since it is not accompanied by a large decrease of pr.

The general F_β-measure (for non-negative real values of β) is another scalar performance measure where

the tpr and pr rates may be combined:

$$F_\beta = (b^2+1)\frac{tpr \cdot pr}{\beta^2 \cdot pr + tpr} \tag{3}$$

where the $b \in [0, infinity]$ is used to control the influence of the tpr and pr separately. When $\beta = 0$ then F_β reduces to the pr, and conversely when $b \to infinity$ then F_b approaches tpr. The well-known F_1-measure assumes that tpr and pr are evenly weighed ($b = 1$):

$$F_1 = 2\frac{tpr \cdot pr}{pr + tpr} \tag{4}$$

The *Precision-Recall Operating Characteristic* (PROC) [26] space allows detector performance to be represented graphically for the data skew in mind. Evaluation considers the entire operating surface, and integrated performance measures are then derived in a similar way to conventional ROC analysis. Unlike ROC analysis, it relies on an inter-class measure, the precision between the positive and negative decisions. Landgrebe *et al.* [26] showed that since precision depends on the degree of skew, an additional dimension, λ, must be introduced for PROC analysis, resulting in a 3-dimensional ROC surface. ROC Isometrics [FLA03] follow a similar approach, where the relationships between a number of performance evaluation criteria were derived with respect to the ROC curve. With PROC analysis, the operating characteristic constitutes a surface of operating points, with each prior resulting in a slice of this surface.

The precision can be defined as:

$$pr = \frac{TP}{TP + lambdaFP}, \tag{5}$$

Given an ROC curve, Equ. 5 allows for performance to be obtained analytically. In the example shown in Figure 11, the precision characteristics are shown to vary significantly for three different prior: *lambda* = 0.5, 0.1, and 0.01. Similarly, the precision-recall characteristic can be integrated across both decision thresholds τ and priors *lambda*, thus obtaining a scalar performance metric AUPREC using the trapezoidal approximation.

Davis and Goadrich [9] describe a methodology to find the PROC achievable curve (analogous to the ROC convex hull) and how to interpolate points to calculate the area under the PROC curve (PROC-AUC). They demonstrated that there is an equivalence between actual operating points in the ROC and PROC spaces. Operating points that belong to the ROC convex hull also belong to the PROC achievable curve. The operating point A in the PROC space is calculated under the true positive (TP_A) and false positives (FP_A) values. Given two operating points A and B apart in the PROC space, the intermediate points between them are calculated by interpolating their TP_A and TP_B, and FP_A and FP_B values. The goal is to find how many negatives examples are necessary to equal one positive, which is the local skew between A and B:

$$lskew_{A,B} = \frac{FP_B - FP_A}{TP_B - TP_B} \tag{6}$$

New $TP_A + x$ values are created for all integer x values $1 \leq x \leq TP_B - TP_A$ and corresponding FP are calculated by linearly increasing the false positives to the local skew for each new point x. Resulting

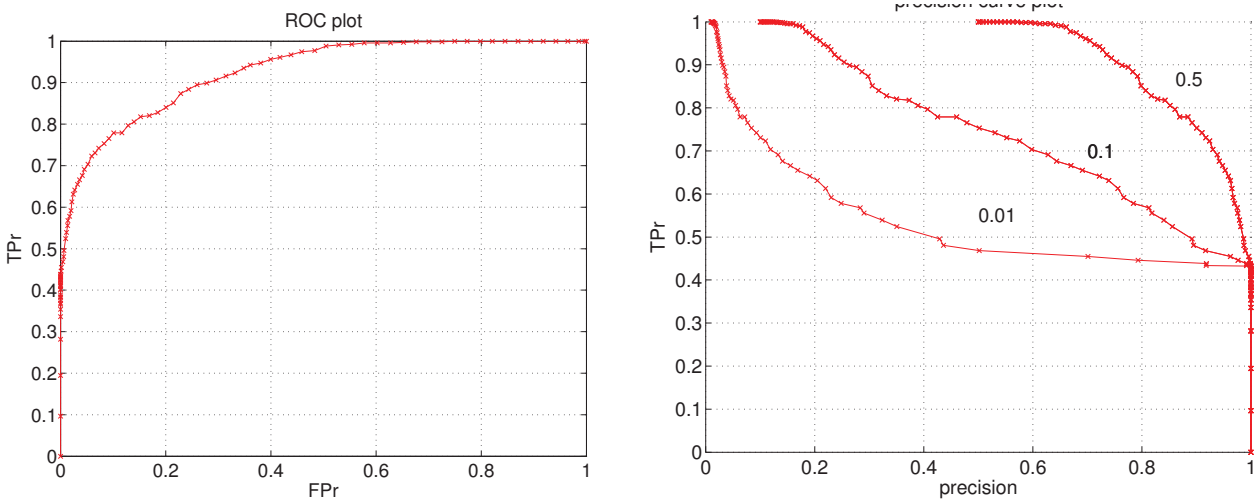


Figure 11: Illustration of (a) ROC curve and (b) corresponding precision-recall characteristics [26].

interpolated PROC points are given by:

$$\left(\frac{TP_A + x}{TP_A + FN_A}, \frac{TP_A + x}{TP_A + x + FP_A + lskew_{A,B} \times x}\right) \tag{7}$$

Once the interpolation between all points in the PROC achievable curve is done, the PROC-AUC may be calculated using trapezoidal integration. Classifier performance comparison in the PROC space is done by comparing the PROC-AUC of all classifiers. Whereas there is a relation between the ROC convex hull and the PROC achievable curve, dominance in the ROC space does not translate as dominance in the PROC space, which is also demonstrated in [9]. This property is related to the local skew used to calculate intermediate operating points in Equ. 7, an information committed by traditional ROC analysis.

As for ROC graphs for binary predictions, where each classifier in the ROC space has its correspondent line in the cost space, we can derive a cost curve correspondent to a ROC curve. The process is very similar to the generation of a ROC curve: by varying the percentage of cases classified as positive from 0 to 100 percent. This percentage is used as a parameter so that each possible parameter value produces a line in the cost space. A set of points in the ROC space is a set of cost lines, one for each ROC point.

3.2.4 Uncertainty of estimates

Given the sources of noise and uncertainty introduced during facial matching, statistical techniques are needed to estimate tpr, tnr, fpr and fnr, over a test set. Many studies report performance in the form of classification score distributions. Parametric and nonparametric (bootstrap) methods may be used to measure the confidence intervals for these distributions, to indicate the significance of the estimates provided for detection quality [6].

Performance estimates will be affected by random noise (natural variations) and by bias in evaluation protocols. Uncertainty arising from random noise declines as the size of the test set increases, and the effects of some protocol bias may also be mitigated. For instance the error rates for an under-represented individual may be validated for consistency with the overall error rates. Experimental trial should also be repeated in different environmental conditions for sensitivity analysis. It is possible to estimate the variance of performance measure and thus confidence intervals under the following assumptions about the distribution of decisions:

- enrolled individuals are representative of the target population;
- attempts by different subjects are independent from one another, and from the threshold;
- error rates vary across the population, i.e., we explicitly allow for goats, wolves, and lambs [11].

3.3 Subject-based analysis

The performance of FR systems is typically evaluated using a transaction-based analysis in the ROC or DET space, and scalar measures over all individuals allow for a first order analysis. However, performance of FR systems may vary drastically from one person to the next[42]. Each individual enrolled to the system is categorized according to the Doddington's zoo subject based analysis [11, 41]. In subject-based analysis, the error rates are assessed with different types of individuals in mind, rather than with the overall number of transactions. An analysis of these individuals and their common properties can expose fundamental weaknesses in a biometric system, and allows us to develop more robust systems. Quantitative methods for dealing with the existence of user variation is an active area of research. User-specific schemes allow us to set user-specific or template-specific thresholds, score normalization, and user-specific fusion [37, 41].

Pattern Specific Error Inhomogeneities (PSEI) analysis [11] has shown that the error rates vary across the entire population (e.g., individuals in a watch list). It has lead to the characterization of target populations (positive) as being composed of sheep and goats. According to this characterization, the sheep, for whom FR systems tend to perform well, dominate the population, whereas the goats, though in a minority, tend to determine the performance of the system through their disproportionate contribution to *FNR*s. Goats are characterized by consistently low classification scores against themselves. In non-target populations (negative), some individuals – called wolves – are exceptionally successful at impersonating many different targets, while some targets, called lambs, are easy to impersonate and thus seem unusually susceptible to many different impostors. The definitions of lambs and wolves are symmetric. Lambs, on average, tend to produce high match scores when being matched against by another user. Similarly, wolves receive high scores when matching against others. For both of these user groups, the match score distributions are significantly different than those of the general population.

Doddington's Zoo analysis [11, 41] categorizes individuals in one of the following four categories: (1) sheeps, easy to identify individuals (positive or negative class), (2) goats, positive class individuals that are difficult to identify, (3) wolves, negative class individuals that impersonate one or more positive class individuals or (4) lambs, positive class individuals that are easy to impersonate.

Category	Positive class	Negative class
Sheep	$frr < 50\%$ and not a lamb	$fpr \leq 30\%$
Lamb	at least 3% of non-target individuals are wolves	–
Goat	$frr \geq 50\%$ and not a lamb	–
Wolf	–	$fpr > 30\%$

Table 5: Doddington's zoo analysis adapted for a sequence of binary decisions over a video track to decide the individual identity. False rejection rate (frr) and false positive rate (fpr) thresholds are applied to each individual detector module.

The traditional way to define the system users likeliness in a Doddington zoo's category is through the classifier output scores for all tested samples [41]. For techniques that provide crisp decisions and the confusion matrix of individual accumulated decisions is used to categorize individuals according to Table 5. This approach is based on the technique used in [29], and considers the fpr and the false rejection rate (frr, the percentage of positive individual rejections) for one individual detector module. The frr and fpr decision thresholds were selected to provide a more conservative categorization.

ROC and DET analysis can be improved significantly if samples from the more difficult cases (e.g., the goats) can be detected automatically and processed differently. The distribution of classification scores in a biometric system will naturally vary across a range of results. Of interest to biometric system for evaluation are individuals that consistently yield poor scores, outside expected random variation. The score distributions for these individuals (goats, lambs, and wolves) are fundamentally different from the distributions of the general population (sheep). The definitions presented above state that match score distributions are user dependent. Once this fact is established, it follows that some users are performing differently than others, and the presence of the animals is established without explicitly labeling users. Doddington *et al.* use a variety of tests at the score level to demonstrate that the animals defined above exist to a statistically significant degree in their biometric system [11].

Results in the literature suggests that contribution made by different individuals to the overall system error is uneven on the [29, 38, 46]. Wittman *et al.* [46] found evidence for the existence of goats, wolves, and lambs in FRGC 2.0 FR data. Their test for *goats* involves analyzing how individual score when matched against themselves. To compute an individual's goat statistic, the worst matching score recorded for each of the individual's input ROIs is compared against all other images from the same individual, and the averaged together. In order to determine a *wolf score*, a statistic is used to represents how well a subject impersonates other individuals enrolled to the system. To compute the wolf score for an individual, they find for each input ROI the best score against a facial model from another subject. Then, these scores are averaged together, resulting in a single wolf score for each of the individuals. The computation of the *lamb statistic* involves two values for each input ROI - the best score of an impostor against the individual's ROI, and the best matching score of the individual's ROI image against an ROI from the same individual. Authors typically use hypothesis testing for the existence goats, lambs, and wolves to demonstrate user dependent match score distributions, to statistically ensure that variations in the data are no random chance. One-way

ANOVA is a method for testing the null hypothesis for differences between independent distributions. The Kruskal-Wallis test is similar to ANOVA except that scores are replaced by ranks, thereby relaxing the assumption of normality.

A recent study by Yager and Dunstone [48] has noted that the four (4) traditional biometric animals are based on only genuine match scores (low for goats) or impostor match scores (high for lambs and wolves). A new class of animals has been defined in terms of a relationship between genuine and impostor match scores. The animals are called worms, chameleons, phantoms, and doves, and have combinations of low/high and genuine/impostor match scores. The new animals differ in that they are defined in terms of a relationship between genuine and impostor match scores. Unlike goats, lambs, and wolves, the specific users who belong to the new animals groups will be identified. For example, chameleons are users who tend to have high genuine match scores and high impostor match scores. The existence test is based on whether or not there are more or less members of an animal group than expected. The authors propose a new group-centric framework for the evaluation of biometric systems based on the biometric menagerie, as opposed to collective statistics.

3.4 Clustering quality of triaging

Clustering or categorization is important in some applications for (1) regrouping per person for more reliable decision in, e.g., in spatio-temporal FR, (2) estimating the number of individuals in a scene, and (3) for reducing computational complexity, by limiting the number of input ROIs to be classified by the FR systems.

A partition of n patterns into K groups defines a *clustering*. This can be represented as a set $A = \{a_1, a_2, ..., a_n\}$, where $a_h \in \{1, 2, ..., K\}$ is the category label assigned to pattern h. The degree of match between two clusterings, say A and B, may be compared by constructing a contingency table. In this figure, c_{11} (c_{22}) is the number of pattern pairs that are in a same (different) cluster in both partitions. The value c_{21} is the number of pattern pairs that are placed in a same cluster by A, but in different clusters by B. The value c_{12} reflects the converse situation. The sum of all elements $m = c_{11} + c_{12} + c_{21} + c_{22} = n(n-1)/2$ is the total number of combinations of two out of n patterns. The four variables within the contingency table have been used to derive measures of similarity between two clusterings A and B [12] [1]. These measures are known in pattern recognition literature as external criterion indices, and are used for evaluating the capacity to recover true cluster structure. Based on a previous comparison of these similarity measures, the Rand Adjusted, defined by:

$$S_{RA}(A,B) = \frac{2(c_{11}c_{22} - c_{12}c_{21})}{2c_{11}c_{22} + (c_{11} + c_{22})(c_{12} + c_{21}) + c_{12}^2 + c_{21}^2} \tag{8}$$

and Jaccard statistic [12], defined by:

$$S_J(A,B) = \frac{c_{11}}{c_{11} + c_{12} + c_{21}} \tag{9}$$

have been selected to assess clustering quality for this study. It is worth noting that variable c_{22} does not appear in $S_J(A,B)$.

Since correct classification results (ground truth) are known for an evaluation data-sets used, their patterns are all accompanied by category labels. These labels are withheld from the system under test, but they provide a reference clustering, R, with which a clustering produced by computer simulation, A, may be compared. Then, variables c_{11} and c_{22} represent the number of pattern pairs which are properly clustered together and apart, respectively, while c_{12} and c_{21} indicate the improperly clustered pairs. In this case, Eqs. 8 and 9 yield *scores* that describe the quality of the clustering produced by a systems. Both the Rand Adjusted and Jaccard measures yield a score ranging from 0 to 1, where 0 denotes maximum dissimilarity, and 1 denotes equivalence. The closer a clustering A is to R, the closer the scores are to 1. Notice the dependence of these scores on the number of clusters in A and R.

3.5 Facial image (ROI) quality

Computation of a quantitative objective image quality measure is an important tool for biometric applications, for quality control to accept, reject, or re-acquire biometric samples, to select a biometric modality, algorithm, and/or system parameters, and as confidence estimators of reliability of decision. For example, selecting high quality images may improve recognition performance, and reduce the overall computation load during feature extraction and matching.

Several face image standards , e.g., ISO/IEC 19794-5 (Biometric Data Interchange Format Face Image Data) and ICAO 9303, have been proposed to establish guidelines to capture facial images, and to assess their quality. In these standards, quality can be measured in term of: (i) image specific qualities such as sharpness, contrast, compression artifacts, and (ii) face specific qualities such as face geometry, pose, illumination, etc. Other universal measures in literature involve comparing each input ROI against the facial models of a person (gallery templates or statistical representation) to measure image variations. The rest of this subsection summarizes two such image-based measures that allow to compute image quality – the distance and distortion between a pair of images.

Image subtraction is a process where pixels values in an image are subtracted from those of another image. This is commonly used to detect changes between two images, for instance to know if an object in an image has moved.

