Survey of academic research and prototypes for face recognition in video

Prepared by:
Eric Granger, Paolo Radtke, Dmitry O. Gorodnichy
Canada Border Services Agency
Ottawa ON
Canada K1A 0L8

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

Contract Report
DRDC-RDDC-2014-C246
September 2014

IMPORTANT INFORMATIVE STATEMENTS

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

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

Canada Border Services Agency
Agence des services frontaliers du Canada

Science and Engineering Directorate

Border Technology Division

Division Report 2014-25 (TR)
July 2014

Survey of academic research and prototypes for Face Recognition in Video

Eric Granger,
Paolo Radtke,
Dmitry O. Gorodnichy

Canada

This page left intentionally blank

Abstract

This report surveys work done by academia in developing Face Recognition solutions for video-surveillance applications. We present an architecture of a generic system for face recognition in video and review academic systems reported in the academic literature suitable for video-surveillance applications. Recommendations on the selection of systems for benchmarking and the analysis of future trends is presented. Techniques presented in this report are those that provide good results on well known reference video data sets and can be used to provide foundations to develop a surveillance system in-house.

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

Communities of Practice: Biometrics and Identity Management, Border and Transportation Security

Canada Safety and Security (CSSP) investment priorities:

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

Acknowledgements

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

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

The feedback from project partners: University of Ottawa (R. Laganiere, S. Matwin), RCMP, TC, CATSA, DRDC, UK HomeOffice, FBI is gratefully acknowledged. Feedback from W. Khreich, M. De le Torre, C. Pagano and R. Sabourin from École de technologie supérieure is also gratefully acknowledged.

Disclaimer

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

Release Notes

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

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

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

Appendices: This report is accompanied by appendices which include the presentations related to this report at the VT4NS 2011 and VT4NS 2013 forums.

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

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

Contents

List of Figures

List of Tables

1 Introduction

The biometric recognition of individuals based on their behavioral or physiological traits, such as the face, finger print, iris, signature and voice, provides a powerful alternative to traditional authentication schemes (e.g., passwords and identification cards) presently applied in a multitude of security and surveillance systems [25]. There are three main types[1] of biometric recognition applications related to face recognition in video (FRiV) [**?**, **?**]: verification applications, identification applications and video-surveillance or screening applications. With *verification applications*, an individual that is enrolled in the system identifies himself and provides a biometric sample. Then, the biometric system seeks to authenticate that the sample corresponds to the specific individual. In contrast, with *identification applications*, an individual provides a biometric sample and the system seeks to determine if the sample corresponds to one of the individuals enrolled to the system. Finally, *video surveillance or screening applications* differ from identification in that the sampling process is performed covertly, and they seek to determine if a given biometric sample corresponds to a restrained list of individuals of interest. This study is focused on video-surveillance applications.

The global market for video surveillance technologies has reached revenues in the billions of $US as traditional analog technologies are being replaced by IP-based digital surveillance technologies. Video surveillance networks are comprised of a growing number of IP-based surveillance cameras, and must transmit or archive massive quantities of data. In this context, video surveillance based on the facial biometric modality is extremely useful. The ability to automatically recognize and track individuals of interest in crowded airports or other public places, and across a network of surveillance cameras may provide enormous benefits in terms of enhanced screening and situation analysis. However, unlike fingerprint and DNA evidence, it is presently very difficult to perform fully-automatic recognition of individuals under surveillance using commercial video-based face recognition (FR) systems.

Biometric systems for automated recognition of faces in video streams have become relevant in a growing number of private and public sector applications. Despite the emergence of many commercial applications, the public sector (government, military, law enforcement, etc.) remains the principal user of biometric technologies for enhanced security. Applications range from open-set video surveillance or screening, where criminals or terrorists included in a watch list must be recognized within dense and moving crowds at major events and airports, to close-set access-control, where individuals enrolled in the system must by identified prior to accessing secured resources. Surveillance applications differ from closed-set identification in that the sampling process is performed covertly, and it seeks to determine if a given biometric sample corresponds to an individual of interest enrolled to a restrained watch list.

1.1 Examined applications

The PROVE-IT(FRiV) project aims at investigating the readiness of technologies for Face Recognition in Video (FRiV), in particular for the following video-surveillance applications:

[1] Other applications include classification or categorization (by gender, race, age) and facial expression recognition, which are not considered in this report.

1. screening of faces (screening against wanted list);
2. fusion of face recognition from different cameras;
3. face recognition-assisted tracking;
4. matching a face/person across several video feeds;
5. multi-modal recognition (e.g., face and voice)
6. soft-biometric based tracking/recognition

These applications are taxonomized into two key categories:

1. **still-to-video** recognition, which deals with matching of facial images in a video stream to still images stored in gallery, as used in *real-time watch list screening*.

2. **video-to-video** recognition, also referred to as **re-identification**, which deals with matching of facial images in a video stream to facial images captured in another video stream, as used in *search and retrieval*, *face tagging*, *video summarization*, and *face tracking* and *re-detection across multiple video streams*.

FR functions are assumed to be embedded as software executing inside some decision support system for intelligent video surveillance. For application scenario (1), an analyst would enroll individuals pre-process a gallery of still images as facial models, with one or more stills per person of interest. For application scenario (2), an analyst would gradually design and adapt facial models of interest over time, possibly during operations, based on videos analyzed from a particular scene or other sources.

This report presents a survey of the state-of-art academic systems that have been proposed for FRiV and serves as the basis for the work conducted within the PROVE-IT(FRiV) project. Techniques presented in this report are those that provide good results on well known reference video data sets and can be used to provide foundations to develop a surveillance system in-house.

The report is organized as follows. First, we present an architecture of a generic systems for FRiV (Section 2). Then, we present a comparative overview of the reviewed techniques and discuss specific academic systems that are found most suitable for FRiV (Section 3). Finally, the analysis of future trends in face recognition for video surveillance applications is presented. The additional information related to way FR technology works and the PROVE-IT(FRiV) project is provided in the Appendix, which contains the presentations from the project kick-off and project final meetings held in Ottawa in 2011 and 2013 under the umbrella of the interdepartmental Video Technology for National Security (VT4NS) forum.

2 A Generic System for Face Recognition in Video

Figure 1 presents a generic biometric systems for automated recognition of faces in video. Assume that video streams are captured using a network of IP cameras with fast a Ethernet interface, and that computer analysis is performed at a distance. Each camera captures streams of 2D images or frames, where each one

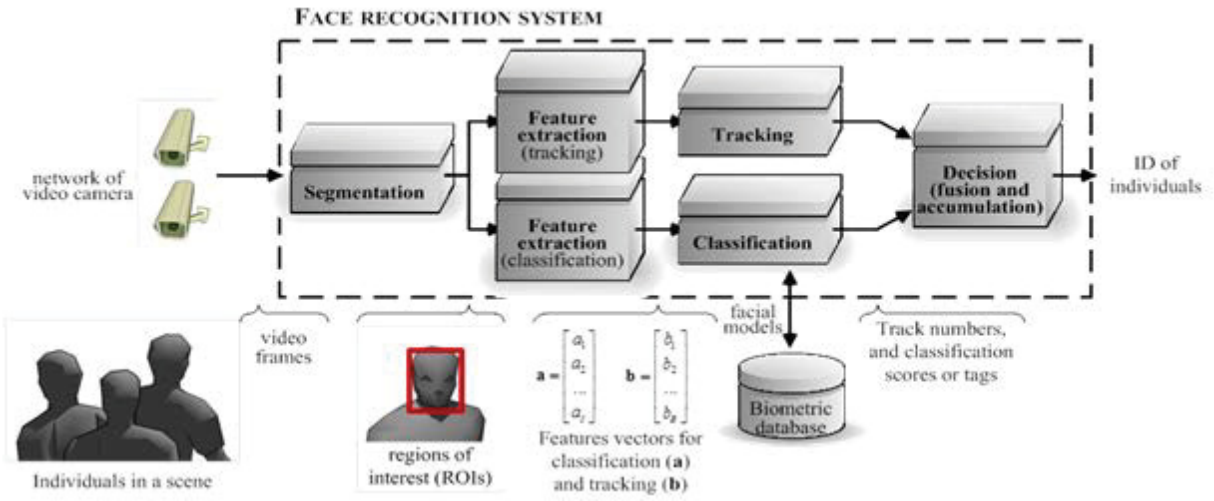


Figure 1: A generic biometric system for video-based face recognition.

provides the system with a particular view of individuals populating the scene. First, the system performs segmentation to isolate regions of interest (ROIs) corresponding to the faces in a frame, from which features are extracted and selected for appearance classification and motion tracking functions. For classification, invariant and discriminant features are assembled into an input pattern, **a**, that corresponds to a spatial vector or an ordered sequence of measurements. Tracking features can be the position, speed, acceleration, and track number assigned to each ROI on the scene. The tracking function follows the movement or expression of faces across video frames, while the classification function matches faces to the models of individuals enrolled to biometric the system.

During the enrollment, one or more reference patterns **a** are captured for an individual, and employed to design his user-specific facial model that is stored in a biometric database. Recognition is typically implemented using a template matcher or using a neural or statistical classifier trained a priori to map the input pattern space to one of N predefined classes, each one corresponding to an individual enrolled to the system. Each facial model may consist of a set of one or more templates (reference captures) for template matching, or consist of a statistical representation of reference captures for neural or statistical pattern classification. With neural network classifiers, for instance, the biometric model of an individual consists of synaptic weights and network architecture. Regardless of the application, face recognition may be modeled in terms of user-specific detection problems [6], each one implemented using one- or two-class pattern classifiers with thresholds applied to classification scores [42, 4, 7].

In still-to-video applications, facial models used for classification are designed using one or more ROIs from reference still 2D images or photographs. Typical law enforcement watch lists are usually comprised of mugshots, which can be used to create a subject gallery for watchlist screening. Many of these products and

researches extend still-to-still face recognition techniques to individual frames in video sequences. Video-to-video applications differ in that facial models are designed using multiple ROIs isolated in reference video streams. Facial models used for classification may be initially designed by capturing one or more reference video streams. These applications use video sequences to generate the gallery and provide a large variation of intra-class samples for each subject, taking advantage of variations in pose and facial motion provided by video sequences. In this context, an analyst may decide to enrol individuals of interest in some video stream, and then recognize and track their activities over multiple video feeds (from various cameras).

During the operation, input patterns **a** are matched against the model of individuals enrolled in the biometric system. The resulting classification score $S_i(\mathbf{a})$ indicates the likelihood that pattern **a** corresponds to individual i, for $i = 1,2,...,N$, and is compared against decision threshold, T_i, to provide an application-specific decision. In surveillance applications, the system outputs a list of all possible identities. To reduce ambiguities during the decision process, some features are also assembled into an input pattern **b** for tracking of an individual's motion or appearance over successive ROIs. The decision module may integrate the responses from tracking and classification modules over time. The following subsections discusses further details of blocks indicated in Figure 1. Relevant approaches that will provide a taxonomy to categorize FRiV academic research and COTS are also indicated.

2.1 Face detection

Faces are detected on images by successively scanning an image for a region (called a *window*) that potentially contains the object, at different sub-sampling levels (which are obtained by resizing the image). The most popular approach to extract facial regions of interest (ROI) from the image is the Viola-Jones algorithm [57]. It is based on the Adaboost classifier using a small set of critical features to train a cascade of classifiers to detect a face, where each new level complexity is increased to provide a high-accuracy result. It is well known for its accuracy and low processing time. Three Haar basis features are used, as depicted in Figure 2. The value of two-, three- and four-rectangle feature is the difference between the sum of pixels inside the white and black areas. This approach is used in [14, 19]. To minimize the calculation of features, only features that best separate negative from positive examples are considered part of the classifier cascade. The goal is to create a structure similar to a decision tree (Figure 3, where at earlier stages many of the rectangle features are rapidly discarded, and at further processing stages, more complex features are used but on fewer features to produce a final decision.

Rowley and Kanade proposed an object detector and demonstrated applications to detect cars and faces[47]. Their approach use multiple neural network classifiers to detect different orientations, successively scanning the image at different locations and sizes. This scanning approach, depicted in Figure 4, is also used by the Viola-Jones algorithm. Each neural network classifier is based on the statistics of localized parts, a transform from a subset of wavelet coefficients to a discrete set of values. Once the classification is done, an arbitration stage takes place to decide overlaps on detected faces. This face detection approach is used in the PittPatt face recognition platform, which was recently acquired by Google Inc.

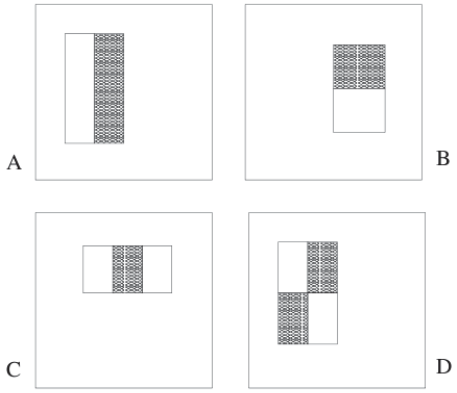


Figure 2: Rectangle inside the detection window. The sum of pixels within the white area is subtracted from the sum of pixels inside the black area. Extracted from [57].

2.2 Feature extraction and selection

The feature extraction and selection process will extract discriminant features from ROIs to use them at the classification. Features are divided in two types, physiological or behavioural. Physiological features relate to the face aspect, and are further categorized as *global* (or holistic) and *local* features. Sometimes both feature types may be combined in a face recognition system. Behavioural features are exclusive to video sequences, creating templates based on the face behaviour, the position of eyes and mouth, for instance, over time. Zhou *et al.* used this approach in [60], which is further discussed in Section 3.2.

