Accelerated Evidence Search Report

Prepared by:
Jean-Philippe Bergeron
David Bissessar
Science & Engineering Directorate, Canadian Border Services Agency
79 Bentley Avenue, Ontario K1A 0L8 Canada

Scientific Authority:
Paul Hubbard
DRDC Centre for Security Science
613-992-0595

Contract Report
DRDC-RDDC-2014-C166
July 2014

IMPORTANT INFORMATIVE STATEMENTS

Accelerated Multi-Camera Evidence Search and Retrieval CSSP Project #: CSSP-2013-CD-1063 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. This project was led by Canada Border Security Agency, Science and Engineering Directorate Border Technology Division

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.

Science and Engineering Directorate

Border Technology Division

Division Report 2014-06
April 2014

CSSP-2013-CD-1063

Accelerated Evidence Search Report

Jean-Philippe Bergeron and David Bissessar

Table of Contents

List of Figures

List of Tables

Executive Summary

This document presents the problem set solutions and findings of the Search and Retrieve project conducted by the CBSA and DRDC.

We tackled the problem of post-incident analysis in historic video footage. In this scenario, an agency must search through accumulated video footage from multiple cameras in an attempt to gather the sequence of events which that led up to an intervention. In environments with several hundred cameras, such a task can take up to two days of manual effort.

Entering this project we suspected that advances in the field of pattern matching in video have made possible new technology which allows image-based queries to be run on the output of selected cameras to return approximate matches from digital surveillance data.

Our study found that the market offerings in this area where either at a low level of maturity and reliability or were not priced such that short term evaluation was possible. In addition, our study identified that publically available datasets for person recognition in indoor environments were not realistic enough to provide a meaningful testbed.

This project produced two important outputs:

1) **An operationally realistic dataset created at an international airport**
   Since no realistic operational data is available publically to measure the effectiveness of software in this task, as a part of this project, a high-quality dataset was compiled in an airport setting. This dataset consists of 160 GB of accumulated video from 93 airport cameras. The passage of 83 volunteers was recorded for a period of approximately 40 minutes: this resulted in 3,720 minutes (or approximately 62 hours) of operational footage. Traveller behaviour is scripted from cooperative to non-cooperative, featuring combinations of people wearing hats or glasses, people averting gaze and persons walking while texting. Specific surveillance points include scenes in disembarkation corridors, at interview counters, in large open areas and chokepoints. Various ground truth data were accumulated including subject presence within the scene, subject behaviour type, camera type and environment type.
2) **A prototype software tool for post event video search**
   We found that a high level of reliability and success could be obtained using face matching techniques rather than general object searching algorithms. A prototype software tool was implemented using custom software to invoke a commercially available face-matching library to search through the accumulated footage of a set of cameras. The software included a map-layout of the surveillance area, which assist the human operator to refine the search. Once compiled, the footage was output into a chronological sequence from the scenes where the person of interest had been identified

As an outcome of the project, we found that the performance of the system was very encouraging in certain conditions, namely, in indoor settings where face recognition quality video was obtained and a trained human operator was using the system. Using the airport dataset, we were consistently able to show that a trained user can navigate the 62 hours of accumulated footage in three to five minutes, resulting in a 788-fold improvement over the brute force method of searching and a 288-fold increase over the observed two day performance rule-of-thumb.

Possible future work includes testing in a variety of operational settings (including outdoor environments), producing a deployment guide for configuring an optimal capture environment and tuning threshold parameters. In terms of algorithm, adding blob-based matching (to complement the current face-matching algorithm) will result in benefits in flexibility and accuracy. Index optimization can yield performance time and licensing cost benefits. Future work in evidence extraction protocols can be beneficial on destined use of the software and court requirements.

Introduction

In operational environments which use video surveillance, it is critical to search footage from multiple cameras to gather video evidence that leads up to significant events, such as a seizure or a health and safety incident. Traditionally it has been necessary to go through footage manually. As deployed systems grow in size, and as demands for video evidence from Crown Prosecutors become more frequent and more exacting, this becomes a daunting task in which we are reaching the limits of being able to effectively process data in required timeframes.

Optimizing technology and processes is an open and ongoing challenge. Advances in video analytics have made possible new technology which allows image-based queries to be run on the output of selected cameras to return approximate matches from IP surveillance data. This may enable previously impossible optimizations of benefit to Border Security and Law Enforcement Communities. This project proposes a technology demonstration that integrates advanced Search and Retrieve (S&R) technologies into existing CCTV camera infrastructure.

The Problem

In the case of incidents, all video footage related, as well as video footage leading up to and following the incident, needs to be gathered and exported. All reported incidents need to be exported and archived locally in case the evidence is required in the legal process. This is a daunting task requiring multiple officers days of work to gather and export all video of interest. The entire story must be exported; it needs to have a certain continuity of evidence where the individual is shown at all times from different cameras.

The Environment

The airport environment is well controlled and indoor. Traveller flow is defined from the arrival corridor to the exit. CCTV equipment is designed to cover a large portion of the CBSA process. The list of cameras in each area is known, as well as the travelling path between areas. There is a list of face recognition capable cameras with a predictable flow in between.

The primary inspection lane (PIL) number and the time of the incident can be retrieved from CBSA databases and can be assumed to be known. Using the flow of people, only a subset of cameras and time ranges need to be searched. Starting from PIL, cameras can be searched going backward to the arrival corridor. A similar process can be done following PIL, into the baggage carousel area, Point area, Secondary area, or the exit.

Definition of Operational Scenario

For all scenarios, video evidence needs to be gathered from the time an individual enters a CBSA-controlled area until they exit.

Scenarios

- Enforcement: This event can happen in a secondary area, for example, when a traveller's possessions are seized.
- Health and safety event: This event can happen anywhere in the CBSA-controlled area, for example, when a traveller falls and is injured.

List of Operational Requirements

- An incident happens with a person of interest - We know the person, location and timestamp. Other information such as the time at PIL and Secondary is known. The system is required to find video evidence from all cameras partially or fully showing the person of interest.
- The system will be given one or more photographs of a person of interest and it needs to successfully find the related video footage.
- Optionally, the system needs to handle the case where a person of interest changes clothing in the process.

Although this project was focused in a CBSA operational context, similar situations exist in other public safety organizations such as Corrections, Via Rail and DFAIT, where certain cameras are capable of facial recognition and the path of a person of interest is predictable.

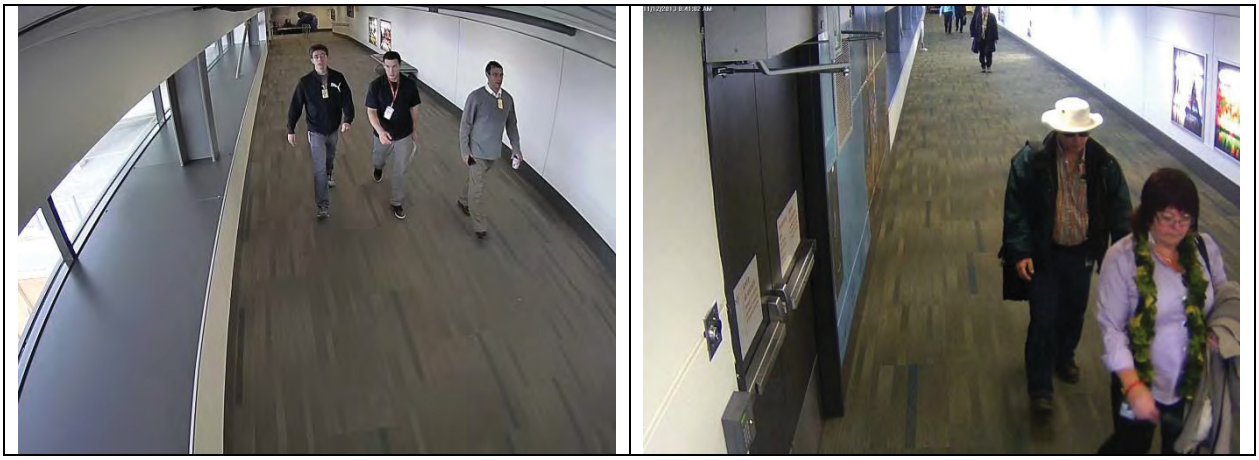


Figure 1 – Sample camera perspectives

The challenge is to use video footage coming from pre-installed CCTV cameras. CCTV cameras are usually installed on the ceiling and set to one side. They have a large field of view pointing down and to the side. Cameras installed are set up for overview monitoring and not for face recognition. Figure 1 is an example of two CCTV cameras to monitor the windows, door and the corridor. The focal length was set as high as possible while conforming to the operational requirements.

As shown in Figure 1, when travellers are far from the camera, their heads have an orientation facing the camera however resolution is minimal and challenging. As a note, resolution for automated facial recognition should range between 32 and 64 pixels between the eyes.

In contrast, to the right of Figure 1, when travellers are closer to the camera, the absolute resolution is higher, however the head angle is poor and some motion blur starts to appear. As seen in Figure 2, while the resulting

images from the second scene have a higher number of pixels between the eyes, due to the angle of capture, the resulting facial images are worse for face recognition.

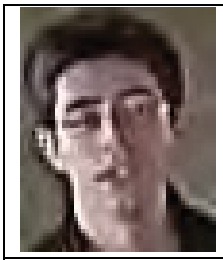
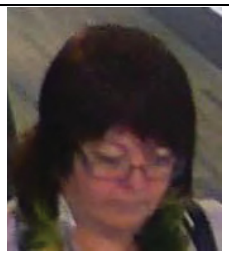
				
15 Pixels	17 pixels	13 pixels	35 pixels	42 pixels

Figure 2 – Samples showing various pixels-between-the-eyes resolutions

There are also issues with hats and sunglasses with regards to the angle of the camera. A hat or sunglasses hide a large portion of the face, making face recognition very challenging.

Datasets

Datasets need to be selected and need to be of a quality representative of the operational scenario. In the field of search and retrieval, video quality makes all the difference. Where a high quality frontal face makes the search trivial, a video with a crowd, poor quality and occlusion makes the search very difficult. According to the operational scenario, the objective is to collect a dataset which is representative of a semi-controlled environment where the flow of people is constrained to a pre-defined path through numerous chokepoints.

Since the idea is to make a storyboard of all appearances of one individual, it must be possible to see the individual at multiple points in the dataset as well as having the ability to link videos together.

i-LIDS / TRECVID

The project team also considered the Gatwick airport dataset which was available as part of the *i-LIDS dataset* [ILIDS2013], (the UK Government's benchmark dataset for video analytics) as well as from Trecvid 2012 [TRECVID2012](courtesy of UK Home Office). This dataset has six scenarios including: sterile zone, parked vehicle, abandoned baggage, doorway surveillance, new technologies (thermal) and multiple camera tracking. The most relevant scenario is multiple camera tracking and it is also used in *TRECVID* as seen in Figure 3.

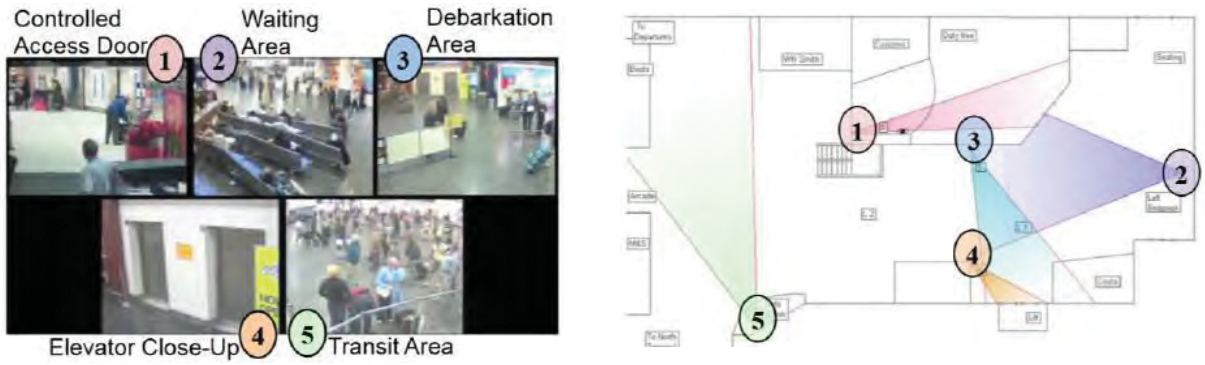


Figure 3 – Sample TRECVID coverage

Pros: It is a real airport environment from the Gatwick airport.

Cons: It is a very difficult dataset to use. Cameras have very little overlap between them and there are a lot of people in the field of view of the cameras. These cameras should be considered overview cameras where no facial details are visible. There is no time synchronization and it is not possible to follow an individual.

Result: Dataset not selected mainly because face recognition is not practical.

PETS (Performance Evaluation of Tracking and Surveillance)

The *PETS* dataset aims to solve the tracking problem and contains a tracking ground truth. The team considered both the 2006 and 2009 datasets [PETS2006] [PETS2009].

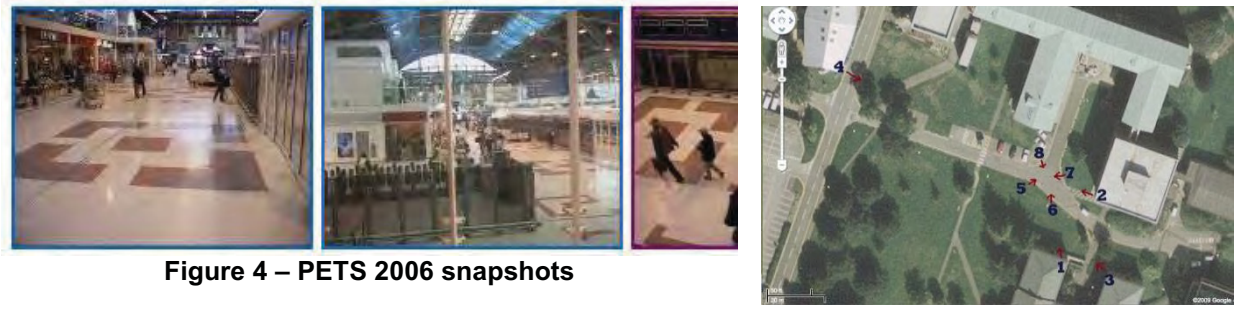


Figure 4 – PETS 2006 snapshots

Figure 5 – PETS 2009 snapshots

Pros: *PETS* 2006 comes from a train station and it might look like a CBSA environment. *PETS* 2009 has more cameras and a map but it is located outdoors, unlike the CBSA airport environment.

Cons: It is not exactly a CBSA environment. The field of view is very large and facial recognition is not possible.

Result: Dataset not selected mainly because face recognition is not practical.

Chokepoint

The *Chokepoint* video dataset is designed for experiments in person identification/verification under real-world surveillance conditions using existing technologies [Won11]. An array of three cameras are placed above several portals (natural chokepoints in terms of pedestrian traffic) to capture subjects walking through each portal in a natural way.

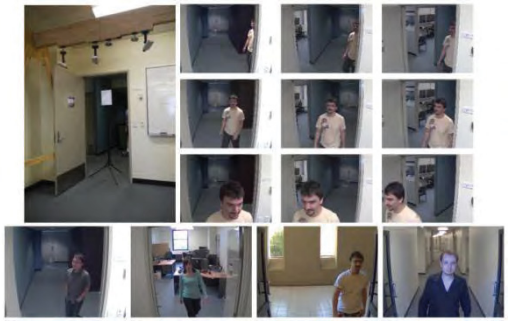


Figure 6 – Chokepoint dataset

Information:

- Camera: AXIS P1343
- Frame Rate: 30 FPS
- Resolution: 800*600 pixels
- Duration: 101 minutes

The video is a collection of 78 smaller clips where 72 have one individual walking at a time, as seen in Figure 6, and six clips have a crowd of people.

In the test sequences, 29 known individuals walk through a chokepoint for a total of 1,281 events.

Pros: The video footage looks like CBSA footage at chokepoints. The facial information is of good quality and is distinguishable.

Cons: There are no overview cameras, no camera overlap or cameras between chokepoints. The image quality and field of view from the Chokepoint dataset is better than the image quality in average airport environments.

Result: Dataset selected

Airport Face Recognition in Video dataset

As publically available datasets were assessed it quickly became obvious that none were sufficiently realistic to meet the project goals. The CBSA Airport Dataset [CBSA2013] was created for the purposes of the project as existing datasets were not adequate.

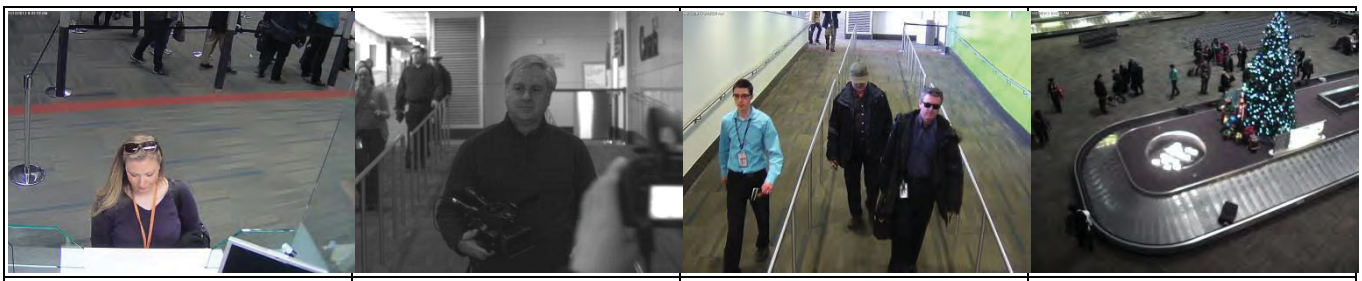
This is real video footage from a CBSA location that was captured using 83 volunteers. It has a real-life implication and realistic camera locations. The cameras were not located specifically for this project which results in the most realistic results for the search software. The field of view of all cameras was set to be a compromise between facial recognition and use of operations. If cameras were too zoomed in, CCTV coverage would suffer. There was an additional seven 1080P camcorders added to see if the addition of cameras would help the results.