Assuming that all facial representations are coherent gray-level images that are scaled to the same 2D size, the basic *point-wise distance*, PWD(**R**,**M**), measures the distance between an input ROI **R** and a facial model **M** (e.g., a gallery template or a mean value of densities in a generative statistical representation). Point-wise distance is defined as the average difference between the gray-level values of all pairs of corresponding pixels (two points in the same location):

$$\texttt{PWD}(\mathbf{R},\mathbf{M}) = \frac{1}{n}\sum_{x\in mask} \mid \mathbf{R}(x)-\mathbf{M}(x) \mid = \frac{1}{LW}\sum_{l=1}^{L}\sum_{w=1}^{W} \mid \mathbf{R}(l,w)-\mathbf{M}(l,w) \mid \tag{10}$$

where n is the total number of pixels in the facial region of length L and width W ($n = L \times W$), and $\mathbf{R}(x)$ or $\mathbf{R}(l,w)$ is the gray-level value of the pixel in location x or (l,w) in image **R**, respectively.

It is assumed in Equ. 10 that facial images were normalized with respect to position and size – gray-level values were therefore compared for corresponding locations of the face. The *regional distance* compensates

for a displacement of up to three pixels of the images in the plane. The regional distance, RD(**R**,**M**), is defined as the average of the minimum difference between the gray-level value of a pixel and the gray-level value of each pixel in the neighborhood of the corresponding pixel:

$$\mathtt{RD}(\mathbf{R},\mathbf{M}) = \frac{1}{n}\sum_{x\in mask}\min_{i\in neighb(x)} \mid \mathbf{R}(x)-\mathbf{M}(i)\mid = \frac{1}{LW}\sum_{l=1}^{L}\sum_{w=1}^{W}\min_{(i,j)\in neighb(l,w)} \mid \mathbf{R}(l,w)-\mathbf{M}(i,j)\mid \qquad (11)$$

where $neighb(x)$ or $neighb(l,w)$ is a square neighborhood of 5×5 pixels around x or (l,w), respectively.

To compensates for uniform affine transformation of the gray-level values of one of the images, the *affine gray-level distance* is often used. It is defined as the minimum Euclidian distance between the gray-level values of one of the image and any affine transformation of the gray-level values of the other image.

The *universal image quality index* (Q) [WAN02] allows to compare different types of image distortions by modeling any image distortion as a combination of the following three factors – loss of correlation, luminance distortion and contrast distortion.

Let $\mathbf{r} = (r_1, r_2, ..., r_N)$ and $\mathbf{m} = (m_1, m_2, ..., m_N)$ be vectorized version of a ROI and a facial model M, respectively. The universal quality index in [WAN02] is defined as

$$\mathtt{Q}(\mathbf{r},\mathbf{m}) = \frac{4\sigma_{rm}\overline{r}\,\overline{m}}{(\sigma_r^2+\sigma_m^2)(\overline{r}^2+\overline{m}^2)} \qquad (12)$$

Statistical features for Equ.12 are measured locally to accommodate space-variant nature of image quality, and then combined to an overall quality measure for the entire image. A local quality index Q_j is calculated with $\mathtt{Q}(\mathbf{r},\mathbf{m})$ by sliding a window of $B\times B$ pixels from the top-left corner to the bottom-right corner of the image. For a total of W steps, the overall quality index is given by:

$$\mathtt{Qtot}(\mathbf{r},\mathbf{m}) = \frac{1}{W}\sum_{j=1}^{W} Q_j \qquad (13)$$

The universal quality index $\mathtt{Q}(\mathbf{r},\mathbf{m})$ of Equ.12 can be written as a product of the three factors – loss of correlation, luminance distortion and contrast distortion:

$$\mathtt{Q}(\mathbf{r},\mathbf{m}) = \frac{\sigma_{rm}}{\sigma_r\,\sigma_m}\cdot\frac{2\overline{r}\,\overline{m}}{\overline{r}^2+\overline{m}^2}\cdot\frac{2\sigma_r\,\sigma_m}{\sigma_r^2+\sigma_m^2} \qquad (14)$$

3.6 Analysis of computational complexity

Measuring the average enrollment and matching time in seconds does not allow for unbiased evaluation of complexity because of the dependency on specific implementations and platforms. Approximating the time and memory complexity of various components of a FR system important in the context of real-time surveillance systems, where complexity scales according to system parameters [28].

A first order approximation of the computational complexity for the algorithms may be obtained by assessing their execution time FR systems on an idealized computer. Thus, the time complexity, T, combined

with a fixed computing area C, allows for comparison of area-time complexities, CT. To that effect, assume that FR techniques are implemented as software programs running on a generic, single processor, random access machine (RAM) [8], where instructions are executed one after the other. This generic machine is capable of no more than one operation per cycle. Using the RAM model avoids the challenging task of accurately determining the performance of specific VLIW or superscalar machines, which is beyond the scope of this analysis.

Time complexity can be estimated from the maximum execution time required by a FR system to process a single ROI. The result is a total worst-case running time formula, T, which summarizes the behavior of a FR system as a function of key parameters: the ROI size, the dimensionality of the input patterns **a** used for classification, M, and the number of reference samples or templates per individual, N. Specifically, T can be defined as the sum of the worst-case running times T_p for each operation p that is required to process a ROI [8]:

$$T = \sum_p T_p = \sum_p o_p \cdot n_p \tag{15}$$

where o_p is the constant amount of time needed to execute an operation p, and n_p is the number of times this operation is executed.

For simplicity, we assume that p can take one out of two values, 1 or 2, where o_1 is the time required to execute an elementary operation such as: $x+y$, $x-y$, $\max(x,y)$, $\min(x,y)$, $1-x$, $x<y$, etc., and o_2 is the time needed to compute a division x/y, a multiplication $x \cdot y$, a square root $\sqrt{x}$, or an exponent e^x. In addition, x and y are assumed to be real numbers represented by an integer with a b bit resolution, where b corresponds to the number of bits needed to represent each system's elementary values (i.e., template and input pattern components) with a sufficient precision. Operations of type $p=2$ are more complex than those of type $p=1$, and their complexity, which is a function of b, depends on their specific implementation. Nevertheless, to simplify the analysis, it is assumed that $o_2 \simeq F(b) \cdot o_1 = F \cdot o_1$, where F remains as an explicit parameter in the complexity formulas. Finally, the *growth rate* is obtained by making the parameters of the worst-case complexity tend to infinity.

In a similar way, memory complexity M may be estimated as the number of 8 bit registers needed during biometric recognition process to store variables. Only the worst-case memory space required for storage during matching phase is considered.

3.7 Time-Based Analysis

Systems for FRiVS typically use different algorithms and techniques to implement their functions such as face detection, matching and tracking. In order to evaluate and compare systems for FRiVS it is important to observe their ability to detect a person of interest globally over time, using all its functions to process a video stream. Besides, decisions taken by an operator take place over a time period that is longer than a frame rate.

In this study, we propose to count the number of positive matching predictions over a moving window of time for input ROIs that correspond to a high quality facial track. Assume for instance a system that produces decisions at a maximum rate of 30 fps once, Each detected ROI is presented to the face matcher,

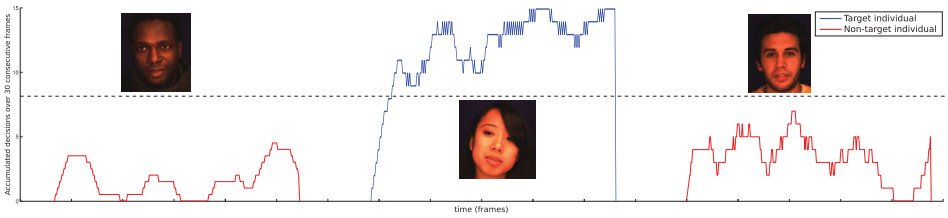


Figure 12: Accumulation of the positive predictions over time for a FRiVS system with 3 individuals in the CMU-FIA data-set.

that produces predictions, true or false, for each person enrolled to the system. Given a high quality face track, the number of positive predictions for a person of interest will grows rapidly. In this was, the operator can more reliably assume the presence of a person of interest. Figure 12 show an example of a time-based analysis on the CMU-FIA data-set. The 2nd person is enrolled to the system as a person of interest, while persons 1 and 3 are unknown to the system. In this example, the operator receives a notification once the number of positive predictions surpasses a rate of 8 positives per second.

4 Benchmarking Protocols

Automated systems designed for detecting individuals of interest involve open-set FR, where the vast majority of individuals interacting with the system are not of interest (not enrolled). Table 6 presents a taxonomy of FRiVS evaluation scenarios for video surveillance applications described in the introduction. For this project, CMU-FIA, Chokepoint and ND-Q0-Flip data-sets are found relevant for evaluation protocols involving mono-modal recognition.

There are a few variants of the screening problem that are evaluated with the same protocol. Depending on the enrollment process, these applications either perform still-to-video (e.g., screening against the watch

Table 6: Taxonomy of FRiVS evaluation scenarios.

Surveillance setup	**Applications**	**data-set**
Type 0 **cooperative biometric setup**	access control, eGate	N/A
Type 1: **semi-constr. setup, one-at-time**	primary inspection lane	CMU-FIA
Type 2: **unconstr. free-flow, one-at-time**	checkpoint entry (portal)	Chokepoint (1 person)
Type 3: **unconstr. free-flow, many-at-time**	indoor airport setting	Chokepoint, ND-Q0-Flip (n persons)
Type 4: **no lighting or structural constr.**	outdoor setting	N/A

Table 7: Chokepoint video sequences selected for performance evaluation. The sequences are captured with one of three cameras when subject are leaving portal number 2.

Data sequences	no. of subjects	type of scenario
1) P2L_S1_C1, P2L_S1_C2, P2L_S1_C3	1	type 2, with different cameras
2) P2L_S2_C1, P2L_S3_C1, P2L_S4_C1	1	type 2, with different recorded sequence
3) P2L_S4_C1, P2L_S4_C2, P2L_S5_C3	24, crowded	type 3, with different cameras

list gallery of still images) or video-to-video recognition (e.g., tracking and re-detecting individuals in multiple video-streams). During the enrollment process, still-to-video recognition involves pre-processing and storing an input pattern extracted from the ROI corresponding to each still image of the watch list. There may be one or more still images per individual, and corresponding patterns form the facial model of individuals of interest in the cohort. Video-to-video recognition involves capturing one or more video sequences for enrolling a person of interest. High quality ROIs are extracted to form the facial model of individuals of interest.

As mentioned, the objective of FR in video surveillance is to accurately and efficiently detect the presence of individuals of interest in semi-constrained and unconstrained environments, leading to either continued surveillance or interdiction. The rest of this report presents two generic benchmarking protocols for evaluation state-of-the-art commercial technologies and academic systems.

4.1 Generic protocol: still-to-video recognition

Still-to-video face recognition for Type 2 and 3 application scenarios can be evaluated using the Chokepoint video surveillance data-set [47] described in Section 2. As mentioned earlier, this data-set is suitable for medium- to large-scale benchmarking of systems for mono-modal recognition and tracking of faces over one or more cameras in watch list applications. It is provided with the ground truth (person ID, eye location and ROIs for each frame), as well as a high resolution mug shot for each individual in the data-set. These still images can be used as facial models of people in a watch list. Table 7 shows the Chokepoint video sequences that were used in our evaluations. Figure 13 shows examples of frame captures from one of those sequences.

Algorithm 1 shown in Figure 4.1 and the dataflow diagram shown in Fig. 15 present an overview of our protocol for still-to-video screening over a single video sequence. In our still-to-video screening protocol, the mugshot images are employed to define the facial model of each person in the Chokepoint video data, and these models are stored in the gallery of templates. In our experiments, each person in a given video takes a turn at being the individual of interest in a watchlist. Given the matching sores and image quality of each person for each ROI in a video. if is possible to assess the performance of a system on many levels. The reader is referred to the next subsection for details on performance metrics. This process should be replicated over several video sequences to assess the confidence of performance evaluations.

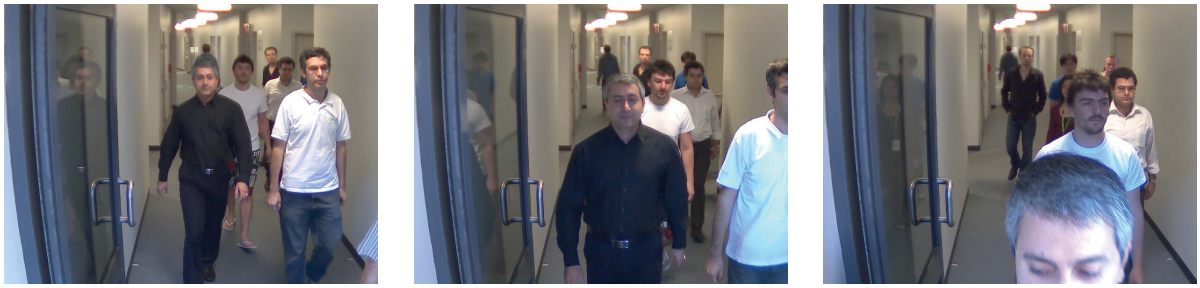


Figure 13: Examples of frames 1003, 1028 and 1059 captured for sequence P2L_S5_C3.

Algorithm 1: Overview of Protocol for Still-to-Video Screening Applications.

{Generate gallery of templates from still images}
for $d = 1, 2, ..., \texttt{nPersonsData}$ **do**
  Apply FD algorithm to the still image of person d in the data set;
  Perform feature extraction and selection to reference ROI_d detected at center of still image;
  `Gallery` $\Leftarrow$ feature-based representation of reference ROI_d, template $\mathbf{m}_d$;
end for

{Repeat experiment for each person in the video sequence}
for $v = 1, 2, ..., \texttt{nPersonsVideo}$ **do**
  Set `Watchlist` = person of interest, v, with his template $\mathbf{m}_v$ from `Gallery`;
  {Repeat for each frame in the video sequence}
  for $f = 1, 2, ..., \texttt{nFrames}$ **do**
    Apply FD algorithm to frame f;
    {Repeat for each ROI in frame f}
    for $r = 1, 2, ..., \texttt{nROIs}$ **do**
      Quality $q_v(r, f) \Leftarrow$ image distortion of $\text{ROI}(f, r)$ with respect to ROI_v
      Perform feature extraction and selection to $\text{ROI}(f, r)$;
      Set input $\mathbf{a}(f, r)$ = feature-based representation of $\text{ROI}(f, r)$ in frame f;
      Score $s_v(r, f) \Leftarrow$ matching score between input $\mathbf{a}(f, r)$ and template $\mathbf{m}_v$;
    end for
  end for
end for

{Multi-level analysis of performance (refer to Subsection 2.3)}
Level-1: transaction-based analysis;
Level-2: subject-based analysis;
Level-3: time-based analysis of a sequence;

Figure 14: Overview of Protocol for Still-to-Video Screening Applications.

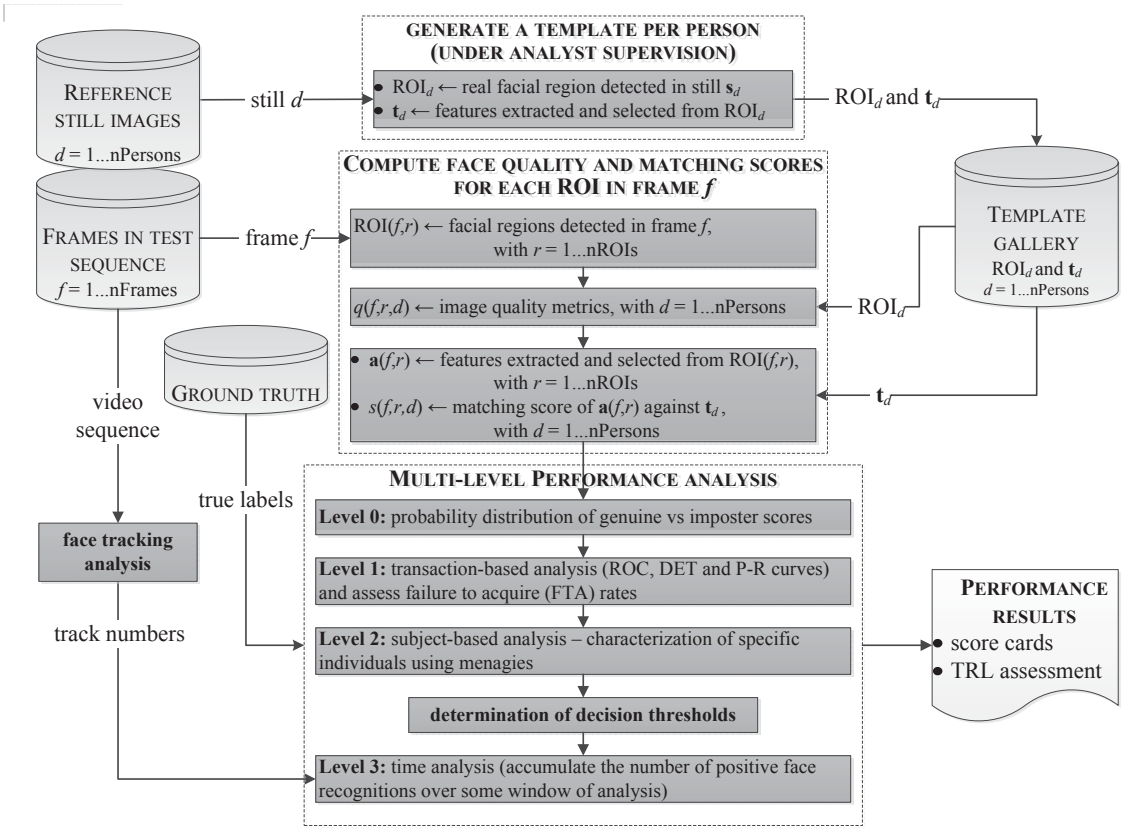


Figure 15: Dataflow diagram representing benchmarking protocol for still-to-video recognition.