The most used global feature is based on the principal component analysis (PCA) and the eigen-faces features introduced by Turk *et al.* in [55]. Images are resized to a common size (fixed for the application) and the resized facial image is decorrelated using gray-scale pixel intensity to produce a new set of pixel values (the PCA eigen values) that is used for actual classification. Feature dimensionality is of concern, as large images will require a very large processing time. In order to reduce feature set size and select only the most discriminant features, a process known as the feature subset selection, the resultant eigen vector from the PCA, is used to sort the features in relevance order to select a subset of relevant features. Both [14, 19], discussed in Sections 3.6.2 and 3.6.3, use the eigen-faces features and the resulting PCA eigen-vector to select discriminant features.

Local features relate to the invariant aspects in the face, such as eyes, mouth and nose position. The most used local feature is used for elastic bunch graph matching [52, 8], which uses a set of landmarks (or anchor points) to define points of interest that are used to build a graph that describes the face. Local features are extracted from jets within each region in the graph. Each jet component is the filter response of a certain Gabor wavelet extracted at a point (x,y) within the region, producing a feature vector that describes the points surrounding each point (x,y) that are selected regarding the graph nodes. This process generates a template that is used for matching. To allow the representation of different poses, one template

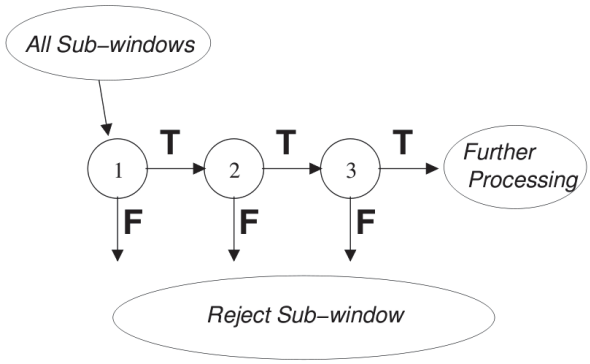


Figure 3: Detection cascade in the Viola-Jones algorithm.

for each different pose is needed. This led to the development of 3D techniques that map 2D faces to a 3D mesh representation to correct the face pose to a pose looking straight to the camera, which is performed by Animetrics on their FaceR technology and by Genex on their SureMatch 3D [20]. The academic system implementing this approach is the CSU Face Identification Evaluation System, described in Section 3.1.

Another well known local feature used in recent academic research is the local binary pattern, proposed by Hadid [24]. The local binary pattern is a lighting invariant feature extractor based on neighbour pixels to a reference point over a window of interest that is used to map the image, as in Figure 6. The windows define a square region, where the center point is is used to calculate the local binary pattern feature according to Figure 7 and Equations 1 and 2. Figure 7 shows the calculation for 8 points to define a circle ($P = 8$) on an area with a 1-pixel wide radius ($R = 1.0$). Each pixel in the circle is subtracted from the center pixel (interpolation is used when pixels are not integer values) according to equation 2 to calculate a threshold level, which is used in equation 1 to calculate the pixel value multiplied by 2 to the power of the current pixel count. Summing all pixel values yields the local binary pattern feature value for this region. Besides being invariant to lighting conditions, the local binary pattern is also very fast to calculate.

$$LBP_{P,R} = sum_{p=0}^{P-1} s(g_p - g_c) 2^p \tag{1}$$

$$s(x) = \begin{cases} 1, if x \geq 0, \\ 0, otherwise \end{cases} \tag{2}$$

If needed, feature selection to use only the most relevant features with the local binary pattern may be performed using PCA through the eigen-vector values, the same approach that is also used with global (holistic) features. The approaches detailed in Sections 3.6.2 and 3.6.3 use an hybrid approach, combining both eigen-faces (holistic) and local binary pattern features and selecting the most relevant features through PCA.

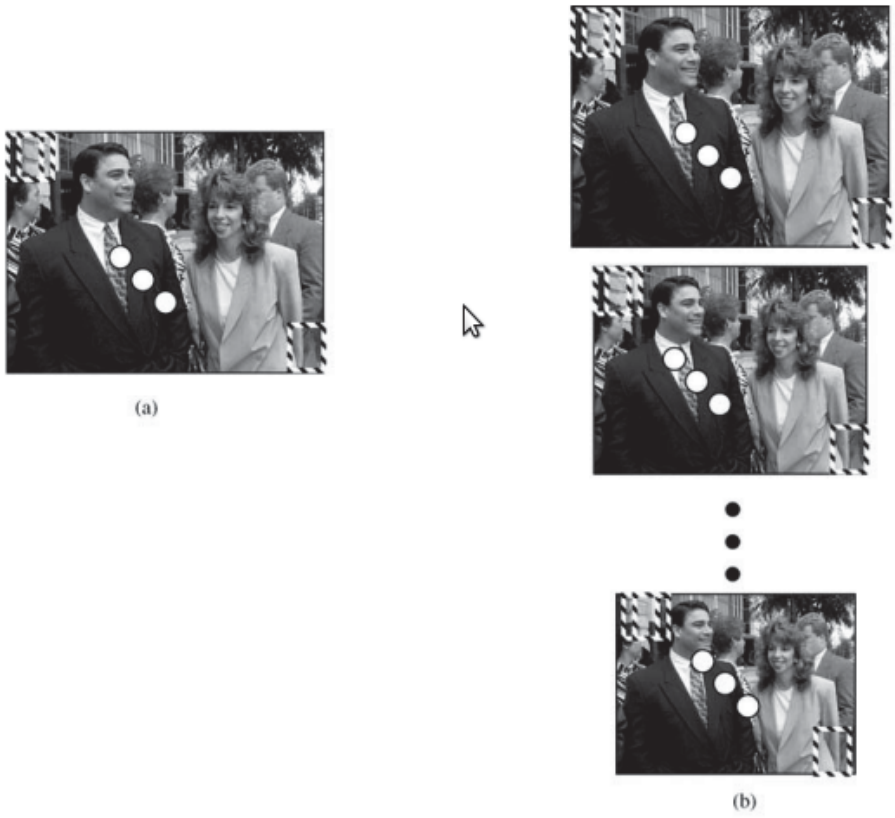


Figure 4: Sampling at position (a) and at different image scales (b), extracted from [47].

2.3 Face matching

Matta and Dugelay discussed in [39] several feature extraction approaches which target either the face physiology or its behaviour over time. Face physiology is a traditional approach used on static face recognition and has been extended to video and is divided in two categories: based on *local features* and based on *appearance features*. A local feature is computed based on the relationship between invariant aspects (landmarks) in the face, such as the eyes and nose to extract feature jets (Gabor jets, for instance) between pairs of landmarks so that each landmark has a vector of jets between every other landmark for comparison. On the other hand, appearance feature based approach uses the pixel intensity of the whole face and is a *holistic* approach. Two appearance features widely used are *Eigenfaces* [55], *Fisherfaces* [5], which uses principal component analysis (PCA) and linear discriminant analysis (LDA), respectively. Another well known, more recent, local feature is the *local binary pattern* (LBP) [24], which has been well adopted for being invariant to lighting conditions.

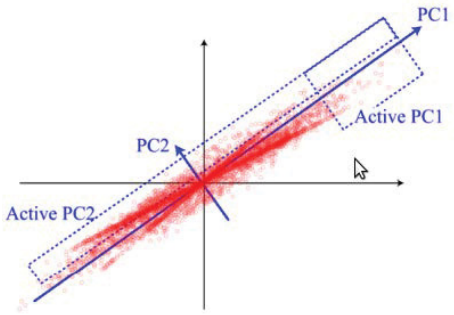


Figure 5: Linear decorrelation through PCA. The 2D feature space is aligned to axes that maximizes discrimination.

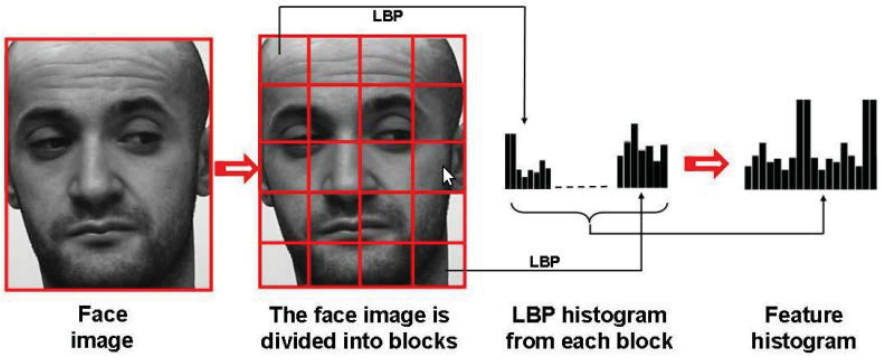


Figure 6: Local binary pattern applied over an image, extracted from [24].

Face behaviour features are specific to video sequences and model the facial movements as a discriminant feature. These features are relevant to video sequences owing to lower quality of video images when compared to full frontal still photos in traditional static face recognition. Facial movements are defined as the motion of specific face parts, such as the nose, eyes and mouth. Liu and Cheng used in [36] a hidden Markov model (HMM), whereas Zhou *et al.* [59] used particle filtering on a stochastic tracking and detection method.

Screening from an ROI extracted from a still image or video stream is finding whether the feature vector v (extracted from the face ROI) represents an individual on the watch list or not, using a template matcher or using a statistical or neural classifier. A template matcher traditionally compares face samples to templates in a gallery by finding facial features using a holistic approach or local approaches. Most commercial products, such as those detailed in report [20], rely on template matching. For the case of statistical or neural classifiers, one relevant approach is to exploit a set of face samples not belonging to individuals of interest,

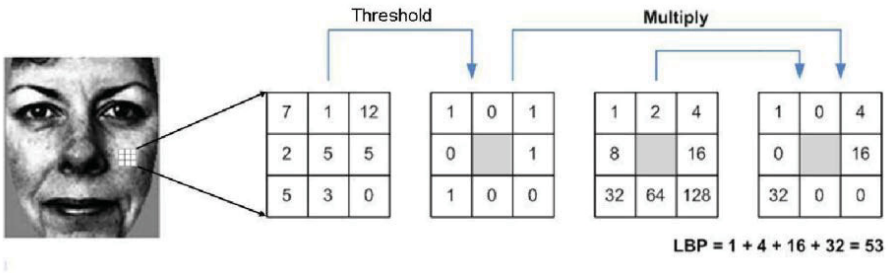


Figure 7: Example of LBP calculation, extracted from [24].

the *universal model* (**UM**). This way, the n–class problem ($n = |WL| + 1$) has to determine the membership in distinct classes $t_i, t_i \in WL$, or to the universal model ($UM \cup WL =$). Another option is to train a n–class classifier and use data from the universal model to determine class specific thresholds in order to have a rejection option for unknown users in the system. A template matcher is usually more time consuming than a classifier, because it must compare the probe samples to all templates in the gallery. Adding or removing templates from a gallery is however a fast procedure compared to re-training a classifier. Using modular classification architectures (one classifier per individual in the watch list) will improve processing time for updating.

Several powerful techniques have been proposed to recognize faces in still images [58]. A common approach to recognize faces in video consists in exploiting only spatial information (*i.e.*, appearance), and applying extensions of static image-based techniques on high-quality face images produced through segmentation. The predominant techniques are appearance-based methods like Eigenfaces, and feature-based methods like Elastic Bunch Graph Matching [58]. More recently, some authors have exploited temporal information contained in video sequences to improve performance of video-based face recognition. For example, track-and-classify systems (as the one shown in Figure 1) combine spatial information with information on motion and appearance of faces in a scene [14]. Regardless, the performance of these techniques may degrade considerably when applied in real-world applications.

2.4 Tracking and decision modules

The tracking module starts following a face after it has been first detected. This module provides information to face annotation applications, enrollment from video applications and may provide relevant information for the decision stage on face identification applications. A face annotation application needs to track each face to consistently assign a unique tag for each unique face across the video sequence. Similarly an application allowing enrollment from a video sequence needs to follow the face over video so that only faces from the correct individual are enrolled. Finally, the decision module can use the tracking information to provide a more accurate decision over time, accumulating a decision score, instead of deciding on a single frame. Moreover, some research specifically targeted for video-to-video FR [34, 60, **?**] tackles tracking and

recognition in the same process using a *tracking-and-recognition* approach.

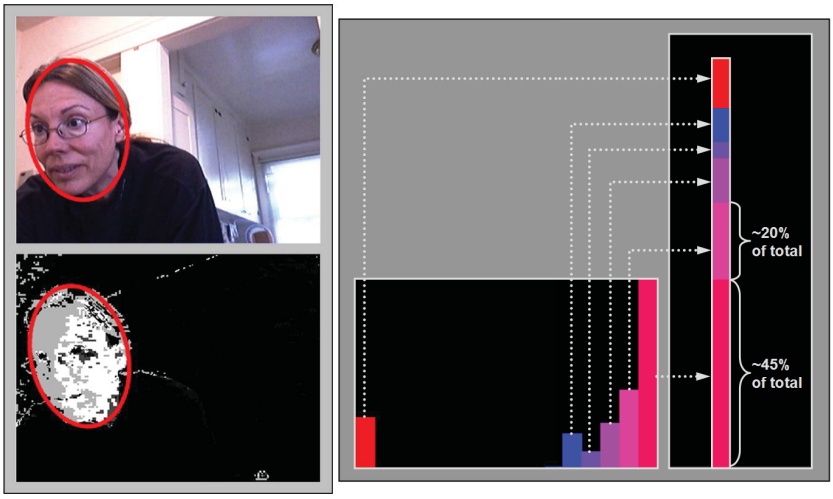


Figure 8: Camshift algorithm: the face image region has it's color histogram extracted, and the updated position is obtained by identifying a region close to the previous frame that matches the same color histogram.

Several face tracking approaches are found in the literature. One well known algorithm is the *Continuously Adaptive Mean Shift* – CAMSHIFT [11], an improved variation of the original Mean Shift algorithm [18, 2], for being implemented in the popular OpenCV library [10]. The algorithm calculates a color histogram to represent a face, so that on incoming video frames it calculates the probability of a pixel representing a face (see Figure 8). Based on those probabilities, the face window is shifted, and its current angle and size is calculated. The algorithm is known to perform reasonably well, but has trouble when the object followed is close to objects with similar colors. Other common approaches for face tracking are to use Kalman Filters [3, 54] or particle filters [1], also known as the CONDENSATION algorithm. Particle filter trackers establish density models to describe the system's state and noise.

