In Proceedings of IEEE SSCI Workshop on Computational Intelligence in
Biometrics and Identity Management (CIBIM), December 2014, Orlando, FL.

Target-based evaluation of face recognition technology for video surveillance applications

Dmitry Gorodnichy and Eric Granger

Abstract—**This paper concerns the problem of real-time watch-list screening (WLS) using face recognition (FR) technology. The risk of flagging innocent travellers can be very high when deploying a FR system for WLS since: (i) faces captured in surveillance video vary considerably due to pose, expression, illumination, and camera inter-operability; (ii) reference images of targets in a watch-list are typically of limited quality or quantity; (iii) the performance of FR systems may vary significantly from one individual to another (according to so-called "biometric menagerie" phenomenon); (iv) the number of travellers drastically exceeds the number of target people in a watch-list; and finally and most critically, (v) due to the nature of optics, images of faces captured by video-surveillance cameras are focused and sharp only over a very short period of time if ever at all. Existing evaluation frameworks were originally developed for spatial face identification from still images, and do not allow one to properly examine the suitability of the FR technology for WLS with respect to the above listed risk factors intrinsically present in any video surveillance application. This paper introduces the target-based multi-level FR performance evaluation framework that is suitable for WLS. According to the framework, Level 0 (face detection analysis) deals with the system's ability to process low resolution faces. Level 1 (transaction-based analysis) deals with the ability to match faces in open-set problems, where target vs. non-target distributions are unbalanced. Level 2 (subject-based analysis) deals with robustness of the system to different types of target individuals. Finally, Level 3 (spatio-temporal analysis) allows one to examine the overall FR system discrimination by means of accumulating the recognition decision confidence over a face track, which can be used for developing more robust intelligent decision-making schemes including face triaging.The results from testing a commercial state-of-art COTS FR product on a public video data-set are shown to illustrate the benefits of this framework.**

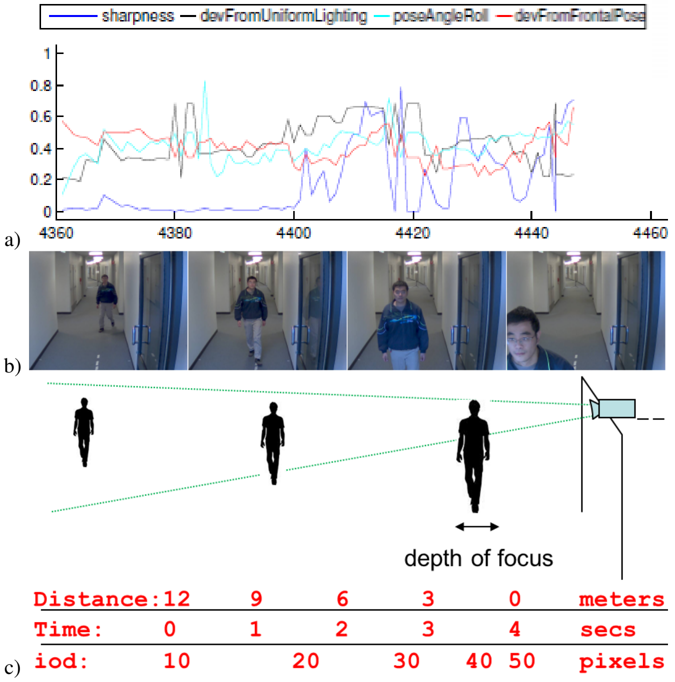


Fig. 1. Low resolution and variation in quality of faces captured by a surveillance camera - demonstrated using sequence P2L-S4-C1.1 from the Chokepoint data-set (frames 4361-4447 corresponding to individual with ID=1 are shown): a) face quality metrics computed by a commercial FR product: face "sharpness" and deviation from uniform lightning and frontal orientation, b) jpeg images extracted from video, c) inter-ocular distance (iod) in pixels vs. time observed and distance to the camera, overlaid on the pictorial representation of the WLS process.

I. Introduction

In watch-list screening recognition, faces observed by a surveillance camera are continuously matched against the faces in a watch-list database (see Figure 1) [1]-[8]. Because the number of travellers is significantly larger than the WLS of criminals[1] in a watch-list database and because of the real-time constraint, which may not permit manual adjudication of the recognition results by a human analyst, the risk of erroneously flagging an innocent traveller due to a false match can be very high.

The problem is further aggravated by the nature of optics[2]. Faces captured by the video-surveillance cameras are in focus only in a small range of about 1-2 feet, or otherwise they are very small (if captured at distance) or blurred (if the range of focus is manually increased by decreasing the camera aperture or shutter speed). This is illustrated in Figure 1, which shows frames from the Chokepoint data-set [8] that simulates an airport chokepoint environment[3], in which an individual is going through a portal observed by an industry standard

Dmitry Gorodnichy is a Senior Research Scientist with Science & Engineering Directorate, Canadian Border Services Agency and an Adjunct Professor at School of Computer Science and Electrical Engineering, University of Ottawa. Eric Granger is a Professor of Systems Engineering with École de technologie supérieure, Université du Québec.

[1] For the simplicity of presentation, we use the airport scenario and the terms like "criminals" and "travellers" for target and non-target individuals. It is understood however that other scenarios for the use of WLS exist. On the complexity scale, the airport scenario presents the easiest possible setup for the WLS, as both lighting and travellers motion pattern can be partially controlled.

[2] http://www.exposureguide.com/focusing-basics.htm.

[3] The Chokepoint data-set can be downloaded for free at http://arma.sourceforge.net/chokepoint.