4.2 Generic protocol: video-to-video recognition

In many applications such as speaker recognition, the use of a "Universal Background Model" is widely used for better discrimination between target voice from all other sounds [7]. This UBM is built by selecting samples of the background sound that characterizes a recording environment, and is used to discriminate between the individual (speaker) of interest and other sounds. In the same manner, the cohort model is a set of selected samples from non-target samples from already known voices to discriminate known individuals from other known speakers. These cohort and UBM models constitute an important source of discriminative information in the training stage of the classifiers.

It is desirable to have several individuals in the data-set to form watch lists. At least 10 individuals are needed as candidates to form the Cohort Model (CM), each one comprised of several samples for design and testing. This allows to collect information about the system's ability to detect individuals of interest, and guarantees enough samples per fold when performing cross-validation. Among the remaining individuals,

several different ones are needed to build the Universal Model (UM) for system design, which guarantees the presence of unknown individuals in testing. In practice, data from many 'unknown' individuals may be available from a UM, for design and testing.

For example, the CMU-FIA data-set is composed by 20-second videos of face data from 221 participants, mimicking a passport checking scenario, in both indoor and outdoor scenarios. Data has been captured over three sessions, with at least a month between each one. On the first session, 221 participants were present, 180 of whom returned for the second session, and 153 for the third. In our experiments, 45 classes have been isolated as candidates for the watch list, as they fulfil the following conditions: (1) the subject is present in every session, (2) at least 150 samples (5 secs of footage) are available in training, and (3), at least 300 samples (10 secs of footage) are available in testing. Among the remaining 176 classes, 88 have been randomly chosen to build the UM for design, which guarantees the presence of unknown individuals (the other 88 individuals) during testing phases.

Independent simulation trials are performed for each individual of interest (assigned to the watch list). Prior each trial, the data-set is partitioned into two parts – data subsets comprised of ROIs extracted from design and test sequences. Different steps of the design phase are always performed using *dbLearn*, where ROIs are captured during the first session, and under the most ideal conditions possible (frontal cameras, high quality, constant indoor lighting, etc.). The order of ROIs in this design subset is not important. Then, the design subset is further divided into folds for *k*-fold cross validation, allowing to establish confidence intervals and statistical significance of results. During each trial, one or more folds are systematically or randomly selected for validation (selection of different systems parameters, most notably the decision threshold), and the remaining folds used, as needed, as reference design samples. Each fold is organized to contain a even number of samples from the individual of interest (positives, under test) and other individuals (negatives, random selection of samples from the UM and CM).

Given the limited data[4] (i.e., positive samples), simulation scenarios follow a 2 x 5-fold cross-validation process for independent 10 replication. At replication 5, the 5 folds are regenerated after a randomization of the sample order for each class. The first step of a simulation scenario is the generation of the *dbLearn* dataset, which is used for system design. *dbLearn* remains unchanged for the two sets of five replications. Dataset *dbLearn* is divided into the following subsets, based on the 2x5 cross-validation methodology:

- *dbTrain*: the training dataset used to represent facial models for different individuals.
- *dbVal*: validation dataset used to set system parameters, most notably the decision threshold.

Again, *dbLearn* is composed of positive reference samples of an individual of interest, as well as the same amount of negative samples randomly selected for other individuals in the scene (UM), and from the other individuals of interest in the watch list (CM).

After a design phase, *dbTest* is presented to the system under evaluation, and performance metrics are computed. Unlike *dbLearn*, the testing data-set *dbTest* remains constant over all tests, and preserves the chronological order of ROIs. The testing phase is performed using *dbTest*, where ROIs are captured under

[4]This procedure assumes at least 5 positive samples for *dbLearn*

variable conditions with, e.g., different lighting and pose, ageing, occlusion, camera angle, etc. As recommended in Section 3, systems are evaluated by estimating the, precision pr, recall tpr (seen in precision-recall curves), F-measure, and complexity over *dbTest*. When the 10 replications are completed, the average value and confidence interval of these estimates are computed. For additional insight on systems under test, the performance of each individual is further analyzed to characterize the type of individual following the Doddington zoo taxonomy.

4.3 Recommendations

For evaluation of mono-modal recognition and tracking of faces over one or more cameras, the authors of this report recommend using the CMU-FIA (for Type 1 setup), Chokepoint (for Types 2 and 3 setups) and ND-Q0-Flip for Types 3 setup) datasets. For application involving multi-modal recognition and tracking, MOBIO (face and voice modalities) and NIST-MBGC (face and iris modalities) data-sets are the most suitable for benchmarking studies. These publicly-available data-sets are suitable for medium- to large-scale benchmarking, and appear commonly in the open literature.

Evaluation of FR systems in video surveillance should be viewed in terms of independent user-specific detection problems. Although, ROC and DET analysis are commonly used to measure the quality in transaction-based analysis, the precision-recall space provides a more informative picture of system performance, when dealing with highly skewed data. Indeed, given the open set recognition problem, we are primarily interested in analyzing performance of the positive class (individual of interest), with imbalanced class distributions, i.e., the number of negative cases heavily outnumbers the number of positive cases. This particular characteristic compresses the area of interest in a ROC graph to a small corner in the lower left side of the ROC space. Other alternatives like Cost Curves or Brier Curves also allow one to observe the impact on performance of systems when operational data are very skewed or when the costs of errors differ.

Given a classifier and its results on a set of test samples, we must assess confidence on estimates of accuracy, precision, recall, or F-measure. For instance, Goutte and Gaussier [21] present a probabilistic interpretation of precision, recall, and F-score to compare performance scores produced by two information retrieval systems. Moreover, several studies have shown that performance varies across a population of individuals. Analyzing the type of individuals and their common properties, using subject-based analysis (Doddington's zoo), can reveal detailed insight of systems under test.

Given the uncontrolled nature of video-surveillance applications, the image and tracking quality are important characteristics of input ROIs, and will have a considerable impact on system performance. Assessing clustering quality (for triaging), using the Rand Adjusted, measure the quality of ROI groupings, and indicates the capacity to reduce computational complexity by grouping input ROIs according to individuals in a scene. To avoid bias, the complexity of different technologies and systems should be estimated analytically whenever possible, and compared in terms of both processing time and memory requirements. Of course, performance results dependent on the application set-up, environment and population. To interpret results correctly, some additional information should be considered, such as:

- type of evaluation: technology, scenario or operational;

- scale of evaluation: number of subjects, number of transactions per subject, etc.
- details on demographics and operational environment;
- time between enrollments and operational transactions;
- quality and decision thresholds used during capture of samples;
- factors potentially affecting performance were controlled.

5 Experimental Results

In the following, the developed evaluation methodology is applied to testing three commercial products. The products are: Cognitec (FaceVACS-SDK 8.5 Release Date: 2011-12-19), PittPatt (Face Detection, Tracking and Recognition FTR SDK 5.2.2, Release Date: 2010) and Neurotechnology (Verilook SDK 5.4, Release Date: 2011).

A multi-level performance analysis presented in Section 4 is applied to evaluate and compare the system performance to one another. First, the level 0 or score-based analysis illustrates the probability distribution of the genuine scores against that of the impostor. Then, level 1 analysis provides a transaction-based performance evaluation of decisions using the ROC, DET and PROC curves. The subject-based evaluation (level 2 analysis) focuses on the categorization of individuals using according to Doddington's Zoo. Finally, a time analysis is performed at level 3 to evaluate systems performance over video streams, by accumulating the positive predictions (and hence an increased confidence in decision making) over a moving time window.

These products are tested using the Chokepoint data-set described in Section 2. For the testing, ten individuals are randomly selected from the Chokepoint dataset as target individuals and included in the watch list. The target individuals include six males and four females, and have the following identification numbers (id) in the Chokepoint dataset: $1, 4, 5, 7, 9, 10, 11, 12, 16$ and 29. For each target individual, the remaining 28 individuals are considered as impostors. The performance is evaluated based on a fixed operating point (face matching threshold) of 5% false positive rate, using three different distances between the eyes: $10, 20$ and 30 pixels. The video streams from Chokepoint dataset portal 2, session 1, camera 1.1 (P2L_S1_C1.1) are considered as a validation set and used to compute matching thresholds for each target individual and each distance between the eyes. These thresholds are then applied to Chokepoint dataset portal 2, session 4, camera 1.1 (P2L_S4_C1.1), to evaluate systems performance for each target individual.

As emphasized earlier, target-specific thresholds allow for improved system performance since some individuals are naturally more difficult to recognize than others and the risk associated with recognition errors varies from one individual to another. However, setting and optimizing a specific threshold for each target individual makes systems comparison more difficult to illustrate and harder to summarize. In particular, averaging the performances among all target individuals (each with a different matching threshold) may not provide meaningful results. Therefore, the detailed results for each system based on each target individual are presented using the concept of a *Report Card*, one Report Card per each target individual.

For each tested FR system, two-page Report Cards illustrating four levels of performance analysis are generated for each target individual in the watch list, showing the performance of the system at three minimum allowed facial resolutions: 10 pixels, 20 pixels and 30 pixels between the eyes.

For the purpose of illustrating the evaluation methodology, this report presents the Report Cards produced for target individual 1 only (shown in the next six pages), followed by a comparative analysis of the performance of all three systems for that particular individual based on the obtained results. A separate report entitled "Results from evaluation of three commercial off-the-shelf face recognition systems on Chokepoint dataset" provides the exhaustive results of this evaluation, showing the two-page reports cards for *all* tested target individuals for each of three tested systems [23].

By analyzing and comparing Reports Cards from three different FR products, important conclusions can be made about tested products and suitability of FR technology for video-surveillance applications in general. This is further discussed next.

Biometric System: Cognitec	**Biometric Modality**: Face
Data Source: Chokepoint – Portal 2, leaving portal setup, session 4, camera 1.1	**Individual Template**: 1

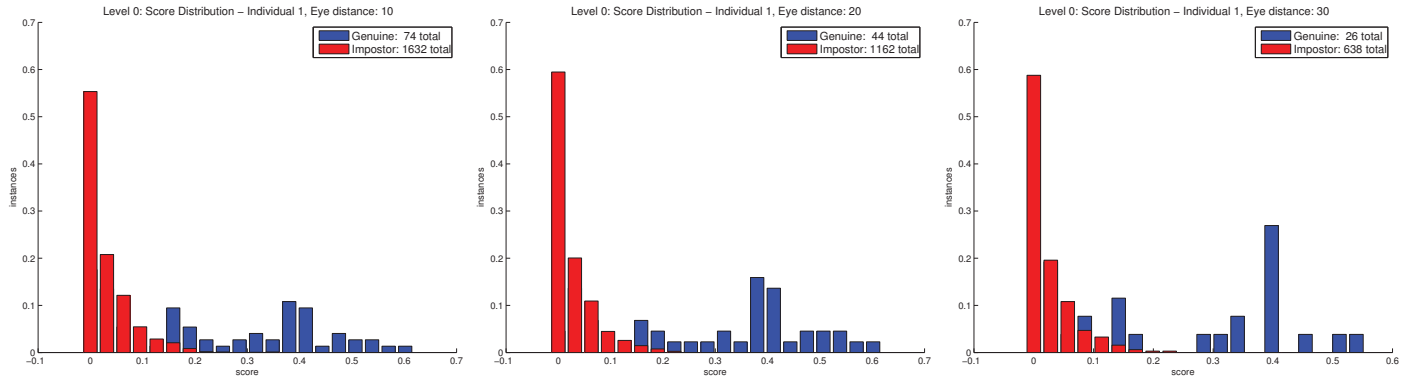


Fig. 1: Level 0 – Class scores distributions.

Level 1 Analysis

The figures bellow detail several performance curves, and the stars indicate the selected operation point for a target $fpr = 5\%$ (for each ed distance between eyes). The table summarizes the number of genuine and impostor samples, as well as face detection related metrics.

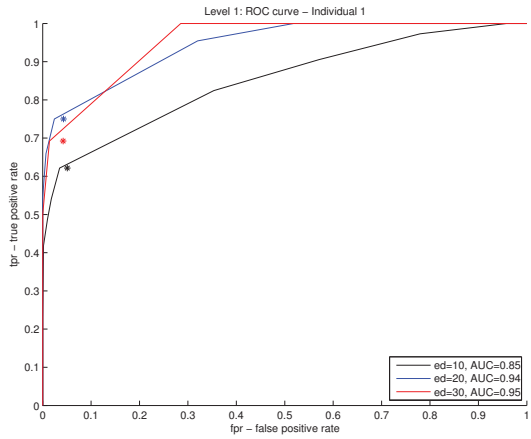


Fig. 2: ROC curve.

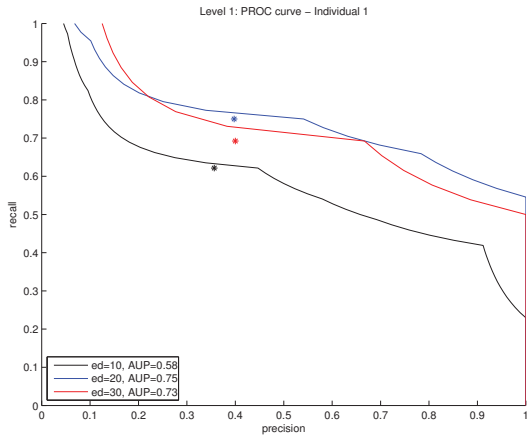


Fig. 3: PROC curve.

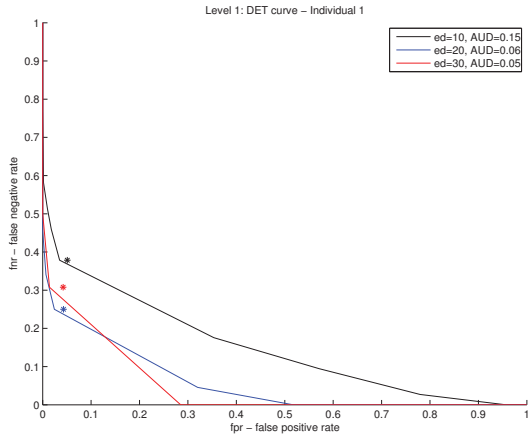


Fig. 4: DET curve.

Measure	Eyes distance (pixels)		
	10	**20**	**30**
Genuine faces (total)	74	44	26
Impostor faces (total)	1632	1162	638
Detection Level			
Falsely detected faces	30.42%	11.65%	18.74%
Failure to acquire rate	2.25%	30.42%	60.96%
Matching Level			
Low quality faces	6.57%	11.72%	19.20%
Operating points	0.1383	0.1315	0.1294
False positive rates	5.09%	4.30%	4.23%
True positive rates	62.16%	75.00%	69.23%

Tab. 1: Test set results for $fpr = 5\%$.