This module is implemented in most commercial products targeting video surveillance applications, but providers usually make no claim as to which specific method is applied for tracking. Academic research traditionally focus on individual modules, thus a system built over academic research has to select from the available research which tracking method to use.

Table 1: Summary of academic systems proposed for FRiV.

Author	Name	Approach	Features	Learning	Set	Tracking	Application
Zhou et al.[60]	Simultaneous Tracking & Recognition	video-to-video	global	classifier independent	closed	yes	1, 3
Ekenel et al.[16]	Local appearance-based face models	video-to-video	local	classifier independent	open	no	1
Kamgar-Parsi et al.[27]	Face Morphing to Boost Training Data	still-to-video	local and global	classifier independent	open	no	1
Li et al.[35]	TCM-kNN	still-to-still	global	batch	open	no	1
Polikar [45]	Learn++ (boosting)	video-to-video	global or local	incremental	open	no	1
Connoly et al. [15]	Adaptive MCSs	video-to-video	global	incremental	closed	no	1
Gomerra et al.[19]	Adaptive MCSs	video-to-video	global	incremental	closed	no	1

3 Overview of academic face recognition approaches

Several powerful techniques have been proposed to recognize faces in still 2D images [58]. Over the past 15 to 20 years, results from the Face Recognition Vendor Tests (FRVT) conducted by NIST [23, 22] have reported two orders of magnitude improvement in performance in face recognition under frontal and controlled conditions. However, the rapid and discrete recognition of faces captures from CCTV or IP-based video cameras remains a very challenging problem, especially in unconstrained and cluttered environments. Faces captured in video frames are typically low quality and generally small. Furthermore, faces acquired from semi- and unconstrained scenes may vary considerably due to limited control over operational conditions (e.g., illumination, pose, facial expression, orientation and occlusion), and due to changes in an individual's physiology (e.g., aging) [39, 59].

Table 1 provides a summary of the publications that were reviewed in the PROVE-IT(FRiV) study. To categorize FRiV systems and approaches, the following properties are analyzed:

- **Approach type** – a desirable FRiV for the tasks described in the Introduction should either use a *still-to-video* approach, to allow the use of still photos on the watch list, or a *video-to-video approach*, to be able to learn short video sequence of targets for identification and tracking. Some systems may also support both approaches.

- **Feature type** – features extracted can be either physiological or behavioral. As physiological features are further divided as local and global features, feature types are categorized as *physiological local*, *physiological global*, *physiological hybrid* (when it combines both local and global features) and *behavioral*.

- **Matching type** – indicates whether a template matching or a classifier based approach (either statistical or neural) is used.

- **Set** – watch list based screening requires the system to deal with an *open* set problem, whereas many systems, both academic and commercial, still work with *closed* set problems, such as identity verification.

- **Tracking** – indicates whether the system/technique has embedded tracking and if it's target for single or multiple faces. If it does not have tracking, an additional module needs implementation to allow face tracking.

- **Applications** – The video surveillance applications of interest defined in the Introduction.

The "set" property indicates suitability for open set applications. Finally, the "applications" property summarizes how the solution can be used in surveillance applications examined in this project.

In the following we discuss specific academic systems suitable for FRiV and provide recommendations on systems for benchmarking.

3.1 CSU Face Identification Evaluation System

Developed at the Colorado State University (CSU) by Beveridge *et al.* the Face Identification Evaluation System[2] is a tool for baseline benchmarking and comparison. Among the available matching algorithms, it includes an implementation of the elastic bunch graph matching algorithm [52, 8].

3.2 Simultaneous Face Tracking and Recognition

Proposed by Zhou *et al.* in [60], this video-based face recognition approach tackles simultaneously the tracking and recognition approach in Figure 1, whereas most research solves both problems as separate issues. The statistical model used has three components, one motion equation to track the face, one identity equation to govern the temporal evolution of the identity variable and an observation equation to link the motion and identity equations, which is used to determine the likelihood to a template in the gallery.

The main contribution of this approach is the use of video sequences to create the template gallery, instead of few still pictures. This approach increases template diversity and includes natural motion information, which is how a target is observed on video. While video is the optimal medium for enrollment, still images can also be used for this task, but with reduced recognition efficiency. The proposed approach is not developed over one specific classification algorithm, thus the learning strategy depends on the actual classifier used on some implementation. Finally, tests detailed in [60] are only on closed set identification problems.

Performance was evaluated on three data sets, two from the University of South Florida and one extracted from the MOBO data set. Recognition rates with the MOBO data set averaged 95%, with values ranging from 88% to 100%. Difficulties faced by the authors are related to subjects looking away from the camera, specially individuals that walked looking down.

3.3 Local Appearance-Based Face Models

Ekenel *et al.* discussed in [16] a two-level approach for a screening system. It first uses a face recorder module (Figure 9), that finds and tracks faces over time. It tries to find faces using skin tone (a local appearance feature) and, once a potential match is found, it tries to find the eyes to confirm that its actually a face. If a face is found, it stores the image and updates the skin model and tracking information to process the next frame. This module combines the functions of the segmentation and tracking modules in Figure 1.

The second module is a face recognition module (Figure 10), which uses images stored by the face recorder (previous module) for actual classification. It performs a more refined face detection technique, which involves aligning the eyes before actual feature extraction of local facial features through discrete cosine transform. Classification itself may use any classifier, and experiments were made using both kNN and a Gaussian mixture model. This module relates to the classification module in Figure 1.

Experimental results on custom collected data provided true positive rates of 70% for the kNN and 80% for the Gaussian mixture model for a 10% false alarm rate. Whereas results are not especially remarkable,

[2] http://www.cs.colostate.edu/evalfacerec/algorithms5.php

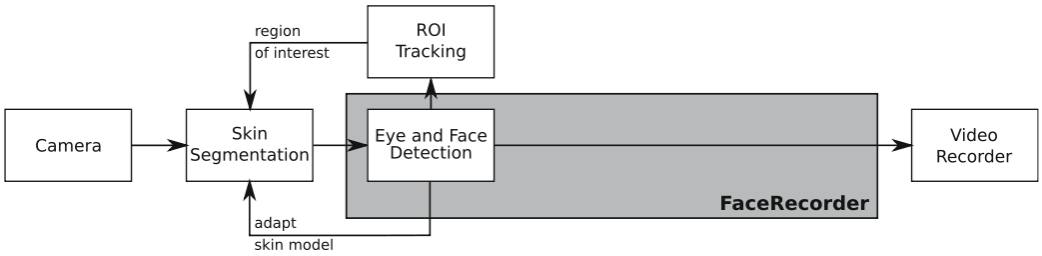


Figure 9: Face recorder module, extracted from [16].

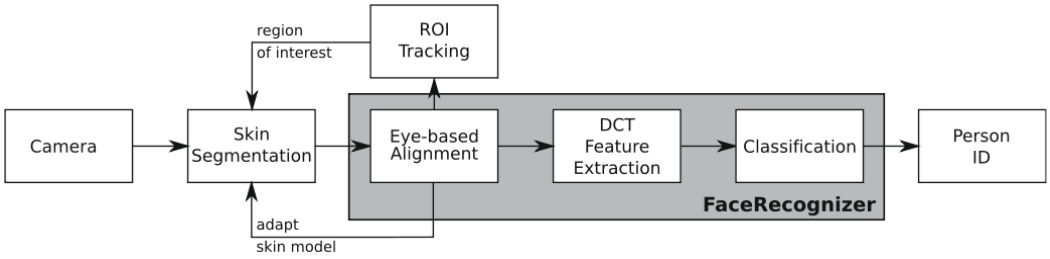


Figure 10: Face recognition module, extracted from [16].

the system demonstrates the typical separation of the tracking and recognition tasks, unlike the approach proposed by Zhou in [60]. More advanced two-level adaptive approaches is discussed later in this section.

3.4 Face Morphing to Boost Training Data

This approach, proposed by Kamgar-Parsi *et al.* in [27], follows the multi-classifier approach outlined in [6], where each user in the system is represented by a distinct classifier. However, the main contribution of this approach is how it artificially enlarges the template samples to improve inter-class separability for classification through morphing. Figure 11 shows the feature points located by the Active Shape Model algorithm, which are the used as anchor points to generate the morphed images in Figure 12, gradually transforming the Jennifer Aniston face to Angelina Jolie's face.

The goal of such morphing approach is to generate what the authors call bordeline accept and borderline reject samples. Considering Jennifer Aniston's face in 12, the original sample, in the leftmost image, is very similar to the first morphed image towards Angelina Jolie, as most facial features of the original sample are still present. This is a borderline accept sample, the farthest from the original sample the classifier should accept. However, the next sample, the borderline reject, is a sample that has facial features of both original samples, but is not similar to any of them and should be rejected. For a two-class classifier, the first two images belong to the positive class, whereas the next three belongs to the negative class.

The system was compared to the TCM-kNN algorithm [35] and provided better results, reaching almost

Figure 11: Features points for morphing, extracted from [27].

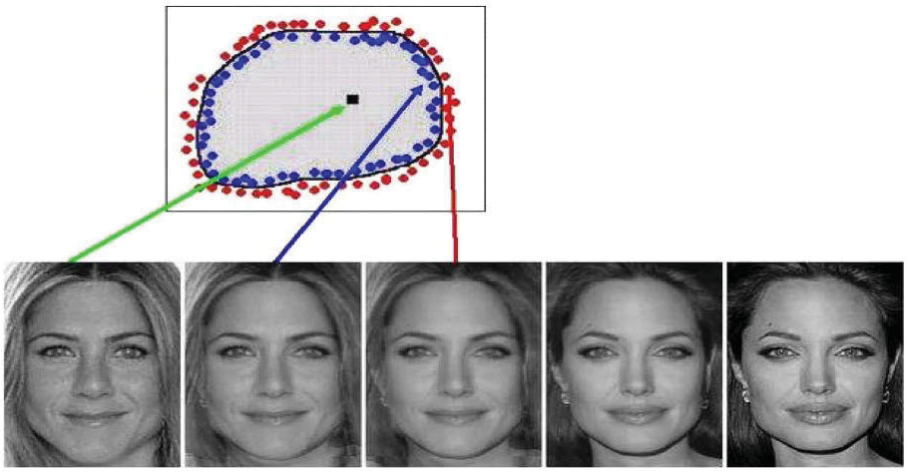


Figure 12: Morphed faces – from Jennifer Aniston to Angelina Jolie, extracted from [27].

100% true positive rate, with very low false positive rate. But the test was performed with data sequences extracted from the NRL and FERET data sets, and no details on how images were distributed is provided in the paper, therefore further investigation with standard data sets is desired. Finally, the system is good for still-to-video classification, but a video-to-video approach is unfeasible owing to the combinatorial explosion of morphing images between samples. Also, enrollment of new (or temporary) users requires more time than traditional approaches to generate the morphed images.

3.5 Transduction Confidence Machine kNN

Li and Wechsler's **Transduction Confidence Machine kNN** (TCM-kNN) is an adaptation of the classical kNN classification algorithm, a typical classification module in Figure 1. Instead of using simple distance measures, the TCM-kNN algorithm calculates a strangeness measure between class samples in the gallery

and probe images. Unlike the original kNN, which is a very fast classifier updating a desired feature for dynamic environments, the strangeness measure is calculated between each training sample and all remaining training samples. Thus, updating the classifier requires recalculating the entire set of strangeness values, not only values associated with new data.

Performance tests on the FERET data set provide classification accuracy of 91.14% for a watch list with 40 individuals. The internal tests conducted by Laboratoire dimagerie, de vision et dintelligence artificielle (LIVIA) at ETS have verified that the TCM-kNN can be used on video-to-video watch-list based applications, but the required time to calculate the strangeness measure for a high number of samples may not be adequate for an actual real-time surveillance system.

3.6 Adaptive Multi-Classifier Systems

Ensembles of classifiers, also called Ensembles of Detectors (EoD), have been often used to improve accuracy of classification and detection. The same approach has been successfully used with face recognition applications. At enrollment, a new classifier or ensemble of classifiers is created for each target face. These approaches are used as the classification module in Figure 1.

Bengio and Mariéthoz discussed in [6] the use of an EoD-based modular approach, where each user enrolled in the system is represented by one classifier, or an ensemble of classifiers. This approach allows for easier modular classifier model updating, as only the model corresponding to the updated target needs to be modified. As also discussed in report [21], EoD-based modular approach is particularly suitable for real-time watch list screening applications, where the recognition decision is based on the accumulation of all data corresponding to a target individual, as this individual is visually tracked over time by a camera.

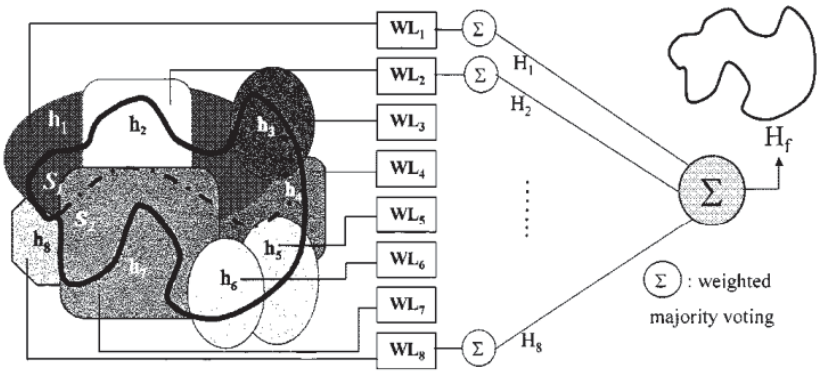


Figure 13: Majority voting combination of classifiers trained with different data, extracted from [45].