DOI: 10.1109/CIBIM.2014.7015451

SVGA (600x800 pixels) 30 fps surveillance camera. The face is observed by the camera for about 3 secs (87 frames), during which the resolution of the face changes from 14 pixels between the eyes (inter-ocular distance iod=14), when the face is first detected, to 64 pixels, when the person passes under the camera. The camera is focused on a distance at which this face is captured with iod=40-50 pixels. This corresponds to 1/3 secs (nine frames: 4433-4441), which the person passes very quickly without looking into the camera.

If a WLS system is designed so that it only processes facial images that are in focus, then a chance of missing a target individual is very high. For example, this occurs if his/her face is not aligned with the camera field of view, or if it is of low quality due to blur or occlusion. On the other hand, if the system uses all facial images including those that are out-of-focus and small, then the risk of falsely matching non-target people increases. To minimize this risk, a FR system should be evaluated and implemented specifically with the WLS problem in mind. Most currently used methodologies for evaluation of FR systems were developed for face identification and verification from still images and do not provide sufficient means for evaluating and improving the performance of WLS systems with respect to this risk. This paper establishes an evaluation methodology to address this risk.

We appreciate the fact that facial images in surveillance video *are meant to be* of low resolution/quality and develop a target-based multi-level FR performance evaluation methodology is specifically tailored to the WLS problem allowing one to design and tune WLS systems with respect to all risk factors playing a role in a surveillance application. In addition to addressing the low face resolution/quality condition of WLS, our methodology also allows one to approach WLS as combination of independent target detection problems [7], which has the advantage in that 1) each detector can be assessed and tuned separately, 2) it accounts for data imbalance, 3) it makes use of the temporal information to improve the performance, and 4) it allows one to measure the performance of a complete system over time, which can be used for developing more robust intelligent decision-making schemes including face triaging.

The paper is organized as follows. Next section provides general considerations related to designing FR systems for video surveillance applications, which includes the development of the video surveillance scenario taxonomy that can be used to facilitate the evaluation of FR systems using public data-sets. The concept of the target-based WLS is introduced in Section III followed by the description of key stages of the target-based evaluation and implementation of WLS systems in Section IV. Multi-level evaluation methodology is described in Section V, illustrated by the results from testing using a commercial FR product. Insights gained from the obtained results conclude the paper.

II. General considerations

Prior to deploying a FR system in a video surveillance application, a FR user/developer needs to have a knowledge of which FR tasks are feasible in which video surveillance scenarios. Such knowledge[4] was obtained in the PROVE-IT(FRiV) study [12]-[15] and is briefly summarized below.

A. Taxonomy of video surveillance scenarios

In evaluation of technologies for video surveillance applications, it is proposed to categorize all possible video surveillance scenarios according to "who-what-where" factor triangle as shown in Table I. The "where" factors relate to the settings in which subjects are captured; they include illumination, camera position and are normally possible to control. The "what" factors relate to the procedure imposed on subject during the capture; they include the direction, diversity of subject motion and can be partially controlled. Finally, the "who" factors relate to the subjects being captured; they include person's orientation, expression and normally cannot be controlled, unless the subject cooperates with the capture as is done at eGates in Automated Border Control applications.

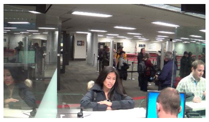
Based on this categorization of factors, five basic types of video surveillance scenario types of increasing complexity are recognized, camera positioning and quality being assumed the best technically possible in each scenario:

Type 0 ("eGate"), cooperative scenario in automated border control.
Type 1 ("Kiosk"), as at passport control or biometric kiosk.
Type 2 ("Portal"), as in a one-way corridor or choke-point portal.
Type 3 ("Halls"), as in airport halls.
Type 4 ("Outdoors"), all other scenarios.

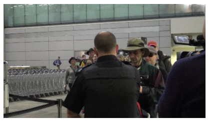
Types 1-3 present three typical scenarios of increasing complexity possible in airport. The images from an operational airport surveillance cameras[5] corresponding to those types are shown in the figure under Table I. Unless a WLS solution is not proven successful in an easier scenario, it should not be contemplated for the harder one.

TABLE I. Taxonomy of in video surveillance scenarios .

Type	"Who" factors (person)	"What" factor (activity)	"Where" factors (setup)
0: eGate	controlled	controlled	controlled
1: Kiosk	semi-controlled	controlled	controlled
2: Portal	uncontrolled	semi-uncontrolled	controlled
3: Hall	uncontrolled	uncontrolled	controlled
4: Outdoor	uncontrolled	uncontrolled	uncontrolled


Type 1

Type 2


Type 3

B. Data-sets

There are several public video data-sets that simulate the defined above video surveillance types and which can be used for evaluation purposes. Of particular relevance are the "Faces

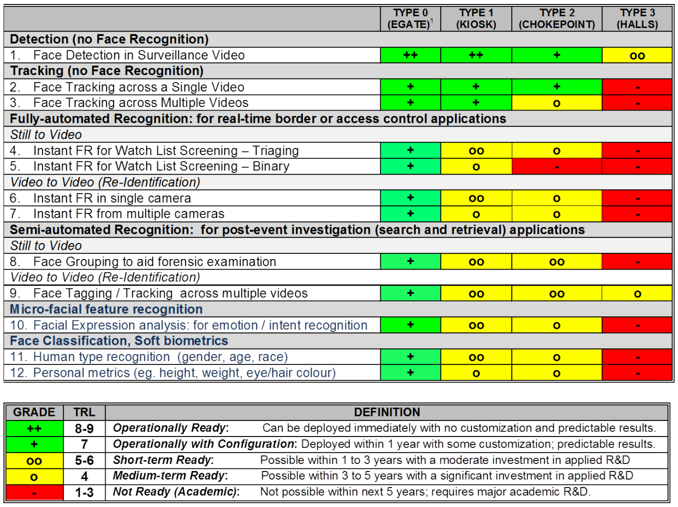
[4] Similar knowledge related to the feasibility of video analytics tasks, including people detection and tracking, was obtained in the PROVE-IT(VA) study and is reported in [17].