Level 2 Analysis

Distance	Ind. 1	Ind. 2	Ind. 3	Ind. 4	Ind. 5	Ind. 6	Ind. 7	Ind. 8	Ind. 9	Ind. 10	Ind. 11	Ind. 12	Ind. 13	Ind. 14	Ind. 15
10 px.	**62.16**%	0.00%	0.00%	1.61%	8.20%	1.18%	14.47%	0.00%	3.39%	3.03%	1.35%	3.03%	7.78%	6.33%	0.00%
20 px.	**75.00**%	0.00%	0.00%	2.63%	4.44%	1.69%	19.67%	0.00%	0.00%	4.00%	0.00%	0.00%	3.08%	5.66%	0.00%
30 px.	**69.23**%	0.00%	0.00%	0.00%	8.33%	0.00%	12.50%	0.00%	0.00%	3.57%	0.00%	0.00%	5.13%	6.67%	0.00%

(a)

Distance	Ind. 16	Ind. 17	Ind. 18	Ind. 19	Ind. 20	Ind. 21	Ind. 22	Ind. 23	Ind. 24	Ind. 25	Ind. 26	Ind. 27	Ind. 28	Ind. 29	Ind. 30
10 px.	5.00%	3.08%	4.23%	2.78%	7.55%	0.00%	0.00%	14.75%	6.17%	3.90%	11.11%	1.37%	3.23%	3.45%	5.63%
20 px.	2.70%	0.00%	5.66%	2.00%	9.52%	0.00%	0.00%	12.50%	5.26%	1.79%	2.78%	2.04%	4.00%	4.65%	6.52%
30 px.	5.26%	0.00%	6.25%	3.70%	12.50%	0.00%	0.00%	14.29%	3.23%	3.12%	0.00%	3.85%	3.57%	7.69%	0.00%

(b)

Tab. 2: Dodington's zoo based analysis for the detection module associated to individual 1. Columns details the individuals in the data set, while lines detail their detection by the module for each value of distance between the eyes. Colors are as follows: green for sheep like individuals (easy to predict), yellow for goat like individuals (difficult to predict), blue for lamb like individuals (can be impersonated by someone else) and red for wolf like individuals (who can impersonate another user).

Level 3 Analysis

Each of the below figures details the performance of systems by accumulating positive predictions over a time-window on the video stream for different distances between the eyes. The tracker is used to separate faces of different persons, and accumulate their predictions. Matching thresholds are set to provide a 5% false positive rate, and positive individual decision takes place after accumulating 20 detections in a 30 frames window (1 sec). Red stars indicate faces that have not been correctly matched to the target individual, while blue stars indicate that the individual captured in the video has been successfully matched to the target individual.

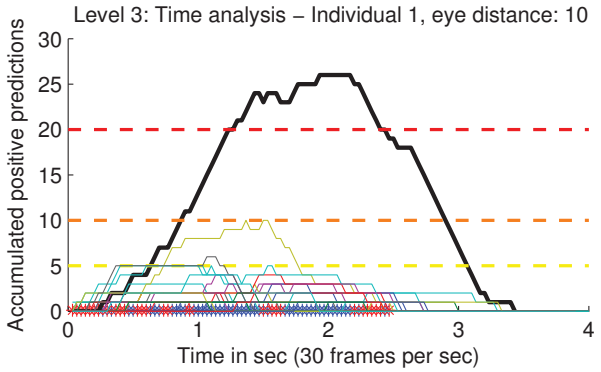


Fig. 5: Accumulated detections for 10 pixels between eyes.

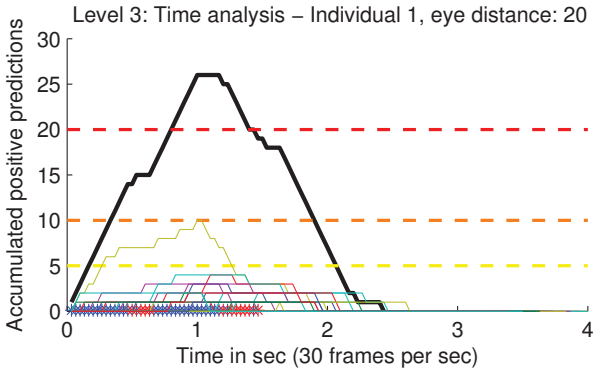


Fig. 6: Accumulated detections for 20 pixels between eyes.

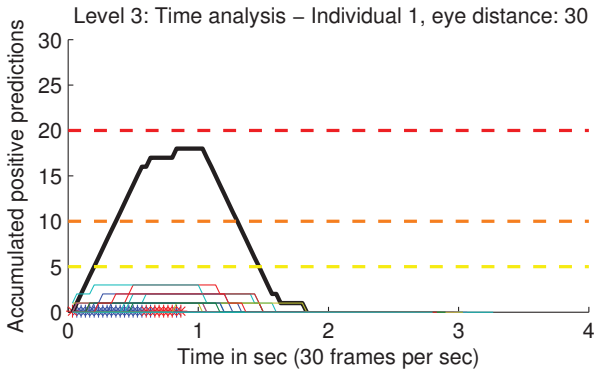


Fig. 7: Accumulated detections for 30 pixels between eyes.

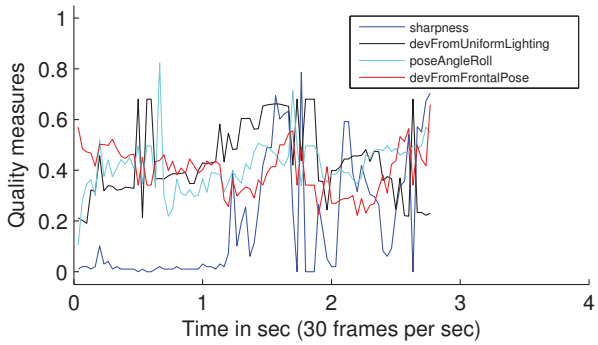


Fig. 8: Variations of quality measures.

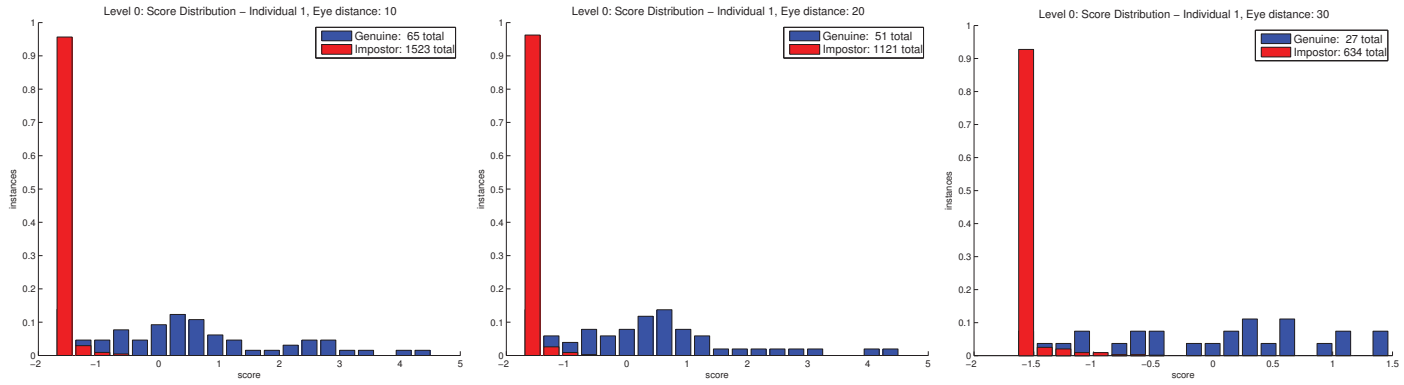


Fig. 1: Level 0 – Class scores distributions.

Level 1 Analysis

The figures bellow detail several performance curves, and the stars indicate the selected operation point for a target $fpr = 5\%$ (for each ed distance between eyes). The table summarizes the number of genuine and impostor samples, as well as face detection related metrics.

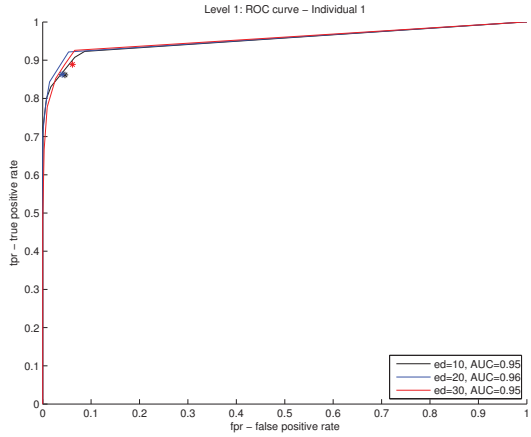


Fig. 2: ROC curve.

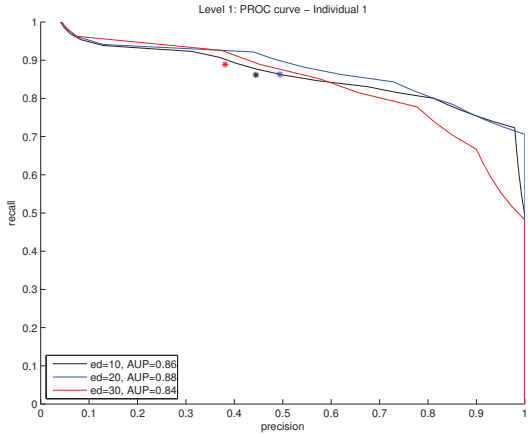


Fig. 3: PROC curve.

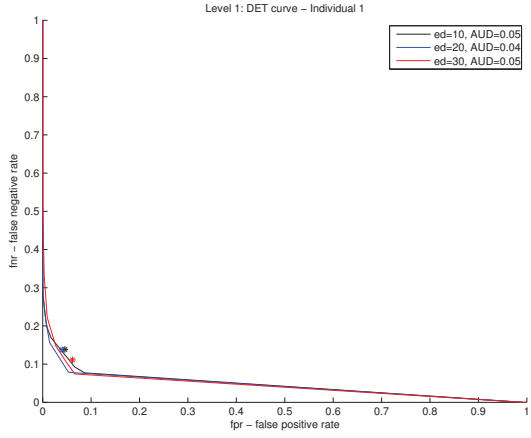


Fig. 4: DET curve.

Measure	Eyes distance (pixels)		
	10	**20**	**30**
Genuine faces (total)	65	51	27
Impostor faces (total)	1523	1121	634
Detection Level			
Falsely detected faces	1.79%	1.10%	1.93%
Failure to acquire rate	10.54%	33.97%	62.76%
Matching Level			
Low quality faces	0.00%	0.00%	0.00%
Operating points	-1.2690	-1.2726	-1.3660
False positive rates	4.60%	4.01%	6.15%
True positive rates	86.15%	86.27%	88.89%

Tab. 1: Test set results for $fpr = 5\%$.

Level 2 Analysis

Distance	Ind. 1	Ind. 2	Ind. 3	Ind. 4	Ind. 5	Ind. 6	Ind. 7	Ind. 8	Ind. 9	Ind. 10	Ind. 11	Ind. 12	Ind. 13	Ind. 14	Ind. 15
10 px.	**86.15%**	0.00%	0.00%	0.00%	18.03%	14.10%	35.82%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	5.41%	0.00%
20 px.	**86.27%**	0.00%	0.00%	0.00%	16.67%	5.08%	42.86%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.41%	0.00%
30 px.	**88.89%**	0.00%	0.00%	0.00%	25.00%	0.00%	72.73%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6.67%	0.00%

(a)

Distance	Ind. 16	Ind. 17	Ind. 18	Ind. 19	Ind. 20	Ind. 21	Ind. 22	Ind. 23	Ind. 24	Ind. 25	Ind. 26	Ind. 27	Ind. 28	Ind. 29	Ind. 30
10 px.	0.00%	1.61%	1.54%	4.35%	3.64%	0.00%	0.00%	5.17%	1.28%	0.00%	0.00%	0.00%	0.00%	0.00%	14.52%
20 px.	0.00%	0.00%	1.96%	5.66%	2.44%	0.00%	0.00%	0.00%	1.72%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
30 px.	0.00%	3.33%	3.33%	3.45%	0.00%	0.00%	0.00%	5.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.69%

(b)

Tab. 2: Dodington's zoo based analysis for the detection module associated to individual 1. Columns details the individuals in the data set, while lines detail their detection by the module for each value of distance between the eyes. Colors are as follows: green for sheep like individuals (easy to predict), yellow for goat like individuals (difficult to predict), blue for lamb like individuals (can be impersonated by someone else) and red for wolf like individuals (who can impersonate another user).

Level 3 Analysis

Each of the below figures details the performance of systems by accumulating positive predictions over a time-window on the video stream for different distances between the eyes. The tracker is used to separate faces of different persons, and accumulate their predictions. Matching thresholds are set to provide a 5% false positive rate, and positive individual decision takes place after accumulating 20 detections in a 30 frames window (1 sec). Red stars indicate faces that have not been correctly matched to the target individual, while blue stars indicate that the individual captured in the video has been successfully matched to the target individual.

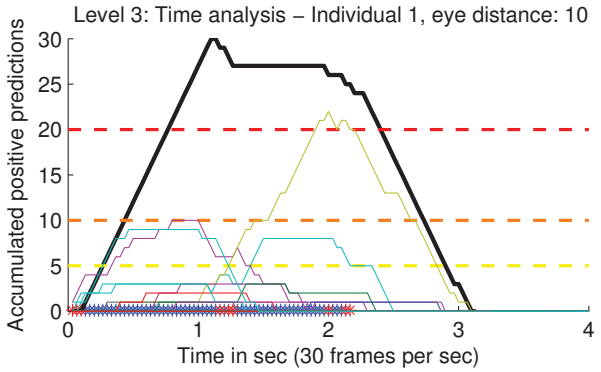


Fig. 5: Accumulated detections for 10 pixels between eyes.

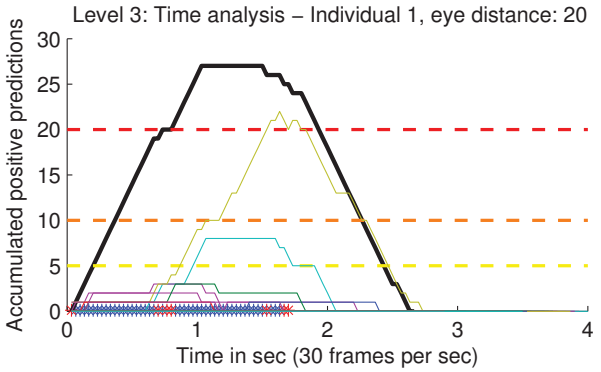


Fig. 6: Accumulated detections for 20 pixels between eyes.

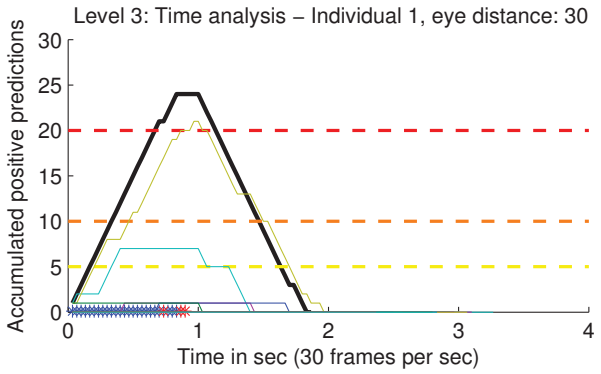


Fig. 7: Accumulated detections for 30 pixels between eyes.

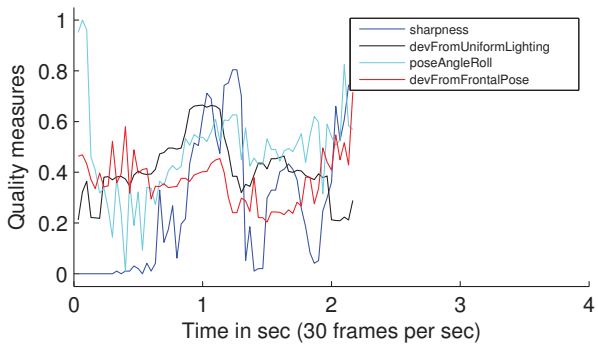


Fig. 8: Variations of quality measures.

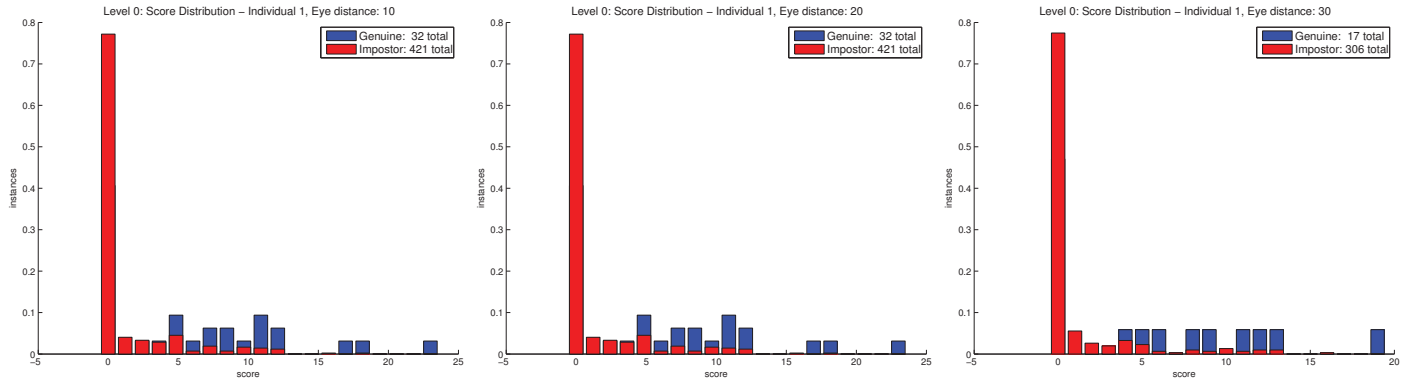


Fig. 1: Level 0 – Class scores distributions.

