Accepted in 13th Conference on Computer and Robot Vision (Victoria, BC, Canada, June 1-3, 2016)

Recognizing people and their activities in surveillance video: technology state of readiness and roadmap

Dmitry O. Gorodnichy[13], David Bissessar[13], Eric Granger[2], Robert Laganiére[3]

[1]Science and Engineering Directorate, Canada Border Services Agency
[2]École de technologie supérieure, Université du Québec
[3]School of Computer Science and Electrical Engineering, University of Ottawa

Abstract—**This paper presents a technology readiness assessment framework called PROVE-IT(), which allows one to access the readiness of face recognition and video analytic technologies for video surveillance applications, and the roadmap for the deployment of technologies for automated recognition of people and their activities in video, based on the proposed assessment framework and the evaluations conducted by the Canada Border Services Agency and its partners over the past five years.**

I. Introduction

As a result of the increasingly growing demand for security, many countries have been deploying closed circuit television (CCTV) video surveillance systems as an important tool for enhancing preventive measures and aiding post-incident investigations. Thousands of surveillance cameras are installed at border crossings, airports, and other public places. Millions of hours of video data are being recorded daily.

Over the years, however, it has been realized that video surveillance systems are not used very efficiently. In the real-time monitoring mode, the problem is that an event may easily pass unnoticed due to false or simultaneous alarms and lack of time needed to rewind and analyze all them. In archival post-event investigation mode, the quantity of video data that need to be processed makes post-incident investigation very difficult. Due to the temporal nature of video data, it is very difficult for a human to analyze video data within a limited amount of time.

The solution to these problems is seen in deploying video recognition technologies that use the advances in facial biometrics and video analytics (computer vision and machine learning) to automatically detect and recognize people and their activities in video [1-9]. The performance of these technologies however varies drastically from one surveillance scenario to another, which is why they are still generally not considered ready for deployment by a majority of CCTV users.

Over the past eight years, with the support from the Defence Research and Development Canada (DRDC), the Canada Border Services Agency (CBSA) has been leading a number of projects aimed at evaluating and advancing these technologies. In 2014 this effort culminated with the development of a technology readiness assessment framework, called PROVE-IT(), which was then applied to prepare recommendations related to technologies that can be developed and deployed for recognizing people and their activities in surveillance video over the next years (technology roadmap). These recommendations led to developing new projects, technologies and pilots by the agency. They also contributed to developing general guidelines related to the use of biometrics and video analytics in surveillance systems, such as those currently prepared by the International Standards Organization Special Committee on Biometrics (ISO SC 37). In the following this framework and the technology readiness assessment results obtained using it are presented.

The paper is organized as follows. In Section 2, general high-level considerations related to recognition in video are presented. Section 3 describes the PROVE-IT() readiness assessment framework. The application of the PROVE-IT() framework for assessing the technology readiness of face recognition in video (FRiV) and video analytics (VA) is presented next in Section 4. The summary of recommendations related to technology development and deployment including a discussion on the importance of developing visual analytic tools and training procedures for CCTV operators is presented in Section 5. Discussions conclude the paper.

II. Strategic understanding of the problem

Through the course of this work, the term “recognition" is used in a wide sense to include any recognition that is possible in video data, whether related to recognition of *an* event (synonymous to the traditional use of the term “detection") or *the* event (synonymous to the traditional use of the term “identification"). The terms “recognition system" and “detection system" are therefore used inter-changeably.

Table 1: Recognition in video: “verbs” vs. “nouns” of the problem.

Objective to Recognize what?	Automated recognition	Manual recognition	Relies on what?
Noun (subject)	biometrics	forensic examination	spatial detail
Verb (activity)	video analytics	CCTV monitoring	temporal detail

A. Two types of events in video: nouns vs. verbs

An automated recognition system aims at automatically recognizing an event in video. As visualized in Table 1 (first introduced in [4]), two types of events are generally observable in video: those related to subjects (nouns) and those related their activities (verbs). When detected automatically, they relate to biometric and video analytics technologies. When detected manually, they relate to the work done by forensic analysts and CCTV operators.

Critically, these two types of events are different from each other in that the former operates mainly on spatial

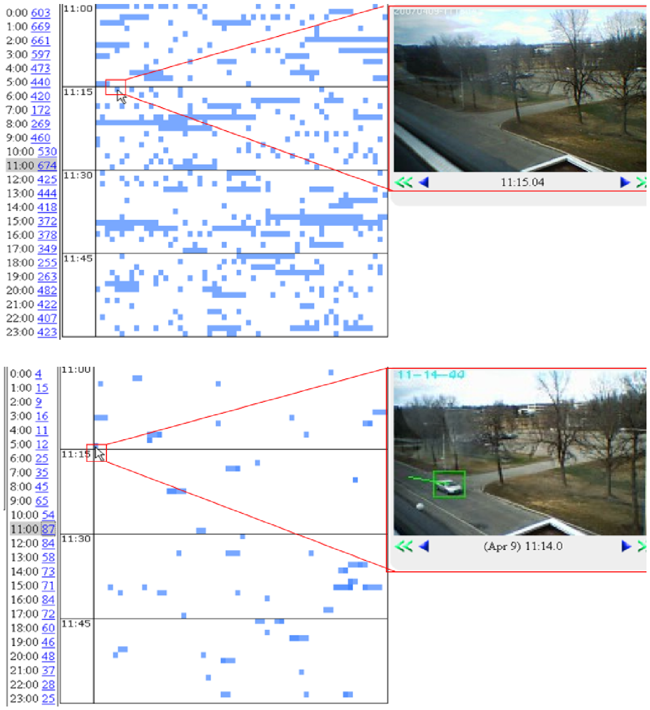


Figure 1: Examples of a "poorly performing" system (top) and a "well performing" system (bottom) (from [2]). Detected events (cars) are shown as blue boxes in a one-hour window rectangle for two systems running at the same time, each line corresponding to a minute in an hour. "Poorly performing" system generates over 90% of False alarms, but may still be useful for certain CCTV applications.

detail of the video information (thus requiring higher resolution of video images), while the latter works on a temporal details of it (thus permitting lower resolution of video images, yet requiring their continuity in time).