[5] The images are taken from the *CBSA Airport* Dataset created by the CBSA Science and Engineering Directorate.

in Action" data-set [9], which is similar to Type 1 scenario, and the Chokepoint and S2V data-sets [8], [10], which simulate Type 2 scenario. It is vital for WLS developers to examine the performance of the FR system on these data-sets prior to testing in real surveillance settings. By doing so they can expose in advance the vulnerabilities of the WLS system and develop the system that deals with those vulnerabilities. At the same time, it should also be noted that public data-sets provide an "optimistic" level of the video surveillance quality, as they do not show artifacts due to bandwidth and motion compression, which are commonly present in operational CCTV systems.

At the moment there is a limited number of public data-sets that simulate real surveillance settings. Following the described taxonomy of the video surveillance setups more public data-sets can be created, further sub-categorized if needed, for example, by density of traffic, camera resolution, or image compressions. Of special value will be the data-sets that are obtained from real life operational surveillance cameras.

TABLE II. ASSESSMENT OF FR TECHNOLOGY READINESS FOR VIDEO SURVEILLANCE APPLICATIONS (FROM [12]).



	TYPE 0 (EGATE)[1]	TYPE 1 (KIOSK)	TYPE 2 (CHOKEPOINT)	TYPE 3 (HALLS)
Detection (no Face Recognition)				
1. Face Detection in Surveillance Video	++	++	+	oo
Tracking (no Face Recognition)				
2. Face Tracking across a Single Video	+	+	+	-
3. Face Tracking across Multiple Videos	+	+	o	-
Fully-automated Recognition: for real-time border or access control applications				
Still to Video				
4. Instant FR for Watch List Screening – Triaging	+	oo	o	-
5. Instant FR for Watch List Screening – Binary	+	o	-	-
Video to Video (Re-Identification)				
6. Instant FR in single camera	+	oo	o	-
7. Instant FR from multiple cameras	+	o	o	-
Semi-automated Recognition: for post-event investigation (search and retrieval) applications				
Still to Video				
8. Face Grouping to aid forensic examination	+	oo	oo	-
Video to Video (Re-Identification)				
9. Face Tagging / Tracking across multiple videos	+	oo	oo	o
Micro-facial feature recognition				
10. Facial Expression analysis: for emotion / intent recognition	+	oo	o	-
Face Classification, Soft biometrics				
11. Human type recognition (gender, age, race)	+	oo	o	-
12. Personal metrics (eg. height, weight, eye/hair colour)	+	o	o	-

GRADE	TRL	DEFINITION
++	8-9	***Operationally Ready***: Can be deployed immediately with no customization and predictable results.
+	7	***Operationally with Configuration***: Deployed within 1 year with some customization; predictable results.
oo	5-6	***Short-term Ready***: Possible within 1 to 3 years with a moderate investment in applied R&D
o	4	***Medium-term Ready***: Possible within 3 to 5 years with a significant investment in applied R&D
-	1-3	***Not Ready (Academic)***: Not possible within next 5 years; requires major academic R&D.

C. *Taxonomy of FR tasks and technology readiness assessment*

FR tasks, which can be potentially executed in video surveillance applications, are categorized from easiest to hardest, as follows:

- by level of performed face processing: face detection, tracking, recognition, classification, facial expression analysis;
- by mode of operation: real-time vs. post-event operation;
- by decision making mode: automated (binary or triaging) vs. semi-automated (as part of an analytic tool or filter);
- by data modality: still-to-video vs. video-to-video.

Based on the in-house evaluations and literature reviews [13], [16], the feasibility of each FR task is accessed for each video surveillance type as shown in Table II.

According to these assessment results, the development of a fully-automated still-to-video instant FR system for WLS presents a higher challenge compared to the development of a semi-automated post-event investigation system. It is feasible for a simple surveillance scenario such as Type 1 (i.e. at the kiosk or passport control, where a person does not move and is in focus for a period of time) and possibly at the Type 2 scenario (a person walking through a choke-point), which is examined in this paper.

Technology readiness assessment (Table II) provides the basis for developing recommendations related to the deployment and further research and development of the FR technology for video surveillance applications.

D. *Selection of FR products for video surveillance application*

A survey of academic solutions and commercial products for face recognition in video surveillance applications is presented in [13], [14]. A critical functionality of these FR product is the ability to process facial images in low resolution and quality. In particular, based on our analysis of face resolutions in operational surveillance cameras (presented in Introduction), it is important that FR solution can detect faces with i.o.d less than 60 pixels and also that it can track the detected faces over time. If a FR product does not provide a face tracking functionality, this functionality can be developed by the integrator using video analytic techniques such as those based on tracking body / head motion and soft biometrics like cloths colour or texture.

III. TARGET-BASED DESIGN FOR WLS

Given a set of one or more still reference images of a target individual, a WLS system seeks to raise an alarm when this individual is detected in a video stream. Two FR methodologies are possible for designing a solution to this problem (see Figure 2): *Cohort-based* (CB) and *Target-based* (TB). In CB design, which evolved from 1-to-N identification paradigm in close-set environments and which is used in most cases for WLS by industry and academia [1], [2], [3], each region of interest (ROI) captured by a face detector is compared on the same basis to *all* faces[6] in the watch-list gallery. In contrast to that, the TB design evolved from the target recognition paradigm in open-set environments, where each target image in the watch-list is processed independently and is independently compared to all ROIs detected in video, while simultaneously tracking all detected ROIs.