3.6.1 Learn++

The Learn++ algorithm, proposed in [45], is a multi-classifier algorithm, in the sense that every user in the system is modelled as a two-class classification module, as in [6], and each module is an ensemble of classifiers combined through weighted majority voting. The major feature in this classifier is the incremental procedure used, which uses new data to train one new base classifier (a MLP classifier) that is added to the existing ensemble and combined through the weighted majority voting approach. This approach is illustrated in Figure 13, where different training data $(h_1,\ldots,h_8)$ is used to train classifiers $(WL_1,\ldots,WL_8)$ and approximate the H_f hyper-space. Tests performed do not include face recognition problems, but the algorithm is appropriate for this type of application, and results demonstrate that its performance is similar to the same base classifier trained on batch mode.

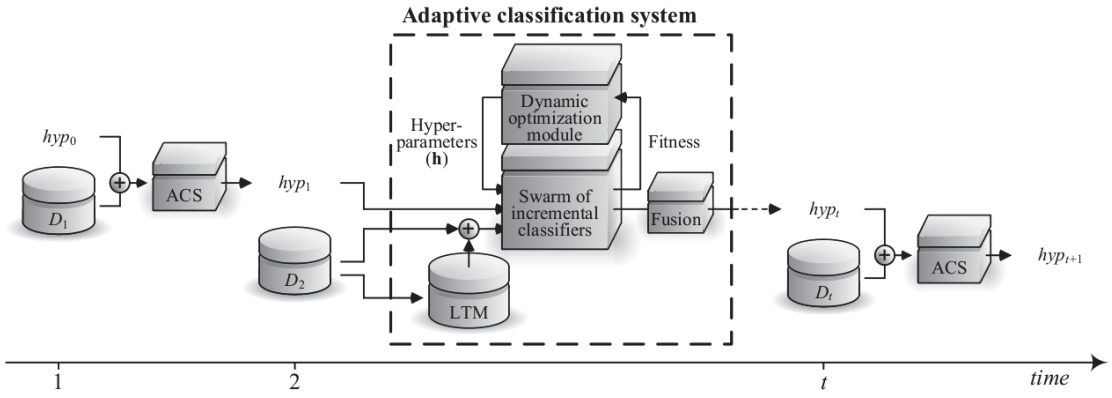


Figure 14: Adaptative classification system, extracted from [14].

3.6.2 Evolving Ensembles Based on Dynamic PSO

Connolly *et al.* proposed in [14] the Adaptive Classification System (ACS), which allows incremental learning without the need to retrain the system on batch mode with all data. Unlike in Learn++, which adds a classifier to a growing ensemble, Connolly's approach uses dynamic niching particle swarm optimization (dnPSO) to update classifiers in an ensemble with new data, as detailed in Figure 14. Each classifier in the ensemble is a particle optimized by the dnPSO algorithm. Instead of selecting the best classifier in the swarm, classifiers that converged to different maxima are combined.

The system uses a long-term memory to avoid knowledge corruption when updating base pFAM classifiers in the ensemble. Instead of using exclusively new data, data stored from previous updates is combined to new data to update classifier hyper parameters. At each incremental update, a portion of the new data block is used to update the long term memory. Experiments are performed with the NRC-IIT [**?**] and

MOBO datasets [3] and provided comparable results to a pFAM classifier trained on batch mode (with all data). All tests were performed on closed-set problems, but the system may (at least theoretically) be used with open-set problems.

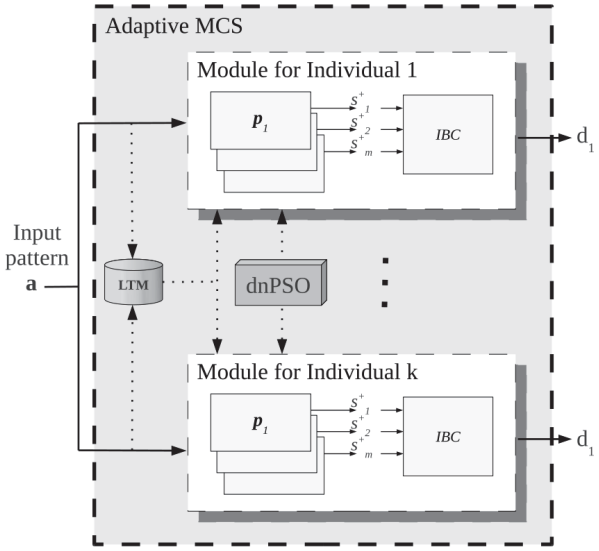


Figure 15: Learn and combine adaptive multi-classification system, extracted from [19].

3.6.3 Learn and Combine Face-Based Video Surveillance

Proposed by Gomorra *et al.* in [19], this adaptive multi-classifier system also uses the modular approach described in [6], with an ensemble of classifiers per module as detailed in Figure 15. It is similar to some extent to Learn++, in the sense that when adapting the system to new training data a new classifier is trained, but, instead of using a weighted majority approach to combine classifiers in the ensemble, classifiers are combined through iterative Boolean combination (IBC) [29], a decision level fusion approach. This provides a broader selection of operational points for selection, which is desirable for actual systems. Figure 15 also indicates the use of a long term memory to avoid classifier knowledge corruption, similar to the approach used in [14].

Experimental results in [19] were obtained with the FIA data set on a watch list based screening application with 10 individuals, and the remaining individuals in the data set used either as negative class samples or as unseen data by the classifiers (also negative class, but only for testing, not training). Results demonstrated similar performance levels to those of batch classifiers, but with fast update time and good compression levels. The system was designed for video-to-video operation, but still-to-video operation is

[3] For the datasets used for FRiV evaluation, see report [21].

possible, and experiments using a very limited number of class samples per update produced promising results as well.

3.7 Conclusions

Through the analysis of academic literature in the last few years it is found that the volume of research on holistic approaches based on neural networks and statistical face classifiers has surpassed the volume of research on pure template matching and local feature based approaches for FRiV. A comprehensive comparative evaluation is required to compare the performance of these two approaches. However, based on the reported performance of academic systems, the template matching based systems (such as used in COTS technologies) should be outperformed by state-of the art holistic pattern recognition techniques such as:

- **TCM-kNN** [35], which is an statistical classifier adapted to open set problems.
- **Adaptive Multi-Classifier Systems** [19], which are modular approaches that design Ensembles of Detector (EoD), one per target face.

Compared to traditional approaches used in COTS products, these two approaches allow for more efficient accumulation of facial data over time, which is very important for video-based applications, where the lack of detail in spatial domain (measured in image quality and resolution) is compensated by the abundance of information coming from the temporal domain (ie. number of frames captured over time) [**?**, **?**]. They also allow the efficient development of target-based screening systems that can be optimized for each target in a watch list and that are robust to the scale of the traffic (i.e. the number of faces captured by the camera).

4 Future Trends

4.1 Modular Architectures

Several specialized architectures have been proposed for FR in video surveillance [43]. Most notably, the open-set Transduction Confidence Machine-*k*NN (TCM-*k*NN) algorithm [35] modified the traditional *k*NN, using transduction to calculate a measure of strangeness between samples. These systems have also been modeled in terms of user-specific detectors, each one implemented using one or more binary (1- or 2-class) classifiers [43]. This modular approach was employed with user-specific SVMs [16] and ensembles of 2-class ARTMAP neural classifiers[43]. Binary ensembles are justified by the limited amount of positive samples for design, and by the complexity of real-world video scenes [43]. However, these architectures do not consider or exploit class imbalance information to enhance performance. This information is relevant in the context of video surveillance, owing to the potentially small number of positive samples w.r.t. the negative ones, and to changing operational conditions.

In other similar biometric applications, the use of a Universal Background Model (UBM), a negative class generated from samples of other unknown sound from which the target voice is discriminated, as well

as a Cohort Model (CM), a negative class representing voices from other people, has also been proposed for open-set applications [12]. The use of such CM can be of a definite interest when using class-modular architecture. Regardless, these architectures do not necessarily provide a high level of performance. They do not consider or exploit class imbalance information positive (target) class samples w.r.t. the negative ones to enhance performance. Moreover, assimilating new samples over time can be a challenging task: if those new samples exhibit significant changes (pose, illumination, ageing, etc.) compared to the previously learned one, the past knowledge might be corrupted.

4.2 Adaptive Biometrics

Adaptive biometrics is an active research area in biometrics and video surveillance. Watch list based screening applications need updating from time to time, as either new individuals may be added to the watch list (WL), existing individuals may be updated with new photos, specially owing to ageing factors. When using template matching, the template gallery may become too large if simply adding new templates to the gallery is performed. Instead, template selection is used to determine which templates are relevant. For instance, clustering techniques have been used in the literature [56, 37] for template selection, to group together templates that are similar and separate those that present large intra-class variability.

Adaptive biometric systems allow for adaptation of its internal parameters and structure in response to new representative reference data. Some adaptive biometric systems have been proposed in the literature to refine biometric models according to the intra-class variations in input samples. Two well known algorithms for semi-supervised adaptation are the self-training and co-updating. In self-update methods, biometric models are first designed using a labeled data set, and then operational samples with a high degree of confidence are integrated for update. The notion of high degree of confidence is subjective, and depends on both the classification algorithm and the application, but in general a threshold on the similarity score is used. Co-update is a self-training version adapted to the use of two (or eventually more) classifiers, that will be trained to improve mutually. For that, each classifier should be trained on different views of the samples. The procedure starts with the design of the two classifiers on the labeled design data set. Once new samples are collected, both classifiers are used to label the samples and those with high degree of confidence (at least by one of the classifiers) are added to WL. A potential advantage of the co-update algorithm is that it can retrieve update samples that are not typical of the distribution of target data, and adaptation to abrupt changes is possible.

In supervised learning strategies, new data samples are assumed to be analyzed and labeled by a human operator with knowledge of the correct intra-class variations. For instance, labeled data becomes available when a system administrator requires multiple (re)enrollment sessions, separated by a given interval of time, or when he can analyst data off-line from operational scenarios. In a human-centric framework, a proficient administrator can gradually create and updates the biometric models of a system over time, as the operational environment unfolds. When using classifiers, the most simple approach is to use *batch learning*, which is to fully train classifiers with the new and previous data (photos and video sequences). This approach is time consuming and is an impediment to security applications that require enrollment from real-time video streams, such as following an individual over time based on a suspicious behavior. One way

to perform this task is by changing classifier hyper-parameters with the new data, which may include the use of a long term memory to avoid knowledge corruption [14]. when using an ensemble of classifiers, another approach is to train new classifiers with the new data, which are then combined with the previous model [28, 19].

4.3 Multi-modal systems

Evidence from several studies suggest that the accuracy and reliability of a biometric system can be improved by integrating the evidence obtained from multiple different sources of information [7, 31, 30, 26, 51, 46, 9, 17, 38, 49], including multiple samplings for a same biometric trait using different sensors, multiple different biometric traits, multiple instances and multiple samplings for the same biometric trait using a same sensor, or multiple feature extraction and classification algorithms processing the same biometric sample [25]. Various studies have also shown that poor quality biometric samples lead to a reduction in the accuracy during operations. Fusion controlled by quality measures has been shown to offer a significant gain in accuracy, but falls outside the scope of this paper [44].

Multi-modal biometric systems can mitigate certain performance and robustness limitations associated with single-modality systems. For instance, a system may capture face and voice data. Multiple biometrics are known to provide higher accuracy than one single-biometric system, especially if false negative rate fnr is an important consideration. While it is true that false positive rate fpr would almost certainly decline in multiple biometric systems, false negative rates fnr may also increase.

The fundamental differentiator in multi-modal system design is the level at which information from different biometric modalities is combined. Biometric sources of information are typically integrated at the feature, score and decision levels [53, 46, 9, 38, 49]. Feature-level multi-modal models utilize feature vectors from different biometric modalities to create a new feature vector, which is then utilized as the basis for matching. Since features extracted from sensor measurements contain richer information content about a biometric modality, feature-level fusion should provide the higher level of accuracy, although commercial systems rarely reveal their feature patterns. The combined feature patterns may also be incompatible and increase system complexity [32]. A single biometric can be easily spoofed, whereas spoofing becomes more difficult with multi-modal systems. Due to non-universality, each biometric modality is prone to failure to enroll of some samples, the use of multiple biometrics will alleviate this issue.

4.4 Fusion techniques

Techniques for score-level fusion are commonly employed in biometrics when scores generated by the different commercial systems may be accessed [26, 51]. They utilize system-specific scores resulting from comparisons from multiple biometric systems to generate a composite score used to differentiate impostor and genuine transactions. The primary advantage of this is that a system designer can specify optimal operating points for multiple systems, assign relative weights, and develop statistical models by which scores from divergent systems can be utilized to differentiate genuine and impostor score distributions. Most commercial biometric systems provide access to score data, such that other commercial algorithms

can be leveraged. Similarity score level fusion relies on the scores generated by each matcher(s) associated with the modalities involved. Scores are processed through a combination of normalization and fusion techniques. Of the three main approaches, score-level fusion provides the strongest balance of performance and commercial viability. The main limitations are the impact of score normalization methods on the overall decision boundaries, and the availability of representative training samples.

Despite reducing information, techniques for decision-level fusion may provide a simple and robust framework for combination, regardless of the specific type of biometric modality and system. Disadvantages include the limitations placed on decision boundaries due to the restricted operations that can be performed on binary decisions, and the need for independent data to design combination rules. Decision-level techniques utilize match decisions from more than one system to render a global decision. Typical decision-level multi-modal system logic includes the following:

- If system A = match and system B = match, then system (A+B) = match
- If system A = match or system B = match, then system (A+B) = match
- If system A = no match or system B = no match, then system (A+B) = no match

In contrast to verification and authentication application, multi-modal application to surveillance is limited. The biometric modalities that can be used for surveillance must be able to capture individual samples at a distance and under limited or no cooperation. In general, gate cues, voice or iris could be combined with face recognition in a semi-constrained surveillance environment. For instance, the combination of face and gait cues for identification purposes has been studied in [50], using both score level (product, sum, min, and max) and decision level (majority voting) fusion, and the Product rule has been shown to provide the best performance. Connaughton *et al.* [13] evaluated several score and rank-level fusion techniques on a multi-modal dataset of face and iris samples and showed that the Borda-count method using an exponential vote-weighting scheme achieved the best performance. Similar to academic research, a number of commercial products offer support for multi-modal biometrics as shown in our evaluation of commercial products [21].