While presenting two different challenges and often dealt by two different communities of developers and users, these two types of events intrinsically belong to the same problem, which is the automated extraction of evidence from video. This is how they are treated in our work: as two sides of the same "*video recognition*" problem. An event that a video recognition system tries to detect is referred to as "*target*". An event that is processed by the system is called "probe". The result of video recognition is either recognizing a probe as a target or not.

B. "*Poorly performing* " vs. "*well performing*" systems

Video-surveillance is used in three modes of operation: active real-time, passive real-time, and archival (through recordings). Active monitoring involves trained personnel who watch video streams at all times. Passive monitoring involves employees who watch video streams in conjunction with other duties. In the third mode, CCTV systems record video data for the purpose of post-event analysis.

For either mode of operation, the performance of an event detection system intrinsically depends on three types of problem complexity: 1) complexity of setup, 2) complexity of the recognition task, and 3) intelligence of the recognition algorithm, and may vary from being "*very poor*" to "*reasonably good*". Both performance extremes are shown in Figure 1 (from [2]), which compares the performance of the basic motion-detection technology included by default in most surveillance systems and an advanced object-detection-based video analytics technology. While this figure shows two particular systems, it is representative of the performance of many other video recognition systems, where by the notion of "*system*", a combination of the setup, recognition task, and recognition algorithm is considered.

While a "well performing" system is an obvious candidate for an operational deployment in either mode of CCTV operation, it is noted that a "poorly performing" system may also become a candidate for deployment, specifically for an archival mode, where it can facilitate manual retrieval of evidence that is being routinely performed by many CCTV users. In the latter case however, it can be said that additional tools (such as those for data filtering and event mining) and human analyst expertise play a more important role than the recognition system itself.

C. *Detection errors, metrics and evaluation results*

Two types of detection errors are possible in a recognition system: Type-I, also called False Alarm, False Positive (*FP*) error, and Type-II, also called Miss, False Negative (*FN*) error. Depending on the application, one error may be more critical than the other. It is also noted that, while a Type-I error is normally measurable, Type-II in most operational settings is not measurable.

Performance of the system is traditionally reported by computing True / False Positive and True / False Negative Rates (*TPR*, *TNR*, *FPR* and *FNR*) at different operational thresholds and constructing error trade-off curves such as the Receiver Operating Characteristic curve (ROC), which plots $FPR = FP\ / (TN+FP)$ vs. $TPR = TP\ / (TP+FN)$, and the Precision-Recall Operating Characteristic (PROC), which plots $Precision = TP / (TP+FP)$ vs. $Recall = TPR = TP\ /\ (TP+FN)$. Because video surveillance is an open-set problem, meaning that the system does not have information about "non-target" events/people and the number of "non-targets" is significantly higher than that of "targets", i.e., n>> p(in Figure 1) and *TN* >> *TP*, PROC curves provide additional value for analysis.

Figure 3 show error trade-off curves that have been reported for state-of-art "noun" and "verb" recognition systems. In Figure 3 (example of "noun" recognition) taken from [16], a commercial FR product (Cognitec) is tested for its ability to detect (recognize) a particular individual (ID=1 from the Chokepoint dataset [19]) walking through a corridor. ROC and PROC curves are computed for three different system configurations. Details of this experiment are provided in [16]. As an outcome of this evaluation, one can observe that at *Recall=TPR=0.6* (marked by dashed line) the system exhibits *Precision =~0.8* (80%) when the system is configured to process faces with at least 30 pixels between the eyes (red curve). This can be considered a "well-performing" scenario.

In Figure 3 (example of “verb” recognition) taken from NIST TRECVID 2012 video analytics competition (described in [8]), systems are tested for their ability to detect a “person run" event. Error detection curves plotting Probability of miss (*Pmiss*) as function of Rate of False Alarm (*RFA*) are shown. As an outcome of this evaluation, one can observe that at False Alarm Rate of less than 1/hour (dashed line), the probability of miss is higher than 80% for all systems. This can be considered a “poorly performing" scenario.