The concept of target-based design for WLS relates to the concept of spatial-temporal open-set recognition (vs. spatial close-set recognition) [1], [6], [5]. The introduction of this concept however allows for evaluation with a target-based framework, which is particular suited for the implementation of an WLS solution for the following reasons:

- Biological system (humans) use TB recognition. We track a person in a crowd until finally deciding whether he or she is the one we are looking for. We do not match every persons we see to everyone we know.

- As a consequence from above, automated decision obtained

[6] In a general still-to-video FR system, a gallery may contain face models designed with one or more reference still images per individual.

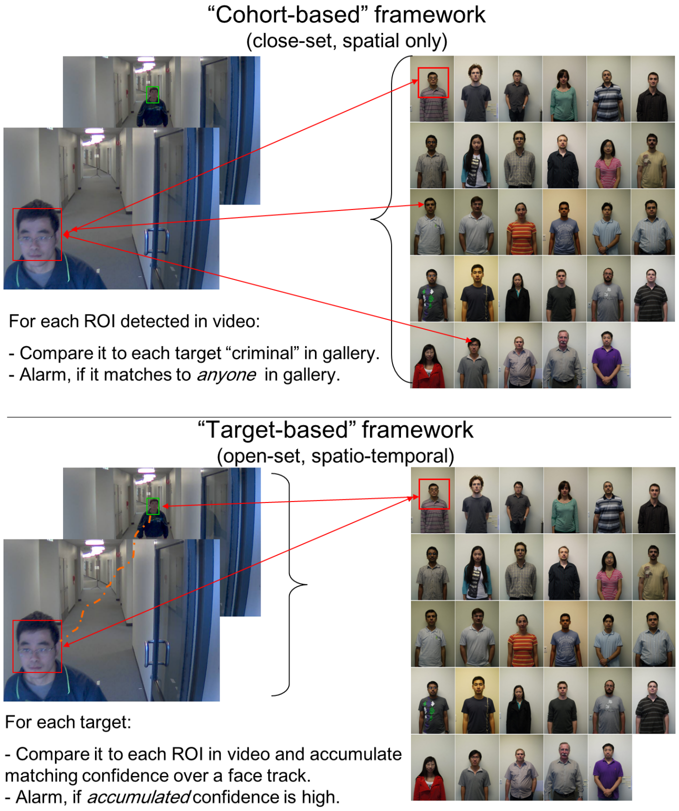

"Cohort-based" framework
(close-set, spatial only)
For each ROI detected in video:
- Compare it to each target "criminal" in gallery.
- Alarm, if it matches to *anyone* in gallery.
"Target-based" framework
(open-set, spatio-temporal)
For each target:
- Compare it to each ROI in video and accumulate matching confidence over a face track.
- Alarm, if *accumulated* confidence is high.

Fig. 2. Cohort-based and target-based methodologies for the WLS problem, illustrated using the images from the Chokepoint dataset.

by TB design can be used in a combination with manual observation by a human analyst.
- In TB design, scores are assigned to a person, as he or she is tracked, and can be fused / updated continuously as more data about him/her are observed. As a result, a level of confidence or risk can be associated with each observed traveller, which is inline with other traveller risk screening techniques used in airport for for border control [19].
- In TB design, target recognition system can be tuned specifically for each target, rather than using the same system parameters (such as score or image quality thresholds) for every target.
- TB design allows one to add more target specific details, which could be both biometric (such as different images at different resolutions) and non-biometric (such as video-analytic / soft-biometric data and general intelligence data coming from other sources).
- TB design is scalable to the number of travellers in video. The same number of processes, or classifiers, equal to the number of targets in a watch-list, are used, regardless of the density of the traffic.
- TB design is also useful for other video-based FR applications, such as: person re-identification (tracking across multiple cameras), video summarization, evidence search and retrieval and others listed tin Table II.

In the next section, we show how target-based design can benefit in the implemention of the WLS systems.

IV. Implementation of the target-based WLS

As summarized in Figure 3, implementation of a target-based WLS system is performed *independently* for each target and consists of the training and testing stages. During the training stage, the parameters of the FR recognition system (for instance, the individual-speciific decision threshold) are tuned so that to maximize the likelihood of recognizing the target at a given acceptable False Match Rate (FMR). This is done based on all information available about the target, which can be biometric (i.e. facial images of the subject) and non-biometrics (e.g. what subject is wearing and whom is s/he is travelling with). Using still facial image(s) of the subject and a subset of video data with and without the subject, the genuine and impostor score distributions[7] are computed by matching the subject to all faces detected in video. Then, based on the measured score distributions, the determination of whether a subject belongs to a harder biometric recognition case is done (so called Doggington Zoo classification: "sheeps" vs "goats" [20]) and the operational threshold corresponding to the selected FMR is obtained for various face resolutions (iod $>$10, 20, 30, 40, and 50 pixels). Other tunable parameters of the system include the lower limit on resolution and quality of captured faces to be accepted by the FR matcher.