Level 1 Analysis

The figures bellow detail several performance curves, and the stars indicate the selected operation point for a target $fpr = 5\%$ (for each ed distance between eyes). The table summarizes the number of genuine and impostor samples, as well as face detection related metrics.

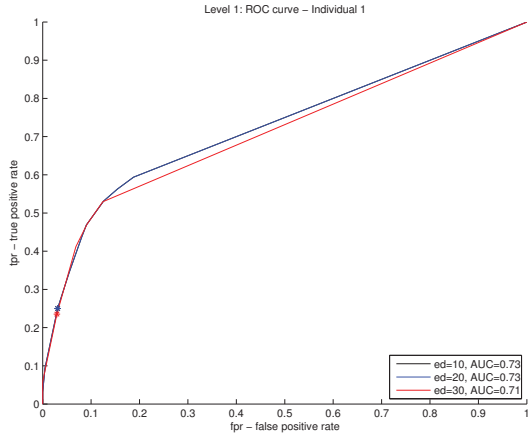


Fig. 2: ROC curve.

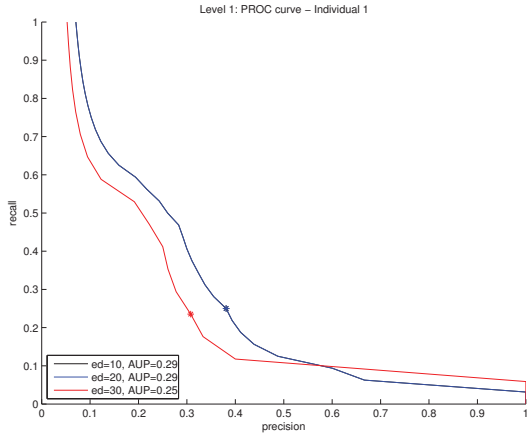


Fig. 3: PROC curve.

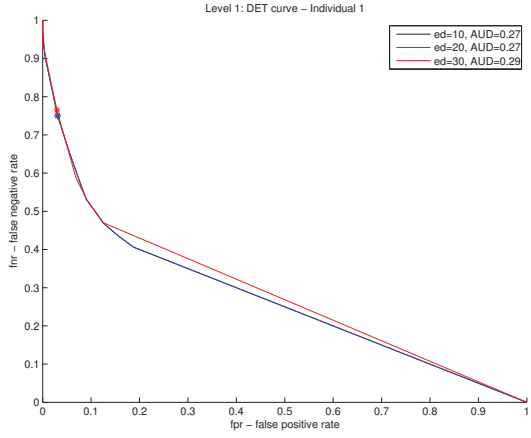


Fig. 4: DET curve.

Measure	Eyes distance (pixels)		
	10	**20**	**30**
Genuine faces (total)	32	32	17
Impostor faces (total)	421	421	306
Detection Level			
Falsely detected faces	0.22%	0.22%	0.31%
Failure to acquire rate	74.48%	74.48%	81.86%
Matching Level			
Low quality faces	0.00%	0.00%	0.00%
Operating points	10.7913	10.7913	10.1951
False positive rates	3.09%	3.09%	2.94%
True positive rates	25.00%	25.00%	23.53%

Tab. 1: Test set results for $fpr = 5\%$.

Level 2 Analysis

Distance	Ind. 1	Ind. 2	Ind. 3	Ind. 4	Ind. 5	Ind. 6	Ind. 7	Ind. 8	Ind. 9	Ind. 10	Ind. 11	Ind. 12	Ind. 13	Ind. 14	Ind. 15
10 px.	**25.00%**	0.00%	0.00%	11.11%	0.00%	0.00%	4.55%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
20 px.	**25.00%**	0.00%	0.00%	11.11%	0.00%	0.00%	4.55%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
30 px.	**23.53%**	0.00%	0.00%	11.11%	0.00%	0.00%	5.88%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

(a)

Distance	Ind. 16	Ind. 17	Ind. 18	Ind. 19	Ind. 20	Ind. 21	Ind. 22	Ind. 23	Ind. 24	Ind. 25	Ind. 26	Ind. 27	Ind. 28	Ind. 29	Ind. 30
10 px.	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9.09%	0.00%	0.00%	0.00%	26.67%	8.70%	0.00%	33.33%
20 px.	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9.09%	0.00%	0.00%	0.00%	26.67%	8.70%	0.00%	33.33%
30 px.	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9.09%	0.00%	0.00%	0.00%	26.67%	7.69%	0.00%	0.00%

(b)

Tab. 2: Dodington's zoo based analysis for the detection module associated to individual 1. Columns details the individuals in the data set, while lines detail their detection by the module for each value of distance between the eyes. Colors are as follows: green for sheep like individuals (easy to predict), yellow for goat like individuals (difficult to predict), blue for lamb like individuals (can be impersonated by someone else) and red for wolf like individuals (who can impersonate another user).

Level 3 Analysis

Each of the below figures details the performance of systems by accumulating positive predictions over a time-window on the video stream for different distances between the eyes. The tracker is used to separate faces of different persons, and accumulate their predictions. Matching thresholds are set to provide a 5% false positive rate, and positive individual decision takes place after accumulating 20 detections in a 30 frames window (1 sec). Red stars indicate faces that have not been correctly matched to the target individual, while blue stars indicate that the individual captured in the video has been successfully matched to the target individual.

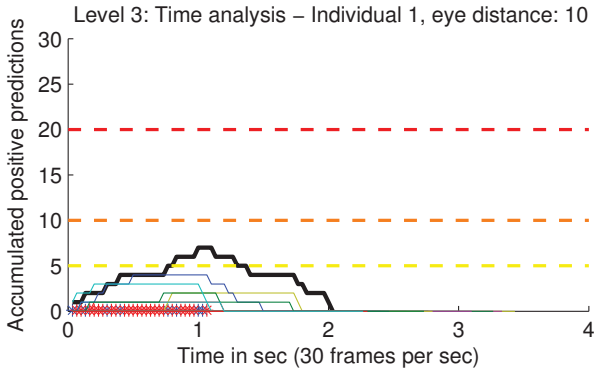


Fig. 5: Accumulated detections for 10 pixels between eyes.

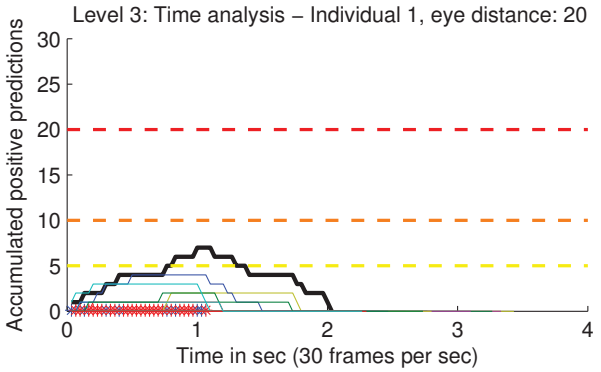


Fig. 6: Accumulated detections for 20 pixels between eyes.

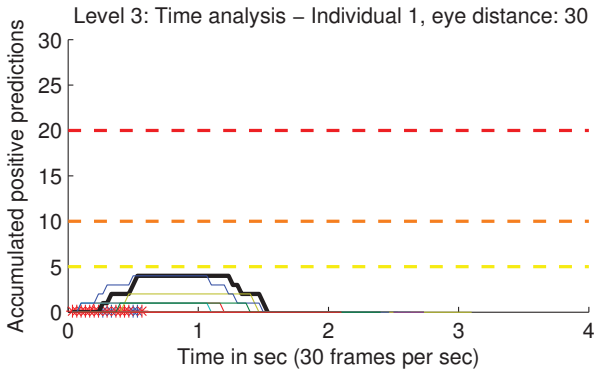


Fig. 7: Accumulated detections for 30 pixels between eyes.

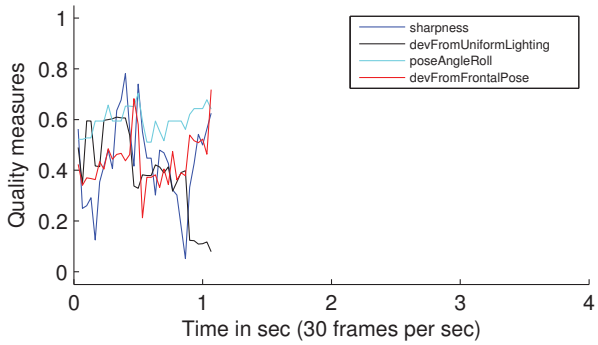


Fig. 8: Variations of quality measures.

5.1 Comparison of COTS Systems

This section provides a comparative analysis of systems performance according to level 1, 2 and 3, based on the distance of 20 pixels between the eyes. The impact of different eye distances on system performance is also discussed.

As shown in previous sections, at the level 0 analysis, the histogram distributions of matching scores are illustrated in Figure 1 of each Report Card for both target and impostor individuals and according to each distance between the eyes. For instance, the middle sub-figure of Figure 1 of each Report Card provides the matching scores produced by each system for 20 pixels between the eyes. Larger overlaps between the two distributions indicate that the system would have a low recognition rate, due to the difficulty in separating the target from impostor classes. On the other hand, less overlap between the distributions indicate easier class separation, and hence the system is expected to achieve a high recognition rate. Both the Cognitec and the PittPatt systems are shown to provide less overlap between the target and impostor distributions compared to that of the Neurotechnology, and therefore expected to provide a higher level of recognition performance as shown in level 1 analysis. The number of genuine and impostor individuals are also presented in the legend of each figure, which provides an indication about the number of frames detected by each system.

The Report Cards also provide other information at the detection level such as the failure to acquire (FTA) rate (the number of faces that are not detected by the system to the total number of faces found in the ground truth), the percentage of falsely detected faces (the number of faces incorrectly detected by the system, which has no match in the ground truth, over the total number of correctly detected faces), and the percentage of low quality faces (the number of correctly detected faces for which the system was not able to provide a matching score, over the total number of correctly detected faces). Table 8 presents the detection performance measures provided by each system using a distance of 20 pixels between the eyes.

As shown shown in Table 8, the Cognitec detector is able to achieve the lowest FTA rate however it also provides the highest percentage of falsely detected faces. On the other hand, the Neurotechnology detector achieves the highest FTA rate but with the lowest percentage of falsely detected faces, which means that the detector drops all frames that have low probabilities to include faces. PittPatt detector seems to be the most accurate providing the best tradeoff between FTA rate and the falsely detected faces. The percentage of low quality faces detected by the Cognitec results from a rejection threshold at the matcher level, which does not provide a matching score for the detected faces that are considered of poor quality. In contrast, the matchers for the other two systems always provide scores for the detected faces regardless of the face quality, therefore the values of the low-quality face measure remain zero.

In surveillance applications, the objective is to detect the presence of an individual of interest with a high

Table 8: Summary of system detection measures for 20 pixels between the eyes.

Measure	Cognitec	PittPatt	Neurotechnology
Failure to acquire rate	30.42%	33.97%	74.48%
Falsely detected faces	11.65%	1.10%	0.22%
Low quality faces	11.72%	0.00%	0.00%

level of accuracy without generating many false alarms. Level 1 analysis evaluates the overall recognition performance – the system capability to correctly match the faces of target individuals on watch list. The transaction-based performance evaluation (level 1 analysis) is illustrated in Report Cards using the ROC, DET and PROC curves. For a given false positive rate, the ROC or DET curve illustrates the proportion of target individuals that has been correctly detected by the system among all target individuals (i.e., the true positive rate or recall), while the precision on the PROC curve evaluates a system's ability to detect target individuals (among all individuals detected as positives) at every recall. For each distance between the eyes, the operating points (matching thresholds) selected during validation (using Chokepoint P2L_S1_C1.1 data) are also shown as stars on the curves achieved by each system using Chokepoint test data (P2L_S4_C1.1).

Larger deviations between the points and the curves indicate large variation between systems performance on validation and test data. Table 9 presents a summary of transaction-based performance evaluation for 20 pixels between the eyes. Since the operating points are computed during validation according to a false positive rate (fpr) of 5%, a high-performing system must therefore maintain a fpr value of 5% or lower during testing, while achieving high-level of true positive rate (tpr), precision ($prec$), $F1$ measure, and Area under the ROC curve (AUC) and $AUC_{0.05}$. Note that $AUC_{0.05}$ means the area under the ROC curve for a $fpr = [0, 0.05]$, which focuses on the low fpr region and is more accurate than the AUC especially when the ROC curves cross.

As shown in Table 9, the three systems are able to maintain a fpr value lower than 5% on average among all target individuals. Neurotechnology system provides the lowest level of recognition performance in terms of all measures (tpr, $prec$, $F1$, AUC and $AUC_{0.05}$). The average level of recognition performance achieved by Cognitec and PittPatt systems is comparable. Although PittPatt system provides a slightly higher average level of matching performance than that of Cognitec, the lower variations in the performance measures achieved by Cognitec system indicate more robustness to target individual variations. This variation in PittPatt system performance can also be seen in the results of Individuals 6 and 11 in Table 9, where the achieved fpr value on the test data is higher than the desired validation value of 5%. Typically, lower variations between the results achieved during validation and testing (along with high level of performance values) provides more confidence in the expected system performance during operations.

Level 2 analysis presents subject-based categorization according to Doddington's Zoo. Each individual can be categorized as a sheep (easy to classify), lamb (a target easy to be impersonated), a goat (a target difficult to classify) or a wolf (an impostor that can easily impersonate someone else). The numbers in Table 2 in the Report Cards present the proportions of samples that have been matched to target individuals, whereas the colors indicate the type of each individual.