The dataset contains the following locations:

- Hall
- Chokepoints
- Waiting line
- Primary inspection lane (PIL)
- Point
- Baggage area
- Secondary
- Cell

The dataset contains the following types of traffic [GG12]:

Type 1 Interview Counter Environment: PIL, interview room
Type 2 Chokepoint Environment: Hallway, Point
Type 3 Multi-Person Chokepoint Environment: Hall, Secondary
Type 4 Large Hall Environment: Waiting line, baggage area



Type 1	Type 2	Type 3	Type 4

Figure 7 – Examples of various environment types in CBSA dataset

Different countermeasures were replicated in the dataset in order to make manual search or assisted search challenging. Techniques were designed to fool either human or software searches. Below are different countermeasure techniques which were used in the dataset:

- Add or remove outerwear in the bathroom
- Spend a long time in a bathroom (no CCTV coverage in bathrooms)
- Look down when walking
- Wear sunglasses while walking
- People loiter for two minutes outside the bathroom
- People walking and texting
- Wear a Hat

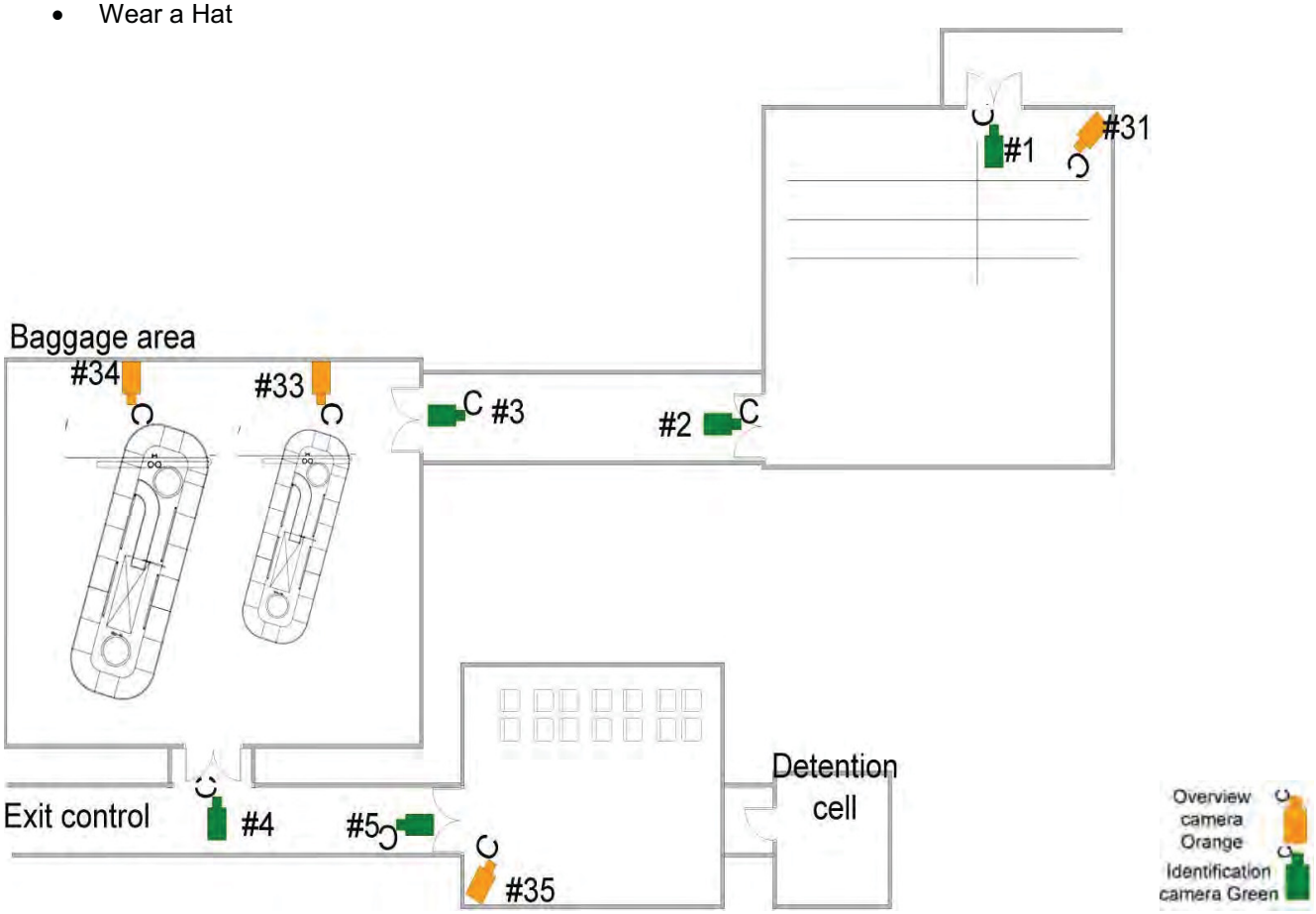


Figure 8 – Sample airport map

Information:

Cameras:

- Panasonic WV-SW355 (1,280 x 960, 30 FPS)
- Samsung SNB-5000 (1,280 x 960, 30 FPS)
- Samsung SNP-5200 (1,280 x 1,024, 30 FPS)
- Sony HDR-CX220 (1,920 x 1,080, 60 FPS)

Camera count: 93 cameras total
Duration: 38 minutes of video per camera for a total of 58.9 hours. The total dataset is 160 GB. Out of the 93 cameras, 30 were good enough for facial identification.

Pros: Dataset with a controlled and predictable flow. Real CBSA environment and precise floor plan. The test subjects are known in order to have a ground truth and mugshot images. Operational data can be reused for other projects.

Cons: It is a mid-scale dataset. There are only 38 minutes of video per camera.

Dataset Chosen

We recall the task at hand: the compilation of the entire video coverage of a person's activity leading up to an intervention or an event. Our survey of datasets led us to select two datasets for use in systems testing: the Chokepoint dataset and the CBSA Airport dataset [CBSA2013].
The i-LIDS dataset [ILIDS2013] was not used as it did not offer sufficient ground coverage of the surveillance area: there were too many blind spots to reconstitute a complete story. In addition the complexity of each scene made this dataset impractical for preliminary testing.

The Chokepoint dataset did not include enough traffic volume, nor was the camera coverage of the scene sufficient enough to highlight the details of a complete trajectory. While the activity and coverage in the Chokepoint dataset were not sufficiently complete or complex to test the full desired functionality, they did provide a useful starting point for preliminary systems evaluation.

The CBSA Airport Dataset provided the required volume, complexity and coverage required for the task at hand. The analyses and results presented in this document are obtained based on the CBSA Airport Dataset.

Technology Survey

As a first step in assembling a technology solution to the problem at hand, the project team conducted a technology survey to see the extent to which there existed a commercially available solution to the problem.

As the survey was conducted, a pattern emerged as to the types of solution that could be used to solve the problem. We propose a categorization into Type A) Traditional video analytics approaches, Type B) 3D tracking from multiple cameras, Type C) 2.5D tracking and Type D) Face recognition based approaches. We will use this categorization to discuss the results of the market study, Type E) Image similarity and Type F) Productivity enhancing user interface.

The selected approach for the prototype implemented in this project was a hybrid of the above which we describe in the close of this section.

Type A: Traditional video analytics such as Bosch, 3VR, Agent VI

This commercial and well known video analytics technique is easy to set up and integrates well with other equipment for search and acquisition. The techniques used are blob tracking, object detection, object crossed line, etc. They are good for tracking one object, but will fail in a crowd. Figure 9 shows sample video analytics from the Bosch IVA. It erroneously merges multiple people in one blob, showing that video analytics is challenging in crowds.

Figure 9 – Images from Bosch IVA

Pros: Easy to configure, usually has a nice user interface and is integrated with CCTV systems

Cons: Might not help solve the problem as multiple people will often be matched in one blob.

Figure 10 – Images from Ipsotek shows vehicle and person detection from Ipsotek. While it successfully detects people and cars, only one person and one car is detected. It does not seem to work on the crowd.

Figure 10 – Images from Ipsotek

Type B: 3D tracking of objects from multiple cameras such as Synesis or IDentrace

3D tracking is a more advanced technique which takes video analytics from multiple cameras and fuses it to find a 3D position[1] [2] [3].

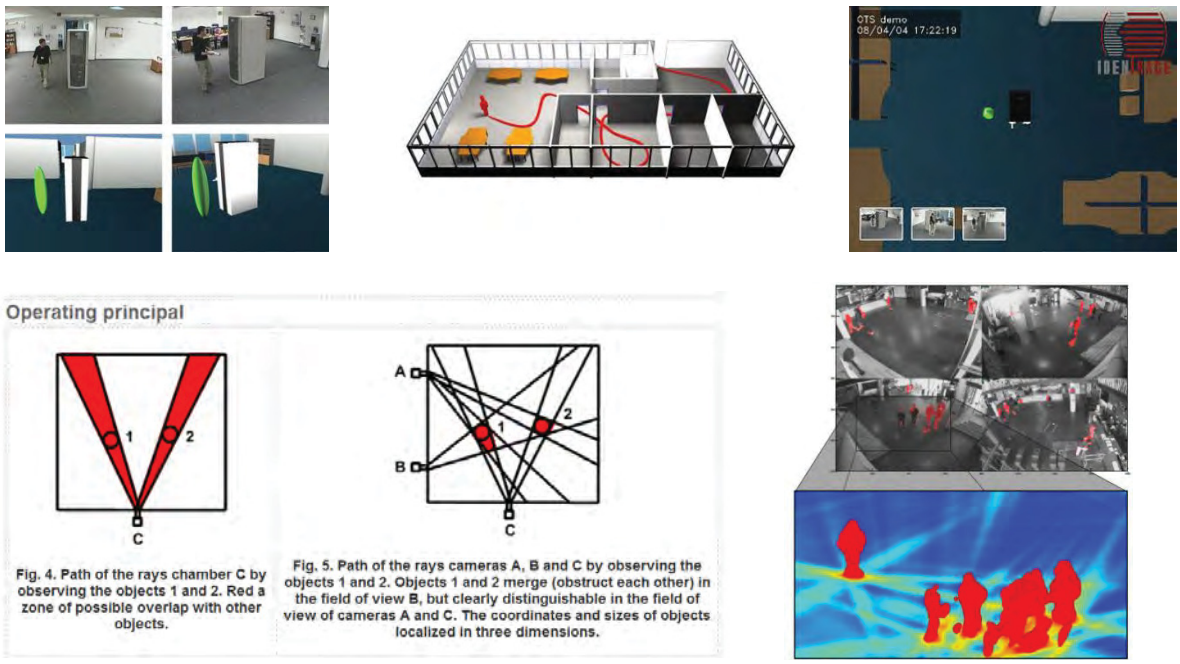


Figure 11 – 3D tracking images

Pros: It can use the potential from multiple cameras in order to have a more accurate position of each object. It solves some of the ambiguity of having object overlap as another camera will have the depth information.

Cons: With multiple people, video analytics still merges multiple people into the same blob. The camera needs to overlap to track people from camera to camera. Cameras need to be very well calibrated knowing their position, field of view and target area. It only works with a limited number of objects.

Type C: 2.5D such as Bosch IVA, IOImage

This type of video analytics tries to measure the size of objects based on the perspective of a very well calibrated camera. From the perspective, the depth and size can be measured.

[1] http://www.identrace.com/products/3d-world-model-server.html
[2] http://en.synesis.ru/en/surveillance/contents/3d-vs
[3] http://babrodtk.at.ifi.uio.no/files/publications/brodtkorb_2013_03_GTC.pdf

Figure 12 – 2.5D sample images

Pros: More information can be captured by calibrating the cameras. The 3D position of an object can be guessed and this would increase tracking.

Cons: The whole object needs to be visible to measure its height. It will be sensitive to occlusion in a crowd.

Type D: Face tracking, detection and recognition such as Cognitec, 3VR

Each face from the video footage would be enrolled and queries would match an individual to all enrolled faces.

3VR

3VR is a video management system (VMS) with video analytics included, such as facial surveillance. A demonstration license for 3VR was acquired and it was tested on the *Chokepoint* and the *Airport Face Recognition in Video* datasets. Great care was taken to set the cameras in an optimal position for FR and for Airport operations. However, it was not capable of detecting faces in our video, even with a minimal face width set to 20 pixels. Seven cameras were imported into 3VR. There were 24 people walking in front of them for a total of 168 events. After the import, only nine events were detected for a detection rate of 5%.

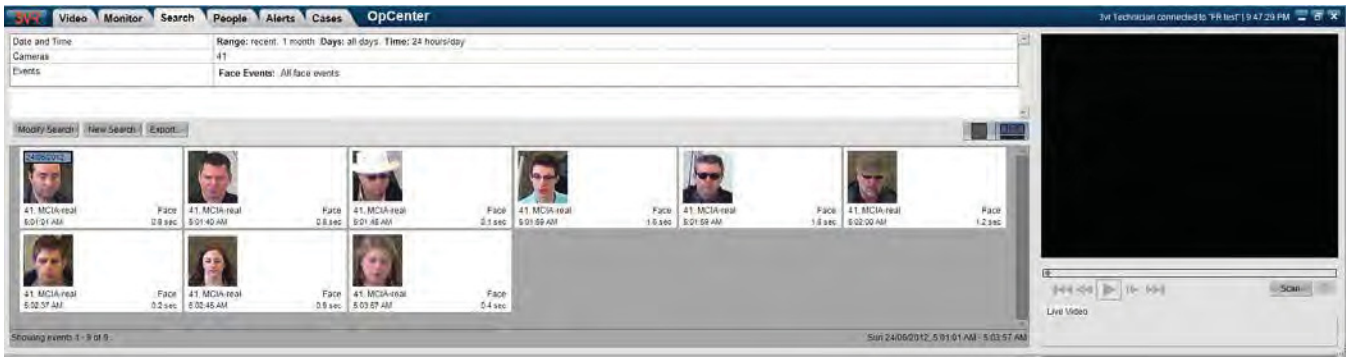


Figure 13 – CBSA dataset images in 3VR

A detection test was run in the *Chokepoint* dataset as well. In *Chokepoint*, there are a total of 1,281 events. A total of 640 events were detected by 3VR for a detection rate of 50%. As a comparison, Cognitec detected 1,276 events for a detection rate of 99.6%. The open source Computer Vision library (OpenCV) detected 1,255 events for a detection rate of 98%.

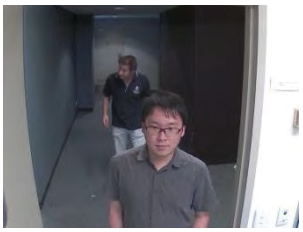


Figure 14 – Chokepoint sample images

Even if the minimum eye distance was set to 20 pixels and the recommended distance is 40 pixels, the image from the event on the right with 44 pixels was missed. This is an example of a missed event; there are many more in the *Chokepoint* dataset.

In Figure 15**Error! Reference source not found.**, subject ID 06 from *Chokepoint* was searched for in all 78 clips. The ground truth contains 47 instances of this subject. The following image shows 23 successful hits as tagged by a human operator. A good idea of the false positives can be seen in the image.

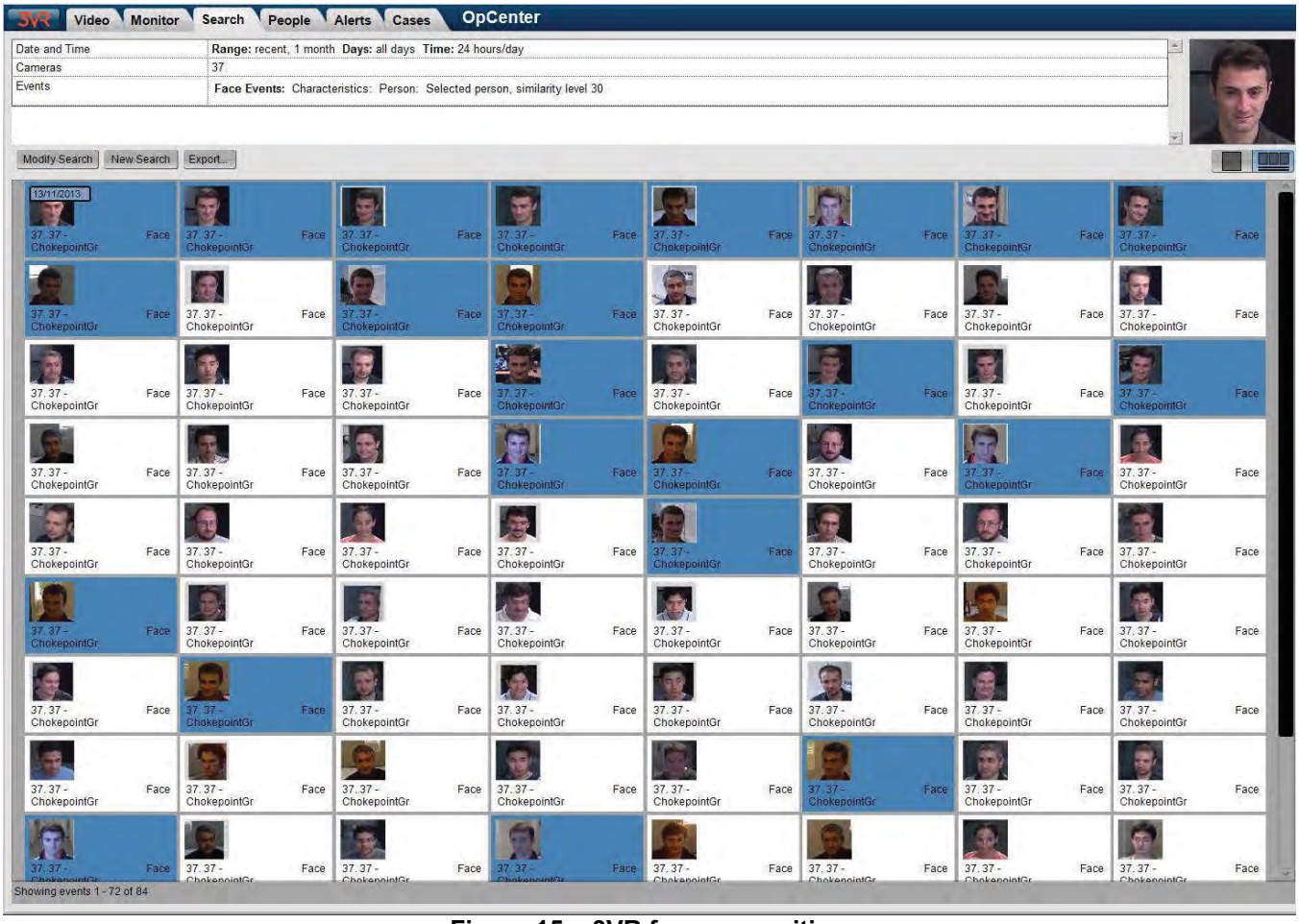


Figure 15 – 3VR face recognition

According to the "*3VR Facial Surveillance Reference Sheet*[4]" the head size must be at least 15% of the width of the image and the event is recommended to last three seconds and must last at least two seconds. However,

4

http://www.3vr.com/sites/default/files/assets_new/3VR%20Facial%20Surveillance%20Reference%20Sheet_5.23.2011.pdf

having an event last two to three seconds requires the individual to walk very slowly or even stop. A normal individual walks two to three meters during this lapse of time and will be outside the field of view. If there is a crowd of people, there will be many occlusions in two seconds. In the *Airport Face Recognition in Video* dataset, in cameras at eye level with a crowd of passengers, some events last as little as 0.1 seconds.

Pros: Would track the right person and works well from multiple camera views. There is a very good chance of finding an individual crossing a chokepoint.

Cons:

- Needs a good facial image of the individual with high resolution and the correct angle, which might be difficult in CCTV videos
- Needs enough FR capable cameras to track an individual

Type E: Feature detection such as Google similar images, piXlogic or imprezzo

Different feature detection software are named content-based image retrieval and all have a similar approach[5]. Feature extraction, clustering and template matching[6] is used. For querying, one training image is given and similar images are extracted. Images are found similar on arbitrary data. The system might find the color or the background to be similar but it might not find the same identity.

For the following images, Google similar search was used to find similar images from the entire web.

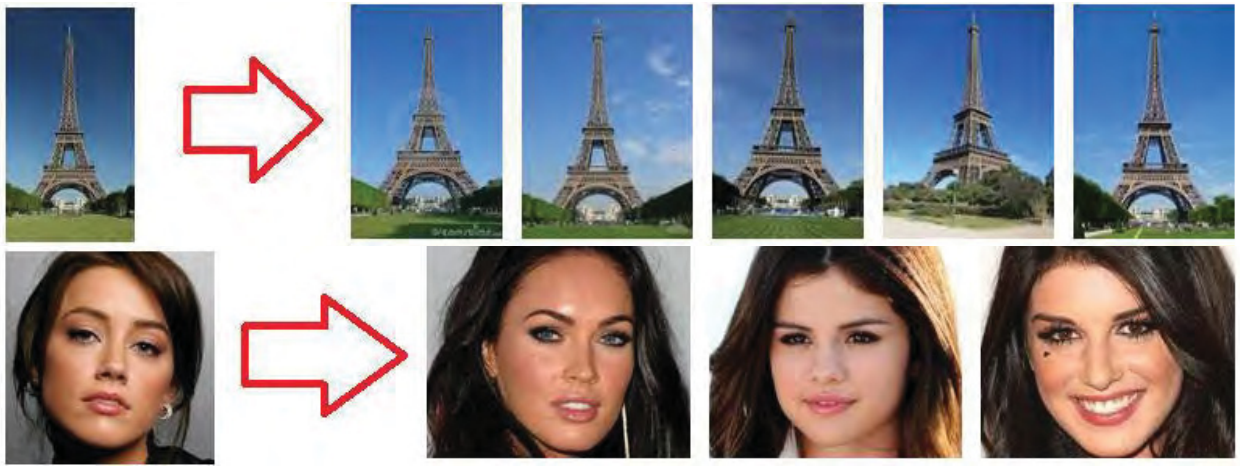


Figure 16 – Google similarity search example

Figure 17 – Immense similarity search example

Pros: This technique does not require calibration of cameras, overlap or a specific field of view. It is a very generic technique which is applicable to most scenarios.