Once the operational thresholds are set, the system performance is evaluated during the testing stage using the multi-level performance evaluation analysis, where each level of analysis addresses a particular risk factor present in the WLS problem:

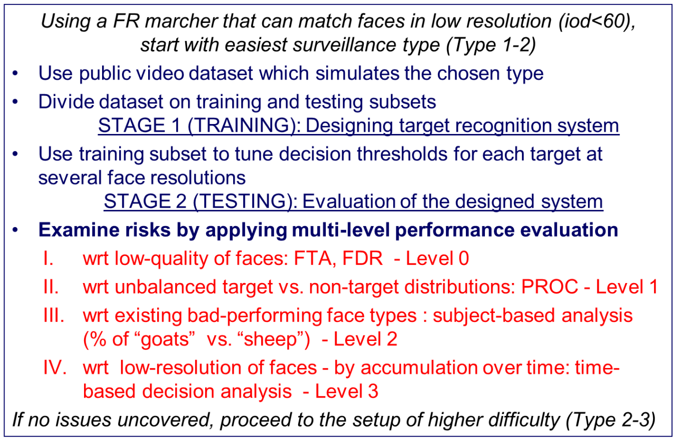
- Level 0 (face detection analysis) addresses the risk related to not being able to capture low resolution / quality images;
- Level 1 (transaction-based analysis) addresses the risk of dealing with a large open-set data, where the likelihood of seeing a criminal is very small;
- Level 2 (subject-based analysis) addresses the risk related to the robustness of the system with respect to all subjects in the watch-list, as some of them can be much harder to recognize than the others;
- Level 3 (time-based analysis) deals with the temporal information present in video and provides means to using this information for developing more robust intelligent recognition decision schemes.

Such testing/training procedure should be done on easier (Type 1 and 2) scenarios first, prior to performing it on the setups of higher difficulty (Type 2 and 3 setups). Public datasets corresponding to those setups can be used.

V. Multi-level performance analysis

The multi-level framework for the evaluation of FR systems in video-surveillance was prompted by the multi-order biometric score analysis originally proposed in [18] to deal with the risks of non-confident matches in biometrics-enabled automated border control systems. In the following, this framework

[7] In the absence of video data with a subject, such data can be simulated by using the available still images.



Using a FR marcher that can match faces in low resolution (iod<60), start with easiest surveillance type (Type 1-2)

- Use public video dataset which simulates the chosen type
- Divide dataset on training and testing subsets

STAGE 1 (TRAINING): Designing target recognition system

- Use training subset to tune decision thresholds for each target at several face resolutions

STAGE 2 (TESTING): Evaluation of the designed system

- **Examine risks by applying multi-level performance evaluation**

I. wrt low-quality of faces: FTA, FDR - Level 0
II. wrt unbalanced target vs. non-target distributions: PROC - Level 1
III. wrt existing bad-performing face types : subject-based analysis (% of "goats" vs. "sheep") - Level 2
IV. wrt low-resolution of faces - by accumulation over time: time-based decision analysis - Level 3

If no issues uncovered, proceed to the setup of higher difficulty (Type 2-3)

Fig. 3. Key stages of target-based WLS implementation. System parameters are target-specific and tuned independently for each target.

is described using the results from testing a commercial FR product on the Chokepoint dataset.

Commercial product: In our experiments, the results are obtained by using a highly acclaimed commercial FR software, the name of which is not disclosed, but which is known to be successfully used in other FR applications such as forensic investigation and automated border control with eGates. It is emphasized however that the objective of our work is not to evaluate any particular FR product, but rather to illustrate the applicability of the evaluation methodology for exposing the vulnerabilities of FR technologies deployed in video surveillance applications. It is understood that the comparative performance evaluation of FR products is reported elsewhere (e.g. by NIST FRVT 2013) and can be used, if needed to extrapolate the results obtained in this paper to other FR products.

The Chokepoint data-set: For the rest of the presentation, the results are presented from evaluating a commercial FR product on Type 2 video surveillance scenario using the Chokepoint data-set. As mentioned above, this data-set is easily accessible and provides an easy of way of validating the applicability of the FR solution for a WLS problem. It contains video sequences with 29 persons walking (one a time and many a time) through several indoor portals captured by three industry standard 30 fps SVGA resolution (800X600 pixels) IP cameras. High quality still photographs of these 29 persons are provided and can be used as watch-list images. There are in total 54 video sequences, 1-3 mins each, stored as collections of jpeg images (30 images per second). Face regions in each sequence are manually labeled, with a total of 64,204 labeled face images. One of the video sequences showing an individual with ID=1 is shown in Figures 1 and 2. The still images of all 29 persons are shown in Figure 2.

For experiments presented below, ten individuals are randomly selected from the data-set and included in the watch-list as target "criminals" – individuals with ID 1,4,5,7,9,10,11,12,16,29 (six males and four females seen in Figure 2). For each of them, the entire video sequence is played, within which all other 28 individuals play the roles of "regular travellers".

During the training stage, a target-based WLS system (target classifier) is constructed using a subset of video sequences designated as a training set. By running a target detector on a training set, the operational thresholds are selected for several face resolutions (iod=10,20,30,40,50) so that to achieve the FMR = 5%. During the testing stage, each constructed target classifier (with individual operational thresholds for each face resolution) is tested on another subset of video sequences designated as a testing set.

The experiments were conducted with all video-sequences of the Chokepoint data-set. In the following, for the purpose of presenting the evaluation methodology and the key insights gained from it, the testing results obtained for the target classifier for individual ID=1 on sequence P2L-S4-C1.1 (shown in Figure 1) are presented only. The results for all target individuals obtained on three different commercial products are provided in [16]. Sequence P2L-S1-C1.1 has been used to tune the operational thresholds of the target classifier.