Finally, the time analysis performed at level 3 provides an unbiased evaluation and comparison of systems performance over video streams. When individuals appear in the video frames, the system attempts to detect individual faces, track their positions over time (by assigning each individual face a unique ID), and then match each face to those in the target watch list. If the final decision (target or non-target) is based on each frame independently, the system is would provide a higher rate of false positive and negative errors. A more deep individual specific analysis can be made by looking at the entire sequences of decisions accumulated on the face trajectories over the entire video frames. When a tracked face is matched to a target individual in the watch list, the positive predictions are accumulated over a moving time window of one

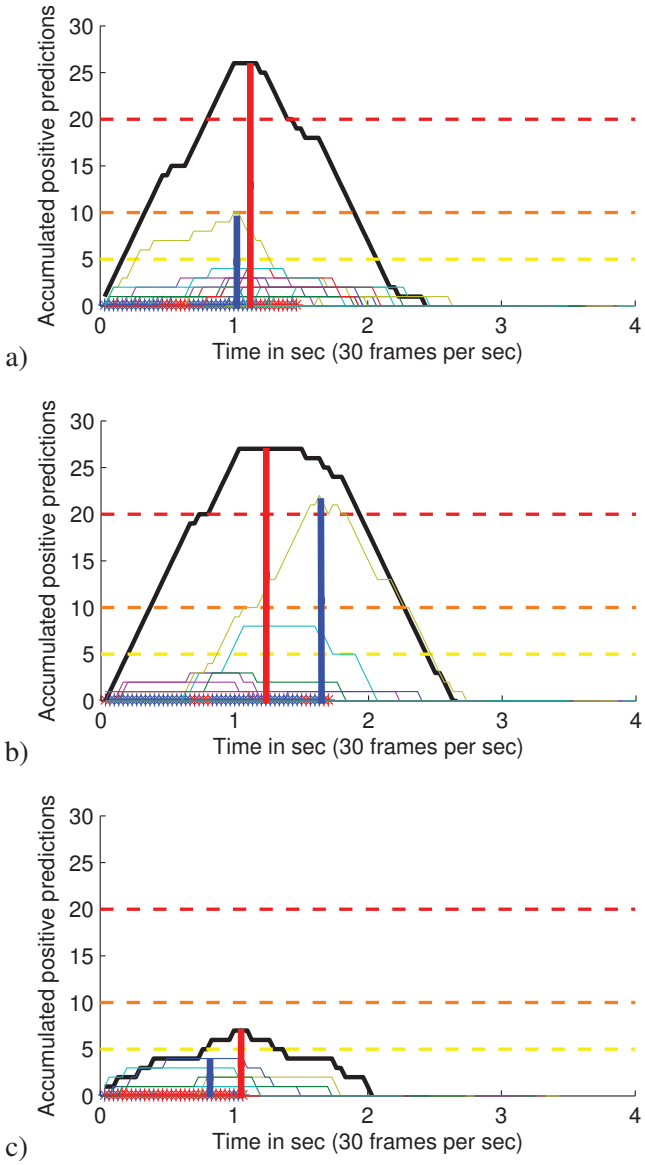
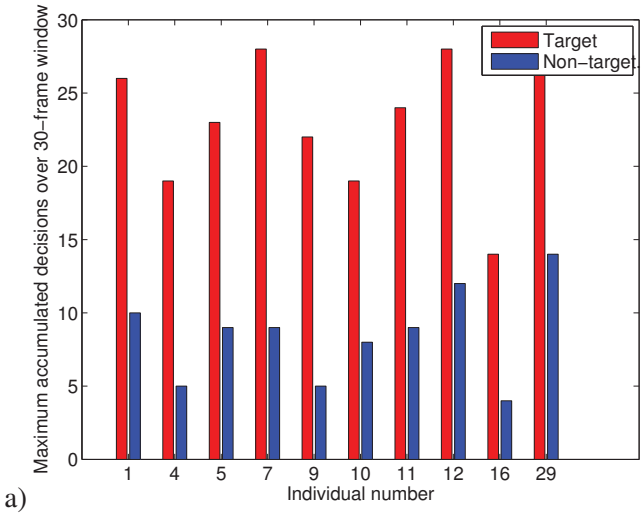
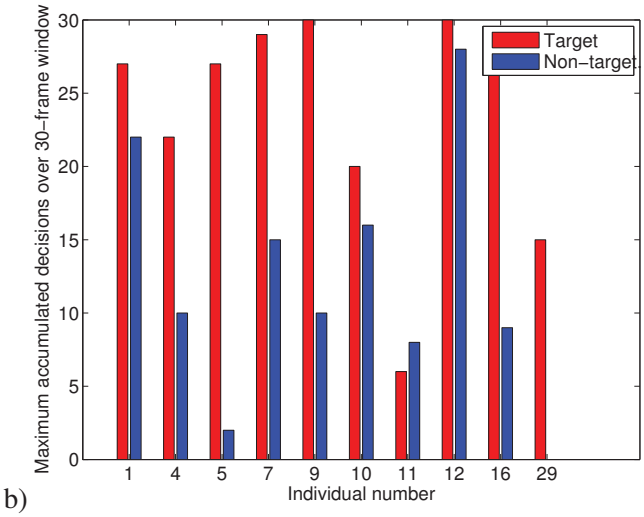


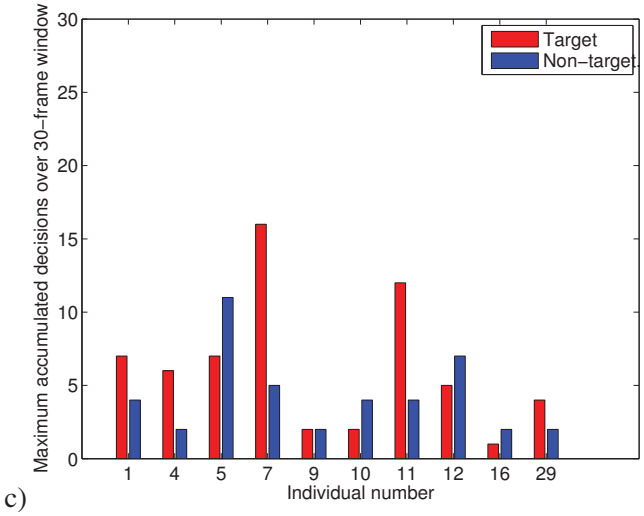
Figure 16: Example of time-based analysis (level 3) for target individual 1 and 20 pixels between the eyes: a) Cognitec, b) PittPatt, c) Neurotechnology.



a)



b)



c)

Figure 17: Summary of time-based analysis (level 3) for 20 pixels between the eyes: a) Cognitec, b) PittPatt, c) Neurotechnology.

second (or 30 frames). Decision thresholds can then be set on these cumulative positive predictions to obtain various levels of confidence in identifying the target individual. The maximum confidence in decision making is obtained when the accumulated positive predictions attain the value of 30 and maintain this value over the face trajectory, which means that the face has been detected, tracked and identified correctly.

Figure 16 presents and example of the time-based analysis produced by each system, for target individual 1 in the watch list using 20 pixels between the eyes. The decision thresholds are set to 5, 10 and 20 (illustrated by the yellow, orange and red dashed lines on the figures) on the accumulated positive predictions, which provide a low, medium and high confidence in the final decision about the target individual. As shown in Figure 16, Cognitec and PittPatt systems are able to achieve a high level of accumulated positive predictions (the black graph exceeded the red threshold), while Neurotechnology system provided a low confidence in identifying the same target individual. The red stars on the figures indicate the detected faces that have not been correctly matched to the target individual, while the blue stars indicate that the individual face captured in the corresponding video frame has been successfully matched to the target individual. These points also provide an indication of the face detection algorithm performance. When more faces are detected over time (as shown in Figure 16 for PittPatt detector), chances are higher to accumulate positive predictions and increase the confidence in identifying a given target individual. The Report Cards evaluate systems time analysis performances for each target individual using the three distances between the eyes.

An important aspect of the time analysis is to illustrate the accumulative positive predictions for the non-target individuals as shown in Figure 16. A large difference between the maximum level of accumulated positive predictions of the target (red bars on figures) and all non-target individuals (blue bars on figures), enforces the system confidence in detecting a specific target and indicates more robustness to variations in environment conditions. This also confirms the importance of setting a subject-specific matching thresholds. Although both the Cognitec and the PittPatt systems achieve a comparable (high) level of accumulated pos-

Table 9: Summary of transaction-based (level 1) analysis for 20 pixels between the eyes.

Product	Measure	Ind01	Ind04	Ind05	Ind07	Ind09	Ind10	Ind11	Ind12	Ind16	Ind29	AVG	STD
Cognitec	fpr	4.30%	3.77%	4.05%	3.84%	5.14%	3.81%	3.73%	5.43%	3.34%	3.10%	4.05%	0.007
	tpr	75.00%	47.37%	68.89%	70.49%	71.05%	62.00%	75.00%	95.56%	43.24%	97.67%	70.63%	0.166
	$prec$	39.76%	29.03%	39.74%	49.43%	31.03%	41.33%	47.56%	40.57%	29.09%	53.85%	40.14%	0.081
	$F1$	0.520	0.360	0.504	0.581	0.432	0.496	0.582	0.570	0.348	0.694	0.509	0.101
	AUC	0.944	0.908	0.936	0.946	0.944	0.941	0.951	0.994	0.945	0.997	0.951	0.025
	$AUC_{0.05}$	0.719	0.443	0.589	0.636	0.567	0.549	0.686	0.885	0.414	0.953	0.644	0.165
PittPatt	fpr	4.01%	3.43%	0.62%	4.48%	1.68%	6.04%	3.30%	11.00%	2.21%	1.75%	3.85%	0.028
	tpr	86.27%	72.22%	91.67%	87.50%	92.11%	48.94%	21.15%	100.00%	84.21%	89.29%	77.34%	0.230
	$prec$	49.44%	40.00%	86.27%	49.49%	64.81%	25.27%	22.92%	26.63%	56.14%	55.56%	47.65%	0.188
	$F1$	0.629	0.515	0.889	0.632	0.761	0.333	0.220	0.421	0.674	0.685	0.576	0.193
	AUC	0.956	0.852	0.968	0.946	0.985	0.725	0.600	0.997	0.916	0.946	0.889	0.123
	$AUC_{0.05}$	0.852	0.613	0.929	0.796	0.945	0.407	0.184	0.948	0.762	0.884	0.732	0.244
Neurotech.	fpr	3.09%	1.80%	13.41%	4.89%	1.35%	2.26%	4.90%	4.76%	1.60%	2.71%	4.08%	0.034
	tpr	25.00%	66.67%	53.85%	45.45%	25.00%	18.18%	50.00%	41.67%	6.67%	40.00%	37.25%	0.173
	$prec$	38.10%	42.86%	10.61%	50.00%	25.00%	16.67%	36.36%	19.23%	12.50%	25.00%	27.63%	0.128
	$F1$	0.302	0.522	0.177	0.476	0.250	0.174	0.421	0.263	0.087	0.308	0.298	0.132
	AUC	0.726	0.978	0.882	0.905	0.866	0.779	0.854	0.805	0.645	0.808	0.825	0.090
	$AUC_{0.05}$	0.206	0.757	0.152	0.402	0.626	0.238	0.400	0.413	0.114	0.356	0.366	0.194

itive predictions, the maximum value of the non-target positive predictions achieved with PittPatt is closer to that of the target individual. This means a slight variation in the operating or environment conditions would make the PittPatt system to incorrectly identify a non-target individual as a person in the watch list.

Figure 17 provides a summary of time-based performance achieved by each system for 20 pixels between the eyes, in terms of the maximum level of accumulated positive predictions of the target and all non-target individuals. In general, both Cognitec and PittPatt systems achieve higher levels of accumulated positive predictions for target individuals than that of the Neurotechnology system. These performance levels are comparable for the Cognitec and the PittPatt systems, however Cognitec seems to have larger differences (and lower variations) between the maximum values of the target and non-target positive predictions than that of the PittPatt. In some cases, such as for individual 11, the PittPatt system provides larger value for the maximum level of accumulated positive predictions of non-target individuals than that of the target individual. This higher level of maximum accumulated positive predictions of non-target individuals is also apparent in the performance of Neurotechnology (as seen in Figure 17) for individuals 5, 10, 12, 16.

Subject-based analysis (level 2) provides more insights on the type of errors committed by the system based on the type of individuals as charachterized by the Doddington's Zoo. An analysis of individuals properties can expose fundamental weaknesses in a biometric system, and allows to develop more robust systems. Level 2 analysis presented in Table 2 in the Report Cards shows that the target individuals discussed above (i.e., individual 11 for PittPatt, and individuals 5, 10, 12, 16 for Neurotechnology system) are assigned to the goat category. Goats like (target) individuals are intrinsically difficult to recognize, they are characterized by consistently low classification scores against themselves, and tend to adversely degrade system performance by increasing the false negative rate. The subject-based analysis shows that the sheep like individuals, which are easy to discriminate from other non-target individuals, dominate the population and the levels of performances provided by each system is relatively high for this category.

In non-target populations, subject-based analysis also show the presence of wolves and lambs (see for example, individual 1 and 12 for PittPatt system). Wolves are exceptionally successful at impersonating other targets, while lambs are easy to impersonate and thereby contributing to a high false alarm rate. The presence of wolves and lambs in non-target individuals, explains the small differences between the maximum level of accumulated positive predictions between the target individuals 1 and 12 and the non-target individuals for PittPatt system, as shown in Figure 16. Subject-based analysis confirms therefore the need for user-specific or template-specific thresholds.

According to Cognitec[5] the sharpness assessment of images quality is computed by applying a $3x3$ mean filter to the image and calculating the distance between original image and filtered one. Larger sharpness values (or larger difference between the original and the filtered image) indicate higher level of noise in the original image, which may lead to lower recognition performances. The sharpness value seems to have an inverse impact on systems performance, as shown in Figure 8 of the Report Cards, and to be positively correlated with deviation from uniform lighting. On the other hand, the deviation from frontal pose and the pose angle roll measures are shown to have a large impact on systems performances. The larger the deviation from frontal pose the lower the recognition performance.

[5] See slide 5: http://biometrics.nist.gov/cs_links/quality/workshopI/proc/weber_bqw.pdf

The impact on system performance of the distance between the eyes is illustrated in Report Cards at level 1-3 analysis. A large distance between the eyes (based on the number of pixels between the center of the eyes) implies a high resolution face image, which is typically captured when the individual is at a small distance from the camera with a close to frontal pose. On the other hand, a small inter-eye distance implies low resolution face image with a large distance from the camera or a close to profile head pose. In surveillance application, the time or the number of frames with large distances between the eyes is typically short, since the individuals approaching the camera would quickly go out of its field of view. In general, the transaction-based performance increases with eye distance, as shown for level 1 analysis in the Report Cards, because the system ability to correctly identify target individuals increases with eye distance. However, the number of genuine and impostor faces decrease (and hence the FTA rate increases) when the distance between eyes increases, because the frames with smaller eye distances than the desired value are dropped. Similarly, the time analysis performance decreases when the eye distance increases due to the short time spent in front of the camera view field. Therefore, the system will not be able to accumulate a high level of confidence in the positive predictions over the small number of captured frames.

The above discussions hold for the two other distances between the eyes of 10 and 30 pixels for each of the evaluated systems. However, Neurotechnology system provides similar and substantially poorer performance for both 10 and 20 pixels between the eyes, which indicates that the system may not be suitable for small distances between the eyes.

6 Conclusions

This report surveys metrics, methodologies and data-sets used for evaluation of face recognition in video and establishes a multi-level evaluation methodology that is suitable for video surveillance applications. The developed methodology allows one to access the vulnerabilities of FR systems when applied to real-time surveillance applications such as screening of faces against wanted list (still-to-video application) and matching a faces across several video feeds (video-to-video application).

The results obtained by using the methodology from the evaluation of three COTS FR products (Cognitec, PittPatt and NeuroTechnology) on the publicly available Chokepoint data-set are presented to illustrate the methodology and to expose the vulnerabilities of each product. The performance results are reported using a two-page Report Card format, one Report Card per each target individual in a Watch List, which summarize the ability of the system to automatically detect and recognize each target individual in a surveillance video-stream. This report showed the Reports Cards obtained only for one individual in the Chokepoint data-set (Individual 1). The complete results obtained for each tested individual from the data-set are presented in a separate report [23].

The obtained evaluation results presented in this report, along with the survey of academic and commercial state of art solutions presented in [20, 24], provided the basis for the assessment of readiness of the FR technology for video surveillance applications. This assessment was the key objective of the PROVE-IT(FRiV) study and it has been reported in [16, 4] and further refined in [17].

References

[1] M. R. Anderberg, *Cluster Analysis for Applications*. Academic Press, 1973.

[2] M. Barreno, A. Cardenas, and D. Tygar, "Optimal ROC for a combination of classifiers," in *Advances in Neural Information Processing Systems (NIPS) 20*, January 2008.

[3] C. Bergamini, L. Oliveira, A. Koerich, and R. Sabourin, "Combining different biometric traits with one-class classification," *Signal Processing*, vol. 89, pp. 2117–2127, 2009.

[4] D. Bissessar, E. Choy, D. Gorodnichy, and T. Mungham, "Face recognition and event detection in video: An overview of prove-it projects (biom401 and bts402)," *CBSA Science and Engineering Directorate Technical Report 2013-04, June 2013*.

[5] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "The relation between the roc curve and the cmc." in *AutoID*. IEEE Computer Society, 2005, pp. 15–20.

[6] R. M. Bolle, S. Pankanti, and N. K. Ratha, "Evaluation techniques for biometrics-based authentication systems (frr)." in *ICPR*, 2000, pp. 2831–2837.

[7] A. Brew and P. Cunningham, "Combining cohort and ubm models in open set speaker detection," vol. 48, no. 1, Van Godewijckstraat 30, Dordrecht, 3311 GZ, Netherlands, 2010, pp. 141 – 159.

[8] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, *Introduction to Algorithms*, 2nd ed. McGraw-Hill Higher Education, 2001.

[9] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 233–240.

[10] B. DeCann and A. Ross, "Relating roc and cmc curves via the biometric menagerie," in *Proc. of 6th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), (Washington DC, USA), September 2013*.

[11] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *International Conference on Spoken Language Processing*, 1998.

[12] R. C. Dubes and A. K. Jain, *Algorithms for Clustering Data*. Prentice Hall, 1988.

[13] H. K. Ekenel, L. Szasz-Toth, and R. Stiefelhagen, "Open-set fr-based visitor interface system," in *Proceedings of the 7th International Conference on Computer Vision Systems: Computer Vision Systems*, ser. ICVS '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 43–52.