[5] http://en.wikipedia.org/wiki/Content-based_image_retrieval#Query_techniques
[6] http://en.wikipedia.org/wiki/Multimedia_Information_Retrieval

Cons:

- Features need to be very discriminative as it worked well in the Eiffel tower case, but less in the facial case.
- It may be hard to find a person in a suit.
- It might give a lot of unrelated results.
- It is optimized for images; searches in videos are more difficult. Most software allows searches in images and not in videos.
- piXlogic was considered but it was not evaluated since there is no possibility of having a testing license. The cost of purchasing the license was too high for this project.

Type F: User interface for better use of user time

A user interface is not a video analytics technique, but it would involve using or creating a better user interface with a map or shortcuts in order to manually find the video quicker. The user interface can be customized with a map of a CBSA location, the path between cameras and the average time taken to walk between cameras. There is a mapping functionality in VMS such as Genetec and Milestone where a camera can be viewed based on the map. However, there is no storyboard showing the trajectory of an individual.

Figure 18 – Productivity enhancing of user interfaces

Pros: It would make the search faster if done properly.

Cons: There is no automated portion.

Selected Approach

The selected approach is a hybrid approach of face recognition Type D and user interface Type F as well as a storyboard.

The face recognition technique is optimal in “chokepoint” type cameras but cannot work in overview cameras. A customized user interface will help the operator find the video footage of a person in the overview cameras between the chokepoint cameras.

Since the map and flow of people is known, there are a limited number of cameras to search through between two chokepoint cameras. It is known that the resolution and quality of facial images in CCTV is low, but by lowering the acceptance threshold and by fusing scores over time, the results will be improved. For normal operations, CCTV systems tend to be built with cameras in chokepoints.

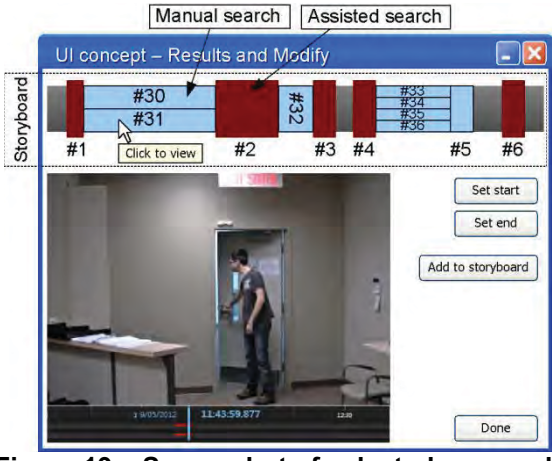


Figure 19 – Screenshot of selected approach

The cons of facial recognition are mitigated by providing a way to find video on cameras which are not face recognition capable.

With the help of a customized interface, an operator will watch a subset of all videos selected by the system for the gaps in time where there is no face recognition enabled cameras. In cases where the system fails to find the individual, it is still possible to use it to view and tag videos to make a manual storyboard.

Pros: The right person would be tracked, works well from multiple camera views. It has a great synergy between automated and manual work.

Cons: Needs limited manual intervention

Solution Concept

Requirement Specifications

The solution aims to increase the efficiency of searching video from multiple CCTV cameras that lead up to an event such as a seizure or a health and safety incident. The goal of the solution is to gather as much video footage as possible from all CCTV cameras relating to one of these incidents.

A semi-automated approach was taken using search and retrieval algorithm software to maximum capacity and using the operator efficiently for the rest of the information gathering process.

The solution was created based on a floor map of the area of interest in order for the software to be aware of the location of different areas such as PIL, Immigration, Point, etc. and the paths between them.

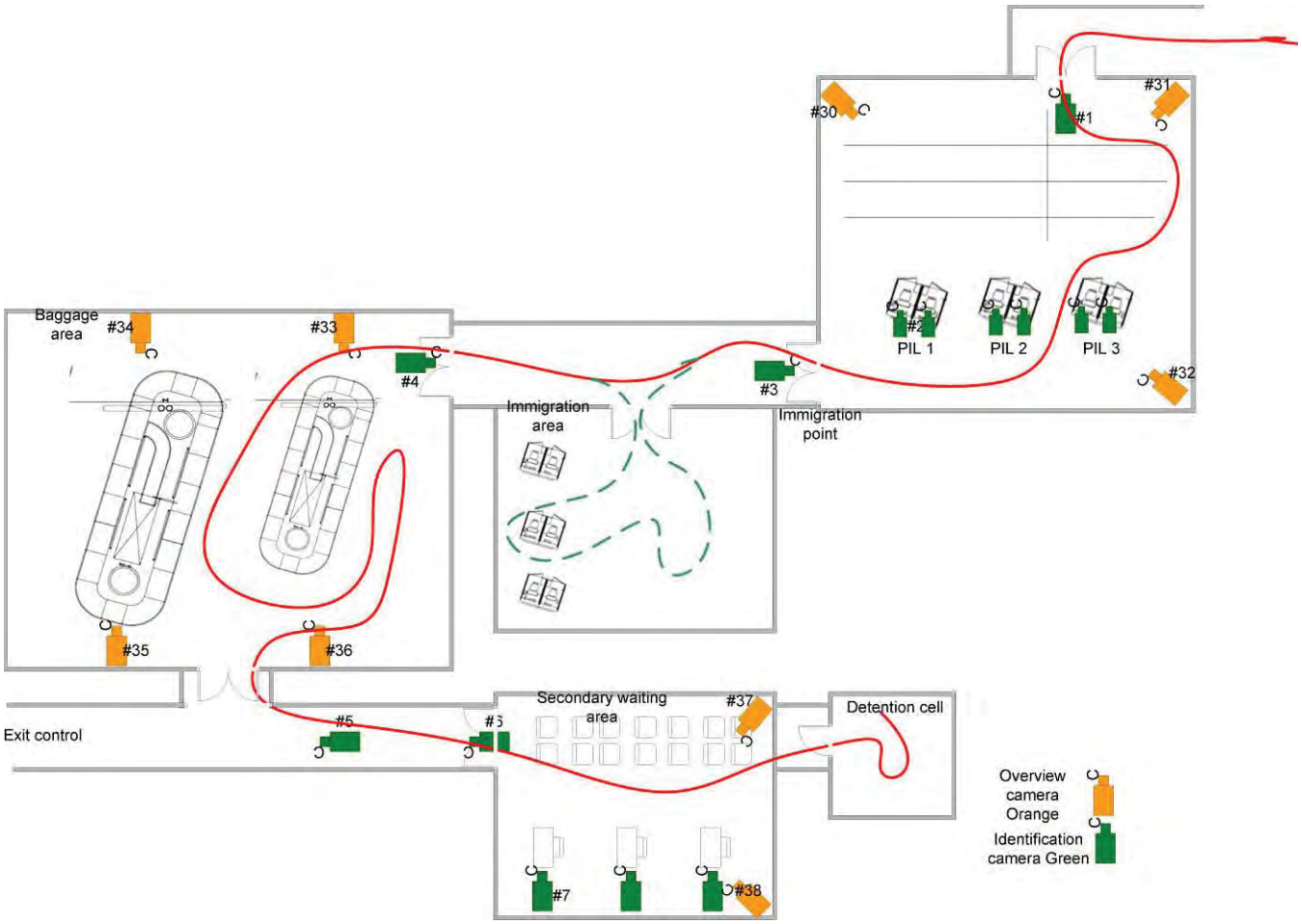


Figure 20 – Notional schematic of an airport

The solution was based on Figure 20**Error! Reference source not found.**, showing the schematics of a sample airport. All chokepoint cameras, in green, could possibly be used to run facial recognition. All overview cameras, in orange, cannot be used by automated search and retrieval and would require an operator to review the footage, but the system would provide clips of interest based on timing of the identification cameras. The semi-automated approach would help the operator to review the overview footage efficiently.

High-Level Approach

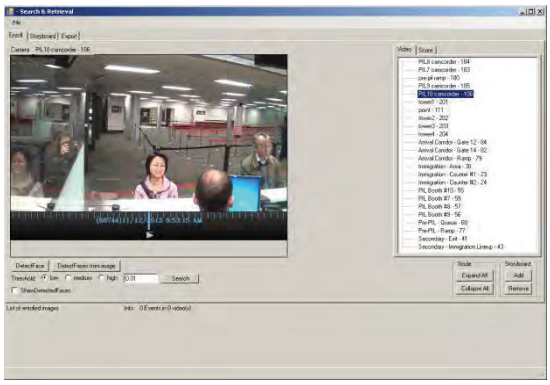
Since the environment was static and contained a predictable path with a known flow of people, the search was optimized to use this information. Video analytics techniques such as facial recognition can be used in controlled environments such as chokepoints. The accuracy of facial recognition improves greatly with a sharp, high resolution facial image. It is unlikely that good results would be obtained from overview cameras.

The approach was to use a commercial facial recognition algorithm to search for the facial image of an individual of interest in a subset of cameras located at chokepoints. An initial facial image was captured from a camera where the individual was known to pass by. This image was searched through previous video footage and a list of matches was filtered by an operator. From each of the matches, more facial images were added to the search query, thus improving the search results. This iterative process found more and more matches as the biometric description of the individual was improved. A manual operator was needed to supervise the facial recognition algorithm in case of errors.

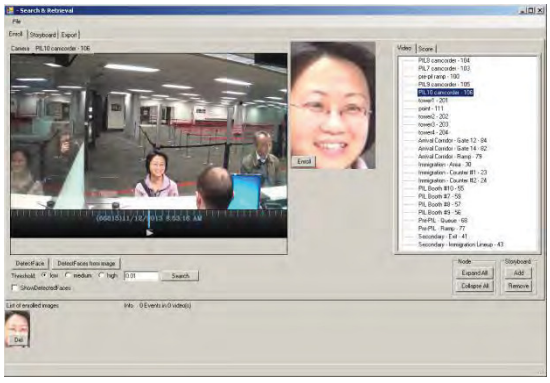
From the list of matches, a storyboard was created with clips of interest when the individual was recognized and gaps, such as when the person was in the waiting line or at the luggage carrousel. The system suggested which camera's video to look at and at what time in order to fill in the gaps. There was a possibility of using blob tracking to follow the individual between cameras. As the person was found in a chokepoint camera, his blob would have been tracked into adjacent cameras.

Example Search

Step 1 - An image was found in video where the face was visible.

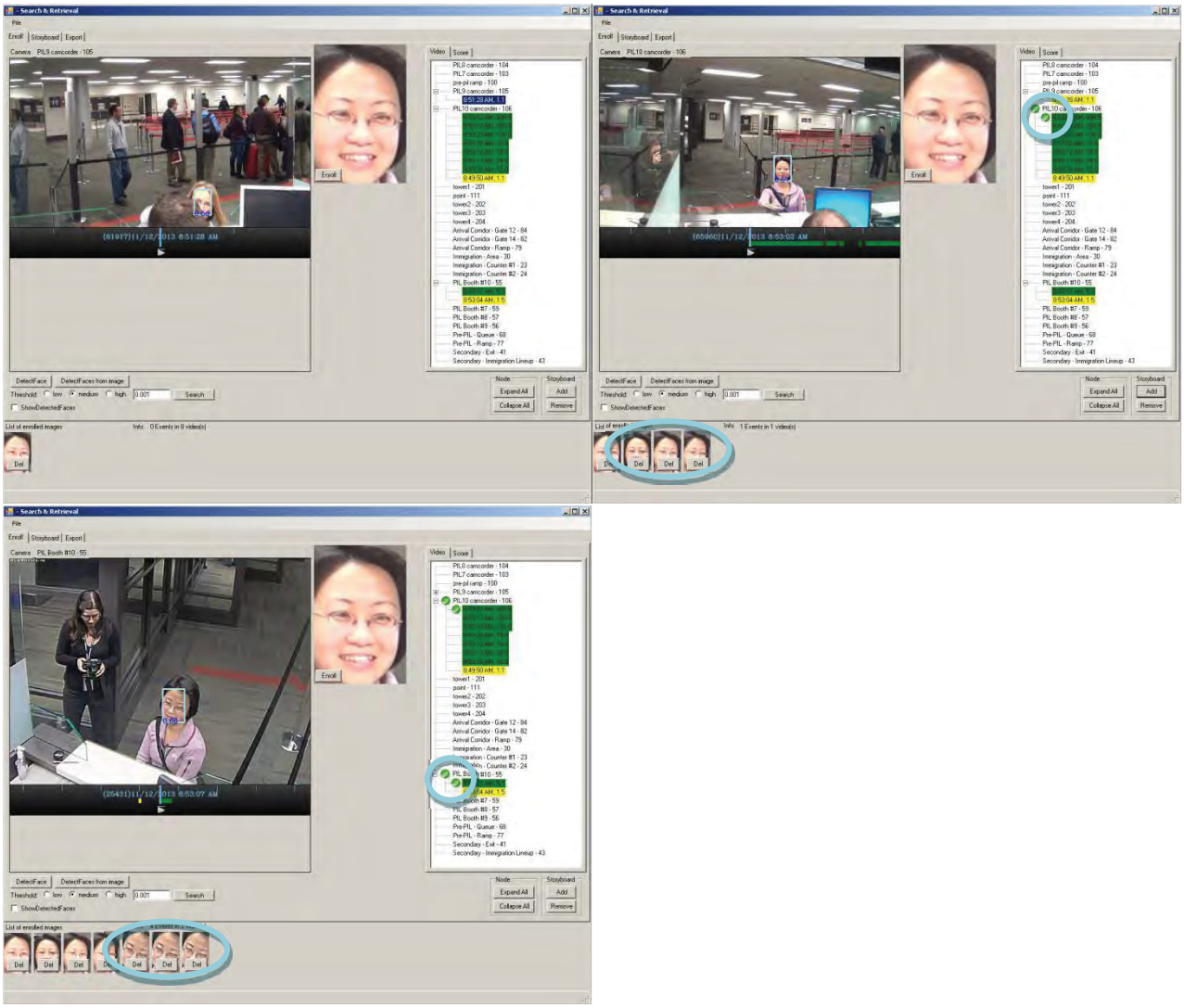


Step 2 - The face was detected and enrolled for searching.

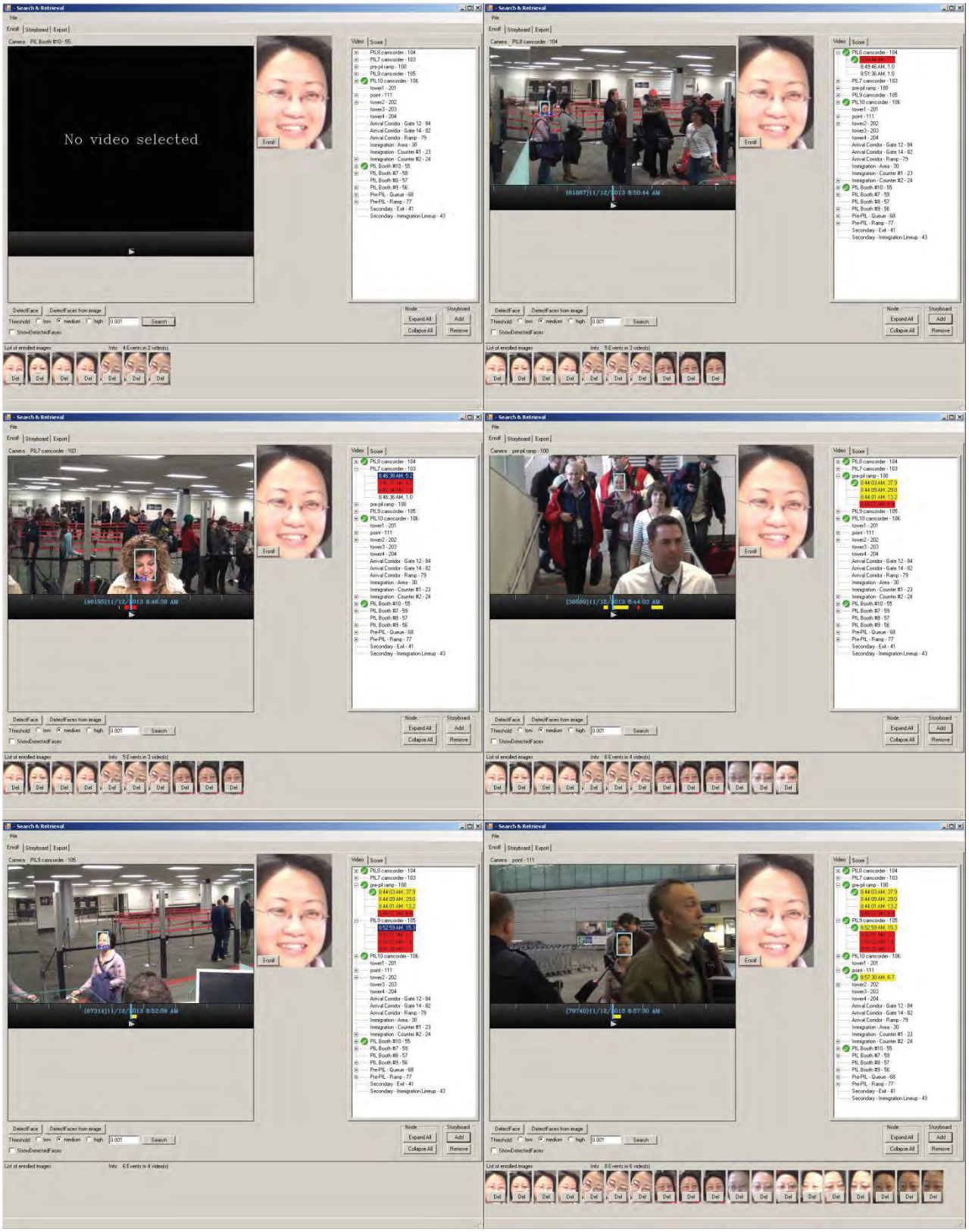


Step 3 - The software found matches in three cameras indicated by the tree view below. The first result was a false positive and the next two results are true positives. In the “PIL 10” camera, only the first result was examined since it was a hit and there is no need to find more than one hit per camera. When the correct results were tagged by the operator, the system added a checkmark (✔) next to the video, and added the first, middle and last face tagged to the list of enrolled images.

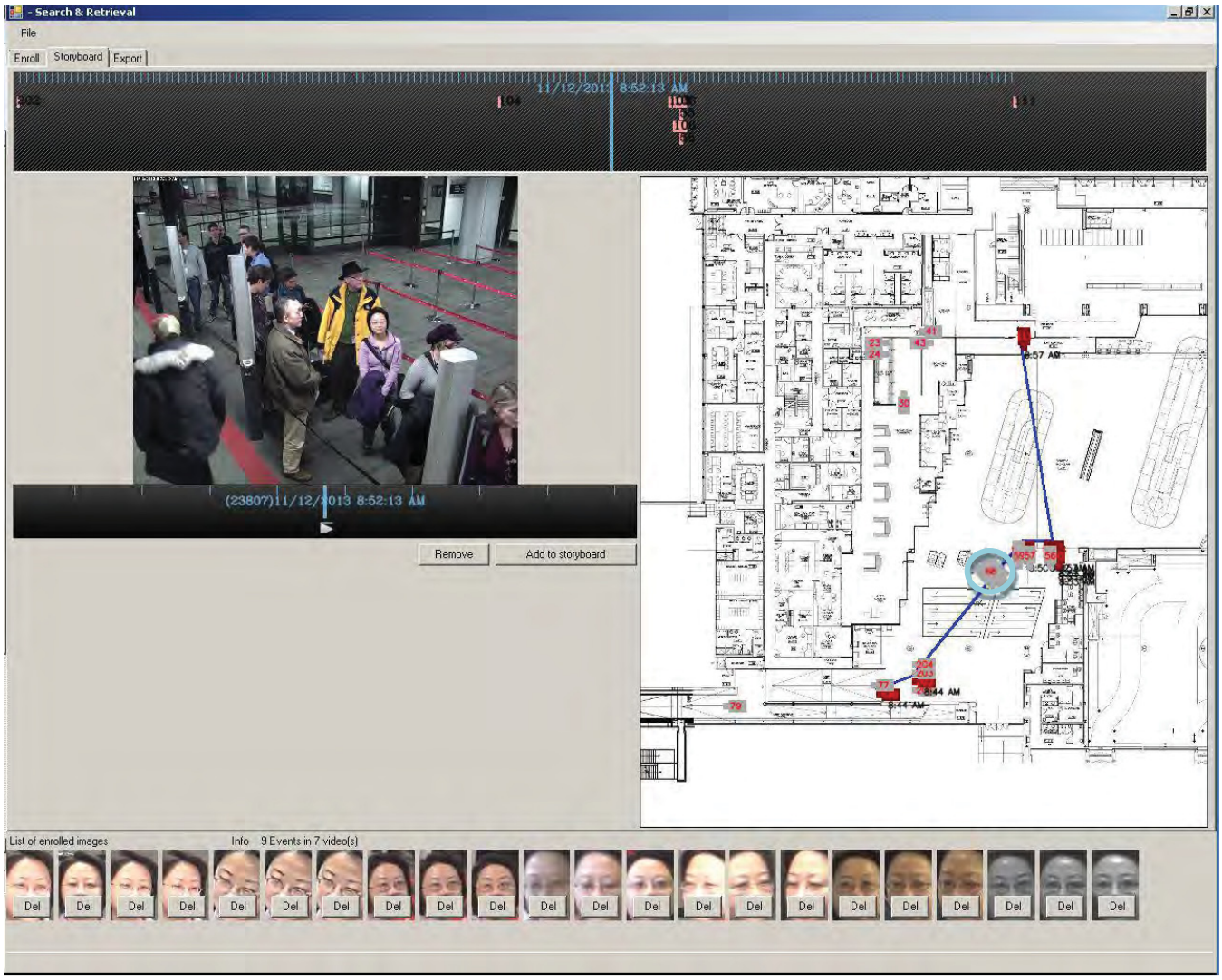
Notice that no other hits were found with the search from the initial image. A second search was done with the added faces to find more hits in an iterative manner.