A. Level 0: Face detection and quality analysis

At Level 0 no recognition performance metrics is measured. Instead Level 0 is used to explore the issues related to the detection of faces in low resolution/quality video, and the ability of the system to enroll and match them. The following metrics are computed: Face Detection Rate (FDR), Failure to Detect (FTD), Failure to Enroll (FTE) due to insufficient quality for different face resolutions. These are shown in Table III. The results for iod over 50 pixels highlight the fact that there are very few faces detected at such resolution, which makes it unpractical to use them for computing curves and averaged metrics.

Additionally, face quality metrics measured by the FR product are also recorded such as: face sharpness, face roll and pitch angles, illumination. These were shown in Figure 1 in the Introduction.

B. Level 1: Transaction-based analysis

Level 1 is the transaction-based analysis traditionally used in event detection systems. It provides an approximate (averaged) outlook of the performance of the system in terms of the false/true positive and negative alarm rates (fpr vs. tpr) and associated ROC and PROC curves that plot $precision = \frac{TP}{TP+FP}$ vs. $recall = \frac{TP}{TP+FN}$, and $fpr = \frac{FP}{TN+FP}$ vs. $tpr = \frac{TP}{TP+FN}$ respectively (FP, TP, FN, TN are the numbers of false/true positive and negatives alarms).

It is emphasized that, because WLS is an open-set problem with number of regular travellers significantly exceeding the number of people in the watch-list, PROC (Precision-Recall Operating Characteristic) curve provides more value for analysis as it allows one to incorporate the knowledge about the skew λ of target vs. non-target population using the $prec = \frac{TP}{TP+\lambda \cdot FP}$ formula [23] .

Table III shows ROC and PROC curves measured for target with ID=1, for three face resolutions: iod>10 (black), >20

TABLE III. LEVEL 0 AND LEVEL 1 ANALYSIS: A) FACE DETECTION AND BASIC FACE MATCHING METRICS, B) ROC AND PROC CURVES.

a)

ID#1.	iod>10	iod>20	iod>30	iod>50
Level 0: Face Detection results				
Target faces detected	74	44	30	8
Non-target faces detected	1632	1162	638	181
False detection rate (FDR)	30%	11%	18%	42%
Failure to Detect (FTD)	2%	30%	60%	~95%
Failure to Enroll (low quality)	6%	11 %	19%	43%
Level 1: Face Matching results				
Precision	~45%	39%	~65%	~32%
False Positive rate	5.1%	4.3%	4.2%	~4%
True Positive (Recall) rate	62%	75%	69%	~38%
Ops. threshold @ fpr=5%	0.138	0.131	0.129	0.114

b)

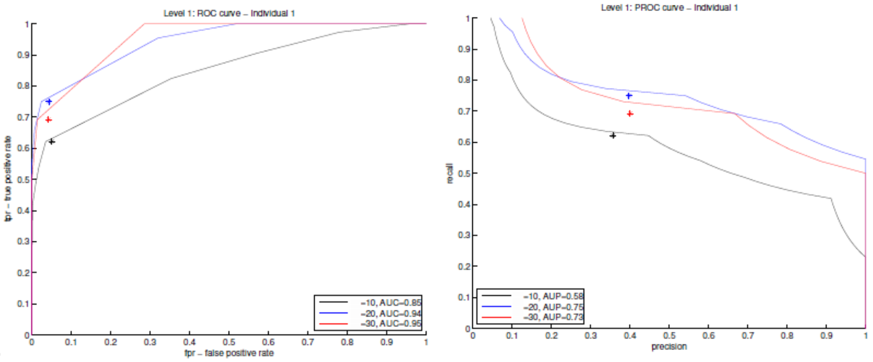


TABLE IV. LEVEL 1 ANALYSIS: SUMMARY OF TRANSACTION-BASED METRICS FOR EACH TARGET IN A WATCH-LIST (DATA FOR IOD>20 SHOWN).

Measure	Ind01	Ind04	Ind05	Ind07	Ind09	Ind10	Ind11	Ind12	Ind16	Ind29	AVG	STD
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

TABLE V. LEVEL 2 SUBJECT-BASED ANALYSIS FOR TARGET ID=1 AND ID=16.

a)

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%

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
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%
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%
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)

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.	4.05%	3.57%	0.00%	4.84%	11.48%	3.53%	3.95%	0.00%	8.47%	10.61%	4.05%	6.06%	2.22%	5.06%	0.00%
20 px.	2.27%	2.17%	0.00%	7.89%	6.67%	0.00%	1.64%	0.00%	5.26%	8.00%	5.77%	4.44%	1.54%	5.66%	0.00%
30 px.	0.00%	0.00%	0.00%	11.76%	4.17%	0.00%	0.00%	0.00%	10.53%	3.57%	6.67%	0.00%	0.00%	0.00%	0.00%

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
40.00%	1.54%	0.00%	4.17%	3.77%	0.00%	0.00%	6.56%	4.94%	7.79%	0.00%	1.37%	4.84%	5.17%	1.41%
43.24%	1.96%	0.00%	4.00%	2.38%	0.00%	0.00%	7.50%	3.51%	5.36%	0.00%	0.00%	2.00%	4.65%	0.00%
26.32%	3.57%	0.00%	0.00%	0.00%	0.00%	0.00%	9.52%	0.00%	0.00%	0.00%	0.00%	7.14%	7.69%	0.00%

(blue),>30 (red). The operating points (corresponding to the operational thresholds) obtained on a training sequence are shown as stars. Larger deviations between the points and the curves indicate large variation between the system performance on training and test data. Table IV shows more detailed summary of Level 1 analysis obtained for all ten target individuals in the watch-list.In addition to fpr, tpr and $precision$, the table shows F_1-measure and area under curves AUG and $AUG_{0.05}$ (part of AUG with $FMR < 0.05$).