4.5 Synthetic face generation

Another point that deserves attention is the that, if only one picture per subject is used for the learning process, then it is impossible for a Machine Learning (ML) system to build an accurate model face model. One promising technique to improve the quality of the face recognition results under such scenario relies on the generation of faces from one single figure, called Candide [48]. Simply put, Candide is a face mask that allows the generation of various faces positions, because it has a full 3D face description. Developers also allowed the generation of face expressions, as presented in the tool developed in [33]. This technology is useful to this problem because it gives a solution for two problems: 1) It provides means for affordable generation of new face positions without the necessity to have a special environment to capture various subject's pictures; and 2) it also allows to generate different facial expressions from the single still image, even if the subject does not cooperate.

Candide also allows face pose generation with partially visible faces. A possible situation that may happen is the case where the surveillance team has a video sequence, with the subject's face partially covered. The frame in question can be used to generate new poses, by adjusting the face mask to the face's angle appearing in the video, and Candide can generate new poses from this information. Of course, if the covered part has some mark, like a scar, Candide will not generate this mark on the unseen face part. These new poses can be added to the ML algorithm to improve its accuracy to detect that particular subject. Report [41] provides more information about Candide.

4.6 Non-binary decision making, triaging and fusion

When the recognition decision obtained on a single measurement or from a single modality is not fully reliable, it can be improved through the use of fusion of decisions, and using non-binary decision schemes such as *triaging*, where a system outputs one of three possible decision outcomes: "green" – for confident non-match, "red" – for confident match, and "yellow" for possible non-confident match.

One technique that can be used for developing traiging recognition systems is examined in [40], which uses the concept of smooth ROC to generate a "smooth" non-binary recognition decision. Another approach is presented in [21], where recognition decisions obtained by tracking a person in video are fused over time to generate the triaging decision.

4.7 Soft Biometrics

When faces are close to cameras, cooperative, constrained, they may be close to ICAO format, making it possible to apply conventional FR approaches. Otherwise, when faces are far, non-cooperative, non-contrained, and are not of sufficient quality for traditional FR technology, they may be suitable for soft biometrics, such recognition of person's age, gender, height, hair colour etc. Such recognized soft biometrics can then be used to further improve the overall performance of the system.

References

[1] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174 –188, feb 2002.

[2] S. Avidan, "Ensemble tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 261–271, February 2007. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2007.35

[3] Y. Bar-Shalom, T. Kirubarajan, and X.-R. Li, *Estimation with Applications to Tracking and Navigation*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

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

[5] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711 –720, jul 1997.

[6] S. Bengio and J. Mariéthoz, "Biometric person authentication is a multiple classifier problem," in *7th International Workshop on Multiple Classifier Systems, MCS*, 0 2007, iDIAP-RR 07-03.

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

[8] D. S. Bolme, "Elastic bunch graph matching," Ph.D. dissertation, Colorado State University, USA, 2003.

[9] K. W. Bowyer, K. I. Chang, P. Yan, P. J. Flynn, E. Hansley, and S. Sarkar, "Multi-modal biometrics: An overview," in *2nd Workshop on Multi-Modal User Authentication (MMUA)*, 2006.

[10] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[11] G. R. Bradski, "Computer Vision Face Tracking For Use in a Perceptual User Interface," *Intel Technology Journal*, no. Q2, 1998.

[12] A. Brew and P. Cunningham, "Combining cohort and ubm models in open set speaker detection," *Multimedia Tools Appl.*, vol. 48, pp. 141–159, May 2010.

[13] R. Connaughton, K. W. Bowyer, and P. J. Flynn, "Fusion of face and iris biometrics from a stand-off video sensor," in *Proceedings of The 22nd Midwest Artificial Intelligence and Cognitive Science Conference (MAICS)*, Cincinnati, Ohio, USA., 2011, pp. 99–106.

[14] J. Connolly, E. Granger, and R. Sabourin, "An adaptive ensemble of fuzzy artmap neural networks for video-based face classification," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*, july 2010, pp. 1 –8.

[15] ——, "An adaptive classification system for video-based fr," *Information Sciences*, 2010.

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

[17] M. Faundez-Zanuy, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal biometric databases: an overview," *Aerospace and Electronic Systems Magazine, IEEE*, vol. 21, no. 8, pp. 29–37, 2006.

[18] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *Information Theory, IEEE Transactions on*, vol. 21, no. 1, pp. 32 – 40, jan 1975.

[19] M. A. Gomorra, E. Granger, P. V. W. Radtke, R. Sabourin, and D. Gorodnichy, "Incremental update of biometric models in face-based video surveillance," in *submitted to the World Congress on Computational Intelligence 2012*, 2012.

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

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

[22] P. Grother and M. Ngan, "Face Recognition Vendor Test (FRVT). performance of face identification algorithms," NIST Interagency Report, Tech. Rep. 8009, 2014.

[23] P. J. Grother, G. W. Quinn, and P. J. Phillips, "Report on the Evaluation of 2D Still-Image Face Recognition Algorithms ," NIST Interagency Report, Tech. Rep. 7709, 2010.

[24] A. Hadid, "The local binary pattern approach and its applications to face analysis," in *Image Processing Theory, Tools and Applications, 2008. IPTA 2008. First Workshops on*, nov. 2008, pp. 1 –9.

[25] A. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *Information Forensics and Security, IEEE Transactions on*, vol. 1, no. 2, pp. 125 – 143, june 2006.

[26] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270 – 2285, 2005.

[27] 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.

[28] M. N. Kapp, R. Sabourin, and P. Maupin, "A dynamic optimization approach for adaptive incremental learning," *International Journal of Intelligent Systems*, vol. 26, no. 11, pp. 1101–1124, 2011.

[29] W. Khreich, E. Granger, A. Miri, and R. Sabourin, "Adaptive roc-based ensembles of hmms applied to anomaly detection," *Pattern Recognition*, vol. 45, no. 1, pp. 208–230, 2012.

[30] T.-K. Kim and J. Kittler, "Design and fusion of pose-invariant face-identification experts," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 16, no. 9, pp. 1096–1106, 2006.

[31] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[32] A. Kumar, V. Kanhangad, and D. Zhang, "Multimodal biometrics management using adaptive score-level combination," in *ICPR*, 2008, pp. 1–4.

[33] H.-S. Le and H. Li, "Face identification system using single hidden markov model and single sample image per person," *Proc. of the IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, vol. 1, no. 1, pp. 330 – 338, July 2004.

[34] B. Li and R. Chellappa, "A generic approach to simultaneous tracking and verification in video," *Image Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 530 – 544, may 2002.

[35] 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.

[36] X. Liu and T. Cheng, "Video-based fr using adaptive hidden markov models," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, june 2003, pp. I–340 – I–345 vol.1.

[37] A. Lumini and L. Nanni, "A clustering method for automatic biometric template selection," *Pattern Recognition*, vol. 39, no. 3, pp. 495 – 497, 2006.

[38] V. Mane and D. Jadhav, "Review of multimodal biometrics: Applications, challenges and research areas," *Int. Journal of Biometrics and Bioinformatics (IJBB)*, vol. 3, pp. 90–95, 2009.

[39] F. Matta and J.-L. Dugelay, "Person recognition using facial video information: A state of the art," *J. Vis. Lang. Comput.*, vol. 20, pp. 180–187, June 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1540663.1541043

[40] S. Matwin, D. Gorodnichy, and E. Granger, "Using smooth roc method for evaluation and decision making in biometric systems," CBSA, Border Technology Division, Tech. Rep. 2014-10 (TR), 2014.

[41] E. Neves, S. Matwin, D. Gorodnichy, and E. Granger, "3d face generation tool candide for better face matching in surveillance video," CBSA, Tech. Rep. 2014-11 (TR), 2014.

[42] I. Oh and C. I. Suen, "A class-modular feedforward neural network for handwritting recognition," *Pattern Recognition*, vol. 35, pp. 229–244, 2002.

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

[44] N. Poh, T. Bourlai, J. Kittler, L. Allano, F. Alonso-Fernandez, O. Ambekar, J. Baker, B. Dorizzi, O. Fatukasi, J. Fierrez, H. Ganster, J. Ortega-Garcia, D. Maurer, A. A. Salah, T. Scheidat, and C. Vielhauer, "Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms," *Trans. Info. For. Sec.*, vol. 4, no. 4, pp. 849–866, Dec. 2009.

[45] R. Polikar, L. Upda, S. Upda, and V. Honavar, "Learn++: an incremental learning algorithm for supervised neural networks," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 31, no. 4, pp. 497 –508, nov 2001.

[46] A. Ross and A. K. Jain, "Multimodal biometrics: an overview," in *Proceedings of 12th European Signal Processing Conference*, 2004, pp. 1221–1224.

[47] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 23 –38, jan 1998.

[48] M. Rydfalk, "Candide, a parameterized face," *Technical Report LiTH-ISY-I-866, Dept. of Electrical Engineering, Linköping University*, vol. 1, no. 1, pp. 330 – 338, 1987.

[49] S. Sahoo, M. Prasanna, and T. Choubisa, "Multimodal Biometric Person Authentication : A Review," *IETE Technical Review*, vol. 29, no. 1, pp. 54–75, 2012. [Online]. Available: http://tr.ietejournals.org/article.asp?issn=0256-4602;year=2012;volume=29;issue=1;spage=54;epage=75;aulast=Sahoo;t=6

[50] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," in *Digital Signal Processing*, 2004, pp. 449–480.

[51] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 450–455, Mar. 2005.

[52] J. Steffens, E. Elagin, and H. Neven, "Personspotter-fast and robust system for human detection, tracking and recognition," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, apr 1998, pp. 516 –521.

[53] S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann, "Review of classifier combination methods," in *Machine Learning in Document Analysis and Recognition*, ser. Studies in Computational Intelligence, S. Marinai and H. Fujisawa, Eds. Springer Berlin / Heidelberg, 2008, vol. 90, pp. 361–386.

[54] P. Turaga, G. Singh, and P. Bora, "Face tracking using kalman filter with dynamic noise statistics," in *TENCON 2004. 2004 IEEE Region 10 Conference*, vol. A, nov. 2004, pp. 575 – 578 Vol. 1.

[55] M. Turk and A. Pentland, "Fr using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, jun 1991, pp. 586 –591.

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

[57] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I–511 – I–518 vol.1.

[58] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Fr: A literature survey," *ACM Comput. Surv.*, vol. 35, pp. 399–458, December 2003.

[59] S. K. Zhou, R. Chellappa, and W. Zhao, *Unconstrained FR*. Springer, 2006.

[60] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 214–245, 2003.

5 Appendices

Appendix A. Presentation "**Face Recognition: inside the 'black box'** " by Dmitry O. Gorodnichy from the PROVE-IT(FRiV) kick-off meeting, Video Technology for National Security (VT4NS 2011) workshop, 23 September 2011.

Appendix B. Presentation "**Face Recognition in Video Surveillance Applications**" by Eric Granger from the PROVE-IT(FRiV) final deliverable meeting, Video Technology for National Security (VT4NS 2013) workshop, 27 March 2013.

Face Recognition: inside the "black box"

Dr. Dmitry Gorodnichy

VT4NS 2011 (Ottawa, 23 Sept. 2011)

Outline: Key take-away messages

- Why are we so interested in FR and esp. in FRiV
  - Most "accessible" + The only that Humans can do
- "Watch List" example
  - What results with 3 COTS state-of-art products show…
  - Looking deeper (behind the output images) - into "live scores"
- FR by computers vs. FR by humans
  - Recognition from Video vs. recognition from Photo
- Many Face Processing tasks of FRiV
  - FR image-based vs. video-based
  - All COTS are image-based!
- FR companies: FR developers (few) vs. FR integrators (many)
- FR developers: featured-based vs. holistic-based - pro & con
- FR "success" stories
  - are all "human" success stories! –
  - Need a human + Good marketing + Good pre/post-processing
- FR vs. Iris / Fingerprint Biometrics
  - FR does not produce biometric distance!
- Way to do: find good FRiV applications + use temporal data
  - Example from 2004: Face annotation in TV

Face in Video is the MOST collectable modality

… and the only modality that can be validated by human!

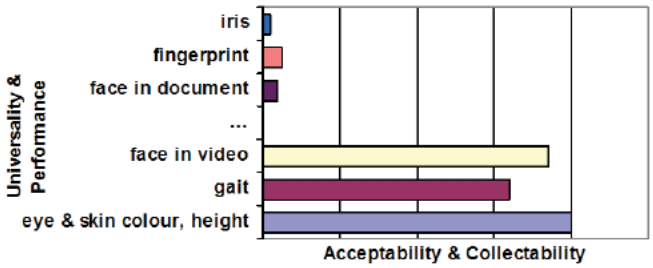


Figure 1: Quality vs availability of different image-based biometric modalities.