Step 4 – The search was performed again. This search yielded more results because more enrolled images were used for the search. The chance of finding a hit was increased.



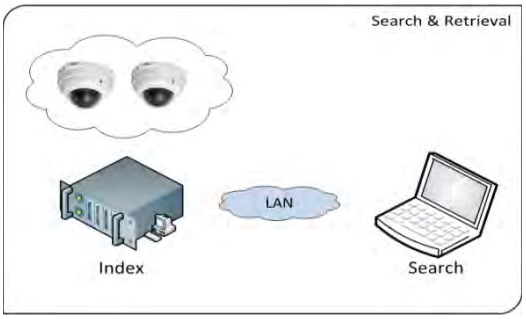
Step 5 – The Storyboard was completed manually. By selecting a camera where the face had been recognized, the operator easily located a camera on the map where there was a high likelihood of finding the person and then tagged it. In the example below, the person of interest was located in the PIL area. The operator selected Camera 68 and quickly found the person waiting in line knowing that she had to wait in line before being interviewed at PIL.



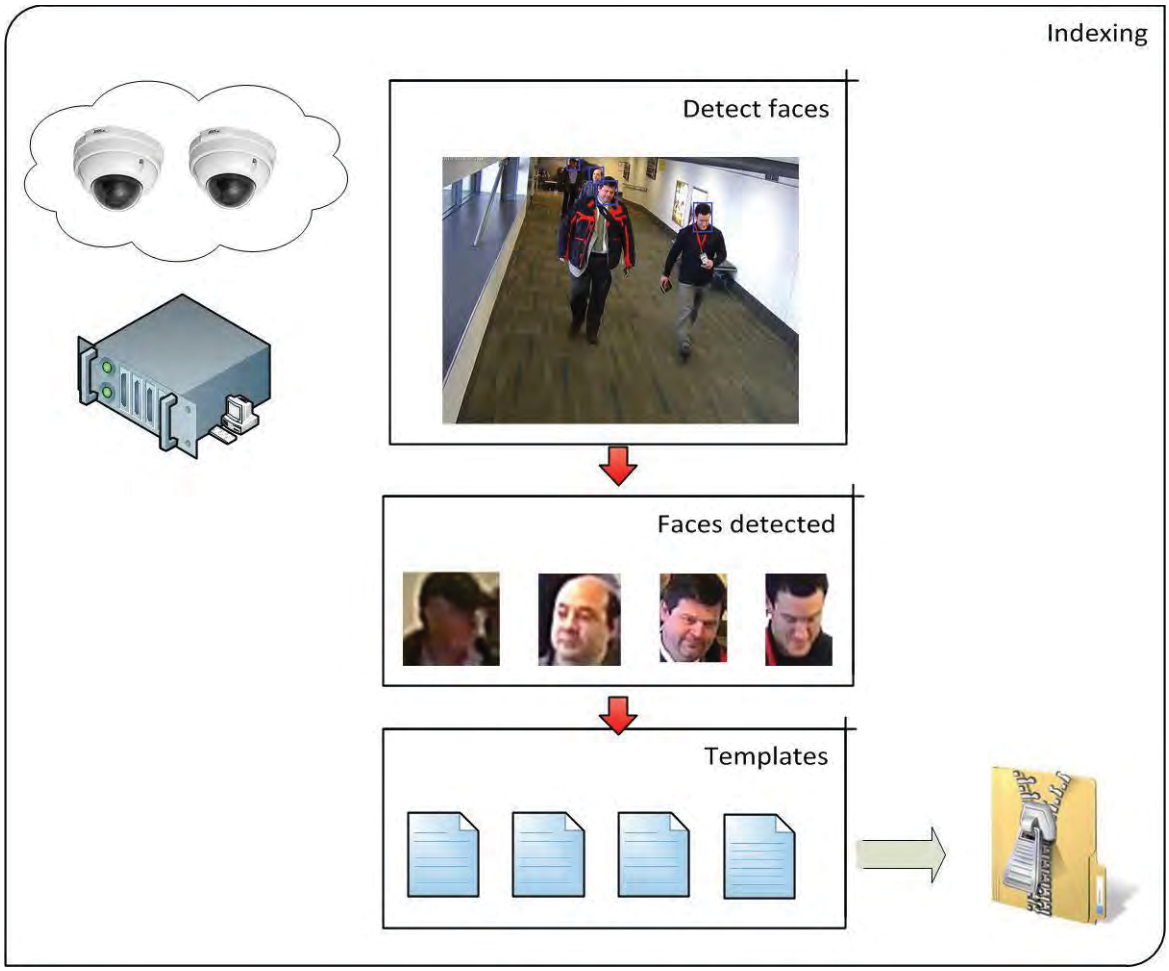
Step 6 – Manually adjust the start and end time of all events in order to capture the person in as much footage as possible.

Step 7 – Export the project as a video file.

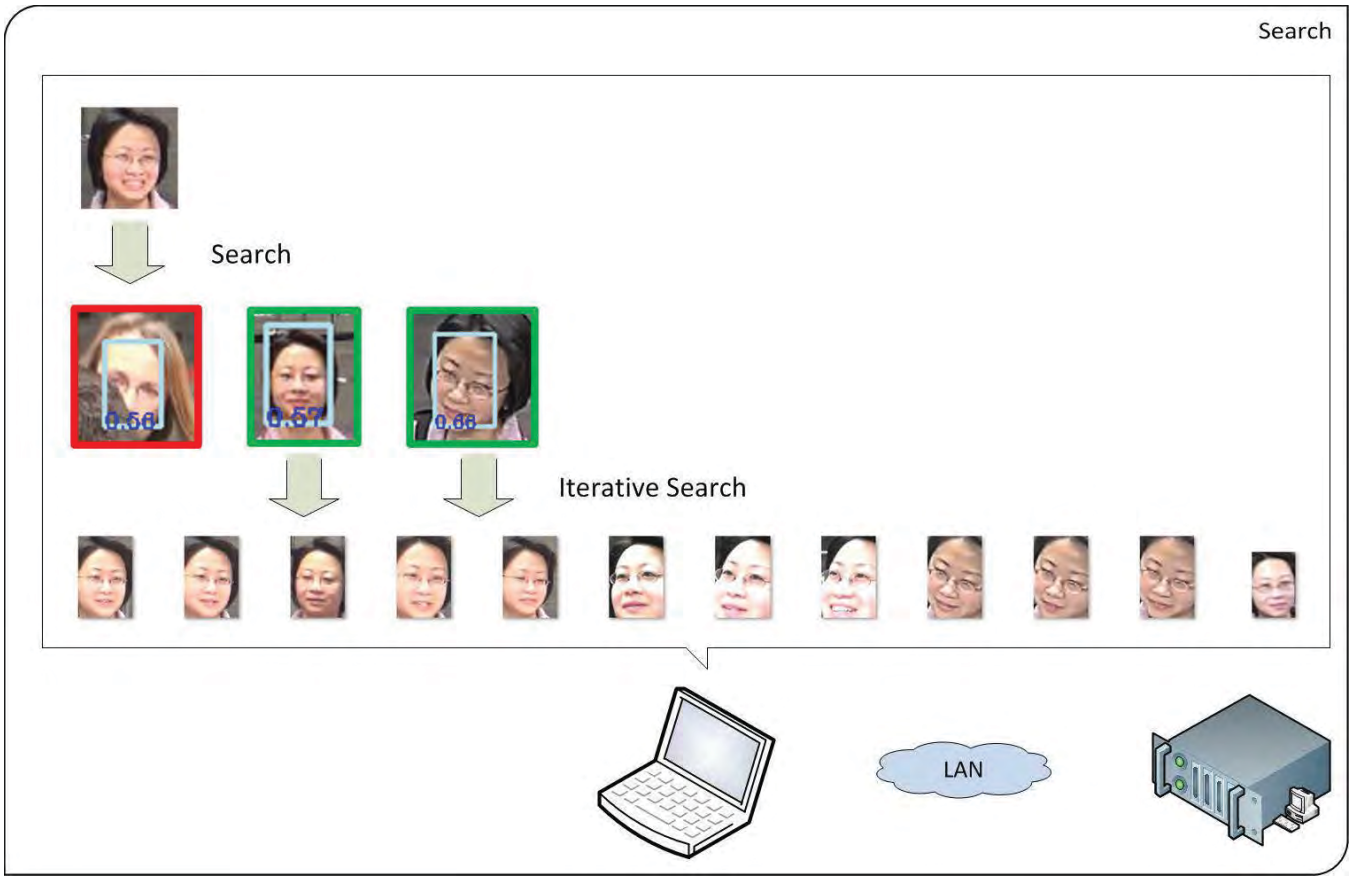
Software Architecture



It is infeasible to search directly in the video footage for all queries as the video captured could consist of days of video from hundreds of cameras. The solution is to index the video and run queries on the index instead. All CCTV cameras would be indexed by an indexing server and the searching machine would query the index on the server.

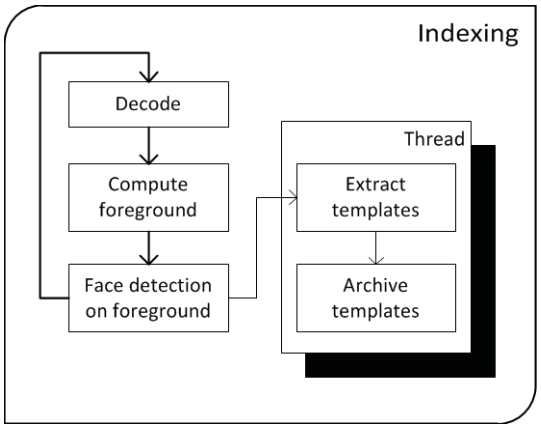


The indexing server connects to all cameras, decodes the frames and detects all faces. For each face detected, a template would be created and then all templates would be stored in an archive, which is called the index.



Indexing Strategy

To index the video, a server loaded video files. In the lab, our server used two powerful Nvidia Geforce 690 cards which decoded video from multiple cameras at 500 FPS. The background detection was done on the graphic card at about 400 FPS and face detection was executed only on the foreground portion. This process was extremely fast when there was no motion. When there was motion, a bounding box was computed around the foreground mask and then faces were detected only on the foreground area. This sped up facial detection considerably, especially when the foreground area was small.



Each detected face was sent to a different thread and the templates were extracted. Template extraction is a CPU intensive process and took about 0.5 seconds on a recent machine. By having the extraction in its thread, the rest of the process did not slow down, but a delay between detection and the time where the template was available to be searched was introduced.

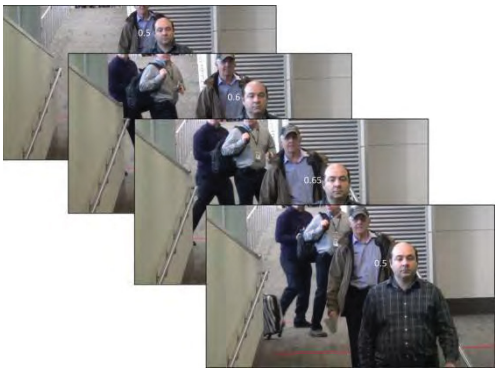
Templates were stored sequentially in zip files split by cameras and time. Storing small files in a zip format is much faster than storing thousands of small files directly on the disk because files can be read and written sequentially. Fragmentation is also reduced.

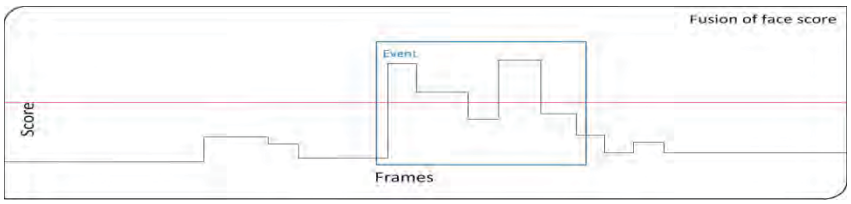
Search Algorithm

The search is iterative. One or more images are used to start the search and the results are filtered by a manual operator. All correct matches are used to execute the next search which will lead to better results. The face recognition algorithm used was Cognitec 8.7.0 and it provided an application programming interface (API) to search for one face against an array of faces. But in this project, the faces to search against are not independent, they came from a video. Also, there was not only one face to search for but a list of faces to be queried against the video. Two fusion techniques were implemented - one to fuse adjacent faces from the video and another one to fuse the results for the different queried images.

Fusion of Faces from Video

When a search is done, every face in the index is assigned a score for the question "Is the face from X?" Each frame from the video is assigned a score. For frames where the score is greater than a certain threshold they are merged into an event.



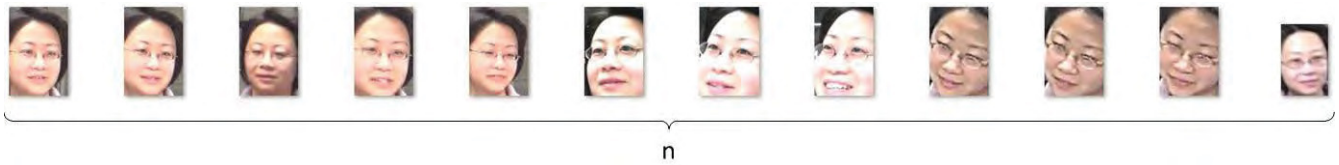


The event will contain multiple frames. The frames are fused together by taking their maximum and their sum (total score). An event is kept if:

- It contains at least two frames
- The maximum combined score is at least 50%

The events are presented in descending order of total score.

Fusion of Faces from Query



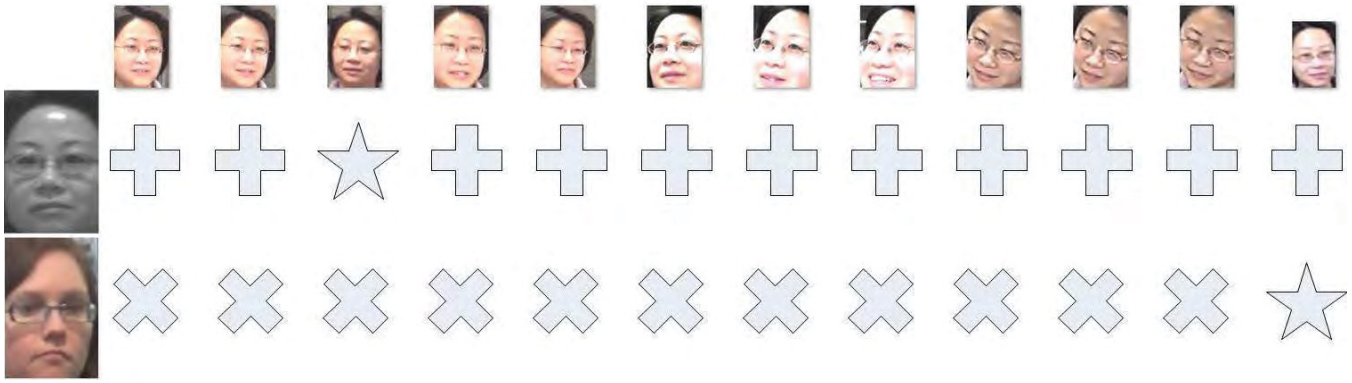
Scores from all faces are fused together in order to improve results. Two fusion approaches are described below:

The initial technique was to run the query "n" times and then take the frames where at least one query matched above a threshold. This increases the "true positive" considerably as having more images increases the amount of orientation of the face and leads to more matches. However, it would increase the "false positive" as well, since only one of the queries has to be above the threshold to pass the event.

The improved technique is still to run the query "n" times, but now it is required for at least one query to be above the threshold as well as the average of queries to be above a lower threshold. It is similar to the question: "Is there one query matching as well as most queries matching?"

In the image below, the star represents matching above the threshold (50%). The plus sign represents matching above a lower threshold (10%), and the X represents no match. In the first row, which is a true positive, one image matches well and all images match fairly well. This will be considered a candidate which can be sent to the operator.

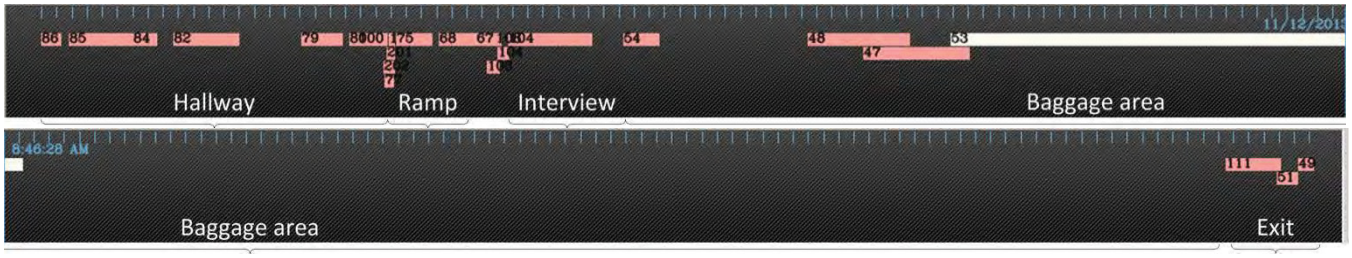
In the second row, which is a false positive, one image matches in error. It may be caused by the poor quality of the image, glasses, orientation, etc., however all other images do not match. It has a high chance of being a false hit, and will be removed from the results. This technique removed a lot of false matches.



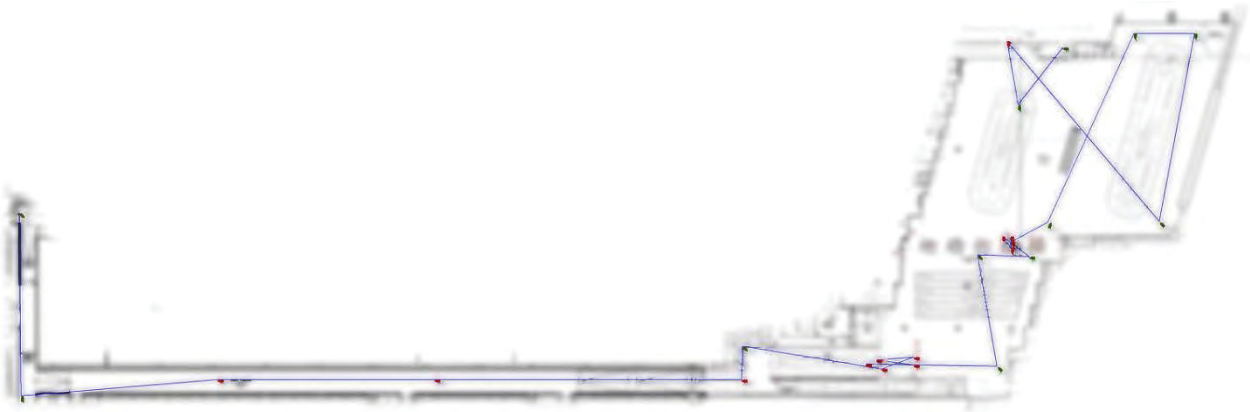
More research could be done on the fusion. This project used a basic technique which worked well with our dataset.