C. Level 2: Subject-based analysis

Performance of the FR system may vary drastically from one person to the next. In Level 2 analysis, each individual enrolled to the system is categorized as one of four biometric types according to the Doddington Zoo subject-based analysis [20], [22]: (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, and (4) "lambs", positive class individuals that are easy to impersonate. 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 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 one to set user-specific or template-specific thresholds, score normalization, and user-specific fusion.

The traditional way to define the Doddington zoo's category of the subjects is through the classifier output scores for all tested samples [20]. For target-based systems with techniques that provide crisp decisions, the confusion matrix of individual accumulated decisions is used to categorize individuals based on the technique used in [5] as shown in Table VI. The results of Level 2 analysis are presented in Figure V, which shows the match rates measured for each person in the video sequence against the two target individuals (with ID=1 and 16), coloured according to the Doddington zoo's categorization: green - easy to recognize "sheep", yellow – hard to recognize "goat" subjects. It can be seen that one the two targets is much harder to recognize than the other.

TABLE VI. DECISION MATRIX FOR SUBJECT-BASED ANALYSIS.

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

D. Level 3: Spatio-temporal analysis

Systems for WLS use different algorithms and techniques to implement their functions, such as face detection, matching and tracking. In order to assess the applicability of the FR systems for WLS, it is important to observe the ability of the system to detect a person of interest globally over time, using all its functions. Besides, decisions taken by an operator take place in a time scale that is longer than a frame rate. For robust intelligent decisions, the number of positive matching predictions over a moving window of time for each ROIs that correspond to a high quality facial track is accumulated. This constitutes Level 3 analysis, the results which are shown in Figure 4.

Figures 4.a-b show the time-based analysis for target individual with ID=1 for faces with iod > 20 pixels. 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 seen in the figures, the system is able to achieve a high level of accumulated positive predictions (the black graph exceeded the red threshold). 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 face captured in the 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, chances are higher to accumulate positive predictions and increase the confidence in identifying a given target individual.

An important aspect of the time analysis is to illustrate the accumulative positive predictions for the non-target individuals, in addition to that of target individuals. These are shown in Figure 4.c. The figure provides a summary of time-based performance achieved by the system for iod > 20 pixels for the target and all (18) non-target individuals observed in a video sequence, measured in terms of the maximum level of accumulated positive predictions. A large difference between the maximum level of accumulated positive predictions of the target (red bars) and all non-target individuals (blue bars) 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 subject-specific matching thresholds.

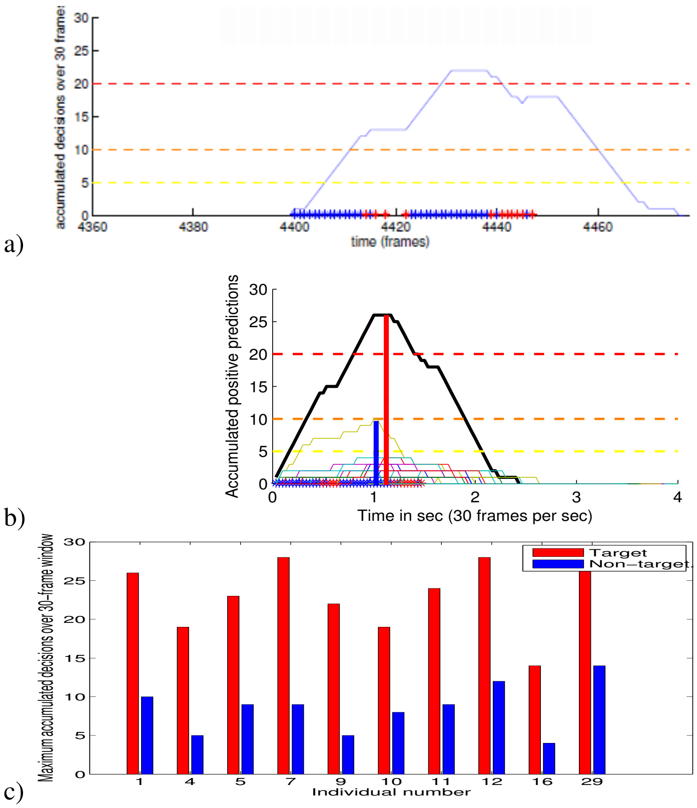


Fig. 4. Level 3: Spatio-temporal analysis: statistics of match predictions accumulated over a face track for target individual against himself (a) and against everybody in a video-sequence (b), the summary of match accumulations for all ten targets (c). Data for matching faces with $iod > 20$ is shown.

E. Face Triaging

Level-3 analysis makes use of the temporal information available in video, which allows one to examine and improve the confidence in decision by accumulating system predictions over a moving time window. This is particularly important when processing low-resolution/quality images and can be used, further combined with face quality metrics obtained in Level 0 analysis, in designing more intelligent WLS solutions. Specifically, this can be used for designing face triaging systems, which generate a "red" alarm (leading to the apprehension of a traveller) only if the matching result is obtained by consistent accumulation of positive matching results. If, on the other hand, the matches are not consistent over time, then "yellow" alarm can be raised, resulting in more careful risk assessment of an individual without making him or her experience any complication to the travel.Such triage-based decision-making schemes for WLS will be very important for the adoption of the technology in public places, so that most people will never have bad experiences because of false matches due to the limitations of automated FR technologies.