[Ref. Gorodnichy, IEEE CRV 2005]

Three main applications

- White list:
  - identify pre-approved travellers (for NEXUS)
    - With IRIS in collaborative / overt mode
    - + Faces ?
- Black list:
  - Identify PDPs (Previously Deported Persons)
    - With FACES in non-collaborative / covert mode
- "Imposter" problem:
  - Based on facial document photos, *to Deny or NOT to Deny* - entry to country

Five "recognition" tasks

1. **Verification (1 to 1, aka Authentification)**
   - Is it "John Doe" (name on his card) ? ➔ If not, follow SOP … (eg. Access Card)
2. **Identification (1 to N,** from "**White List", N is large and growing)**
   - Who is he? ➔ If not identified, follow SOP … (eg. Pre-approved NEXUS traveler)
3. **Screening** (1 to M, from **"Black List", M is fixed and not large)**
   - Who is he? ➔If identified, follow SOP … (eg. Previously Deported Person)
4. **Classification / Categorization** (1 to K, **K – small**)
   - What is his type (eg. Gender, Age, race) ? ➔ Soft biometrics
   - Whom of K people he resembles most ? ➔ tracking/matching
5. ***Similarity quantifier*** (for forensic investigation)
   - Which of L photographs belong to the same person ?
   - Final decision made by Analyst
   - "Imposter" problem: Based on facial document photos, to Deny or NOT to Deny - entry to country

Case study: Watch-List Screening

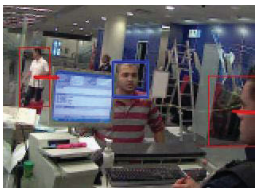
Type 1: Constrained setup
- CBSA PIA , CATSA body check

Type 2: unconstrained Free-flow, one-at-time
- CBSA POE exit/entry , CATSA luggage check

Type 3: unconstrained Free-flow, many-at-time
- CBSA / CATSA Airport



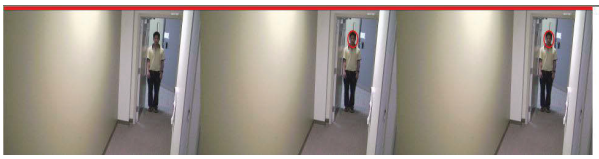


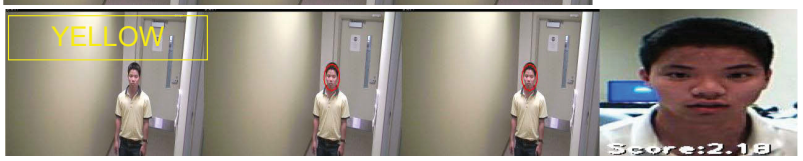


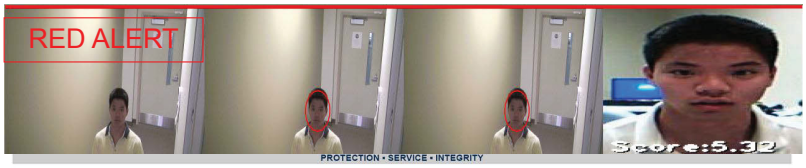

"Watch-List": 60 (CBSA WANTED) + 6 (VSB)

You want to get this…

for a guy in Watch List

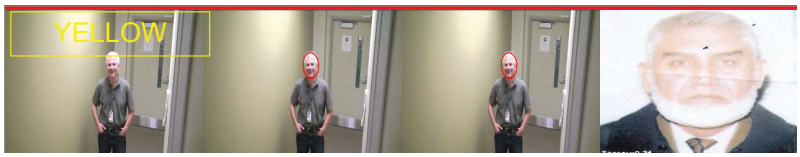






But you may get this …

for a random traveller

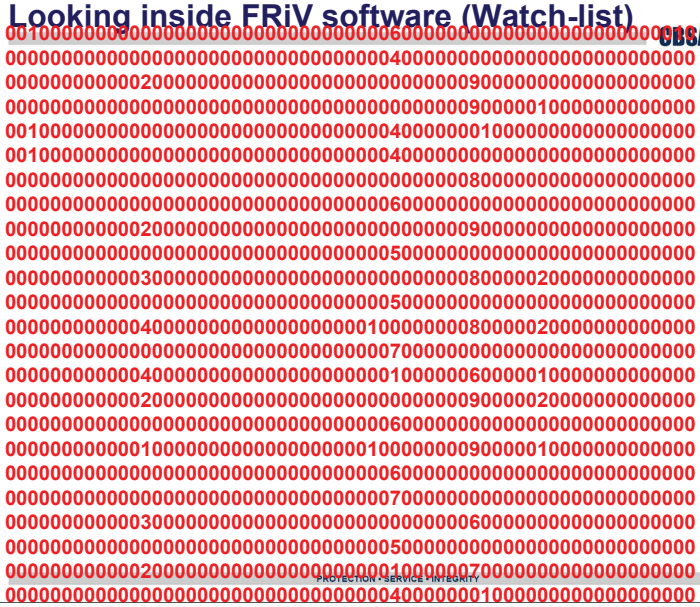


What is the COST of it ?? (with 1 traveller every second!)

vs.

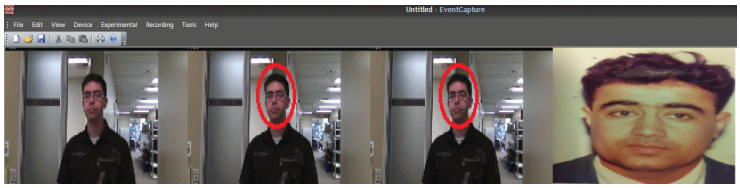
The COST of missing a criminal (in this very rare event)

Looking inside FRiV software (Watch-list)



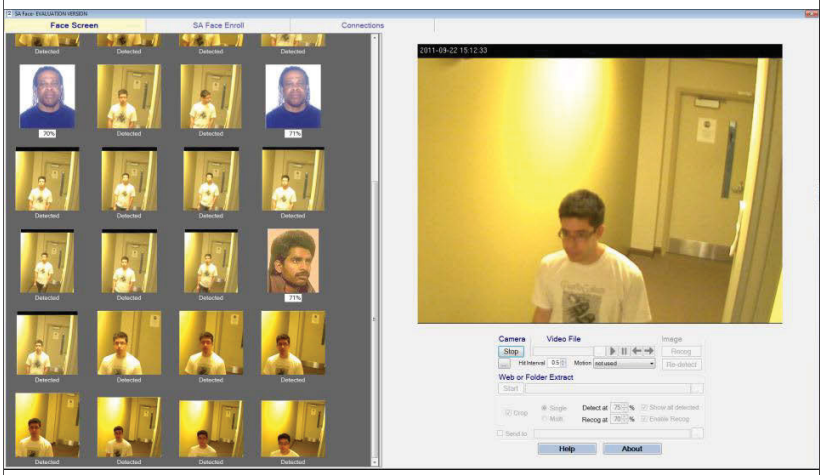
and this…

A person from the stored Gallery may be not recognized.



NB: this is one of the best COTS products (tested by NIST)!

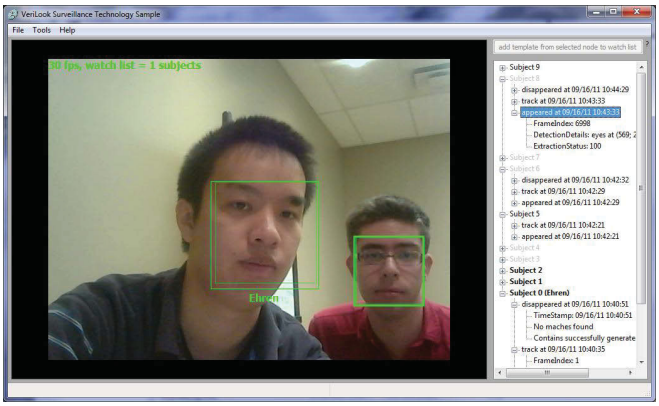
Facial Forensics



Neurotechnology

CBSA ASFC

Works ok at close range (eye-level), but fails in CCTV settings



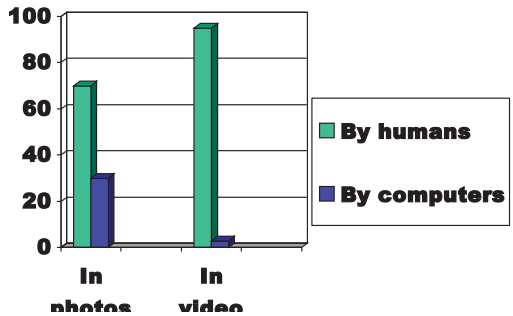
So why computers are so bad with FRiV?...

CBSA ASFC

FR by computers vs. FR by humans

CBSA ASFC

Ref. Gorodnichy, NATO Biometrics workshop, Ottawa, Oct. 2004)



FR from TV

CBSA ASFC

Humans don't have problem recognizing people & activities

Despite "bad" resolution + orientation, expression, occlusion

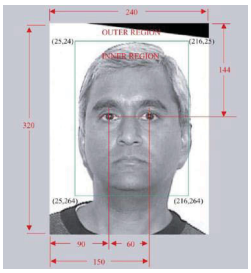
PLAY VIDEO

Computers do!

Video vs. photographs

CBSA ASFC

- Faces stored on documents are much easier to recognize than faces from videos. Photos taken in a controlled environment provide:
  - **Canonical face model adopted by ICAO'02 for passport-type documents**
  - high resolution - 60 pixels **i.o.d.** (intra-ocular distance)
  - high quality
  - face "nicely" positioned
- Videos taken in much less constrained/less controlled environment, e.g. "hidden" camera, where people do not usually face the camera, result in:
  - Poor illumination
  - Blurriness, bad focus
  - Individual frames of poor quality

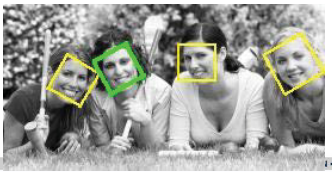


Good news: Faces can be detected!

CBSA ASFC

Although False detections & Misses are possible!

- 2002: computers can detect faces
  - with i.o.d >= 10 pixels
  - in poor illumination,
  - with different orientations: +/- 45º
  - different facial expressions
- 2007.



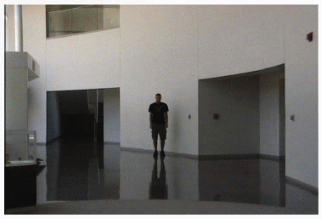
FR by computers vs. FR by humans (from NIST)

CBSA ASFC

Ref. NIST 2010 preliminary results

MBGC

Walking - Conversation



- Human subject raters respond…
  - 1. sure they are the same person
  - 2. think they are the same person
  - 3. not sure
  - 4. think they are not the same person
  - 5. sure they are not the same person

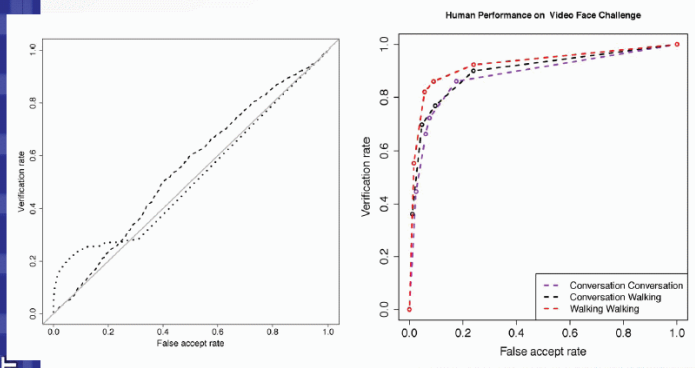
NIST

FR by computers vs. FR by humans (from NIST)

CBSA ASFC

MBGC

There is Head Room

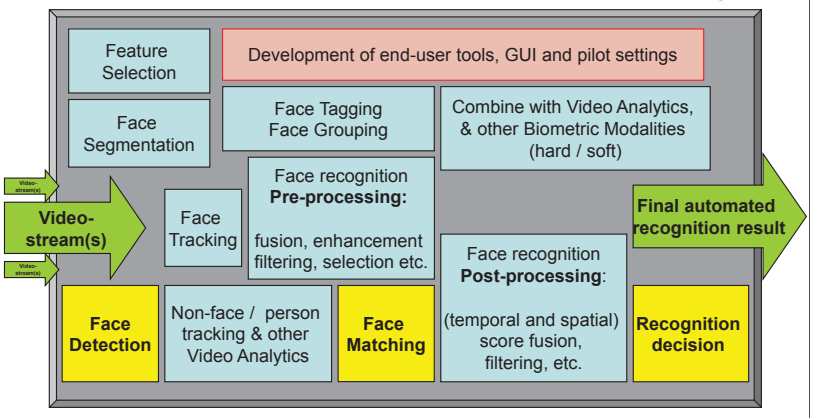


Ref. NIST 2010 preliminary results

NIST

"Face Processing" tasks of FRiV

CBSA ASFC



- Modules available in COTS FR SDK products
- Modules developed by integrators
- visible to end-user

New term: "Face Processing"

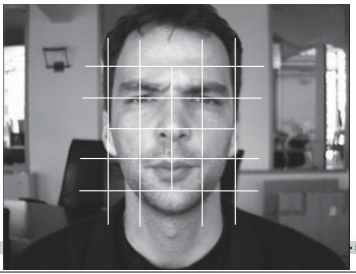
CBSA ASFC

- introduced in 2003 (by Gorodnichy) - http://www.visioninterface.net/fpiv04/preface.html
  - at IEEE CVPR Workshop on **Face Processing in Video**, 2004
  - at Special Issue of **Image and Computer Journal** on **Face Process in Video**
- refers to all intermediate tasks that need to be executed with Facial data:

- from detecting facial image(s) in video-stream(s): Face Detection
- to tracking and grouping them together (tagging them with unique tags: Face Tagging)
- to feeding them to the recognition modules and obtaining the matching score(s): Face Matching (either image-based or video-based), with or without fusion (where fusion can be done either at score level or image/feature level) → for FRiV application #2
- to converting the matching score(s) to a biometric/ recognition decision
- to combining them with other biometric and visual data (such as voice, iris, person's height, cloth colour etc) →for FRiV application #4-6

Applicability of regular TV video (320x240) for FRiV tasks

CBSA ASFC



Face size	¼ image	1/8 image	1/16 image	1/32 image
In pixels	80x80	40x40	20x20	10x10
Between eyes-IOD	40	20	10	5
Eye size	20	10	5	2
Nose size	10	5	-	-
FS	√	√	√	b
FD	√	√	b	-
FT	√	√	b	-
FL	√	b	-	-
FER	√	√	b	-
FC (1-to-K)	√	√	b	-
FM / FI	√	√	-	-
Person detection	√	√	√	√

√– good
b – barely applicable
- – not good

Segment, Detect, Track, Localize, Expression, Classify, Memorize/Identify

FR products / companies

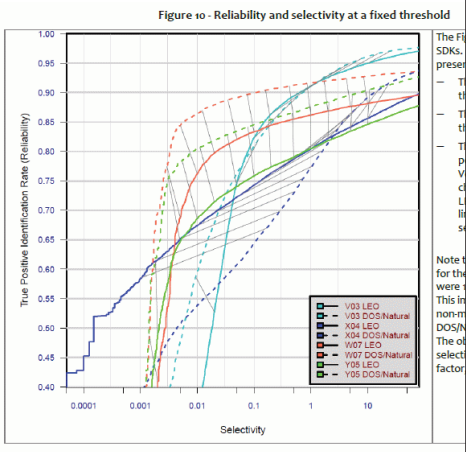
CBSA ASFC

- There are many FR integrators
  - Including Govn't (CBSA)
- Much less FR developers
  - Including Academic and GoC
    - Dmitry Gorodnichy
    - Eric Granger
    - Qinghan Xiao
  - Of which only a few have a proved good (eg. By participating at NIST tests)
- MOST deal with still image Matching only!