Export

Below is an example of a completed storyboard. The numbered blocks represent video footage in a multi-track environment similar to video-editing software. In the hallway, cameras 79, 82 and 84 were found using face recognition software as highlighted by “red” cameras on the map. Cameras 85 and 86 were manually added. In the ramp, as well as the interview area, all nine cameras were found using face recognition. Overview cameras 68 and 75 were manually added. In the baggage area, a few cameras were manually added until it was impossible to follow the individual any longer due to low resolution and occlusion. Thanks to the face recognition software, the individual was found back at the exit point, and could be manually tracked to the exit.



The software produces a storyboard, as well as a map, which can be stored in XML however access to all of the original video footage is required.

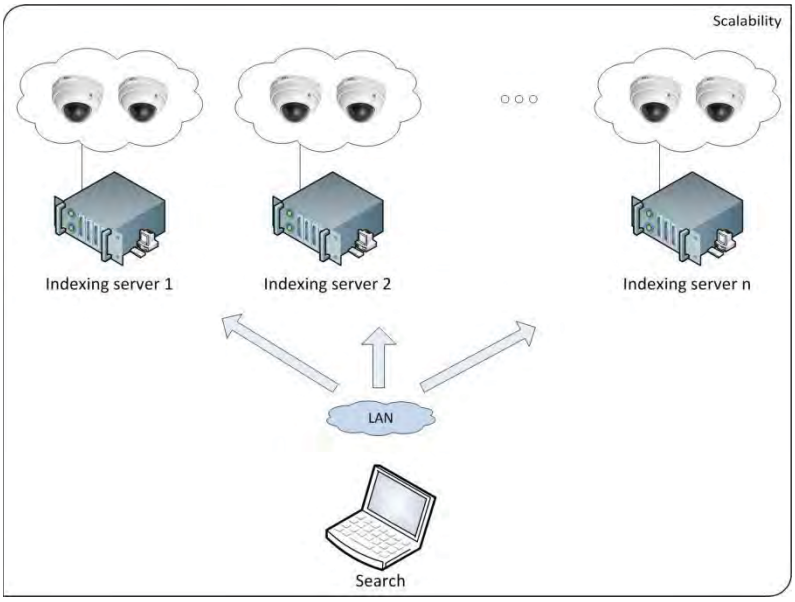


There is a possibility of exporting the project as an AVI. It concatenates video from all cameras and indicates the camera number and timestamp at all times. In cases where multiple videos are overlapping in time, they will be concatenated in the export.

More research needs to be done to make the export ready for court. Currently, the original aspect ratio and frame rate is not kept since the output video cannot change the aspect ratio, resolution or frame rate during re-encoding. Videos would have to be exported in their original form and a standalone player would need to be distributed to play them synchronized in their original format.

Scalability

In our proof of concept, only one server was used. It could potentially scale by having multiple indexing servers and a distributed index. The machine running the search would send the query to all indexing servers and would merge the results coming back before display. Using this technique it could easily scale to hundreds of cameras being indexed in real-time.



Indexing Server Count

The question is to know how many indexing servers would be required. If the indexing servers have a powerful graphic card (GPU), decoding could be done "for-free[7]" and the background subtraction could be run very efficiently on the GPU. For the decoding and background subtraction, 15 cameras at 1,080P and 30 FPS could easily be processed on one machine. Face detection would depend on the size of the faces being searched for and the amount of motion in the video. As seen in the following table, and assuming 100% motion, face detection speed ranges from 7 FPS to 33 FPS depending on the size of the face being searched for.

More research could be done to improve performance or to predict the correct number of indexing servers needed however a rough estimate would say that assuming 25% motion and a face size of 60 pixels between the eyes, a machine running 8 x CPU and 1 x GPU could index between four and eight cameras.

Using the GPU to detect faces using OpenCV[8] was also tried. It has a lower performance than Cognitec, however it only uses 12% of the CPU versus 100% for Cognitec. Having less CPU usage leaves more CPU available for creating templates. This method was not chosen as the OpenCV face detection misses more faces, especially when searching for small faces.

Pixels Between the Eyes	FPS Using Cognitec	FPS Using OpenCV (GPU)
Resolution	1,080P 30 FPS	
CPU Usage	100%	12%
15	7	X
20	9	X
25	11	9
30	13	9
35	17	9
50	21	15
70	31	20
100	33	23

Table 1 – Resolution vs FPS for face detection in video

Only cameras capable of face recognition would be indexed. For example, a system with 100 cameras may have only 20 capable of face recognition so only those 20 would be indexed, reducing the server load.

Index Size

As the index consists of one template per face detected per frame per person per camera, the index grows very quickly. For example, in our dataset, there were 83 people going through 30 cameras capable of face recognition and the index had 350,000 templates. The license model of Cognitec is to pay for a pre-defined number of templates. At the speed at which the index grew, it would be expected to have millions of templates created per day in a real environment.

Limit Template Creation

Currently, one person passing in front a camera for two seconds creates 60 templates on a camera running at 30 FPS. An optimization could be to limit the number of templates to the best three templates. It would reduce the size of the index, but at the same time it would reduce the quality of the search results. It happens quite often that only one of many templates matches as the quality of the image is generally poor using CCTV cameras.

[7] Decoding h.264 video on a recent graphic card is done in a special chip and takes neither CPU nor GPU power.
[8] http://docs.opencv.org/modules/gpu/doc/object_detection.html

Split Templates

Templates are divided per camera in blocks of 30 minutes. If the search is targeted on a limited number of cameras and for a certain time range, the number of templates to load in memory could be limited and it could fit within the Cognitec license. If the number of templates is still too large, a subset of the templates could be loaded and searched. Then, that subset could be unloaded to make space for the next subset. This would significantly slow down the searches.

Face Detection Capable Cameras

Using face detection capable cameras such as Panasonic, cameras could run the face detection embedded and send the faces over the network. This would greatly reduce the burden on the indexing server, however the accuracy of facial detection should be tested to see how small a face can be to be detected.

Results/Findings

Pre-Processing Time

The entire video for all cameras was indexed to extract all faces and create templates. It was a relatively slow process and having a fast enough indexing strategy made it possible to index cameras in real-time. The total duration for all cameras was 59 hours and indexing took a total of 45 hours.

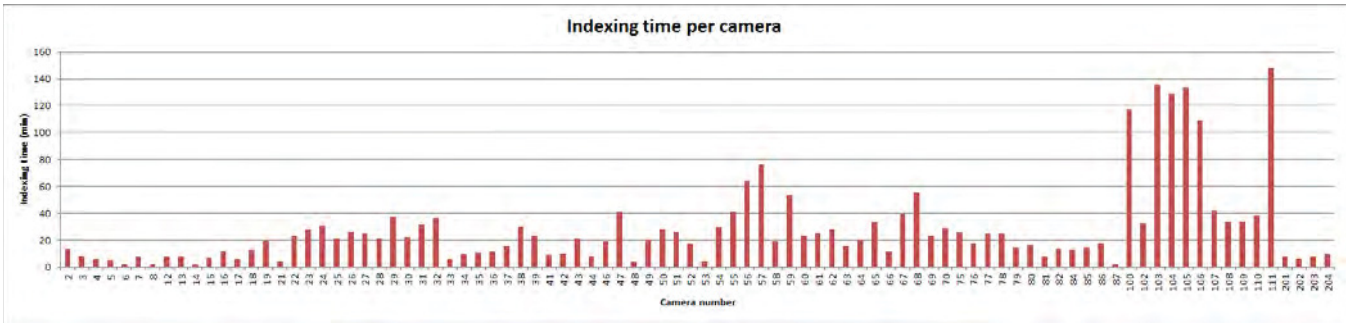


Figure 21 – Indexing time per camera

Indexing time is proportional to the amount of motion and to the number of faces seen. As seen in Table 2, in cameras where there is no motion, 38 minutes of video can be analyzed in about five minutes. On average, video was processed in 29 minutes with the slowest being 147 minutes[9].

	Video Length (minutes)	Average Indexing Time (minutes)	Frames Per Second
No motion video	38	5	400
Average	38	29	40
High motion video	38	147	8

Table 2 – Indexing time

The variance in processing time is due to the fact that face detection is applied only to areas where there is motion. In cases where there is little motion, processing time is very fast. It drops to about 8 FPS where the amount of motion is high because Cognitec face detection runs at about 8 FPS. This slow processing would be an issue when running in real time from real cameras as the feeds are coming at 30 FPS and in the worst case can be processed at a maximum of 8 FPS. An option would be to skip frames or use advanced buffering.

All of the videos were searched for faces having a resolution of at least 20 pixels between their eyes and faces of up to 15 pixels between their eyes were found. This is a very small resolution of faces to be searched. Searching for small faces takes considerably more time than searching for large faces and the face size to search for should be carefully chosen based on the requirements of the project.

Recognition Accuracy for *Chokepoint*

The software was run on the *Chokepoint* dataset for preliminary evaluation. It contains 1,281 events in the ground truth. Cognitec detected 1,276 events for a detection rate of 99.6%.

[9] Intel I7 3770K, 3.5GHz, 16GB RAM, Windows 7 ultimate with 2 Geforce 690

Subject	Events Recognized	Events in Ground Truth	Matches	Time to Search
ID 01	47	47	100%	-
ID 02	24	24	100%	-
ID 03	47	48	98%	-
ID 05	47	48	98%	-
ID 06	47	47	100%	4.5 min
ID 07	48	48	100%	3.5 min
ID 09	47	47	100%	2.5 min
ID 19	48	48	100%	-

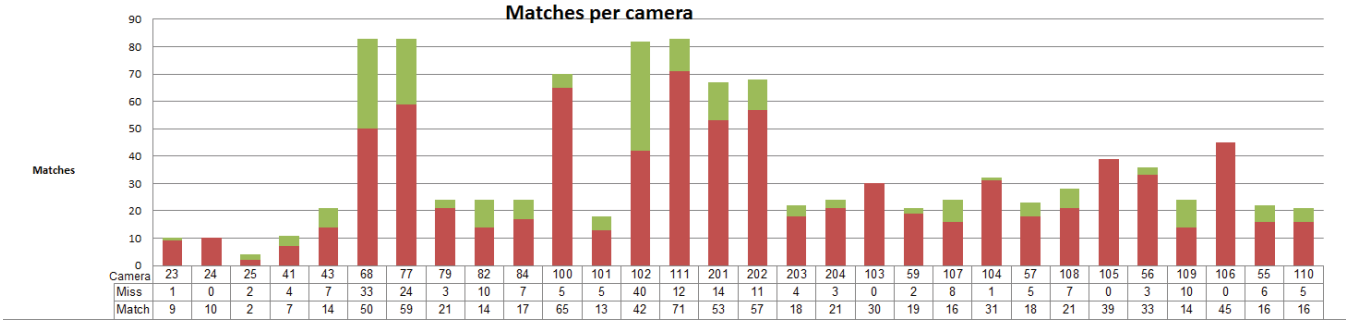
Table 3 – Matches per subject in Chokepoint

Recognition Accuracy for *Airport Face Recognition in Video* Dataset

Subjects Matched by Face Recognition Capable Cameras

Appendix 1 shows which subjects were matched by which face recognition capable cameras. Each row represents a different subject and each column represents a camera. The column where there is a red cell represents a face recognition capable camera. The results are summarized in the following section.

Performance per Camera

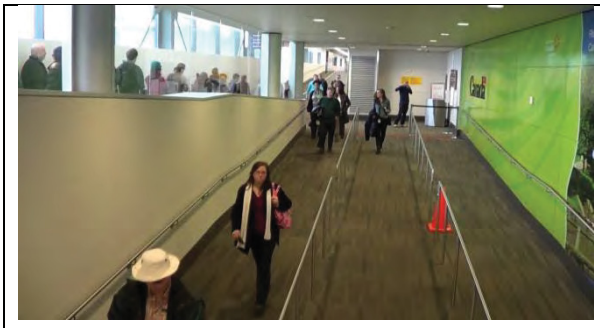


Camera	23	24	25	41	43	68	77	79	82	84	100	101	102	111	201	202	203	204	103	59	107	104	57	108	105	56	109	106	55	110
Miss	1	0	2	4	7	33	24	3	10	7	5	5	40	12	14	11	4	3	0	2	8	1	5	7	0	3	10	0	6	5
Match	9	10	2	7	14	50	59	21	14	17	65	13	42	71	53	57	18	21	30	19	16	31	18	21	39	33	14	45	16	16

Figure 22 – Matches per face recognition capable camera

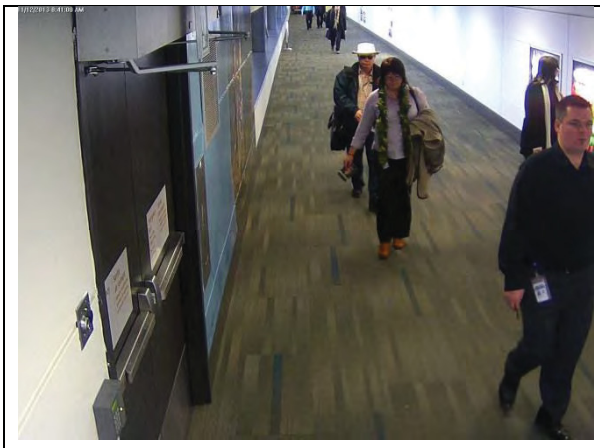
The results in Appendix 1 show that some cameras perform much better than others. In Figure 22, the number of matches per camera is indicated in red and the number of misses is indicated in green.

Different camera locations and settings are shown in the table below in order of match rate. Camera location, camera field of view and facial orientation are elements which greatly affects the results.



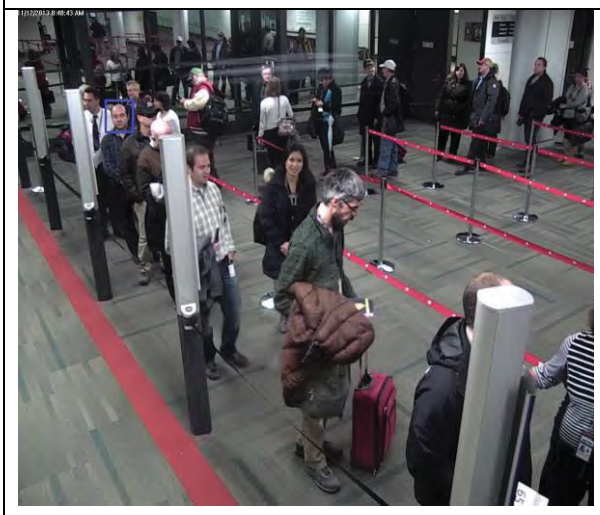
Camera 102

Match: 42 people
Miss: 40 people
Match rate: 51%
Px between eyes: 8 – 20
The camera's field of view is too wide and the image is interlaced. However, it is surprising to obtain a 50% match rate in such hard conditions. It might be because the camera angle is good and the passengers can be seen for a long distance.



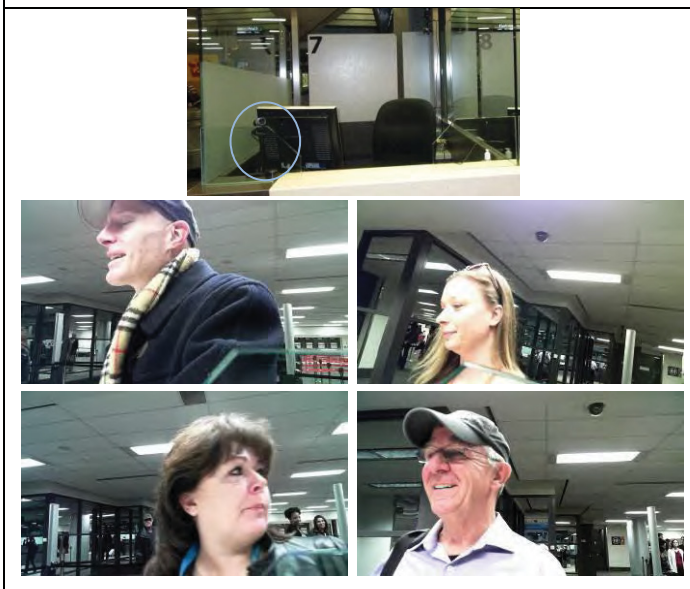
Camera 82

Match: 14 people
Miss: 10 people
Match rate: 58%
Px between eyes: 15 – 35
The camera is not in a good location since the camera is located on the left and most people walking on the right are missed.



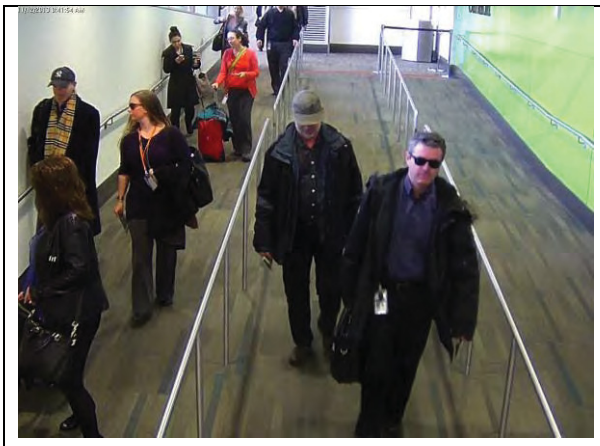
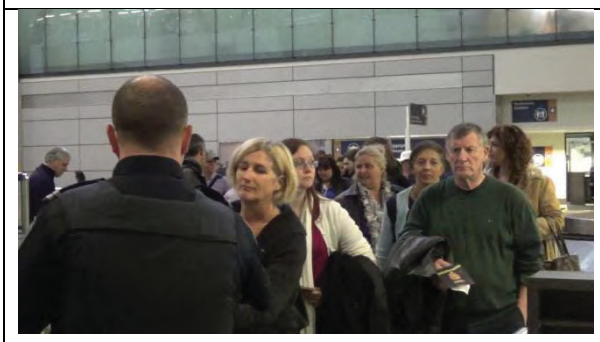
Camera 68

Match: 50 people
Miss: 33 people
Match rate: 60%
Px between eyes: 18 - 20
The camera is not located in an ideal position as many individuals do not look directly at the camera and the resolution is low. The match rate is relatively high since there is a lot of data as the passengers wait in line and have a high chance to have the perfect pose.



Cameras 107, 108, 109, 110

Match: 67 people
Miss: 30 people
Match rate: 69%
Px between eyes: 100 – 300
The camera is in a very bad location. It is at eye level and the resolution is very good, however it misses a lot of people because of the orientation. It was impossible to have the camera centered as it would have been in the officer's way. Having the camera close to the traveller but on the side leads to a large angle.

	**Camera 77** Match: 59 people Miss: 24 people Match rate: 71% Px between eyes: 20 – 30 The camera is in a good location because it is located at the bottom of a down sloping ramp. Most faces are towards the camera. Resolution is low and when it is higher because the person is closer to the camera, the image quality becomes blurry.
	**Cameras 55, 56, 57, 59** Match: 86 people Miss: 16 people Match rate: 84% Px between eyes: 60 – 80 The camera is above eye level and this makes capture challenging. Because of the angle, wearing a hat makes recognition impossible.
	**Camera 202** Match: 57 people Miss: 11 people Match rate: 84% Px between eyes: 50 – 70 The camera is at eye level and has a very small field of view. There is a large crowd of 83 people walking in front of it and there are significant occlusions. The match rate is good but most people are found with only 3 - 10 frames. People are missed mostly because they are hidden by another person.
	**Camera 111** Match: 71 people Miss: 12 people Match rate: 86% Px between eyes: 40 – 60 The camera is at eye level and has a small field of view. There is a large crowd of 83 people walking in front and there are significant occlusions. The match rate is good but most people are found with only 3 - 10 frames.