VI. CONCLUSIONS AND FUTURE WORK

A target-based multi-level FR evaluation methodology is developed to deal with the key sources of risks that are intrinsically present in video surveillance applications. The application of the methodology has been illustrated by evaluating a commercial FR product on the Chokepoint public data-set that simulates Type 2 video surveillance scenario (Table 1). The result of the evaluation showed that tested FR product showed small deviation in its performance from one target individual to another, which is an important indicator of the robustness of the system to new biometric probes. At the same time, the results have also allowed us to expose the vulnerabilities of the system. Specifically, the following key source of risks of using the system in real-life settings are identified: a) the existence of difficult cases "goats", b) sub-optimal face detection and tracking performance, c) limited use of face quality metrics, and d) general upper bounds of the recognition performance. This is consistent with the assigned technology readiness level as "yellow" in Table II, meaning that the technology is not ready for immediate deployment, but is potentially viable in short or medium future under constrained settings.

For improving the performance of FR systems in video-surveillance applications the following three directions are recommended: 1) the development of more advanced face and person tracking pre-processing techniques [24], including person tracking based on video analytics, the survey of which is presented in [17]; 2) the development of more advanced post-processing techniques to accumulate decisions over time, combined with face quality metrics for more meaningful and robust binary and triaging recognition decisions, and 3) combination of FR technologies listed in Table II with video analytics technologies for improved person/event alarm detection and general video data mining, search and retrieval; Finally, the re-assessment of readiness of all FR technologies in video surveillance applications presented in Table II is recommended on annual basis, ideally in a community-driven effort open to all FR developers and users. The methodology described in this paper can serve as the basis for such re-assessment.

Acknowledgment *– This work was supported by the Defence Research and Development Canada (DRDC) Public Security Technical Program (PSTP-03-401BIOM).*

REFERENCES

[1] L. Best-Rowden, B.Klare, J. Klontz, A.K. Jain. "Video-to-Video Face Matching: Establishing a Baseline for Unconstrained Face Recognition", Proc. IEEE BTAS, 2013.

[2] Joshua C. Klontz, Anil K. Jain, "A Case Study on Unconstrained Facial Recognition Using the Boston Marathon". Bombings Suspects, Technical Report MSU-CSE-13-4, 2013.

[3] J. R. Barr, K. W. Bowyer, P. J. Flynn, S. Bismas, "Face recognition from video: a review. International Journal of Pattern Recognition and Artificial Intelligence, April 20, 16:2, 2012.

[4] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Toward development of a FR system for watchlist surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 1925–1937, 2011.

[5] 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.

[6] 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*, 2009,

[7] C. Pagano, E. Granger, R. Sabourin, D. O. Gorodnichy, "Detector Ensembles for FR in Video Surveillance", Proc. of IEEE IJCNN 2012.

[8] 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*.

[9] R. Goh, L. Liu, X. Liu, and T. Chen, "The CMU Face In Action (FIA) Database," Analysis and Modelling of Faces and Gestures, Lecture Notes in Computer Science Volume 3723, 2005, pp 255-263

[10] Z.Huang et al. "Benchmarking Still-to-Video Face Recognition via Partial and Local Linear Discriminant Analysis on COX-S2V Dataset". Proceeding of Asian Conference on Computer Vision, ACCV 2012.

[11] ISO/IEC SC 37 19795-2:2007 Biometric performance testing and reporting - Part 2: Testing methodologies for technology and scenario evaluation.

[12] Dmitry Gorodnichy et al. "PROVE-IT(FRiV): framework and results", Proceedings of NIST International Biometrics Performance Conference, IBPC 2014, April 1-4, 2014, Gaithersburg, MD.

[13] E. Granger, P.Radtke, and D. Gorodnichy, "Survey of academic research and prototypes for face recognition in video", CBSA Science and Engineering Directorate, Division Report 2014-25 (TR).

[14] D. Gorodnichy, E. Granger, and P. Radtke, "Survey of commercial technologies for face recognition in video", CBSA Science and Engineering Directorate, Division Report 2014-06 (TR).

[15] E. Granger and D. Gorodnichy, "Evaluation methodology for face recognition technology in video surveillance applications", Border Technology Division, Division Report 2014-27 (TR).

[16] E. Granger et al. "Results from evaluation of three commercial off-the-shelf face recognition systems on Chokepoint dataset", Border Technology Division, Division Report 2014-29 (TR).

[17] D. Macrini, V. Khoshaein, G. Moradian, C. Whitten, D.O. Gorodnichy, R. Laganiere, "The Current State and TRL Assessment of People Tracking Technology for Video Surveillance applications", CBSA Science and Engineering Directorate, Division Report 2014-14 (TR).

[18] Dmitry O. Gorodnichy. "Further refinement of multi-order biometric score analysis framework and its application to designing and evaluating biometric systems for access and border control". Proc. IEEE SSSI CIBIM Workshop, April 2011, Paris

[19] D.O. Gorodnichy, S. N. Yanushkevich, and V. P. Shmerko. "Automated Border Control: Problem Formalization", Proc. IEEE SSSI CIBIM Workshop, December 2014, Orlando

[20] A. Rattani, G. Marcialis, and F. Roli, "An experimental analysis of the relationship between biometric template update and the Doddington's zoo: A case study in face verification," in *ICIAP 2009*,

[21] 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 CVPRW '06*.

[22] 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, 2012.

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

[24] Dewan, M., Granger, E., Roli, F., Sabourin, R., Marcialis, G.L.: "A comparison of adaptive appearance methods for tracking faces in video surveillance". In: The 5th International Conference on Imaging for Crime Detection and Prevention (December 16-17, 2013)