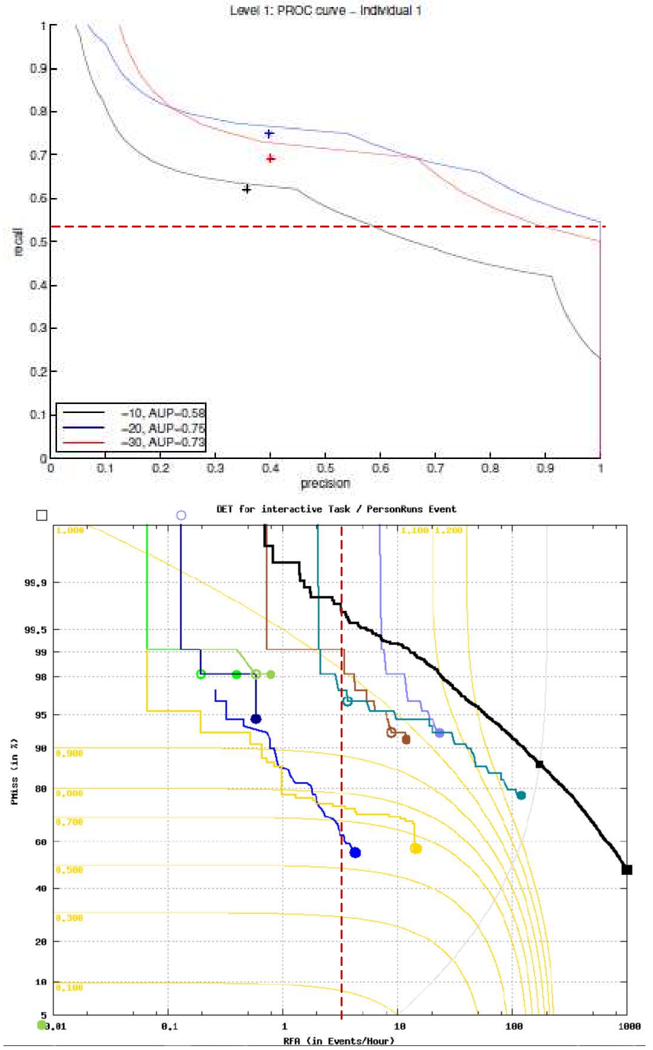


Figure 3: Error trade-off curves reported for state-of-art video recognition systems: a) for a commercial FR product showing the ability of the system to recognize an individual in a chokepoint corridor (top, from [16]), b) for VA solutions presented in TRECVID competition showing the ability of the systems to detect people running in airport halls (bottom, from [8]).

It can be observed that, while above mentioned metrics and curves are very useful for comparing one product to another as well as for monitoring and tuning the performance of a particular system, they may not be easily converted to the recommendations related to the state of readiness of these technologies for deployment in operational scenarios. This is why operational agencies rely on the concept of the Technology Readiness Level (TRL).

D. Technology Readiness Level assessment

The TRL assessment is adopted by many agencies as a risk management tool [11]. It provides a common scale of science and technology exit criteria and allows one to estimate the cost/investment required for deploying a system. According to the TRL assessment framework, the readiness level in the range from Level 1 to Level 9 is assigned to a technology follows:

Level 1: Basic principles observed and reported,
Level 2: Technology concept and/or application formulated,
Level 3: Analytical and experimental critical function and/or characteristic proof of concept,
Level 4: Component validation in laboratory environment,
Level 5: Laboratory-scale similar system or component validated in relevant environment,
Level 6: Pilot-scale similar prototypical system or component validated in relevant environment,
Level 7: Full-scale prototypical system demonstrated in relevant environment,
Level 8: Actual system completed and qualified through test and demonstration,
Level 9: Actual system successfully operated in the field over the full range of expected conditions.

Proper TRL assessment requires access to real environments and real end-users, an approved protocol and team of experts as well as a sufficiently long period of time for conducting the analysis. In certain cases however these may not be available to researchers, in particular in an academic environment or within a limited amount of time or funding allocated for the analysis. Applying a full nine-grade scale may not be appropriate in these cases, as it may give a false impression of the level of detail of the conducted analysis. Additionally, a formal TRL assessment process is often focused on a particular application, with the objective to test and prepare a technology for this particular application. In contrary to that, the objective of many smaller technology evaluation projects is to probe the entire technology landscape in order to identify the areas of focus for further research and investment. This is why a different technology readiness assessment framework is desired that will be suitable for use by a wider community of users (who may not have capacity or capability to conduct comprehensive TRL) as well as convenient for preparing the recommendations related to the technology deployment and best investment opportunities. Such framework, called PROVE-IT(), has been developed by the CBSA and is described below.

III. PROVE-IT() ASSESSMENT FRAMEWORK

A. Assessment scale

The PROVE-IT() assessment framework was developed to provide a light-weight alternative to the conventional nine-point TRL assessment. It uses a semaphore-like three-point scale (“green" or “+"- proved ready; “yellow" or “o" -

possibly ready with additional R&D ; and “red" or “-" - proved not ready for deployment in the nearest future). The relationship between the PROVE-IT() assessment grades and traditional TRL scale is shown in Table 2. Two sub-grades within the “ready" grade and “possibly ready" grade can be introduced to permit additional level of assessment detail when such information is available.

B. Technology landscape map template

Being an approximate measure of readiness, PROVE-IT() assessment can be used to estimate the technology readiness in the entire spectrum of possible deployment conditions and scenarios, using the following three steps (See Figure 5).

Step 1: Define taxonomy of possible operational conditions (scenarios) {Sj}: ordered from simplest to most difficult;

Step 2: Define taxonomy of possible technology application variations {Ti}: ordered from simplest to most difficult, thereby establishing a two-dimensional technology landscape map template.

Step 3: Assign technology readiness colour (green, yellow, red) for each technology application variation at each PROVE-IT(Ti|Sj), using a three-phase performance assessment process described below, thereby completing the technology landscape map template.