[14] T. Fawcett, "An introduction to roc analysis," *Pattern Recogn. Lett.*, vol. 27, pp. 861–874, June 2006.

[15] R. Goh, L. Liu, X. Liu, and T. Chen, "The CMU Face In Action (FIA) Database," 2005, pp. 255–263.

[16] D. Gorodnichy and E. Granger, "Evaluation of face recognition for video surveillance," *Proceedings of NIST International Biometrics Performance Conference (IBPC 2012)*.

[17] D. Gorodnichy, E. Granger, J.-P. B. abd D. Bissessar, E. Choy, T. Mungham, R. Laganiere, S. Matwin, E. Neves, C. Pagano, M. D. la Torre, and P. Radtke, "Prove-it(friv): framework and results," *Proceedings of NIST International Biometrics Performance Conference (IBPC 2014)*.

[18] D. Gorodnichy, "Further refinement of multi-order biometric score analysis framework and its application to designing and evaluating biometric systems for access and border control," *IEEE SSCI Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM), 2011, Paris*.

[19] D. O. Gorodnichy, "Video-based framework for face recognition in video," *Proc. of Second Canadian Conference on Computer and Robot Vision (CRV'05), Workshop on Face Processing in Video*, vol. 1, no. 1, pp. 330 – 338, May 2005. [Online]. Available: http://www.videorecognition.com/FRiV

[20] D. Gorodnichy, E. Granger, and P.Radtke, "Survey of commercial technologies for face recognition in video," CBSA, Border Technology Division, Tech. Rep. 2014-22 (TR), 2014.

[21] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *in: Proceedings of the 27th European Conference on Information Retrieval*, 2005, pp. 345–359.

[22] E. Granger and D. Gorodnichy, "Evaluation methodology for face recognition technology in video surveillance applications," CBSA, Tech. Rep. 2014-27 (TR), 2014.

[23] E. Granger, D. Gorodnichy, E. Choy, W. Khreich, P.Radtke, J. Bergeron, and D. Bissessar, "Results from evaluation of three commercial off-the-shelf face recognition systems on chokepoint dataset," CBSA, Tech. Rep. 2014-29 (TR), 2014.

[24] E. Granger, P. Radtke, and D. Gorodnichy, "Survey of academic research and prototypes for face recognition in video," CBSA, Border Technology Division, Tech. Rep. 2014-25 (TR), 2014.

[25] R. Gross and J. Shi, "The CMU motion of body (MoBo) database, 2001," Carnegie Mellon University, Tech. Rep. CMU-RI-TR-01-18, 2001.

[26] T. Landgrebe, P. Paclik, R. Duin, and A. Bradley, "Precision-recall operating characteristic (p-roc) curves in imprecise environments," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4, 0-0 2006, pp. 123 –127.

[27] N. Lavrac, H. Motoda, T. Fawcett, R. Holte, P. Langley, and P. W. Adriaans, "Introduction: Lessons learned from data mining applications and collaborative problem solving," *Machine Learning*, vol. 57, no. 1-2, pp. 13–34, 2004.

[28] S. Lawrence, C. Giles, A. C. Tsoi, and A. Back, "Fr: a convolutional neural-network approach," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 98 –113, jan 1997.

[29] F. Li and H. Wechsler, "Open set fr using transduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 11, pp. 1686 –1697, nov. 2005.

[30] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Fifth European Conference on Speech Communication and Technology*, vol. 97, no. 4, 1997, pp. 1895–1898.

[31] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The det curve in assessment of detection task performance." in *EUROSPEECH*, G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds. ISCA, 1997.

[32] C. McCool and S. Marcel, "Mobio database for the icpr 2010 face and speech competition," *IDIAP Communication Report, Idiap-Com-02-2009*. [Online]. Available: http://www.idiap.ch/scientific-research/resources/mobio

[33] U. of Notre Dame Biometrics Data Sets, "http://www3.nd.edu/ cvrl/cvrl/data_sets.html."

[34] C. Pagano, E. Granger, R. Sabourin, and D. O. Gorodnichy, "Detector ensembles for fr in video surveillance," in *Accepted for publication in the IJCNN 2012 proceedings*, 2012.

[35] J. N. Pato and E. W. B. C. N. R. C. Lynette I. Millett, *Biometric Recognition:Challenges and Opportunities*. The National Academies Press, 2010.

[36] N. Poh and J. Kittler, "A unified framework for biometric expert fusion incorporating quality measures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 3 –18, jan. 2012.

[37] ——, "Incorporating model-specific score distribution in speaker verification systems." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 3, pp. 594–606, 2008.

[38] N. Poh, R. Wong, J. Kittler, and F. Roli, "Challenges and research directions for adaptive biometric recognition systems," in *Advances in Biometrics*, ser. Lecture Notes in Computer Science, M. Tistarelli and M. Nixon, Eds. Springer Berlin / Heidelberg, 2009, vol. 5558, pp. 753–764.

[39] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *In Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 445–453.

[40] A. Rattani, "Adaptive biometric system based on template update paradigm," Ph.D. dissertation, University of Cagliari, Italy, 2010.

[41] A. Rattani, G. Marcialis, and F. Roli, "An experimental analysis of the relationship between biometric template update and the doddingtons zoo: A case study in face verification," in *Image Analysis and Processing ICIAP 2009*, ser. Lecture Notes in Computer Science, P. Foggia, C. Sansone, and M. Vento, Eds. Springer Berlin / Heidelberg, 2009, vol. 5716, pp. 434–442.

[42] E. Tabassi, "Image specific error rate: A biometric performance metric," in *Proc. International Conference on Pattern Recognition*, 2010, pp. 1124–1127.

[43] U. Uludag, A. Ross, and A. K. Jain, "Biometric template selection and update: a case study in fingerprints," *Pattern Recognition*, vol. 37, no. 7, pp. 1533–1542, 2004.

[44] S. D. Walter, "The partial area under the summary roc curve." *Statistics in Medicine*, vol. 24, no. 13, pp. 2025–2040, 2005.

[45] G. Weiss, "The effect of small disjuncts and class distribution on decision tree learning," Ph.D. dissertation, Rutgers University, 2003.

[46] M. Wittman, P. Davis, and P. J. Flynn, "Empirical studies of the existence of the biometric menagerie in the frgc 2.0 color image corpus," in *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, ser. CVPRW '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 33–.

[47] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," *Computer Vision and Pattern Recognition Workshops*. [Online]. Available: http://itee.uq.edu.au/~uqywong6/chokepoint.html

[48] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, pp. 220–230, 2010.

Government of Canada / Gouvernement du Canada

PROVE-IT(FRiV)

Final Deliverable

Face Recognition in Video Surveillance Applications

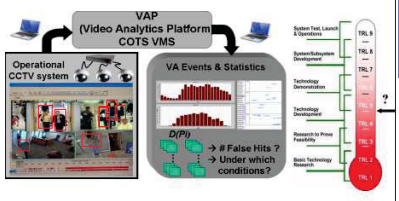
E. Granger, P. Radtke and W. Khreich
Laboratoire d'imagerie, de vision et d'intelligence artificielle
École de technologie supérieure (Université du Québec)

D. Gorodnichy
Science and Engineering Directorate
Canada Border Services Agency

VT4NS'13 (March 27, 2013)

ÉTS Le génie pour l'industrie | LIVIA Laboratoire d'imagerie, de vision et d'intelligence artificielle | Canada

Project number: PSTP 03-401BIOM



PROVE-IT (FRiV)
Pilot and Research on Operational Video-based Evaluation of Infrastructure and Technology:
Face Recognition in Video

Lead: Canada Border Services Agency
Contributing Partners: uOttawa(VIVA,TAMALE), uQuébec-ÉTS
Observing Partners: RCMP, DRDC, DFAIT, CATSA, TC, PCO HomeOffice, FBI, NIST
Start-End: Sep, 2011 – March 2013
Funds: $200,000 (PSTP), In-Kind: $400,000. Total: $600,000
Synergy project: PROVE-IT (VA), Lead: CBSA

Objectives:
- Develop methodology for evaluating FRiV solutions (using sets, mockups, and pilots)
- Assess the applicability of FRiV technologies for the following video surveillance applications :
  1. **Triaging of faces** (screening against Wanted List);
  2. Fusion of face recognition from different cameras
  3. Face recognition-assisted tracking;
  4. Matching a face/person across several video feeds;
  5. Multi-modal recognition (eg face and voice or iris);
  6. Soft-biometric based tracking/recognition
- Investigate, develop and test the **Face Processing** (FP) components that are required for these applications:
  1. Pre-processing, Post-processing, Fusion
  2. Face Detection, Face Tracking, **Face Tagging**

Knowledge:
- Use CBSA's developed Video Analytic Platform (VAP) and to integrate commercial and academic FR and FP codes into operational and mock-up IP-camera based surveillance systems.
- Leverage **CBET** and VT4NS initiatives and portals

Outputs:
1. Identify environmental and procedural constraints under which Instant Face Recognition (iFR) is feasible (has **Technology Readiness Level** TRL> 5)
2. Report findings including recommendations for the deployment of iFR and FRiV technologies by the GoC
3. VT4NS workshop with demonstration of FRiV technology

Impact:
- Establish the foundation for incremental enhancement of in-house knowledge and capacity in the field of FRiV, which will allow GoC to deploy FRiV technologies in operational CCTV environments
- Insure 1) that the delivered results are both technically sound and relevant to GoC needs and 2) that the expertise obtained though this study is retained within the GoC.
- Establish a partnership with Canadian Academia and International federal departments in addressing the challenging problems related to FRiV.

Canada

Summary

CBSA ASFC

With ETS and U. Ottawa (TAMALE Lab)

1. Overview of the FRiV market / solutions
2. Developed methodology for evaluating FRiV solutions (using sets, mockups, and pilots)
3. Investigated, developed and tested the **Face Processing** (FP) components
   1. Pre-processing, Post-processing, Fusion
   2. Face Detection, Face Tracking, **Face Tagging**
4. Identified environmental and procedural constraints under which Instant Face Recognition (iFR) is feasible (has **Technology Readiness Level** TRL> 5)
5. VT4NS workshop with demonstration of FRiV technology

DISCLAIMER:

CBSA ASFC

The results presented in this report were produced in experiments conducted by the CBSA, and should therefore not be construed as vendor's maximum-effort full-capability result. In no way do the results presented in this presentation imply recommendation or endorsement by the Canada Border Services Agency, nor do they imply that the products and equipment identified are necessarily the best available for the purpose.

OVERVIEW

ETS Mandate in PROVE-IT(FRiV)

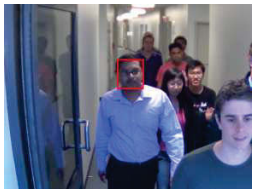
1. **Survey Face Recognition in Video Sruveillance:**
   - commercial technologies and patents
   - academic systems and software
2. **Evaluation Methodologies for FRiVS:**
   - public data sets for medium- to large-scale evaluation
   - experimental protocols for video surveillance scenarios
   - performance metrics and analysis
3. **Case Studies – Evaluate in Applications**
   - unmanned screening of faces against a wanted list
   - fusion of face recognition across cameras, etc.

OVERVIEW

1) Background – FR in Video Surveillance
2) Academic and Commercial Solutions
3) Evaluation Methodologies for FRiVS
4) Case Studies – Screening and Fusion
5) In-House Evaluations of Technologies
6) Conclusions and Recommendations

1) FR in Video Surveillance

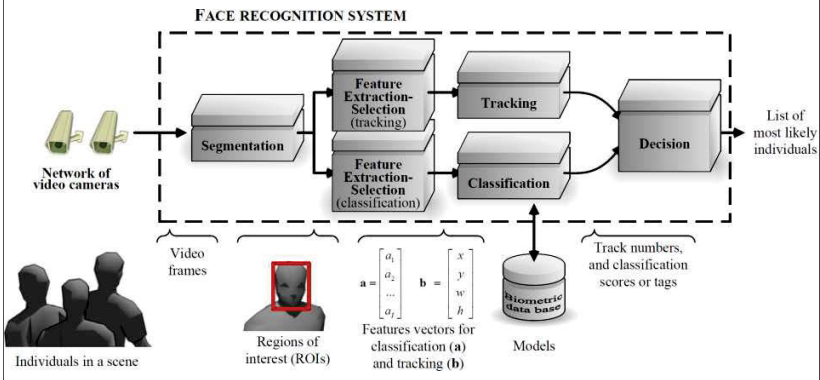
Objectives: enhanced screening and situation analysis across a network of surveillance cameras



- automatically recognize and track individuals within semi- and unconstrained environments
- determine if facial regions captured in video streams correspond to individuals of interest populating a restrained list

1) FR in Video Surveillance

A Generic System for Video-Based FR



1) FR in Video Surveillance

Recognition Scenarios

- **ROIs extracted from video frames are matched against the *facial model* of individuals of interest**
- **Still-to-video recognition:** facial model of each individual are extracted from 1+ gallery of stills

  Typical application: screening against various watchlists
- **Video-to-video recognition:** facial model of each individual are extracted from 1+ video sequences

  Typical application: person re-identification (recognize and track an individual over a network of cameras)

1) FR in Video Surveillance

Taxonomy of Surveillance Setups

Type 0: cooperative biometric setup (for access control, eGate)

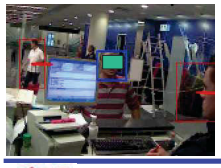
Type 1: semi-constrained setup (for primary inspection lane)

Type 2: unconstrained free-flow, one-at-time (CATSA chokepoint entry and other portals)

Type 3: unconstrained free-flow, many-at-time (airports, train stations and other indoor public spaces)

Type 4: no lighting or structural constraints (outdoor scenes)

1



2

3

1) FR in Video Surveillance

Challenges

- **Complex and changing environments :**
  - low quality and resolution of video frames
  - limited control of acquisition conditions – variation in poses, expressions, illumination, scale, blur, occlusion…
  - ageing and variation of interaction between individual–sensor
  - facial models: poor representatives of real faces because they are designed during enrollment with limited reference data
  - imbalanced data distributions: very few positives (from individuals of interest) w.r.t. negatives (from open world)

1) FR in Video Surveillance

Challenges

- **Computational resources:** video surveillance networks are comprised of a growing number of IP-based cameras
  - **transmit or archive massive quantities of data**
  - **memory requirements**: storage and retrieval of facial models
  - **processing time**: face detection, and matching ROIs against facial models

OVERVIEW

1) Background – FR in Video Surveillance

2) Academic and Commercial Solutions

3) Evaluation Methodologies for FRiVS

4) Case Studies – Screening and Fusion

5) In-House Evaluations of Technologies

6) Conclusions and Recommendations

3) Evaluation Methodologies

- **Objective: benchmark state-of-the-art commercial technologies and academic systems for FRiVS:**
  - public **data sets** for medium- to large-scale evaluation
  - **performance measures**:
    - transaction- subject-based analysis
    - time analysis over tracks
  - experimental **protocols** for different types of surveillance applications, e.g.,
    - screening of faces against watchlist list
    - matching a face across several video feeds
    - fusion of face recognition from different cameras
  - still-to-video and video-to-video **recognition scenarios**

3) Evaluation Methodology

Public Data Sets for FRiVS

DATASET	TARGET APPLICATIONS
CMU MOBO: [GRO01] Carnegie Mellon University Motion of Bodies	subjects performing different walking patterns on a treadmill
CMU FIA: [GOH05] Carnegie Mellon University Faces in Action	subjects mimicking passport checkpoint at airport
Chokepoint [WON11]	video-surveillance subjects walking though portals
MOBIO: [MCC10] EC FP7 Mobile Biometry	m-modal unconstrained authentication on mobile device
ND-Q0-Flip: [BAR11] Notre-Dame Crowd Data	detection of questionable observers that appear often in crowd videos
NIST-MBGC: [PHI09] National Institute of Standards and Technology - Multiple Biometric Grand Challenge	m-modal verification of subjects walking through portal or access control checkpoint (still- and video-to-video)
NRC-IIT: [GOR05] National Research Council – Institute for Information Technology	user identification for secured computer login
XM2VTS: [MAT03] Multi-Modal Verification for Teleservices and Security Applications	multi-modal verification for tele-service and security

3) Evaluation Methodologies

Public Data Sets for FRiV - summary

- **datasets have been characterized according to:**
  - **demographics:** distribution of individuals per session and in the entire dataset;
  - **complexity in scene:** the systematic variations of illumination, motion, occlusion, expression and/or pose for some target application;
  - **capture properties:** the number and type of cameras, duration of video sequences, frame rate and resolution**.**