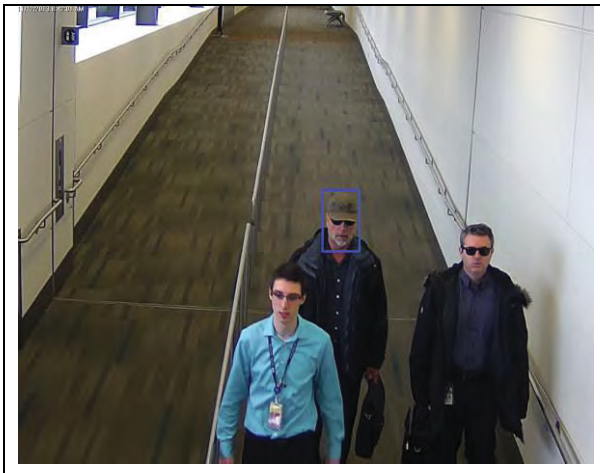
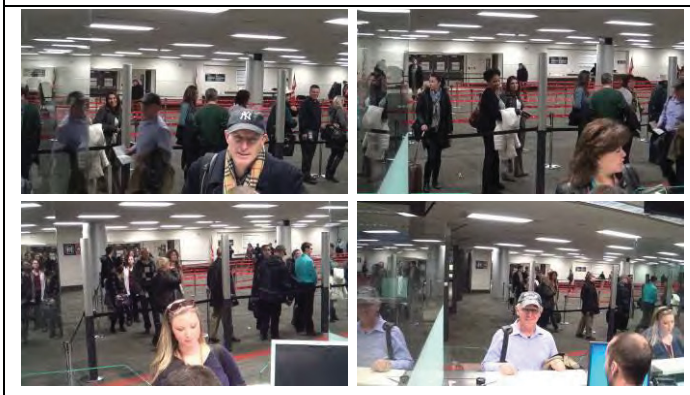
	**Camera 79** Match: 21 people Miss: 3 people Match rate: 87% Px between eyes: 20 – 30 The camera is in a good location as most people face the camera. The match rate is quite high.
	**Cameras 103, 104, 105, 106** Match: 82 people Miss: 1 person Match rate: 99% Px between eyes: 70 – 100 The camera is almost at eye level and has a very small field of view. The resolution is excellent. This would be the perfect location and field of view for a camera in an interview location.

Figure 23 – Vantage points and match results for selected cameras

Performance per Individual

As seen in Figure 24**Error! Reference source not found.**, the match rate varies a lot per individual. Some individuals such as ID 4, 6, 18, 24, 26, 29 etc. are matched at more than 90%. Some other individuals are matched lower rate such as ID 10, 20, 28 etc.

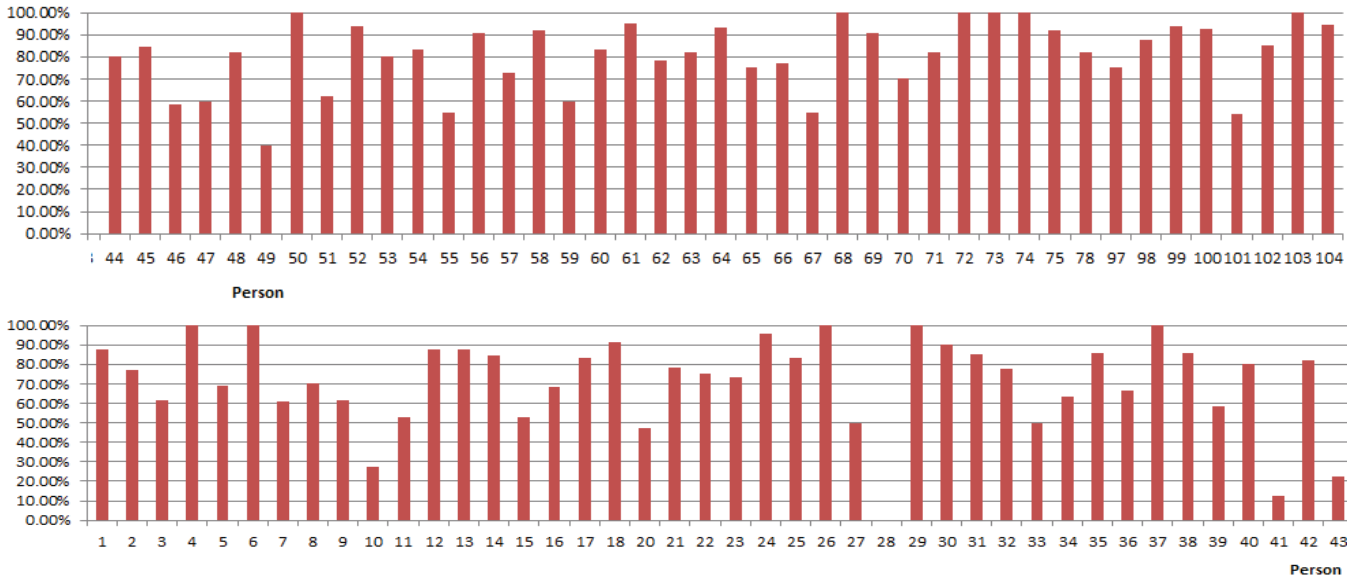


Figure 24 – Matches per person

Effect of Facial Accessories

The effects of facial accessories such as sunglasses and hats have a large impact on performance. In Figure 25 individuals in light green are wearing either a hat or sunglasses and individuals in dark green are wearing both a hat and sunglasses. Having both a hat and sunglasses reduces matching below 50%, and even to 0% in the case of ID 28.

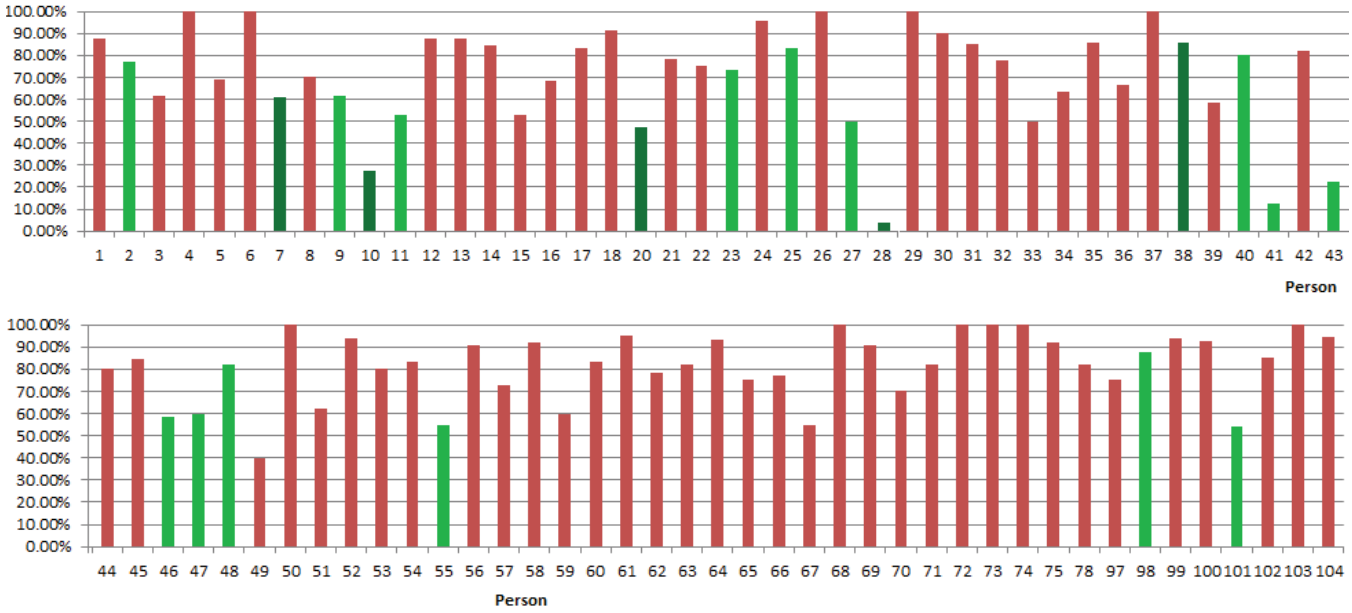


Figure 25 – Per subject match rate given presence of facial adornments

The following table shows the number of individuals categorized by whether or not they were wearing facial accessories:

Adornments	**Number**	**Percent**
Hat (no sunglasses)	**11**	**13%**
Sunglasses (no hat)	**6**	**7%**
Hat and Sunglasses	**5**	**6%**

Neither hat or sunglasses	**61**	**73%**
Total	**83**	**100%**

Table 4 – Numeric breakdown of subjects with facial adornments

Note: Table 4 Error! Reference source not found.**shows that significantly more persons in the dataset wear neither hat nor sunglasses. Though this proportion is not necessarily an accurate representation of reality, a certain bias does exist, and an assumption was made to simulate airport conditions.**

As shown in Figure 26, the match rate is significantly higher for subjects wearing no adornments compared to those wearing both hat and sunglasses (83% and 44% respectively). This is to be expected. Since our matching algorithm relies on face recognition, match rates are correlated with the visibility of facial features. The higher matching rate of "hat only" compared to "sunglasses only" (65% vs. 55% respectively) is also consistent with our expectations since, from a biometric perspective, the ocular region is feature-rich. What is somewhat surprising is the relatively high level of match success when the subject is wearing both hat and sunglasses. While in our experiments, we obtained a match rate of 44%, this is dependent on environmental and subject-based considerations**.**

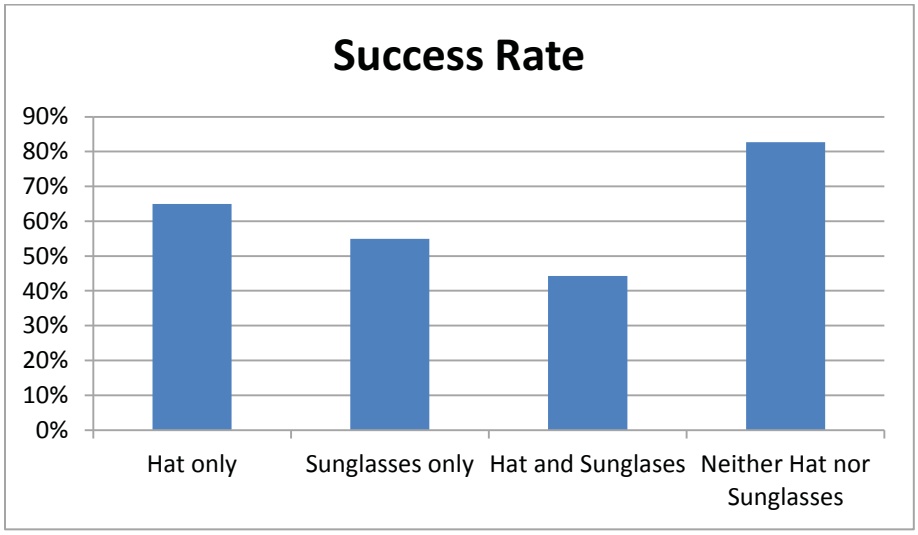
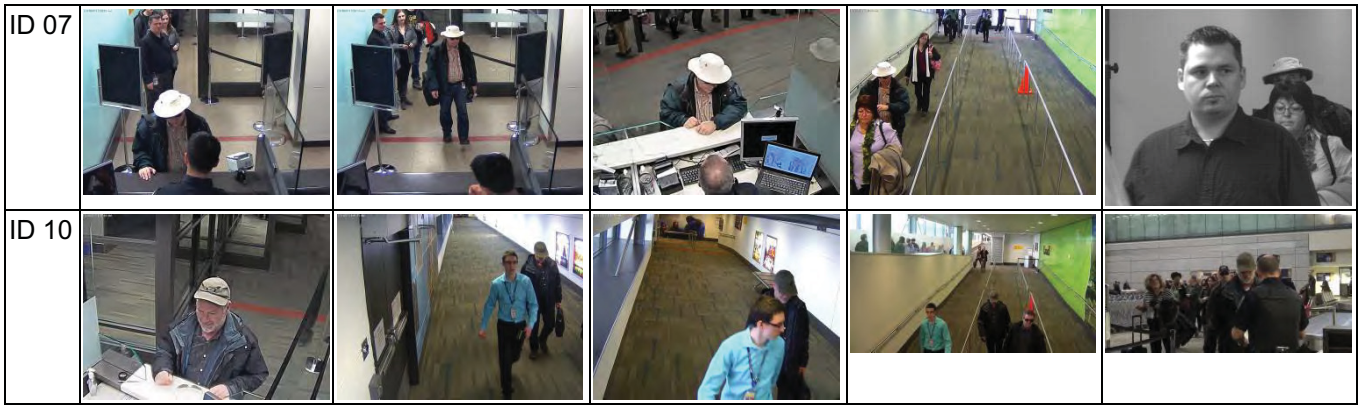


Figure 26 – Match rate given presence of facial adornments

Sample of misses due to sunglasses or hats:



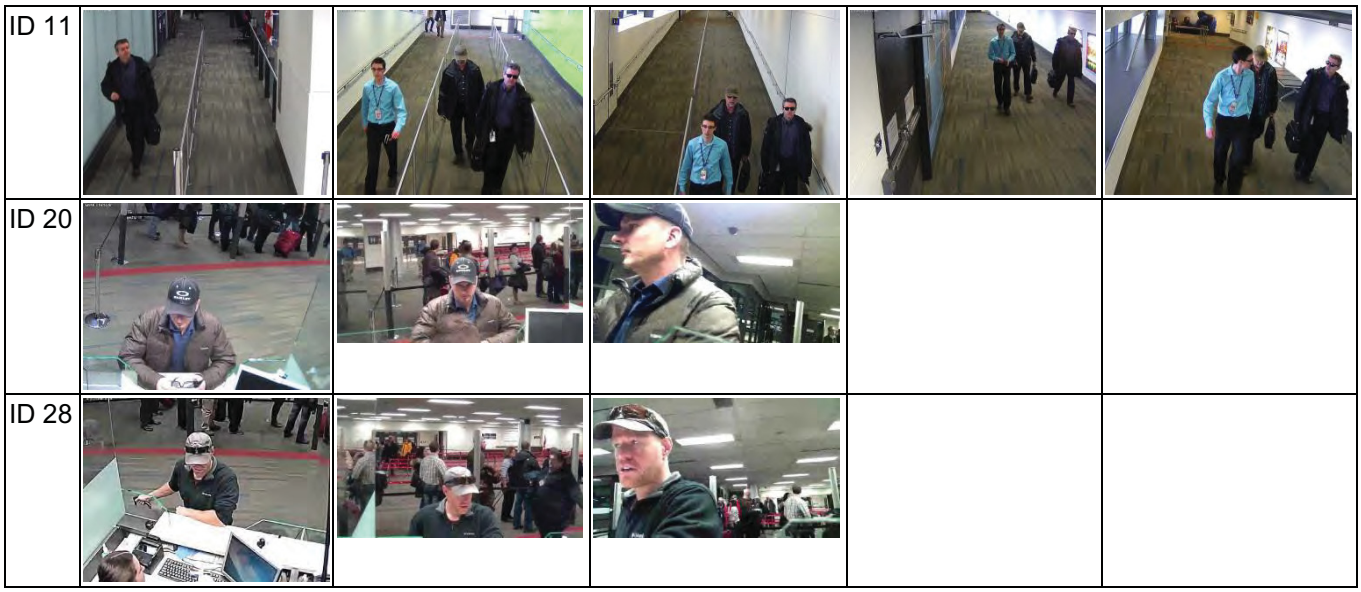
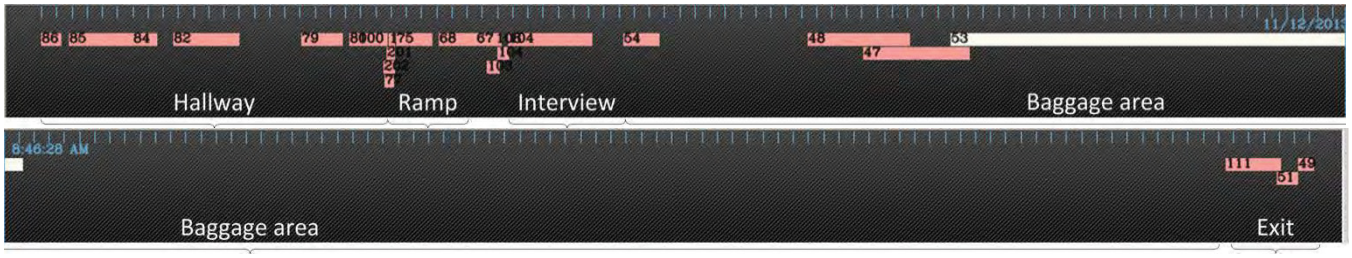


Figure 27 – Examples of misses (per person) with facial adornments

Time to Gather All Video Exports

On average, it takes a trained operator five to ten minutes to find video related to an individual, to complete the storyboard and to export it. An example of a complete storyboard is shown below.

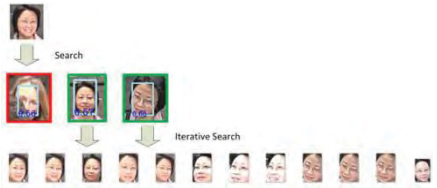


There are gaps in the baggage area video because it is very difficult even for a manual operator to find an individual in a large crowd.

Figure 28 – Overview of a crowd

In the case where a person is found with face recognition, it is simple to manually track them. However, if the person moves out of the field of view or is hidden by other people or objects, it can be difficult to find them again by looking at the video. In the picture above, the only person who it is easy to get an entire storyboard for is the person with the light blue shirt.

Effect of the Quality of the Initial Picture



The searches are iterative - there is an initial search with either a good quality picture taken from a professional still camera or from an image taken from video. The results are then manually filtered and the faces from the correct matches are added to the query in order to improve the results.

Figure 29 shows that, on average, the first iteration of a search from one image yields 40% of the matches. As more images are added, the average match increases to 75% in the fifth iteration. Six iterations were tested but the match rate did not increase.

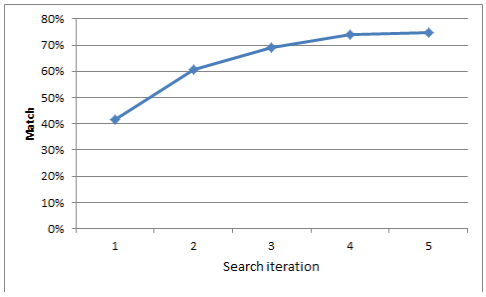


Figure 29 – Average matches per iteration

Adding more images to a search increases the match rate since it adds more information about orientation. The initial image tends to be a frontal view since it comes from a still picture or from a frontal camera during the interview at PIL. The first few hits will also be somewhat frontal because they were matched to the frontal image however they tend to have an angle since they come from chokepoint cameras. As the process iterates, more images are added with different orientations and matching confidence is increased.

The quality of the initial image is important. The effect is shown in Figure 30 where matches from high quality mugshots are compared to matches from a lower quality video. In the first iteration, the search from a high quality image yields 50% correct matches compared to 33% from a lower quality image. The search is noticeably faster from a high quality image where 70% of the matches are achieved in two iterations. It takes four iterations to have 70% of the matches from a lower quality image. Searches from mugshots have 5% more matches, even at iteration five where the search from video could not be further improved. The match rate does not reach 100% because there are people who do not look at the camera and who wear hats and sunglasses. There are instances where the face is clearly visible but are not matched because of algorithm failure.