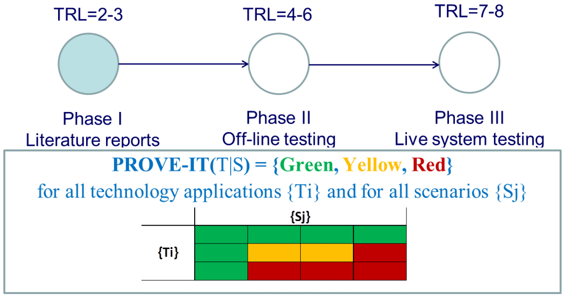


Figure 5: The PROVE-IT() framework: three-phase evaluation process and two-dimensional technology landscape map template.

C. Three-phase assessment process

Following the formal TRL definition described above, the following three key technology assessment phases are defined (see Figure 5).

Phase I: Literature and market review (testing for up to TRL=3). This includes surveying of scientific and industry literature, including company offerings and patent analysis, for the purpose of identifying and harmonizing the lexicon and technology definitions as well as for obtaining the preliminary high-level overview of possible options and solutions; and selection of solutions and scenarios that are believed to be ready for off-line testing for further assessment

Phase II: Off-line testing (testing for up to TRL=6). This includes testing of the solutions on pre-recorded datasets corresponding to different CCTV scenarios, and measuring detection error trade-off metrics.

Phase III: Live system testing (testing for up to TRL=8). This phase requires further customization and refinement of the technologies and scenarios tested in the previous phase for further testing in a live environment with real operational surveillance cameras and CCTV users.

TRL higher than 8 would normally not require additional investigation as it assumes that the technology is already well established and has a substantial deployment history.

D. Taxonomy of video surveillance setups

In the 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 3. 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, four main types of video surveillance scenario types of increasing complexity are recognized as shown in Table 3. The images from an operational airport surveillance cameras corresponding to Types 1-3 are shown in Figure 6. Camera positioning and resolution is assumed to be the best technically possible.

Table 2: Taxonomy of video surveillance scenarios.

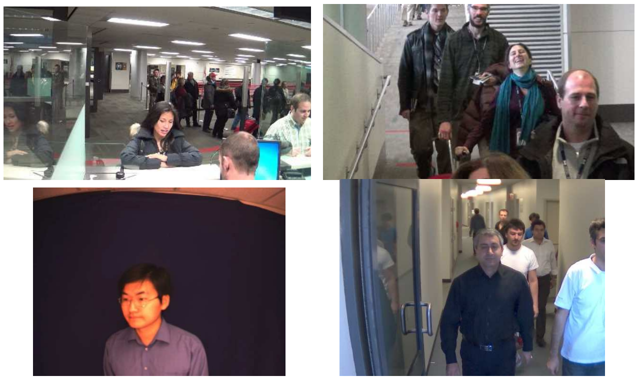
TYPE	“WHO” PERSON FACTORS	“WHAT” ACTIVITY FACTORS	“WHERE” SETUP FACTORS
1 Stationary	semi-controlled	controlled	controlled
2 Portal	uncontrolled	semi-controlled	controlled
3 Hall	uncontrolled	uncontrolled	semi-controlled
4 Outdoor	uncontrolled	uncontrolled	uncontrolled

TYPE	EXAMPLES
1 Stationary	In front of a passport control, kiosks or entrance door
2 Portal	In narrow corridors, chokepoint entries (one or several at time)
3 Hall	In airport halls with controlled lighting (free flow, many at time)
4 Outdoor	Outdoor environments

There are several public video data-sets that simulate the defined above video surveillance types and which can be used for evaluation purposes. It is vital for VA and FRiV potential users to examine the performance of the systems on these data-sets prior to testing in real surveillance settings.

By doing so they can expose in advance the vulnerabilities of the system and develop the strategies to deals with them. 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.

The number of public data-sets that simulate real surveillance settings is growing.. 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, such as the i-Lids and FRL2011 datasets from Home Office [19] and the “People in Airport" dataset that has been created by the CBSA [17].



Type 1 Type 2

Figure 6: Images taken by surveillance cameras corresponding to different setups according to the taxonomy in Table 3: from the CBSA “People in Airport" dataset [17] (top), from public datasets [19,20] (bottom).

There are several public video data-sets that simulate the defined above video surveillance types and which can be used for evaluation purposes. It is vital for VA and FRiV potential users to examine the performance of the systems on these data-sets prior to testing in real surveillance settings.

By doing so they can expose in advance the vulnerabilities of the system and develop the strategies to deals with them. 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.

The number of public data-sets that simulate real surveillance settings is growing.. 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, such as the i-Lids and FRL2011 datasets from Home Office [19] and the “People in Airport" dataset that has been created by the CBSA [17].

IV. ASSESSMENT RESULTS

Since 2008, following the transition of the related technology and knowledgebase from NRC [1,2], the CBSA has taken the lead within the Canadian government in investigating video recognition technologies (VA and FRiV) for video-surveillance applications. A video analytic platform and test bed (VAP) has been developed to allow the integration and testing of third-party VA and FR libraries with the operational CCTV systems [5]. A number of end-user search and retrieval tools (Event Browsers) have been developed to allow the users to browse efficiently through the detected events in search for the evidence, and various mock-up and on-site testing of the technology has been conducted [3,5,8,17]. Feedback related to operational CCTV needs and constraints has been regularly obtained from other government agencies through inter-departmental workshops on Video Technologies for National Security (VT4NS) [3]. At the same time, the project team has been gaining experience and knowledge related to advances in CCTV cameras and Video Management Software, developing recommendations for new CCTV installations across the agency.