FR algorithms at NIST evaluation

Multiple-Biometric Evaluation (MBE) 2010: Report on the Evaluation of 2D Still-Image Face Recognition Algorithms

- P = PITTPATT
- R = SURREY U.
- S = TSINGHUA U.
- T = TOSHIBA
- U = DALIAN TECH U.
- V = NEC
- W = L1 IDENTITY
- X = COGNITEC
- Y = SAGEM
- Z = NEUROTECHNOLO



Taxonomy of FR approaches

From http://www.face-rec.org:

- **Image-Based Face Recognition Algorithms**
  NB: Used in COTS products
  PCA | ICA | LDA | EP | EBGM | Kernel Methods | Trace Transform | AAM | 3-D Morphable Model | 3-D Face RecognitionBayesian Framework | SVM | HMM

- **Holistic vs. feature based approaches**

Need high-resolution (>60 bte)	Work on lower resolution (~20 bte)
More robust to orientation / expression changes	Less robust to orientation / expression changes

NB: None works well with Video

- **Alternative Approach: Video-Based Face Recognition Algorithms**
  NB: So far is mainly done by academic labs

Strategy 1: Find good FRiV applications

For image-based (COTS) products:

While general automated 1-to-N identification may not (never) be possible …

there are many other FRiV applications that are possible!

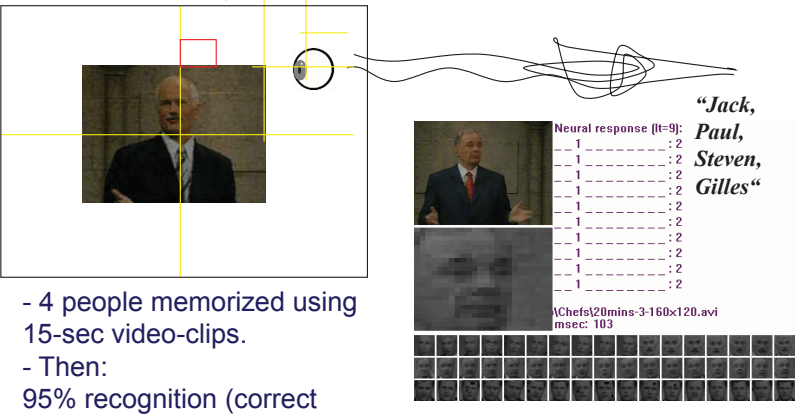
Strategy 2: R&D in Video-based FRiV

Examples:

1. 2002-2007 (www.videorecognition.com/FRiV) from NRC (Gorodnichy)
2. 2004-present from ETS (Granger)

Successful example: Face Annotation (1-to-K)

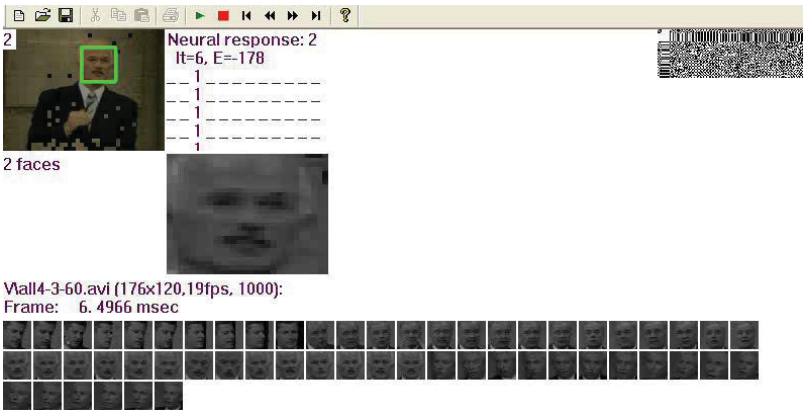
Ref: Gorodnichy, NATO Biometrics workshop, Ottawa, 2004



- 4 people memorized using 15-sec video-clips.
- Then:
95% recognition (correct annotation) achieved on entire 20-min video.

See demo from 2004 (Video)

NB1: dynamic nature of output (first four digits of neural response)!
NB2: the same dynamic nature of output in our today's demo!



Government of Canada / Gouvernement du Canada

PROVE-IT(FRiV)

Final Deliverable

Face Recognition in Video Surveillance Applications

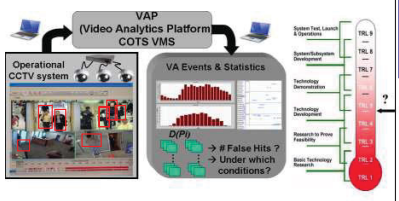
E. Granger, P. Radtke and W. Khreich
Laboratoire d'imagerie, de vision et d'intelligence artificielle
École de technologie supérieure (Université du Québec)

D. Gorodnichy
Science and Engineering Directorate
Canada Border Services Agency

VT4NS'13 (March 27, 2013)

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

Project number: PSTP 03-401BIOM



PROVE-IT (FRiV)
Pilot and **R**esearch on **O**perational **V**ideo-based **E**valuation of **I**nfrastructure and **T**echnology:
Face Recognition in Video

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

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

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

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

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

Canada

Summary

CBSA ASFC

With ETS and U. Ottawa (TAMALE Lab)

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

DISCLAIMER:

CBSA ASFC

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

OVERVIEW

ETS Mandate in PROVE-IT(FRiV)

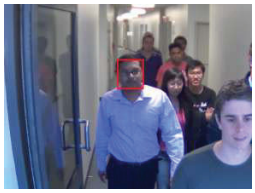
1. **Survey Face Recognition in Video Sruveillance:**
   - commercial technologies and patents
   - academic systems and software
2. **Evaluation Methodologies for FRiVS:**
   - public data sets for medium- to large-scale evaluation
   - experimental protocols for video surveillance scenarios
   - performance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

OVERVIEW

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

1) FR in Video Surveillance

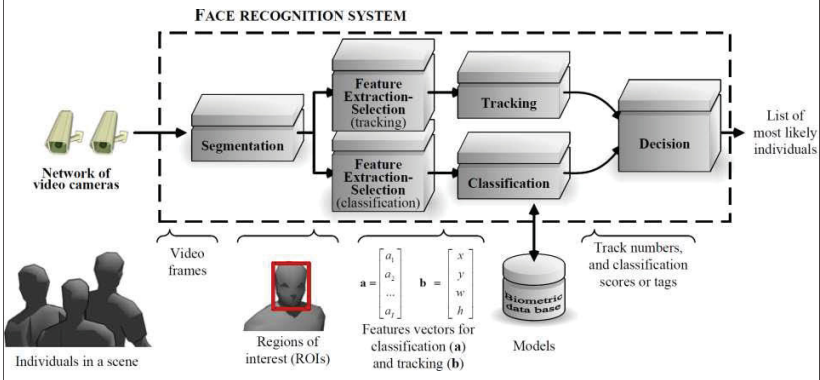
Objectives: enhanced screening and situation analysis across a network of surveillance cameras



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

1) FR in Video Surveillance

A Generic System for Video-Based FR



1) FR in Video Surveillance

Recognition Scenarios

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

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

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

1) FR in Video Surveillance

Taxonomy of Surveillance Setups

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

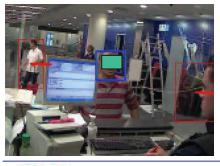
Type 1: semi-constrained setup (for primary inspection lane)

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

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

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

1



2

3

1) FR in Video Surveillance

Challenges

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

1) FR in Video Surveillance

Challenges

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

OVERVIEW

1) Background – FR in Video Surveillance

2) Academic and Commercial Solutions

3) Evaluation Methodologies for FRiVS

4) Case Studies – Screening and Fusion

5) In-House Evaluations of Technologies

6) Conclusions and Recommendations

2) Academic and Commercial Solutions

Distinctive features

- **Recognition type** – still-to-video *vs* video-to-video
- **Facial descriptors** – local features (relies on facial geometry and anchor points) *vs* holistic (relies on face appearance)
- **Facial model:** 1+templates (for template matching) or statistical representation (for a trained classifier)
- **Operational environment:** open *vs* closed-set recognition
- **Spatio-temporal:** recognition assisted by face-body tracking
- **Applications:** see 6 applications of interest for GoC-CBSA

2) Academic and Commercial Solutions

Author	Description	Recognition	Tracking	Applications
Beveridge 2003	CSU Ellastic Graph Bunch Matching	open-set, still-to-video, local	No	watch list screening
Zhou 2003	Simultaneous Face Tracking and Recognition	closed-set, still- and video-to-video, hollistic	Yes	access control
Li 2005	Transduction Confidence Machine *k*-NN	open-set, still-to-video, hollistic	No	watch list screening
Ekenel 2007	Local Appearence-Based Face Models	open-set, video-to-video, hollistic	No	access control
Stallenkamp 2008	Local Appearence-Based Face Models	open-set, video-to-video, hollistic	No	face screening
Connolly 2010	Evolving Ensembles using Dynamic PSO	closed-set, video-to-video, holistic	No	access control
Kamgar-Parsi 2011	Face Morphing to Boost Training Data	open-set, still-to-video, local	No	face screening
Pagano 2012 Gomerra 2012	Adaptive Ensemble of Detectors	open-set, video-to-video, hollistic	Yes	face re-identification

2) Academic and Commercial Solutions

Distinctive features

- **Transaction matching speed**
  - number of templates matched per second
- **Memory consumption**
  - memory required to store one facial model
  - maximum number of individuals in watch list
- **Maximum head rotation**
  - head rotation (looking right or left)

2) Academic and Commercial Solutions

Technology	Vendor	Track	Approach	Applications
Verilook Surveillance SDK	Neurotechnology	Multiple	Still-to-video, video-to-video	- Face anotation, watch list screening, video enrollment, multi-modal biometrics - 60k template matches per second
NeoFace Suite SDK	NEC	Multiple	Still-to-video, video-to-video	- Face anotation, watch list screening, video enrollment, multi-modal biometrics - 1000k template matches per second
FaceR SDK	Animetrics	No	Still-to-still	- Watch list screening, enrollment from video
FaceIT SDK	L1 Identity Solns	No	Still-to-still	- Watch list screening, multi-modal biometrics
PittPatt SDK	Google*	Multiple	Still-to-video, video-to-video	- Face anotation, watch list screening, enrollment from video
FaceVACS SDK	Cognitec	Multiple	Still-to-video, video-to-video	-Face anotation, watch list screening, video enrollment video, multi-modal biometrics - 144k template matches per second. - ISO19794-5 quality assessment
Acsys FRS SDK	Acsys	Multiple	Still-to-video, video-to-video	- Face anotation, watch list screening, enrollment from video - 25k (100k) template matches per second video (still)
SureMatch 3D	Genex	No	Still-to-still	- Watch list screening
Notiface II	FACE-TEK	No	Still-to-still	- Watch list screening

2) Academic and Commercial Solutions

Patents Related to CBSA Applications

- **Mostly for these applications**:
  - access control applications on checkpoints.
  - video analytics applications.
  - 2D face conversion to 3D models to correct perspective

Technology	Assignee
Open set recognition using transduction	George Manson IP Inc.
Combined face and iris recognition system	Honeywell International Inc.
Method and system for automated annotation of persons in video content	Google Inc.
Automatic Biometric Identification Based on Face Recognition and Support Vector Machines	Group of individual inventors

OVERVIEW

1) Background – FR in Video Surveillance

2) Academic and Commercial Solutions

3) Evaluation Methodologies for FRiVS

4) Case Studies – Screening and Fusion

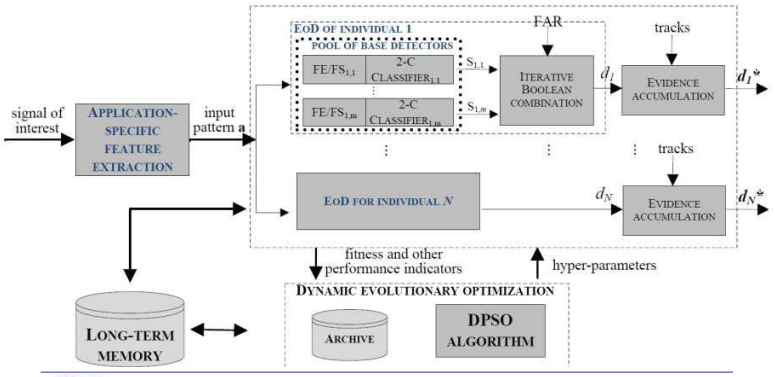
5) In-House Evaluations of Technologies

6) Conclusions and Recommendations

4) Case Studies

Adaptive Multi-Classifier Systems

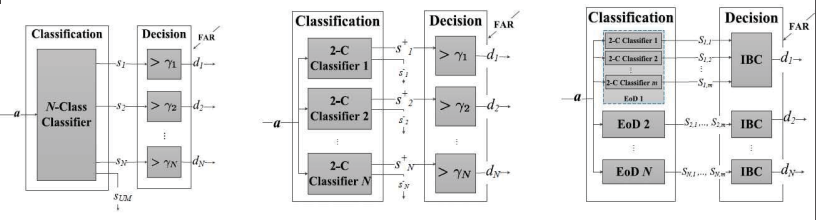
(Pagano *et al.*, IEEE IJCNN 2012)



4) Case Studies

Adaptive Multi-Classifier System

- **Specialized architectures for detection in FRiVS**



monolithic architecture with UM (b) modular architecture (c) modular architecture with EoDs

4) Case Studies

Adaptive Multi-Classifier System

- **Transaction-level accuracy on CMU-FIA data**

pAUC(5%) for 10 individuals

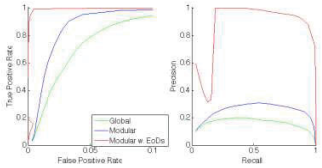
Architecture	Individuals									
	2	23	58	106	147	151	176	188	190	209
Global	0.25± 0.04	0.41± 0.10	0.50± 0.10	0.89± 0.08	0.30± 0.11	0.67± 0.07	0.54± 0.09	0.74± 0.08	0.35± 0.07	0.83± 0.10
Modular	0.27± 0.04	0.42± 0.12	0.68± 0.08	0.82± 0.07	0.56± 0.19	0.72± 0.08	0.45± 0.07	0.77± 0.08	0.51± 0.05	0.83± 0.12
Modular w. EoDs	0.35± 0.03	0.60± 0.11	0.79± 0.06	0.83± 0.11	0.76± 0.13	0.81± 0.09	0.62± 0.06	0.82± 0.07	0.60± 0.06	0.95± 0.04

Table 1: Average pAUC accuracy for 10 individual in interest.