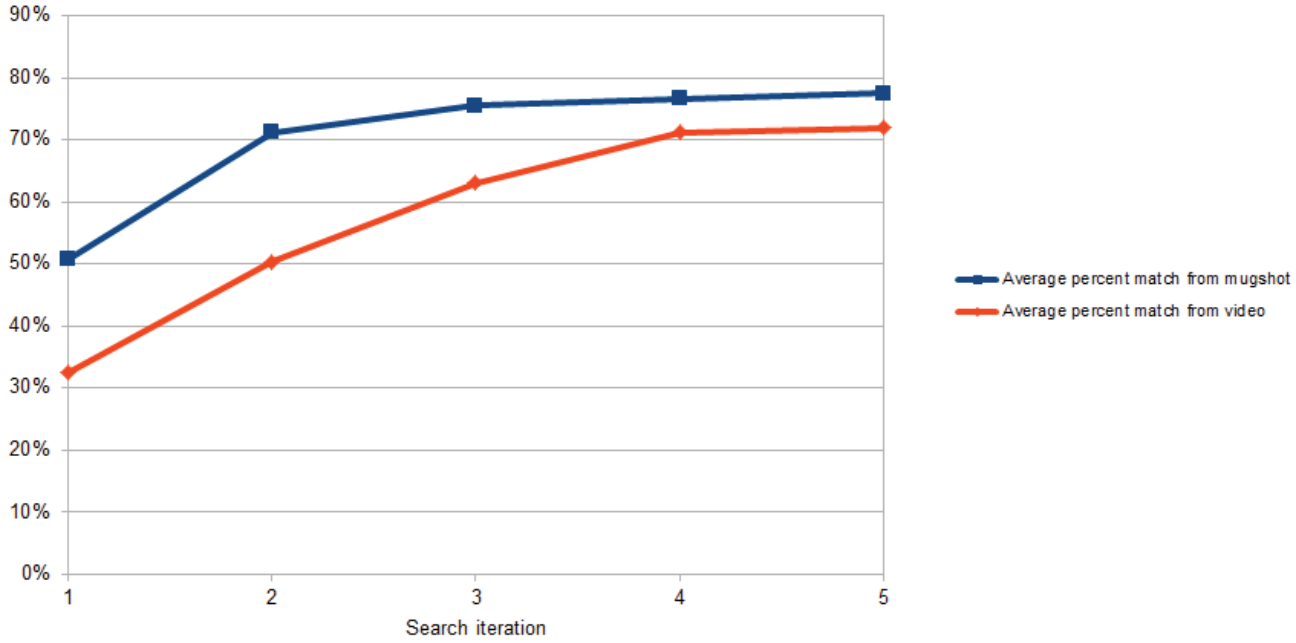


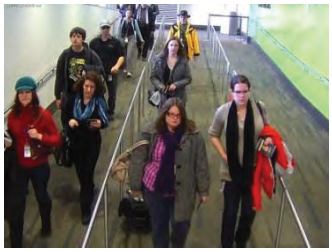
Figure 30 – Matches from mugshots and from video

Mug Shot Quality	Video Quality	
		
		

Table 5 – Comparison of mugshot to video quality

Searching in a Single Chokepoint Camera

It is known from Figure 30 that search accuracy increases with more search iterations. What if there is only one chokepoint camera available? A search for every individual was made from a mugshot picture for Camera 77 only. A sample image from that camera is seen on the right. With one image to search with, the match accuracy is 33%. Only 27 people were found out of 83 compared to a match accuracy of 69% for Camera 77 when searched iteratively with more chokepoint cameras. Thus, having more cameras improves results.



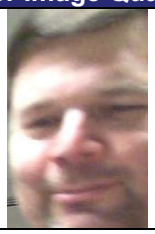
Match for Camera 77

Iterative search	**69% match**

Optimal Camera Placement for Face Recognition

The main challenge with face recognition from video is where to place the camera in order to get the best frontal image. The success of facial recognition improves significantly with better image quality. For a facial image to be of high quality, it needs:

- Frontal pose: < 10 degrees in any direction
- Sufficient resolution: at least 20 pixels between the eyes
- Minimal motion blur

OK Image Quality	Poor Image Quality
	

Note: More resolution is not always better. Often matches only happen in the middle of the image because of facial orientation and motion blur. Below, two non-matches are shown at 15 and 22 pixels, due to the lack of resolution. And three non-matches are shown at 39, 68 and 71 pixels due to poor orientation and motion blur. As a person moves closer to a camera, their relative speed increases and this creates motion blur.

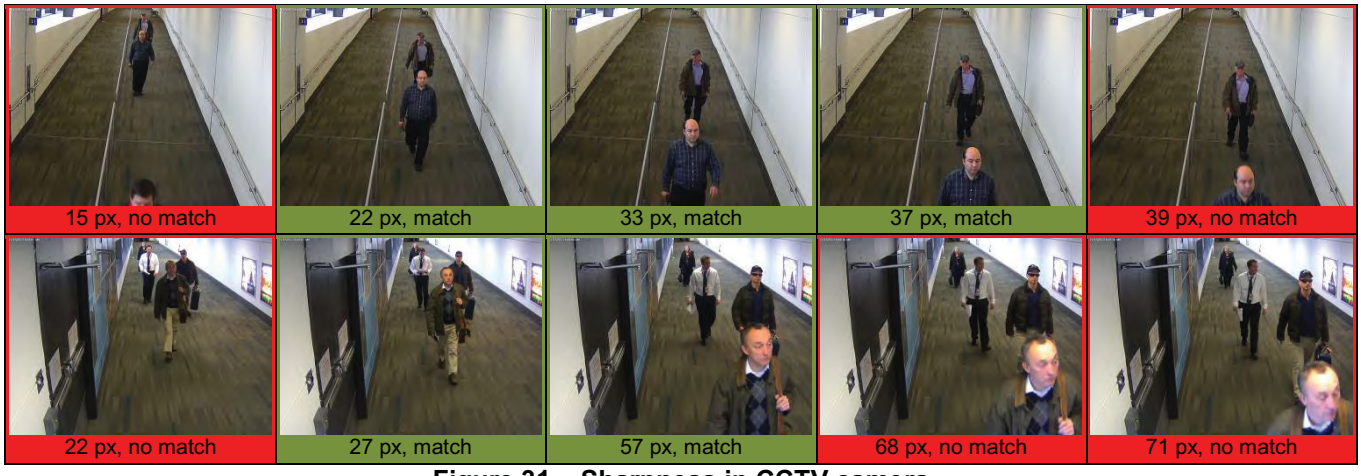


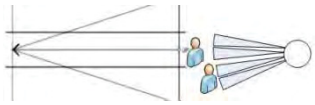
Figure 31 – Sharpness in CCTV camera

Eye-Level vs. On Ceiling

Most CCTV cameras are installed on the ceiling but if there is a need for facial recognition they could be installed lower or at eye-level. Is it better for facial recognition to have the camera installed high on the ceiling or at eye-

level? While many face recognition companies recommend eye-level, it may not always be the optimal position in a crowded environment.

Eye-Level



Our dataset contains five cameras located at eye-level and while they give great results, they miss a few cases because the occlusion is greater and the duration of events is much shorter. We had a tower with four cameras located as shown in the drawing on the right. With a resolution of 1,280 x 1,024 and a focal length of 25 mm, the facial images were of a great quality. However, as seen below it has an issue with occlusion where one passenger was hidden during its entire transaction. If we count the people matched on any of the four cameras, 71 people are matched out of 82 for 87% accuracy. The occlusion problem is significant.

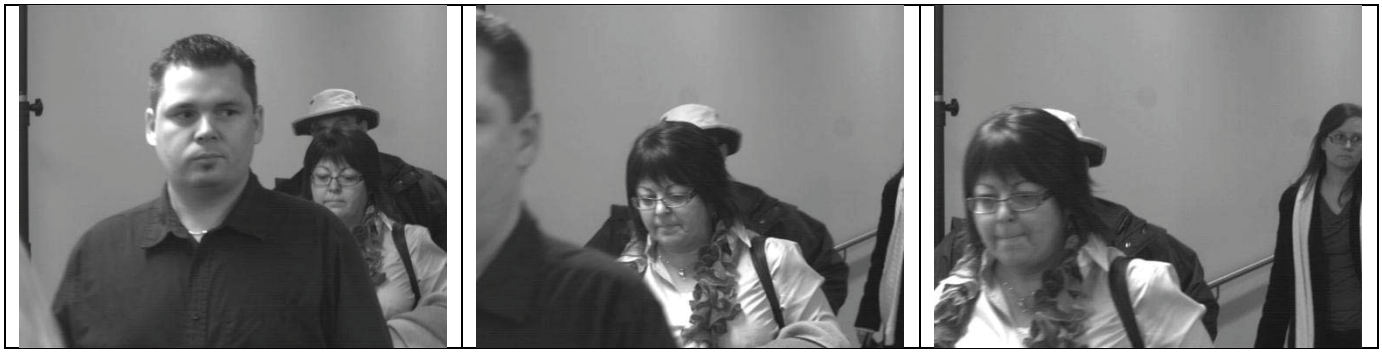


Figure 32 – Occlusion in the tower camera

Figure 33 – Occlusion in Point camera

There was another eye-level camera at the exit of the baggage area and a queue of 83 people passed quickly in front of it. On this camera, 71 people are matched out of 83 for an accuracy of 86%. While there is major occlusion as seen in the image below, it does match a large number of people. Most of the matches were 0.1 - 0.5 seconds in length from this camera. They remained visible for only a few frames and are then occluded again.

Ceiling

Cameras mounted on the ceiling have an angle looking down, and sometimes a horizontal angle. On the positive side, they have much less occlusion as different travellers will be at different heights.

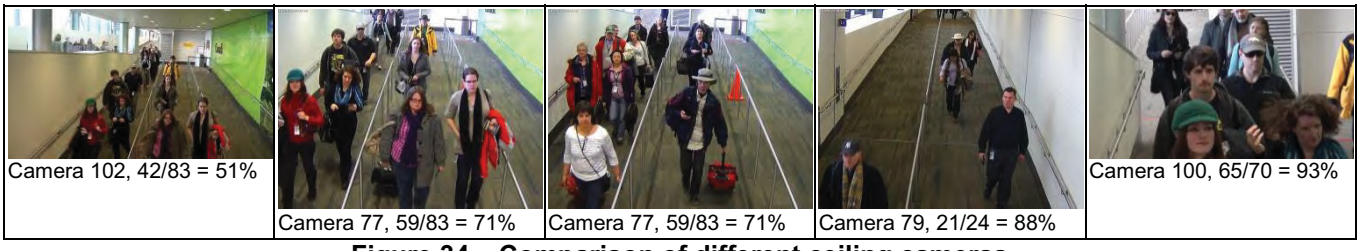


Figure 34 – Comparison of different ceiling cameras

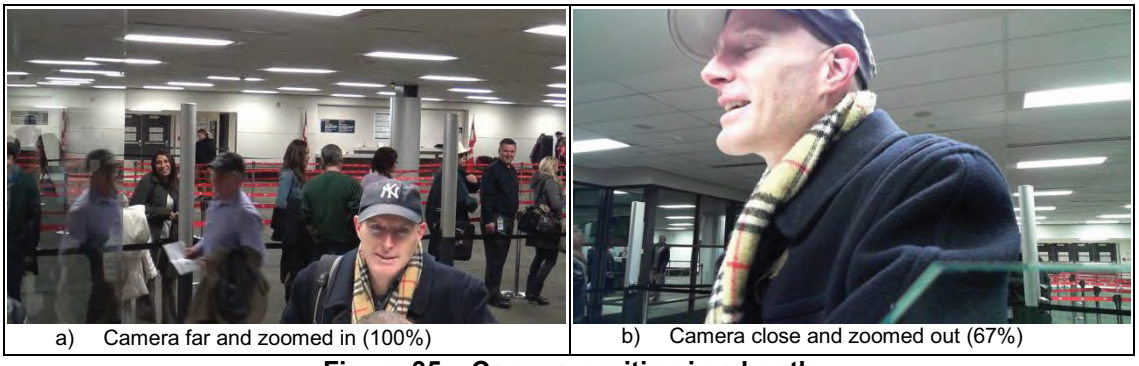
In Figure 34, multiple ceiling cameras are shown. There is minimal occlusion in the facial area, especially when passengers are walking down the ramp. The events last longer and the occlusion is reduced in camera 100 where 65 people are matched out of 70 for an accuracy of 93%.

Optimal Field of View

Wide angle cameras are a poor choice for face recognition. As can be seen in Figure 31, with wide angle cameras the resolution is poor in the center of the image and significant motion blur is present as people move closer to the camera. Figure 34 shows five cameras with different fields of view from left to right. A wider field of view leads to a matching rate of 51%. As the field of view gets narrower, the matching rate increases. The figure on the right achieves a matching rate of 93% because it has a very narrow field of view. The optimal field of view would be as narrow as possible to cover one lane without missing any people. In some scenarios, additional cameras might be needed to cover multiple lanes.

Interview Camera

When a camera is located in a booth, it is better to have the camera further from the subject and zoomed in rather than close to the subject and zoomed out, as can be seen in Figure 35. Having the camera near the subject distorts the face, and it is difficult to have a frontal shot. The field of view in Figure 35 a) is more optimal than the field of view in Figure 35 b).



a) Camera far and zoomed in (100%)

b) Camera close and zoomed out (67%)

Figure 35 – Camera position in a booth

Optimal Overview Cameras

Overview cameras are used for a manual operator to have an overview of the surveillance area. It is likely that a person who is being tracked will be lost in the crowd when they are out of the field of view of the camera or is in a blind spot. To help the operator, overview cameras with a wide angle would be optimal. Having a 90 degree angle camera in corners, a 180 degree angle camera on a wall, or a 360 degree camera in the center of the large room would help. In Figure 28, where a crowd is shown, it is relatively easy to follow someone who is in the field of view, however it becomes a difficult task as the person moves out of the field of view since they have to be manually recognized again. It is much easier to track someone than to recognize someone.

Benefit of the User Interface

The map and storyboard makes searching much easier. In a standard VMS, searches and exports are usually done on multiple cameras for a time range. For example, exports from 14:10 to 14:20 from cameras A, B and C. This format lacks the ability to make a storyboard and follow an individual when they move from camera to camera. In an airport, the storyboard of one traveller contains about 25 cameras and can last 30 minutes. It is easy to get lost in a VMS and forget at what time the person was in each camera.

The export from a VMS would be either:

1. Export from all cameras for the time where the suspect enters the first camera until he exits the last camera.
2. Multiple exports, one per camera at the time where the suspect is visible.

In Option 1, the export needs to be attached with a document indicating the time the subject is in each camera while Option 2 is a cumbersome process to export and then review.

The user interface is optimized for a storyboard. A timeline indicates at what time and in which camera the subject appears. It is easy to see the path, as well as setting the start and end times in each camera. The storyboard shows the gaps where the suspect is not visible.

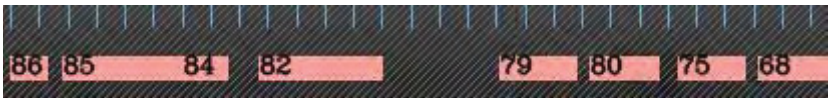
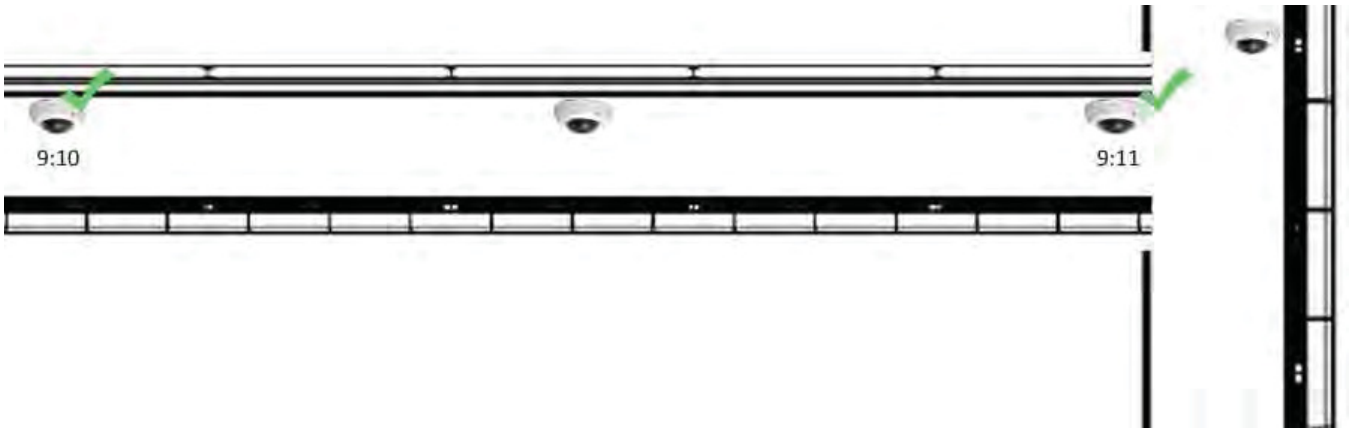


Figure 36 – Storyboard

The map helps to quickly find footage and follow a person. In the map below, the subject was found at 9:10 and 9:11 in two cameras with facial recognition. The camera indicating 9:10 can be selected to find the subject in that camera and selecting the camera on the middle will show the video at 9:10. When the video is forwarded slightly, the subject will have to pass by that camera and then he can be tagged. The same principle can be applied by selecting the camera indicating 9:11 and then clicking the camera on the other hallway to show the individual.



It takes an average of five minutes using the user interface to find all video evidence from 93 cameras totaling 58 hours. It was not tried to find all video evidence using a VMS, but it is estimated that it could take half a day.

Prototype

A demonstration standalone prototype was created and is available for viewing. The searching software runs well on a laptop with a pre-computed index of a pre-recorded video dataset.

Discussion

Image Quality

The quality of the image greatly influences the search accuracy. The main challenge is to place cameras to have a clear frontal view of everyone.

In order to make a facial recognition project successful, the following suggestions should be considered:

- Zoom in / add cameras in order to obtain good quality faces
- Use existing CCTV cameras and add post-processing and a manual operator

In a post-event search scenario where false-positives are acceptable, it is possible to use low-quality images. As a watch-list, it is impossible to use low-quality images. The number of false-positives would activate alarms that would be distracting and would impact travellers.

While it is possible to search for low-quality faces, a significant amount of server resources are required. It might be more cost effective to add a face-recognition camera in addition to existing CCTV cameras in order to reduce server and operator resources. The server would spend less time searching and indexing faces and the operator would obtain more accurate results.

Camera Position

Positioning of cameras is a challenge. The goal is to have a camera capture an individual for the longest duration while keeping a good resolution on the facial area. In our experiment, it was possible to recognize 60 - 80% of people with as much as eight people shoulder to shoulder. This corresponds to as low as 21 pixels between the eyes. While it is possible, it is not the optimal setting.

In an un-crowded area, the optimal camera placement would be at eye-level with a lens zoomed in enough to see a maximum of three people shoulder to shoulder, or about 64 pixels between the eyes.

In a crowded environment, the optimal camera placement would be located at the bottom of a down sloping ramp. This significantly reduces the occlusion between individuals. The field of view should be the same as in the un-crowded area.

The best position for a camera is when it is located far from the subject and zoomed in as opposed to being near the subject and zoomed out. Distortion is reduced when it is further back.

Occlusion

Occlusion is a major problem for an automated system, as well as the operator.

Automated System

For an automated face recognition system, occlusion from hat and sunglasses significantly impacts results. Occlusion from people in a crowd impacts the results, hidden persons will not be matched. To mitigate the effect of hats and sunglasses, cameras should be installed at eye-level. Additional lighting or infrared lighting would improve the results as hats often reduce top-down lighting. To mitigate the effect of occlusion in crowds, having a camera higher on a ramp reduces the occlusion.

Manual Operator

Occlusion impacts a manual operator in the overview area. As it is much easier to track someone than to recognize them, having overview cameras with wider fields of view would help the operator. Another option is to have an array of cameras pointing down with software stitching the images. Having a bird's eye view image of the entire area would greatly improve the efficiency of the operations.

FR License

Cognitec and other face recognition algorithms have a license per person enrolled. The problem is that each person will have many images enrolled in the database. As an example, with 30 face recognition capable cameras and an event lasting two seconds at 30 FPS, a single person would use 1,800 licenses for facial images. With optimizations reducing the number of images to three per event, a single person would use 90 licenses and the match accuracy would be reduced. To mitigate this challenge, a special license could be negotiated with face recognition companies or searches could be limited to a certain time and camera range.

Production Implementation

The current implementation is not ready for production. Files are read from disk in order to index them. This makes the process much easier as there is always data available and data can wait when the server is busy. In order to make the technology production ready, it would need to be connected directly to the cameras or the video management system. When connecting to a camera, the camera will not slow down when the server is busy. The solution is to either buffer the video, process it later or to skip frames. Buffering leads to memory issues and delay in processing while skipping frames leads to data loss.