Since 2011 with additional funding from the DRDC, this effort of CBSA and its partners has converged into development of a comprehensive technology readiness landscape assessment and the deployment roadmap. These are presented below, further extended and revised from previous publications [10,11].

A. PROVE-IT(FRiV) results

A taxonomy of FRiV application variations of increasing complexity has been developed using the following categories:

- **by level of performed face processing** (from easiest to hardest): face detection, face tracking (using video-analytic techniques), face classification, facial expression analysis, identification (identity recognition);
- **by mode of operation :** archival post-event operation vs. real-time operation;
- **by decision making mode** (from easiest to hardest): fully automated (binary) vs. semi-automated (triaging) vs. not-automated (as part of an analytic tool or filter);
- **by data modality** (from easiest to hardest): video-to-video vs. still-to-video.

Following the survey of academic literature [11] and commercial solutions and patents [12], feasible surveillance scenario were assessed (Type 1 and Type 2) and a number of commercial and academic FR solutions have been selected for further testing in those scenarios

Based on the in-house evaluations and literature reviews [11-14], the feasibility of each FRiV application is accessed for each video surveillance type. Table 4 shows the result. The “Faces in Action" [20] and the Chokepoint data-sets [21] (shown in Figure 6) were used to simulate Type 1 and

Type 2 surveillance setups to prove the “yellow" grade readiness of technologies. Other datasets recommended for off-line testing of FRiV applications are: the Still-2-Video dataset [22] the FRL2011 from UK HomeOffice [23] and the “People in Airport" dataset from CBSA (also by request), the image of which are shown in Figure 6.

B. *PROVE-IT(VA) results*

Compared to FRiV, VA technologies operate on a much wider spectrum of possibilities in visual representation of objects. In contrast to generic face detectors that are used to facilitate face recognition, there is no generic object (or person) detector capable of recognizing / detecting particular objects (or persons). This limits considerably the range of fully-automated applications that can be performed with VA. The following taxonomy of applications has been developed for VA technologies:

- detection of people;
- recognition of people activities - at a personal level;
- recognition of people activities - at a crowd level;
- recognition of objects left by or associated with people;
- general detection of camera tampering and intrusion detection.

Based on the in-house evaluations and literature reviews [6,7], the feasibility of each VA application is accessed for each video surveillance type as shown in Table 5 .The following datasets have been used for off-line evaluation: PETS 2006, AVSS, and iLids, which simulate Types 2, 3 and 4 surveillance setups. Of a particular value is the i-Lids dataset, which has the following event detection scenarios: (a) sterile zone, (b) parked vehicle, (c) abandoned baggage, and (d) doorway surveillance. In addition, there is one dataset with a multiple camera tracking scenario. All the scenarios are recorded in a real airport. A subset of this dataset is used at NIST TRECVID competitions.

V. RECOMMENDATIONS

Two main possibilities of using video surveillance technology for recognition of people and their activities are envisaged. The first possibility deals with video cameras used in combination with other sensors and point-and-shoot cameras. For example, RFID readers can be installed in airports to facilitate tracking of people. Sensors can also be used to trigger the capture of video data, in particular of high resolution. Similarly, point-and-shoot cameras can be installed to capture high-resolution high-quality facial images triggered by video analytics and other sensors. The second possibility deals with the traditional use of cameras in video surveillance applications, when multitudes of IP-based surveillance cameras are connected to centralized storage, streaming continuously video data that is stored and processed by video management software. The following recommendations are developed for the latter.

A. *Long-term research and development*

First, it is emphasized that, by the nature of optics and because of the compression required for transmitting video images over IP-networks, faces in surveillance video are “meant to be" of low effective resolutions, where *effective resolution* (also referred to as *informative resolution* [10]) refers to the number of discernable pixels between the eyes. It particular, experiments show that capturing focused non-blurred faces of moving people with more than 60 discernable pixels between the eyes is close to impossible with current state-of-art IP-cameras. Mega-pixel cameras increase the resolution of the image, but they are shown to not increase the effective resolution of faces. That is, even when captured at high resolution, facial images of moving people remain of the same effective resolution, which is proved by sub-sampling the image to a lower resolution and then super-sampling it back to original resolution. This is because objects 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 increasing the shutter speed). See [16] for more detail on the detailed analysis of this phenomenon. Hence only those FR techniques that can process low resolution will be suitable for surveillance applications. For reference, most current COTS FR products required face resolution to be higher than 60 pixels between the eyes.

For improving the performance of person recognition systems in video-surveillance applications the following two main directions are foreseen: i) the development of more advanced face and person tracking pre-processing techniques, including person tracking based on video analytics, the survey of which is presented in [7], and ii) the development of more advanced post-processing techniques that accumulate decisions over time, combined with face quality metrics for more meaningful and robust binary and triaging recognition decisions. In doing that, a higher level of combination of FR technologies with VA technologies is expected.

Recognition and detection technologies may never be expected to be error-free. Hence, an important requirement for enabling the deployment of these technologies is the development of end-user tools for human operators, which operate in support to the current human operator’s work. This includes the development of target-based systems such as those described in [16] and event filtering tools based on advanced computer-human interfaces and the science of visual analytics, which employs the natural efficiency of the human brain in processing visual information.