4) Case Studies

Adaptive Multi-Classifier Systems

- **Transaction-level accuracy on FIA:** impact of the database size and class imbalance



Classification Architecture	10 → 20 individuals pAUC	Compression
global	0.55 ± 0.08 → 0.58 ± 0.09	7.7 ± 2 → 5.2 ± 1.5
modular	0.60 ± 0.09 → 0.58 ± 0.08	13 ± 1.3 → 11 ± 1.2
modular w EoDs	0.71 ± 0.08 → 0.71 ± 0.07	1.4 ± 0.2 → 1.3 ± 0.096

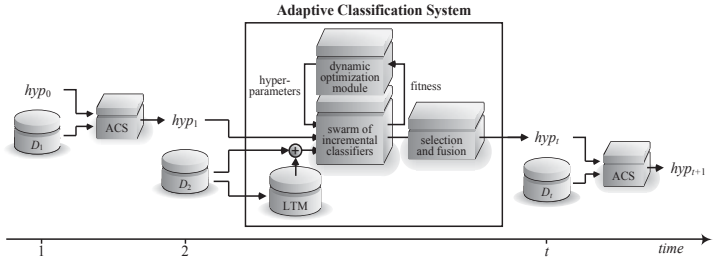
Table 2: Average overall performance (pAUC and compression) for 10 and 20 individual in interest.

4) Case Studies

Adaptive Multi-Classifier System

(de la Torre *et al.*, IJCNN2012)

- **Framework for incremental learning of new reference samples**:

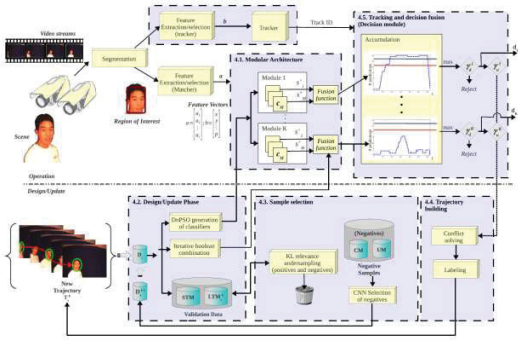


4) Case Studies

A Modular System for Self-Updating

(de la Torre *et al.*, submitted IF 2013)

- Partially-supervised learning from facial trajectories in video FR



4) Case Studies

A Modular System for Self-Updating

- **Transaction-level performance**

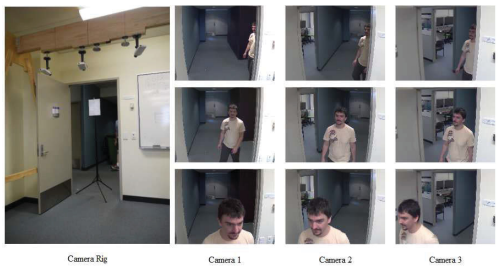
Table 3. Average performance of the system on 10 individuals over 10 experiments. The operations point at fpr = 1% was selected using the ROC space on the validation dataset D_2^*.

Test on	D_2				D_3				D_4			
Approach	fpr	tpr	$prec$	F_1	fpr	tpr	$prec$	F_1	fpr	tpr	$prec$	F_1
No update (Batch learning on D_1)												
OS TCM-kNN	0.2013 ±0.0419	0.9065 ±0.1425	0.0495 ±0.0188	0.0935 ±0.0339	0.2474 ±0.0499	0.5486 ±0.3298	0.0291 ±0.0208	0.0550 ±0.0389	0.1888 ±0.0533	0.4903 ±0.4007	0.0579 ±0.0534	0.1023 ±0.0921
Single PFAM	0.0095 ±0.0184	0.8084 ±0.2048	0.6297 ±0.2236	0.6648 ±0.1930	0.0094 ±0.0201	0.3288 ±0.3440	0.3230 ±0.3270	0.2804 ±0.2862	0.0082 ±0.0178	0.3735 ±0.3911	0.4863 ±0.3812	0.3578 ±0.3456
Learn++ (PFAM)	0.0060 ±0.0068	0.1690 ±0.2365	0.2166 ±0.1800	0.1613 ±0.1669	0.0062 ±0.0083	0.1136 ±0.2055	0.1889 ±0.1651	0.1111 ±0.1260	0.0056 ±0.0064	0.1213 ±0.2223	0.2736 ±0.2397	0.1387 ±0.1824
EoD (PFAM)	0.0062 ±0.0091	0.7702 ±0.2104	0.6756 ±0.2108	0.6789 ±0.1771	0.0064 ±0.0101	0.2675 ±0.2990	0.3801 ±0.3426	0.2553 ±0.2533	0.0053 ±0.0089	0.3185 ±0.3443	0.5256 ±0.3897	0.3366 ±0.3215
Supervised Update (Supervised incremental learning)												
Learn++	0.0060 ±0.0068	0.1690 ±0.2365	0.2166 ±0.1800	0.1613 ±0.1669	0.0057 ±0.0038	0.1187 ±0.1804	0.2025 ±0.1692	0.1278 ±0.1368	0.0119 ±0.0108	0.2057 ±0.2780	0.2419 ±0.1908	0.1917 ±0.1953
EoD L&C (PFAM)	0.0062 ±0.0091	0.7702 ±0.2104	0.6756 ±0.2108	0.6789 ±0.1771	0.0102 ±0.0086	0.4705 ±0.3580	0.3766 ±0.2922	0.3891 ±0.2974	0.0277 ±0.0291	0.7176 ±0.3197	0.4527 ±0.2382	0.5055 ±0.2403
Self Update (Semi-supervised cases)												
AMCS$_d$ (EoD, L&C, PFAM) no LTM	0.0062 ±0.0091	0.7702 ±0.2104	0.6756 ±0.2108	0.6789 ±0.1771	0.0096 ±0.0090	0.4425 ±0.3602	0.3853 ±0.2957	0.3727 ±0.2982	0.0155 ±0.0221	0.5139 ±0.3947	0.4622 ±0.3176	0.4279 ±0.3176

4) Case Studies

Fusion of FR from Multiple Cameras

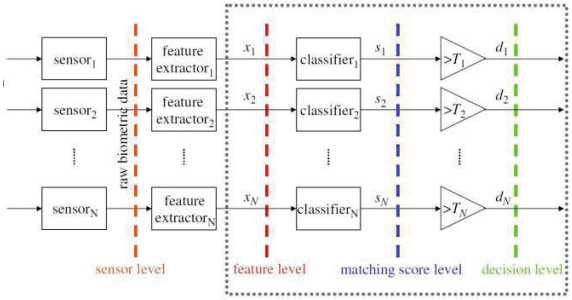
- **Improved accuracy and robustness by fusing FR responses over m-frames, m-cameras and m-videos**



4) Case Studies

Fusion of FR from Multiple Cameras

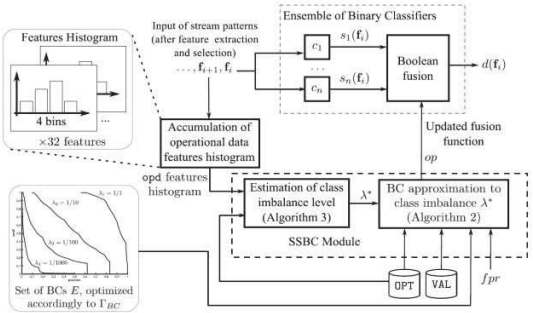
- Dynamic fusion of matcher, system, and/or camera responses
- **Boolean Combination (BC):** threshold-optimized decision-level technique



4) Case Studies

Skew Sensitive BC (Radtke *et al.*, IF 2013)

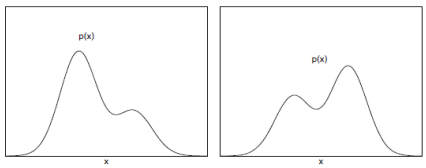
- Adaptive Fusion of Classifiers for Imbalanced Class Distributions
  - periodically estimate class proportions from incoming data, and efficiently adapt the selection of ensembles to reflect operational conditions



4) Case Studies

Estimate class imbalance using the Hellinger distance (González-Castro et al. 2010).

- **Given two different data subsets:**



Validation data **Test data**

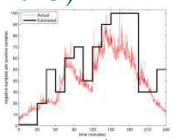
- **Objective:** change the validation data set, so that the Hellinger distance is minimized

$$H_f(T,U) = \sqrt{\sum_{i=1}^{b}\left(\sqrt{\frac{|T_{f,i}|}{|T|}} - \sqrt{\frac{|U_{f,i}|}{|U|}}\right)^2}$$

4) Case Studies

Skew Sensitive BC (Radtke *et al.*, IF 2013)

- Transaction-level accuracy on FIA data



Approach	Measure	Update Period $t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
SSBC	*fpr*	**4.89%** (0.024)	**1.20%** (0.008)	**1.65%** (0.008)	**1.85%** (0.012)	**1.16%** (0.006)	**1.09%** (0.008)	**0.66%** (0.005)	**0.70%** (0.006)
	tpr *recall*	65.58% (0.299)	49.66% (0.329)	54.53% (0.247)	55.67% (0.308)	53.42% (0.261)	51.00% (0.306)	47.52% (0.394)	49.85% (0.399)
	precision	43.68% (0.225)	55.09% (0.315)	41.15% (0.209)	45.33% (0.187)	41.99% (0.177)	47.17% (0.198)	53.59% (0.335)	67.93% (0.314)
	F_1	**0.492** (0.217)	**0.518** (0.255)	**0.446** (0.187)	**0.479** (0.212)	**0.450** (0.191)	**0.470** (0.221)	**0.498** (0.332)	**0.550** (0.344)
BC w/ RUS	*fpr*	4.89% (0.024)	4.32% (0.021)	5.82% (0.025)	5.93% (0.027)	4.65% (0.025)	4.57% (0.025)	3.45% (0.020)	3.63% (0.024)
	tpr *recall*	65.58% (0.299)	67.40% (0.292)	69.71% (0.186)	69.87% (0.231)	69.01% (0.153)	66.06% (0.241)	61.68% (0.320)	64.02% (0.319)
	precision	43.68% (0.225)	38.94% (0.211)	23.37% (0.127)	29.23% (0.109)	23.93% (0.107)	27.04% (0.108)	34.25% (0.184)	54.43% (0.237)
	F_1	0.492 (0.217)	0.470 (0.195)	0.319 (0.136)	0.382 (0.113)	0.332 (0.129)	0.349 (0.134)	0.414 (0.212)	0.550 (0.237)

4) Case Studies

Skew Sensitive BC (Radtke *et al.*, IF 2013)

- Time analysis on FIA data

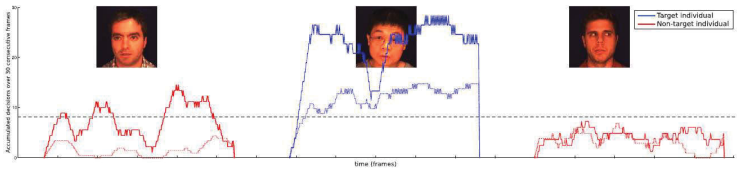


Figure 15: Time analysis for the module of individual 151. The blue line is for a target individual of interest, while the red line is the typical accumulation of a non-target individual. Solid lines are the BC with RUS approach, while dotted lines are the proposed SSBC. Positive detections are made after accumulating at least 8 positive detections over a sliding window of 30 consecutive video frames.

OVERVIEW

1) Background – FR in Video Surveillance

2) Academic and Commercial Solutions

3) Evaluation Methodologies for FRiVS

4) Case Studies – Screening and Fusion

5) In-House Evaluations of Technologies

6) Conclusions and Recommendations

6) Conclusions and Recommendations

Technology Readiness Assessment

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

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

6) Conclusions and Recommendations

Technology Readiness Assessment

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

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

6) Conclusions and Recommendations

Recommendations

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

6) Conclusions and Recommendations

Recommendations

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

6) Conclusions and Recommendations

Recommendations

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