3) Evaluation Methodology

CMU – FIA (mono-modal, 1 face)

- **PIL:** subjects mimicking passport checkpoint at airport

3) Evaluation Methodology

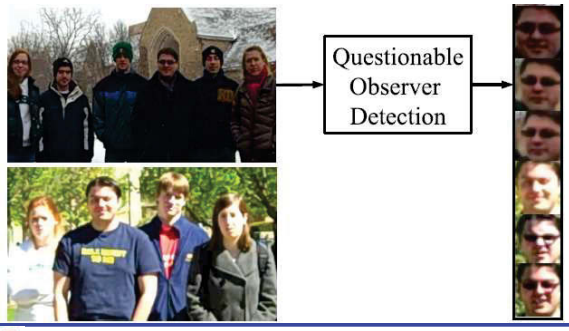
Chokepoint (mono-modal, 1 to 24 faces)

- **CATSA checkpoint:** subjects walking though portals

3) Evaluation Methodology

ND-Q0-Flip (mono-modal, 4 to 12 faces)

- **detection of questionable observers in crowded scenes**

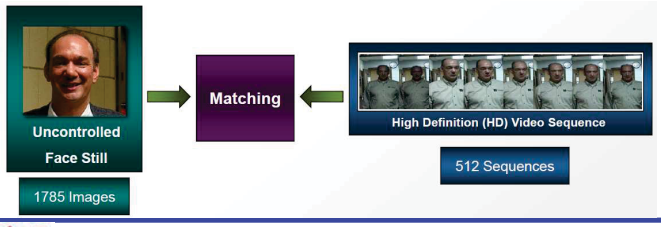


3) Evaluation Methodology

NIST-MBGC (multi-modal, 1 person)

- **portal or access control checkpoint:** unconstrained authentication from face and iris (still- and video-to-video)

http://www.nist.gov/itl/iad/ig/mbgc.cfm



3) Evaluation Methodology

MOBIO (multi-modal, 1 person)

- **PIL:** unconstrained authentication from face & voice

http://www.idiap.ch/dataset/mobio

3) Evaluation Methodology

Laboratory mock up data

- **Watchlist:** 60 individuals from CBSA wanted list + 6 persons from CBSA-VSB group

3) Evaluation Methodology

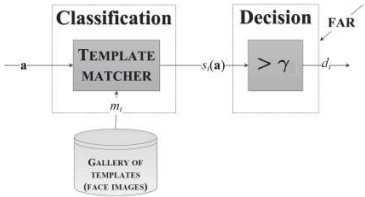
Evaluation of mono-modal scenarios

Evaluation Scenario	CBSA application	Data set
Type 0: **cooperative biometric setup**	access control, eGate	N/A
Type 1: **semi-constr. setup, one-at-time**	primary inspection lane	CMU-FIA
Type 2: **unconstr. free-flow, one-at-time**	CATSA checkpoint entry (portal)	Chokepoint (1 person)
Type 3: **unconstr. free-flow, many-at-time**	indoor airport setting	Chokepoint, ND-Q0-Flip (n persons)
Type 4: **no lighting or structural constr.**	outdoor setting	N/A

3) Evaluation Methodologies

Performance metrics for FRiVS

- **Fundamental task under evaluation:**
  - user-specific analysis – detection of an individual of interest among a restrained list of individuals
  - ability to accurately and efficiently detect the presence of an individual's face under various operational conditions



3) Evaluation Methodologies

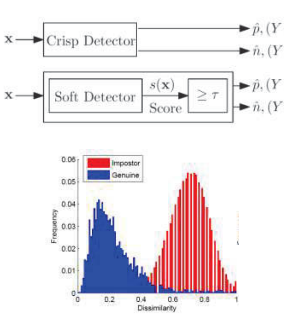
Performance metrics for FRiVS

- **Open-set FR problem with imbalanced class distributions (majority of individual seen are not of interest)**
  - decision spaces to analyze performance of 'positives'
- **Complex environments and ill-defined of facial models**
  - quality of facial captures (ROIs) and tracks
- **Performance varies across a population of individuals, and some individuals are harder to recognize**
  - menageries – statistical tests to characterize individual
- **Growing complexity of surveillance networks**
  - analysis of time and memory complexity

3) Evaluation Methodologies

Metrics – Transaction-Based Analysis

- **1:1 classification (detection) systems:**



Predicted Class \ True Class	p	n
$\hat{p}$	True Positive (TP) Correct detection	False Positive (FP) Type I error
$\hat{n}$	False Negative (FN) Type II error	True Negative (TN) Correct rejection
	$P = TP + FN$	$N = FP + TN$

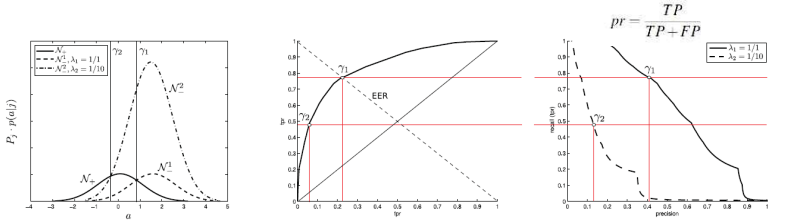
$$\Pr(\hat{p} \mid p) = P_{1|1} \approx tpr = \frac{TP}{P} = \frac{TP}{TP+FN}$$
$$\Pr(\hat{p} \mid n) = P_{1|0} \approx fpr = \frac{FP}{N} = \frac{FP}{FP+TN}$$
$$\Pr(\hat{n} \mid n) = P_{0|0} \approx tnr = \frac{TN}{N} = \frac{TN}{FP+TN}$$
$$\Pr(\hat{n} \mid p) = P_{0|1} \approx fnr = \frac{FN}{P} = \frac{FN}{TP+FN}$$

$$acc = \frac{TP+TN}{P+N}$$
$$LR^{+} = \frac{tpr}{fpr}$$
$$LR^{-} = \frac{fnr}{tnr}$$

3) Evaluation Methodologies

Metrics – Transaction-Based Analysis

- **Evaluation of detectors** – count correct and incorrect decisions over a test set, and express performance trade-offs using…
  - **Traditional:** ROC or DET curves (scalar metric: accuracy, AUC, pAUC)
  - **With imbalances classes:** precision-recall (scalar metric: F-score)



3) Evaluation Methodologies

Metrics – Transaction-Based Analysis

- **P-R curve:** with imbalanced class distributions (very few positive detections from a restrained cohort)
  - unbiased towards the majority (negative) class as skew grows
  - measures *the proportion of correctly predicted positive ROIs* out of the total number of ROIs predicted as belonging to an individual of interest
  - scalar metric that combines ***pr*** and ***tpr*** :

$$F_{\beta} = (\beta^2 + 1)\frac{tpr \cdot pr}{\beta^2 \cdot pr + tpr}$$

3) Evaluation Methodologies

Metrics – Subject-Based Analysis:

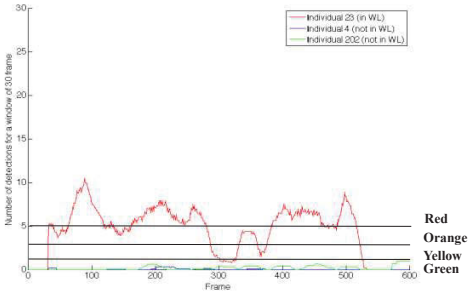
- **Doddington's zoo** – performance is assessed with different types of individuals in mind
  - performance of face recognition systems may vary drastically from one individual to the next
  - an analysis of these individuals and their common properties can:
    - expose fundamental weaknesses in a FR system
    - schemes for user-specific thresholds, score normalization and fusion

positives

3) Evaluation Methodologies

Metrics – Time analysis

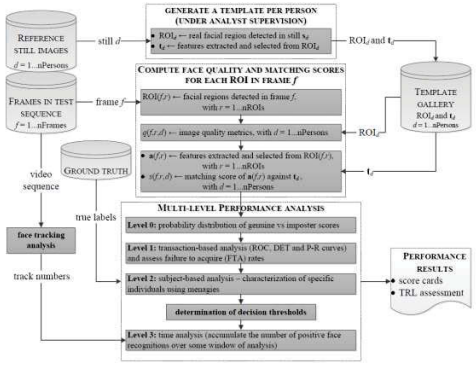
- group ROIs corresponding to high confidence tracks
- accumulate positive predictions over fixed time window:



3) Evaluation Methodologies

Generic protocol: still-to-video recognition

- type 1 and 2 application scenarios using the Chokepoint



OVERVIEW

5) In-House Evaluations

Still-to-Video Results with Chokepoint

- **Example of transaction-based analysis:**

Table 10.2: Summary of transaction-based (level 1) analysis for 20 pixels between the eyes.

Product	Measure	Ind01	Ind04	Ind05	Ind07	Ind09	Ind10	Ind11	Ind12	Ind16	Ind29	AVG	STD
Cognitec	*fpr*	4.30%	3.77%	4.05%	3.84%	5.14%	3.81%	3.73%	5.43%	3.34%	3.10%	4.05%	0.007
	tpr	75.00%	47.37%	68.89%	70.49%	71.05%	62.00%	75.00%	95.56%	43.24%	97.67%	70.63%	0.166
	prec	39.76%	29.03%	39.74%	49.43%	31.03%	41.33%	47.56%	40.57%	29.09%	53.85%	40.14%	0.081
	F1	0.520	0.360	0.504	0.581	0.432	0.496	0.582	0.570	0.348	0.694	0.509	0.101
	AUC	0.944	0.908	0.936	0.946	0.944	0.941	0.951	0.994	0.945	0.997	0.951	0.025
	$AUC_{0.05}$	0.719	0.443	0.589	0.636	0.567	0.549	0.686	0.885	0.414	0.953	0.644	0.165
PittPatt	*fpr*	4.01%	3.43%	0.62%	4.48%	1.68%	6.04%	3.30%	11.00%	2.21%	1.75%	3.85%	0.028
	tpr	86.27%	72.22%	91.67%	87.50%	92.11%	48.94%	21.15%	100.00%	84.21%	89.29%	77.34%	0.230
	prec	49.44%	40.00%	86.27%	49.49%	64.81%	25.27%	22.92%	26.63%	56.14%	55.56%	47.65%	0.188
	F1	0.629	0.515	0.889	0.632	0.761	0.333	0.220	0.421	0.674	0.685	0.576	0.193
	AUC	0.956	0.852	0.968	0.946	0.985	0.725	0.600	0.997	0.916	0.946	0.889	0.123
	$AUC_{0.05}$	0.852	0.613	0.929	0.796	0.945	0.407	0.184	0.948	0.762	0.884	0.732	0.244
Neurotech.	*fpr*	3.09%	1.80%	13.41%	4.89%	1.35%	2.26%	4.90%	4.76%	1.60%	2.71%	4.08%	0.034
	tpr	25.00%	66.67%	53.85%	45.45%	25.00%	18.18%	50.00%	41.67%	6.67%	40.00%	37.25%	0.173
	prec	38.10%	42.86%	10.61%	50.00%	25.00%	16.67%	36.36%	19.23%	12.50%	25.00%	27.63%	0.128
	F1	0.302	0.522	0.177	0.476	0.250	0.174	0.421	0.263	0.087	0.308	0.298	0.132
	AUC	0.726	0.978	0.882	0.905	0.866	0.779	0.854	0.805	0.645	0.808	0.825	0.090
	$AUC_{0.05}$	0.206	0.757	0.152	0.402	0.626	0.238	0.400	0.413	0.114	0.356	0.366	0.194

5) In-House Evaluations

Still-to-Video Results with Chokepoint

- **Example of time-based analysis:**

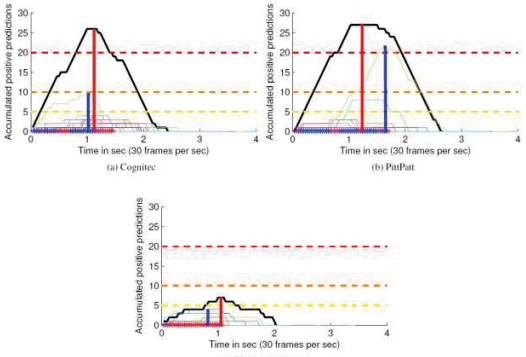


Figure 10.1: Example of time-based analysis (level 3) for target individual 1 and 20 pixels between the eyes.

5) In-House Evaluations

Still-to-Video Results with Chokepoint

- **Example of time-based analysis:**

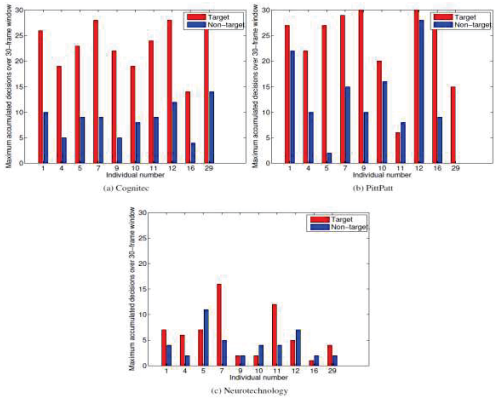


Figure 10.2: Summary of time-based analysis (level 3) for 20 pixels between the eyes.

OVERVIEW

6) Conclusions and Recommendations

Technology Readiness Assessment

- **Five-level scale used in this report:**

Symbol	Years to deploy	TRL	Additional Applied R&D requirement
++	0 (can be deployed immediately by any operational agency with no R&D capacity)	TRL=8-9, complete COTS system deployed and proved useful by many users	no development effort is required to deploy it
+	<1 (by most operational agencies with minimal Applied R&D capability)	TRL=7-8, compete COTS system deployed somewhere	some minor development effort is required to fit business requirements
oo	1-2 (only by operational agencies that have substantial Applied R&D capability)	TRL=5-6, system validation in mock-up or pilot	solid development effort is required
o	2-3 (only by operational agencies that have access to major to Applied R&D)	TRL=4, component validation in relevant 24/7 environment	major development effort is required
-	>3 (not foreseeable for deployment in near future)	TRL=1-3	significant academic / industry R&D required

6) Conclusions and Recommendations

Technology Readiness Assessment

- **A preliminary TRL assessment according to 5 levels.**

FRiV technology	Type 0 (eGates)	Type 1 (PIL)	Type 2 (Portal)	Type 3 (halls)
Face tracking (in consecutive frames)	+	+	+	-
Face-person matching (across multiple feeds)	+	+	+	-
Face Detection	++	++	+	o
Face Grouping / Tagging	+	oo	o	-
Face Fusion (from multiple frames and cameras)	+	oo	o	-
Video-to-video face matching	+	oo	o	-
Visual Analytics tools (post-event search/retrieval)	oo	oo	oo	o
Instant "Watch List" Screening: Binary	oo	oo	-	-
Instant "Watch List" Screening: Triaging	oo	oo	o	-
Post-event forensic examination from snapshots	++	++	+	o
Face expression analysis	+	o	o	-
Face to improve Voice / Iris Biometrics	+	o	-	-
Soft biometrics (eg. height)	o	o	o	-
Gender / Age / Race recognition	o	o	o	-

6) Conclusions and Recommendations

Recommendations

- **Current COTS and Academic products can be found useful for many FRiV applications, but not for all of them!**
- **Post-processing and pre-processing (inc. Video Analytics) are critical for success**
- **Potential for new video-based (eg Biological Vision driven) techniques, as opposed to status-quo still-image-based.**
- **There's no all-inclusive evaluation methodology for FRiV**
  - **FMR/FNMR metric can be misleading**
  - **For operational agency, TRL-based evaluation should prevail**
- **Ultimate metric - satisfaction of the end-user Border Officer!**

6) Conclusions and Recommendations

Recommendations

- **State-of-the-art commercial that implement core FRiVS functions – face detection, grouping, matching and tracking.**
- **They cannot by themselves perform automated FR with a high level of performance in semi- or unconstrained environments:**
  - difficulties capturing high quality ROIs (typically poor quality and low resolution),
  - complex environments, that change during operations,
  - face models are designed a priori using limited number of reference samples.

6) Conclusions and Recommendations

Recommendations

- **For robust and accurate performance in real-word environments:** incorporate the proven academic techniques within state-of-the-art commercial technologies, in particular:
  - modular and ensemble-based classification architectures
  - fusions of multiple sources over different templates and frames, an array of cameras, etc.
  - exploit soft biometric traits and contextual information
  - adaptive biometric to refine facial models over time
  - spatio-temporal recognition – exploit face-person tracking to accurately recognize by accumulating evidence