B. *Near-term deployment and pilots*

Face Detection has become a mature technological solution capable of detecting faces with 10 pixels between the eyes over a wide range of face rotations (± 30° in all axes of rotation) - producing FPR and FNR of less than 1%. This makes it suitable for deployment in many scenarios (TRL>7). This also enables performing many other face processing tasks listed in Table 4.

Two main opportunities for deploying person recognition systems are observed (marked by rectangles in Table 4). The first opportunity addresses archival applications and aims at facilitating the existing post-event search procedures for evidence retrieval from video. A critical example of this opportunity is using face detection and face grouping at low resolutions to improve Search and Retrieval of evidence related to a particular person or incident. This is focus of the work in [17].

The second opportunity addresses real-time applications and aims at developing tools for improved situational awareness and decision making. Examples of such tools are border wait time estimation, traffic control and traditional protection of limited access areas. Another example is Faces on the Move technology where faces of travellers captured in Type 1 and Type 2 setups are matched against a watch list to generate flags that can be used by border officers for triaging travellers. To enable this application, an additional set or array of cameras need to considered to increase the chance of capturing an eye-aligned focused a face in Type 2 (Portal) settings. This is focus of one of the current DRDC CSSP projects [18].

C. Face Triaging

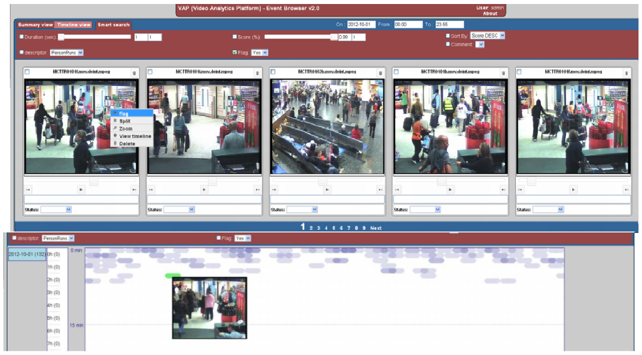
Face Triaging is a new concept related to the use of FR in surveillance applications identified and studied by the CBSA in its studies. It is a particular case of semi-automated face watch-list screening technology that is suitable for the applications where there is a high traffic of people that needs to be processed in real-time, as in border control, where negative consequences for a person who is falsely matched need to be minimized and where there is no possibility (or time) for a human operator to examine the output of the FR system.

The core principle of Face Triaging technology is that “looking similar” to a criminal should never result in treating a person as more risky. Therefore, a new label for the FR system outcome is introduced called “looks similar” (yellow), in addition to traditional “matched” (red) and “non-matched” labels. “Looks similar” label must not bring about any negative connotation about a person and is provided to a triaging officer purely as a flag that the officer may ask additional questions to a traveller within the Standard Operational Procedure as he/she would normally do with other travellers within given flexibility and service standards. This is in contrast to “matched” (red) flag, in which case the triaging officer needs to direct a person to further examination, where his/her identity will be validated using as much time as needed though interrogation and/or additional biometric measurements.

Our analysis of technology readiness indicates that Watch List Screening using Face Triaging has better chance of being deployed for real-time applications compared to traditional binary Watch List Screening (Table 4).

D. End-user tools and training

Figure 3 showed two possible outcomes of applying a video recognition system: with few False alarms and with many False alarms. Either application may be found valuable for end user, as long as proper data processing/filtering tools are developed and training to use these tools is provided. One of the key recommendations therefore made from conducted technology assessment, is that the use of video recognition technologies will require the development of tools for filtering, searching, and mining of events detected by the recognition system. Several such tools have been prototyped and tested by the CBSA [5,8]. The use of these tools (shown in Figure 7) was also instrumental for the TRECVID competition [8]. It is emphasized that such tools should be designed based on best practices in software usability. Finally, training programs should be developed for operators to train them to use innovative video-recognition tools.



VAP (Video Analytics Platform) - Event Browser v2.0

User: admin

On: 2012-10-01 From: 00:00 To: 23:55

□ Duration (sec): 1 1

□ Score (%): 0.99 1

□ Sort By: Score DESC

□ Comment:

□ descriptor: PersonRuns

☑ Flag: Yes

MCTTR0101f.mov.deint.mpeg

MCTTR0101f.mov.deint.mpeg

MCTTR0102b.mov.deint.mpeg

MCTTR0101b.mov.deint.mpeg

MCTTR0101f.mov.deint.mpeg

□ descriptor: PersonRuns

□ Flag: Yes

2012-10-01 (132)

Figure 7: End-user Search and Retrieval tools (Event Browser) used for NIST TRECVID ``Running Event'' detection competition [8]: Annotated snapshot view (top) and TimeLine view (bottom). The alarms detected by Video Analytics, of which majority (over 90%) are False alarms, are filtered out using the user interface designed using the principles from Computer Human Interface and Visual Analytics domains.

VI. CONCLUSIONS

It is not uncommon in business culture to present a technology as ready for deployment. In reality however, while a technology may work under certain conditions, it may not work under different conditions. This is especially True for video-surveillance applications where the lighting and setup conditions in an operational environment may differ drastically from those where technology was demonstrated. The PROVE-IT() assessment framework presented in this paper is a tool that allows one to distinguish and report the applications and conditions in which the technology works and in which it does not. This facilitates developing specifications for the technologies that have been “proved" ready for deployment. This also permits the development of the roadmap for technologies that will be ready in the nearest future. Finally, it addresses privacy related concerns – such as those that impede the development and deployment of face recognition / video analytics technologies in the fear of their recognition power, which may be reported by vendors or observed in science fiction movies, but which is not there in real technologies and applications.