The indexing could be sped up by using a powerful server with arrays of graphic cards (GPU) to detect faces quickly. Every face needs to be converted to a template using the face recognition software API, and this process can only be sped up by paralleling it on multiple CPUs/cores.

There are CCTV cameras available which have embedded face detection such as the Panasonic WV-SP306, WV-SP305 or WV-SC385. With a face detection capable camera, only the faces are sent to the server. This would considerably reduce the burden on the indexing server. The conversion of faces to templates must still be done on the server. These cameras were not tested but they should be evaluated to measure how small faces can be in order to be detected. If many faces are missed, accuracy will be lower.

Conclusion

Using face recognition on CCTV cameras will help an operator perform post-event search and retrieval.

This project is about having a predictable path in a semi-constrained environment. The use of chokepoints is a very good way to constrain a crowd in order to have them pass in front of a camera. To have success in a face recognition project, the quality of the image and the position, location and synchronization of cameras are crucial as match accuracy improves when the number of faces linked to an individual increases. A blurred picture or a person looking away from the camera reduces the likelihood of a match. To improve results, it is better to have as many chokepoint cameras as possible. This helps the face recognition algorithm to build a better model of the face. Accuracy is expected to be lower as the number of people increases. In order to maintain a high level of accuracy, it is expected that the search would be limited to a specific time range in order to limit the amount of data to search through.

A map is used to search camera footage where there is no information from face recognition cameras. The user interface and map helps the operator find all related video footage. With this hybrid search and retrieval method, it becomes easy to follow a person from camera to camera and create the storyboard of events.

The system was tested on a mid-scale database from real CCTV cameras within an airport border clearance operation. For static, predictable CCTV installs, all subjects except one were found in at least one camera. On average, all video footage was found in five to ten minutes. The results were not perfect and there is a reasonable amount of false-positives, however these can be filtered quickly and the overall process of extraction and export is accelerated.

Next Steps

Open Environment

The experiment was done in a semi-controlled environment moving people through chokepoints, standing in front of a booth and following a predictable path. As a next step, it could be tried in a less constrained environment such as outdoors. There would be few chokepoint areas, lighting would not be uniform and there would be no predictable flow of people. It would be a challenge for facial recognition.

Improve Search

The searching algorithm could be improved to give more accurate results using less iteration. Research on a better fusion algorithm could be done to give more weight to correct events using less training data.

Currently, the search is done through all footage from each camera. There are more chances of false hits when searching through a lot of video footage. Information from the first successful matches could be used to rank the results. Assuming a match is tagged at 9:00 in Camera A, knowing that Camera B is five minutes from Camera A, it makes sense to rate the matches in Camera B higher if they are around 9:05. Similarly, the results from Camera C around 9:20 should have more weight.

By fusing the scores from face recognition with the expected timestamps, no events would be lost. Instead the events would be ordered more accurately.

Remove the Threshold Selection

There is currently a threshold selection in order to indicate how strict the software should be. Lowering the threshold leads to more matches as well as more false matches. On the other hand, increasing the threshold would reduce false positives at the risk of missing an event.

To make the threshold selection easier, the following three presets were added:

	FAR
Low	**0.05%**
Medium	**0.1%**
High	**0.01%**

Even with presets, it is difficult for an operator to change the threshold. When set at “low”, the operator would review 10 - 100 false positives while setting it to “medium” would reduce false positives to one to two. Therefore, having an automatic threshold with a possibility of overwrite would be beneficial.

Results could be limited to a fixed number per cameras to save manual processing time, however this is an artificial constraint and does not truly limit the amount of false positives.

Another solution would be to automatically set the threshold to “high” and let the user review the results. When done, the threshold could be lowered to “medium” and start the review again.

Still another solution would be to let the user rank a few results as positive and negative and as the process proceeds, change the threshold to get the positive events as well as masking the negative events.

Improve Indexing

During indexing, face detection is a lengthy process. Face detection is currently run on every frame where there is motion. In order to speed up indexing, face detection could be run every five to ten frames and face tracking could be used to follow the face in the flow. This way, the face detected and the faces tracked would be converted in templates and be searchable. Face tracking is much faster than face detection.

Blob Search

This automated search is done using face recognition and it is successful at finding people in chokepoints while their faces are visible. However, it is not capable of finding someone in a crowd from an overview camera. It also has challenges finding people with hats and sunglasses. Below are suggestions of solutions in order to track someone with visual search instead of facial search.

- Partner with MetScan which created the software TagStream integrated with Milestone.
- Use scale-invariant feature transform (SIFT) points to search for points of interest. The design of a hat or clothing could be searched on. This is efficient when searching for sharp objects with corners but it has issues with faces and the human form because there are no sharp points of interest.
- Use histogram of oriented gradients (HOG) for human detection then tracking. HOG is successful at detecting a human form in an overview camera. Face recognition could be used to find the person of interest in a chokepoint and then map it to a calibrated overview camera. HOG and blob tracking could be used to follow the person from the chokepoint in the overview camera.

In a large, crowded room, the best placement of an overview camera to eliminate occlusion would be a camera placed on the ceiling, pointed downward. If the ceiling is too low, an array of cameras every meter could be installed. Then the images would be stitched together to have a bird's eye view of the room.

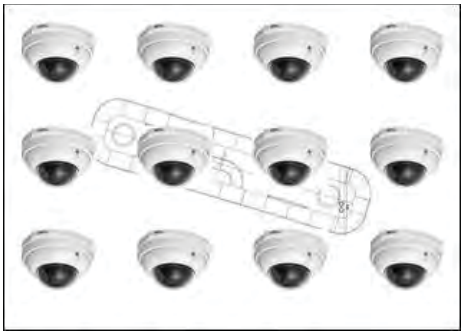


Figure 37 – Array of camera

Video Management System (VMS) Integration

This system is currently standalone and is not integrated in any VMS. Integration in VMS would have the following benefits:

- Seamless searches from the VMS, and
- Access to all cameras and video recorded from the VMS

Milestone has a video analytic framework as seen in Figure 38. Video is captured in the Milestone system and is then sent to the Video Content Analytics (VCA) server running the analytics. This server would populate the face recognition index. Alerts would be formatted and sent back to the Milestone system to be shown to the user.

In general, VCA systems are distributed in one of the following ways:

- As a complete, standalone server system (server-based)
- As a library which must be built into software (library-based)
- As part of the firmware on the device (edge-based)

In our application, we would choose to have server-based video analytics. This does not stop the server from being integrated in the Milestone client.

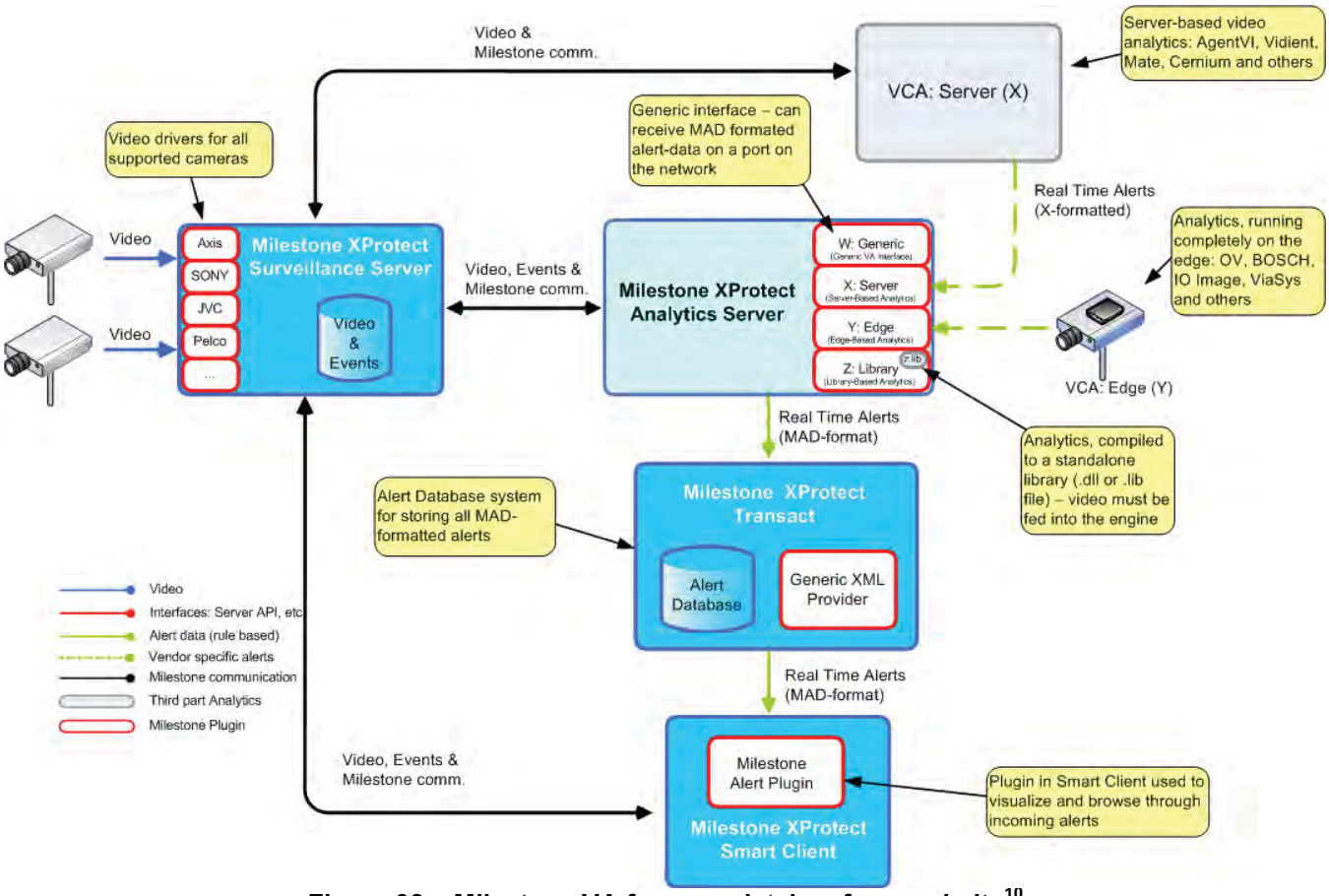


Figure 38 – Milestone VA framework taken from website[10]

Evidence Export

There is a possibility to export the project as an AVI. It concatenates the video from all cameras as well as indicating the camera number and timestamp at all times. More research needs to be done to make the export ready for court. Currently, the original aspect ratio and frame rate is not kept since the output video cannot change aspect ratio, resolution or frame rate during re-encoding.

10

http://clouddownload.milestonesys.com/XProtect%20Analytics%202.2a/Manuals/MilestoneXProtectAnalytics_Generic_VA_Interface_Developers_Manual_en-US.pdf

In order for the exported video to be ready for court, the videos should be exported in their original form, and a standalone player should be distributed to play them synchronized in their original format.

In case where there is integration with a VMS, the exporting features of the VMS could be used. The VMS can export video in the original format as well as watermark the video to prove it hasn't been tampered with.

APPENDIX 1

The appendix show which subjects was matched by which face recognition capable cameras. In *Appendix 1: Matches per individual and camera,* each row represents a subject and each column represents a camera. In the first line, the column where there is a red cell represents a face recognition capable camera. The results are summarized in the following section. The results also represent the ground truth as the results were reviewed by an operator in order to indicate in which cameras the person was but not recognized automatically.

In *Appendix 1: Matches per individual and camera*, the second column, "C", represents the action taken in the video. The codes are explained below.

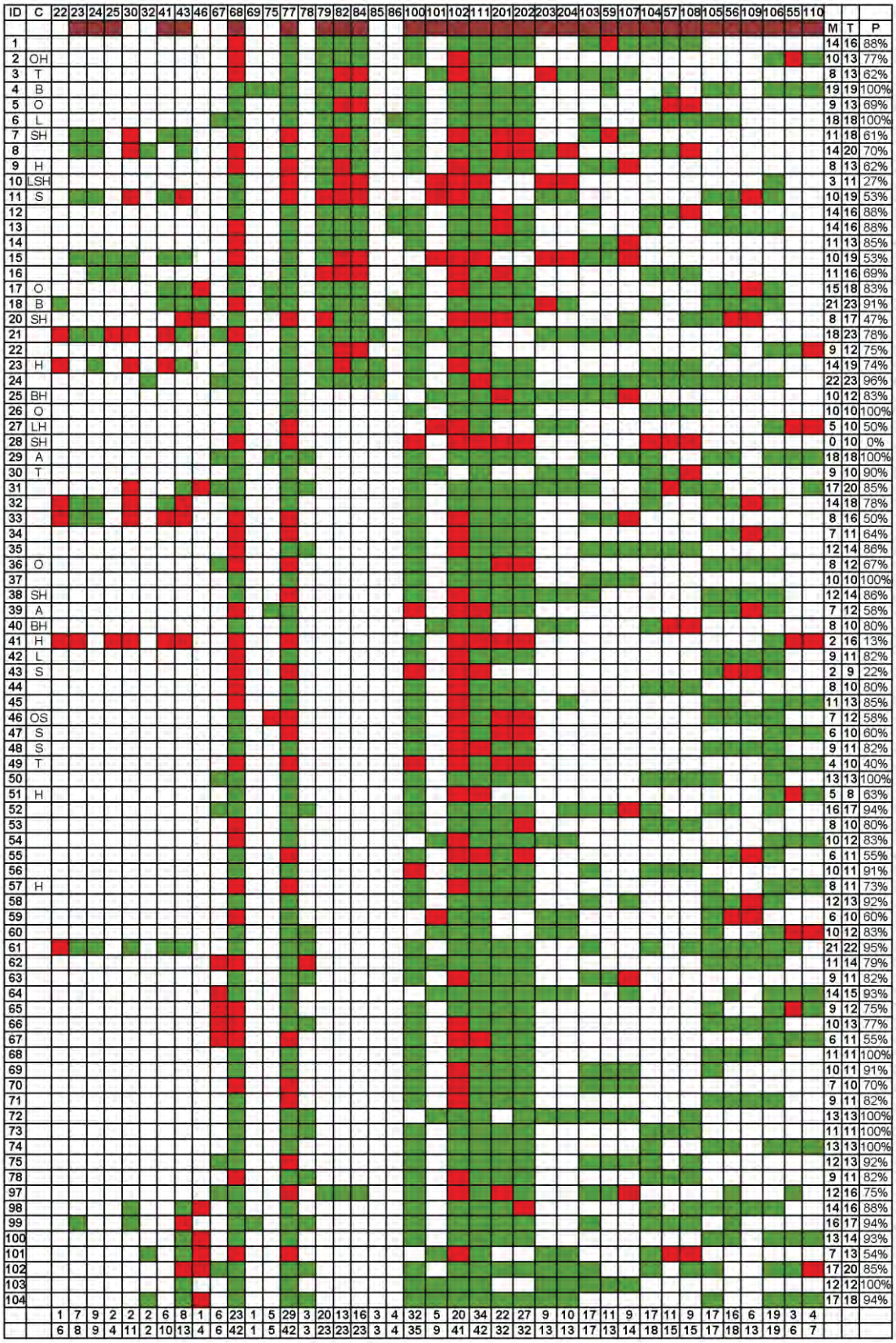
Action Taken in the Video (C)	Code
Add or remove outerwear in the bathroom	**O**
Spend a long time in a bathroom	**B**
Look down when walking	**L**
Have sunglasses during walking	**S**
Loiter for two minutes outside the bathroom	**A**
Walking and texting	**T**
Wear a Hat	**H**

Table 6 – Script codes

In *Appendix 1: Matches per individual and camera*, the color of the cells is explained below. It represents the results per person and cameras, as well as the ground truth.

Cell Color	Meaning
White	The person was not visible in the camera
Red	The person was visible in the camera, but not recognized automatically
Green	The person was visible in the camera, and was recognized automatically

Table 7 – Color legend for person presence matching



ID	C	22	23	24	25	30	32	41	43	46	67	68	69	75	77	78	79	82	84	85	86	100	101	102	111	201	202	203	204	103	59	107	104	57	108	105	56	109	106	55	110	M	T	P
1																																									14	16	88%
2	OH																																										10	13	77%
3	T																																										8	13	62%
4	B																																										19	19	100%
5	O																																										9	13	69%
6	L																																										18	18	100%
7	SH																																										11	18	61%
8																																										14	20	70%
9	H																																										8	13	62%
10	LSH																																										3	11	27%
11	S																																										10	19	53%
12																																										14	16	88%
13																																										14	16	88%
14																																										11	13	85%
15																																										10	19	53%
16																																										11	16	69%
17	O																																										15	18	83%
18	B																																										21	23	91%
20	SH																																										8	17	47%
21																																										18	23	78%
22																																										9	12	75%
23	H																																										14	19	74%
24																																										22	23	96%
25	BH																																										10	12	83%
26	O																																										10	10	100%
27	LH																																										5	10	50%
28	SH																																										0	10	0%
29	A																																										18	18	100%
30	T																																										9	10	90%
31																																											17	20	85%
32																																											14	18	78%
33																																											8	16	50%
34																																											7	11	64%
35																																											12	14	86%
36	O																																										8	12	67%
37																																											10	10	100%
38	SH																																										12	14	86%
39	A																																										7	12	58%
40	BH																																										8	10	80%
41	H																																										2	16	13%
42	L																																										9	11	82%
43	S																																										2	9	22%
44																																											8	10	80%
45																																											11	13	85%
46	OS																																										7	12	58%
47	S																																										6	10	60%
48	S																																										9	11	82%
49	T																																										4	10	40%
50																																											13	13	100%
51	H																																										5	8	63%
52																																											16	17	94%
53																																											8	10	80%
54																																											10	12	83%
55																																											6	11	55%
56																																											10	11	91%
57	H																																										8	11	73%
58																																											12	13	92%
59																																											6	10	60%
60																																											10	12	83%
61																																											21	22	95%
62																																											11	14	79%
63																																											9	11	82%
64																																											14	15	93%
65																																											9	12	75%
66																																											10	13	77%
67																																											6	11	55%
68																																											11	11	100%
69																																											10	11	91%
70																																											7	10	70%
71																																											9	11	82%
72																																											13	13	100%
73																																											11	11	100%
74																																											13	13	100%
75																																											12	13	92%
78																																											9	11	82%
97																																											12	16	75%
98																																											14	16	88%
99																																											16	17	94%
100																																											13	14	93%
101																																											7	13	54%
102																																											17	20	85%
103																																											12	12	100%
104																																											17	18	94%
		1	7	9	2	2	2	6	8	1	6	23	1	5	29	3	20	13	16	3	4	32	5	20	34	22	27	9	10	17	11	9	17	11	9	17	16	6	19	3	4			
		6	8	9	4	11	2	10	13	4	6	42	1	5	42	3	23	23	23	3	4	35	9	41	42	32	32	13	13	17	13	14	18	15	15	17	18	13	19	6	7			

Appendix 1 – Matches per individual and camera

References

[CBSA2013] CBSA Airport Dataset, Available by permission.
CBSA Science and Engineering Directorate.

[GG12] Dmitry Gorodnichy, Eric Granger, "Evaluation of Real-Time Face Recognition Technologies for Video-Surveillance Applications", NIST International Biometric Performance Conference (IBPC 2012), Gaithersburg, March 5-9, 2012. Online:
http://biometrics.nist.gov/cs_links/ibpc2012/presentations/Day3/347_granger_IBPC.pdf

[ILIDS2013] https://www.gov.uk/government/collections/i-lids
(accessed Apr. 17, 2014)

[TRECVID2012] http://www-nlpir.nist.gov/projects/tv2012/tv2012.html
(accessed Apr. 17, 2014)

[PETS2006] http://www.cvg.rdg.ac.uk/PETS2006/data.html
(accessed Apr. 17, 2014)

[PETS2009] http://www.cvg.rdg.ac.uk/PETS2009/a.html
(accessed Apr. 17, 2014)

[Won11] Y. Wong, S. Chen, S. Mau, C. Sanderson, B.C. Lovell, Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition, IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops, pages 81-88. IEEE, June 2011.