The PROVE-IT() framework has been applied for face recognition in video (FRIV) and video analytic (VA) technologies. The outcome is a set of practical recommendations for FRiV and VA developers and CCTV users related to the best investment in these technologies, and the technology roadmap for the deployment of technologies capable of automatically recognizing people and their activities in surveillance video expressed using two-dimensional technology landscape maps shown in Tables 4 and 5.

In conclusion, it is recommended that the readiness of all technologies presented in Tables 4 and 5 be re-assessed on a regular basis, ideally in a community-driven effort open to all FR/VA developers and CCTV users. The methodology described in this paper can serve as the basis for such re-assessment. A new ViSTER (Video Surveillance Technology Evaluation and Research Group) portal [23] has been set up to facilitate this process.

ACKNOWLEDGMENT

This work was supported by the DRDC Centre for Security Sciences: VT4NS, C-BET, PSTP08-0109BIOM, PSTP-03-401BIOM, PSTP-03-402BTS, CSSP-2013-CD-1063, CSSP-2014-CP-2000 projects. Feedback from Marek Rejman-Greene (HomeOffice), Richard Vorder Bruegge (FBI) and John Garofolo (NIST, DHS) is gratefully acknowledged.

REFERENCES

[1] D. Gorodnichy, M. A. Ali, E. Dubrofsky, K. Woodbeck. Zoom on Evidence with the ACE Surveillance, CRV International Workshop on Video Processing and Recognition (VideoRec'07), May 28-30, 2007. Montreal. Online: http://www.computer-vision.org/VideoRec07/program.html

[2] D. Gorodnichy and T. Mungham. Automated video surveillance: challenges and solutions. ACE Surveillance (Annotated Critical Evidence) case study. NATO SET-125 Symposium "Sensor and Technology for Defense against Terrorism", 2008. Online: https://www.researchgate.net/publication/229040125_Automated_video_surveillance_challenges_and_solutions_ACE_Surveillance_Annotated_Critical_Evidence_case_study

[3] D. Gorodnichy, J.-P. Bergeron, D. Bissessar, E. Choy, J. Sciandra, "Video Analytics technology: the foundations, market analysis and demonstrations", Technical Report DRDC-RDDC-2014-C251. http://cradpdf.drdc-rddc.gc.ca/PDFS/unc167/p801081_A1b.pdf

[4] D. Gorodnichy. "Recognition in Video" , University of Toronto IPSI Public Lecture Series, November 2009. Online: Appendix E, ibid (http://cradpdf.drdc-rddc.gc.ca/PDFS/unc167/p801081_A1b.pdf)

[5] D. Gorodnichy and E. Dubrofsky. VAP/VAT: Video Analytics Platform and Test Bed for Testing and Deploying Video Analytics. In Proceedings of SPIE Volume 7667: Conference on Defense, Security and Sensing, 2010

[6] D. Gorodnichy, D. Macrini, R. Laganiere, "Video analytics evaluation: survey of datasets, performance metrics and approaches ", Technical Report DRDC-RDDC-2014-C248. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc167/p801081_A1b.pdf

[7] 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", Technical Report DRDC-RDDC-2014-C293. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc161/p800731_A1b.pdf

[8] C. Whiten, R. Laganiére, E. Fazl-Ersi, F. Shi, G. Bilodeau, D. O. Gorodnichy, J. Bergeron, E. Choy, D. Bissessar . VIVA-uOttawa / CBSA at TRECVID 2012: Interactive Surveillance Event Detection. Online: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.12.org.html

[9] D. Gorodnichy and E.Granger, "Evaluation of Face Recognition for Video Surveillance ". NIST International Biometric Performance Conference (IBPC 2012), Gaithersburg, March 5-9, 2012. Online: http://www.nist.gov/itl/iad/ig/ibpc2012.cfm .

[10] D. Gorodnichy and E. Granger " PROVE-IT(FRiV): framework and results ". NIST International Biometrics Performance Conference (IBPC 2014), Gaithersburg, MD, April 1-4, 2014. Online: http://www.nist.gov/itl/iad/ig/ibpc2014.cfm .

[11] D. Bissessar, E. Choy, D. Gorodnichy, T. Mungham, "Face Recognition and Event Detection in Video: An Overview of PROVE-IT Projects (BIOM401 and BTS402)", Technical Report DRDC-RDDC-2014-C167. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc157/p800402_A1b.pdf

[12] D. Gorodnichy, E.Granger, and P. Radtke, "Survey of commercial technologies for face recognition in video ", Technical Report DRDC-RDDC-2014-C245. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc159/p800510_A1b.pdf

[13] E. Granger, P. Radtke, and D. Gorodnichy, "Survey of academic research and prototypes for face recognition in video ", Technical Report DRDC-RDDC-2014-C246. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc167/p800522.pdf

[14] E. Granger and D. Gorodnichy, "Evaluation methodology for face recognition technology in video surveillance applications", Technical Report DRDC-RDDC-2014-C249. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc167/p800519_A1b.pdf

[15] 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", Technical Report DRDC-RDDC-2014-C247. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc167/p800520_A1b.pdf

[16] D. Gorodnichy and Eric Granger, Target-based evaluation of face recognition technology for video surveillance applications , Proc. of IEEE SSCI CIBIM 2014 workshop, Orlando, December 2014.

[17] J.-P. Bergeron and D. Bissessar, "Accelerated Evidence Search Report", DRDC-RDDC-2014-C166. Online: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc159/p800470_A1b.pdf

[18] DRDC website: "Government of Canada invests in Canada's Safety and Security" (January 29, 2014), http://www.drdc-rddc.gc.ca/en/dynamic-article.page?doc=government-of-canada-invests-in-canada-s-safety-and-security/hr0e3lxs

[19] iLids (The Imagery Library for Intelligent Detection Systems) : www.homeoffice.gov.uk/science-research/hosdb/i-lids.

[20] 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 .

[21] 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

[22] 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.

[23] ViSTER (Video Surveillance Technology Evaluation and Research Group) Portal. Online: http://sites.google.com/site/vistercanada

Table 3: Technology readiness assessment grades according to PROVE-IT() framework.

GRADE	TRL	Definition and required proof	Years to deploy and R&D effort required
++	8-9	Unambiguously **proved ready** through deployments and pilots in operational settings	*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	**Possibly ready**, may be proved ready if additional evidence is provided	*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	Unambiguously **proved not ready** for given operational settings	*Not Ready*: Not possible within next 5 years; requires major academic R&D.

Table 4: PROVE-IT(FRiV) results. The readiness assessment of face recognition for video surveillance applications.

Face Recognition In Video technologies	Type 0[1] (eGate)	Type 1 (Stationary)	Type 2 (Portal)	Type 3 (Hall)
Detection (no Face Recognition)				
1. Face Detection in Surveillance Video	++	++	+	oo
Tracking (no Face Recognition)[2]				
2. Face Tracking across a Single Video	+	+	+	-
3. Face Tracking across Multiple Videos	+	+	o	-
Semi-automated Recognition[34]: for post-event investigation (search and retrieval of evidence)				
Video to Video (Re-Identification)				
4. Face Grouping, Tagging, Tracking across multiple videos	+	oo	oo	o
Still to Video				
5. FR to aid manual forensic examination	+	oo	oo	-
Fully-automated Recognition: for real-time interdiction (border / access control)				
Video to Video (Re-Identification)				
6. Instant FR in single camera	+	oo	o	-
7. Instant FR from multiple cameras	+	o	o	-
Still to Video				
8. Instant FR for Watch List Screening – Triaging	+	oo	o	-
9. Instant FR for Watch List Screening – Binary	+	o	-	-
Micro-facial feature recognition				
10. Facial Expression analysis: for emotion / intent recognition	+	oo	o	-
Soft and multiple biometrics				
11. Human attribute recognition (gender, age, race)	+	oo	o	-
12. Personal metrics (height, weight, eye/hair colour)	+	o	o	-
13. FR to improve voice or iris biometrics	+	o	-	-

Notes:

1. The readiness of FR applications for cooperative scenario at eGate (Type 0) is provided as point of reference to contrast the performance of the same FR applications in non-cooperative scenarios (Types 1-3).
2. See assessment results for person detection and tracking from PROVE-IT(VA) evaluation.
3. Type 4 scenario (outdoors) is not included in the FRiV assessment since there is no evidence that the technology works in easier setups.
4. The applications marked by boxes have been recommended for pilots. See [17,18] for more details.
5. The references to the academic research/prototypes and commercial technologies that were used in the assessment are provided in [12-15].

Table 5: PROVE-IT(VA) results. The readiness assessment of video analytics for video surveillance applications.

Video Analytics technologies	Type 1 (Kiosk)	Type 2a (Portal)	Type 2b (Portal)	Type 3 (Halls)	Type 4 outdoor
Person Detection and Tracking (without Face Recognition)					
a. Person counting	++	+	oo	o	o
b. Person tracking in single camera	++	+	oo	o	o
c. Person matching in single camera	oo	o	o	-	-
d. Person matching in multiple cameras	o	o	-	-	-
Person Event Detection					
a. Improper standing place	++	++	+	o	o
b. Opposite flow detection	++	++	oo	o	o
c. Running detection [1]	++	++	oo	-	-
d. Tail-gating detection	++	++	oo	-	-
e. Loitering detection	++	+	-	-	-
f. Fall detection	++	oo	-	-	-
Crowd Analysis					
a. Density estimation	n/a		oo	oo	oo
b. Rapid dispersion			oo	oo	oo
c. Crowd formation			oo	oo	oo
d. Crowd Splitting			o	-	-
e. Crowd Merging			o	-	-
Baggage Detection and Tracking					
a. Static Object (>n sec)	+	+[1]	o[1,2]	-	-
b. Object removal	o[2]	o[2]	-	-	-
c. Dropping Object	o[2]	o[2]	-	-	-
d. Abandoned Object	o[2]	o[2]	-	-	-
e. Unattended Object	o[2]	o[2]	-	-	-
f. Carried Object	-	-	-	-	-
Person-Baggage Association Analysis					
a. Person-Baggage Association	o	-	-	-	-
b. Owner change	-	-	-	-	-
Camera Tampering Detection					
Occlusion, Focus moved, Camera moved	++	++	++	++	+
Physical Security					
Virtual trip-wire, intrusion detection	++	++	++	++	+

Notes:

1. For low traffic only.
2. For large objects only.
3. The references to the academic research/prototypes and commercial technologies that were used in the assessment are provided in [6-8].