Protocols for Evaluation of an Interactive Video Tracking System utilizing Face Recognition

by

Jonathan Wong, **B.Eng.**

A thesis submitted to the

Faculty of Graduate and Postdoctoral Affairs

in partial fulfillment of the requirements for the degree of

Master of Applied Science in Biomedical Engineering

Ottawa-Carleton Institute for Biomedical Engineering

Department of Systems and Computer Engineering

Carleton University

Ottawa, Ontario

January, 2016

Abstract

The Search and Retrieve prototype was developed as an interactive video tracking prototype to find and follow targets in a multi-camera surveillance system using face recognition. The overall potential benefits of using interactive face recognition for video tracking are unknown. We developed an evaluation protocol for the Search and Retrieve program using human-computer interaction and video tracking metrics. The protocol included three tracking cases: manual tracking, automated tracking using face recognition, and interactive tracking using both. We demonstrate that adding the operator's skill through interaction to the face recognition tracking present in the Search and Retrieve program can measurably increase recall by an average of 8 times for automated tracking and 39% over manual tracking in a limited time span of 20 minutes and without any prior training of the system for the user. The system's precision remains constant over the three cases.

Acknowledgments

I would like to thank David Bissessar, Kathryn Mills, Tony Mungham, Jean-Philippe Bergeron, and the Canada Border Services Agency as a whole, and my supervisor Andy Adler for their help and support in making this project possible.

This thesis builds upon the foundation set by the 2013 Defense and Research Development Canada Search and Retrieve project whose funding is gratefully acknowledged.

Contents

List of Tables

List of Figures

Nomenclature

AMCV	Airport Multi-Camera Video (dataset)
CBSA	Canada Border Services Agency
CDT	Correctly Detected Track
COTS	Commercial off the shelf
FN	False Negative
FOV	Field of View
FP	False Positive
FNR	False Negative Rate
FPR	False Positive Rate
FPS	Frames per Second
FR	Face Recognition
GT	Ground Truth
HCI	Human Computer Interaction
IP	Internet Protocol
PIL	Primary Inspection Lane

PTZ	Pan-Tilt-Zoom
S&R	Search and Retrieve
TF	Track Fragmentation
TP	True Positive
UI	User Interface

Chapter 1

Introduction

1.1 Motivation

Modern surveillance systems provide operators with large volumes of video data. However, large scale surveillance systems are often limited by the ability of operators to observe and interpret the video in a reasonable timeframe. With an increasing prevalence of large surveillance deployments by governments and private organizations, there is demand for the ability to manage and sort recorded footage.

Incidents, such as law enforcement actions or health and safety incidents are regularly captured by surveillance systems. In the case of such incidents, all video footage related, as well as video footage leading up to and following the incident, needs to be gathered and forms the complete story of the incident. The relevant footage may be needed to be exported to DVD or other formats for distribution and is also archived locally in case the evidence is required in a legal process. For the Canada Border Services Agency, this is a daunting multi-camera video tracking challenge. The entire story must be exported; a certain continuity of evidence where the individual is shown at all times from different cameras is typically required. According to the CBSA, anecdotally the process of finding and exporting footage for an incident can take multiple

officers several days. An improved method of exporting achieved surveillance footage was required.

To help alleviate the resource requirement on CBSA officers, the Search and Retrieve (S&R) software tool prototype was developed. The tool combines a commercial off the shelf (COTS) face recognition library to search accumulated footage from a surveillance system. The system operator can increase the accuracy of the face recognition by verifying any matches produced by the face recognition library. By itself, face recognition was unlikely to create the full story required as few cameras in a surveillance system are optimized for face recognition. Therefore a map-layout of the surveillance area, which assists the human operator to refine the search, was included in the prototype. The map-interface allows the operator to narrow the search area for the tracking target spatially using face recognition found instances of the target as a starting point for the search temporally. The S&R tool was designed as an interactive video tracking system requiring the frequent intervention of the operator to classify faces and fill in missing footage between face recognition capable cameras.

The S&R prototype has potential to save CBSA officers time in exporting video. With a highly customized interface, the inclusion of face recognition, and skilled operators intimately familiar with the surveillance system deployment area, the benefit of the S&R prototype's individual features are unclear. The goal of this work is to evaluate the S&R prototype's effectiveness as an interactive video tracking system by breaking down the benefits of its individual features, and determining whether combining the map-interface and automated face recognition tracking provides any advantages. By better understanding where the greatest potential benefits of the S&R prototype exist, future development can be better directed.

1.2 Problem Statement

The Search and Retrieve interactive video tracking prototype relies on both the effectiveness of face recognition and the skill of its operator to create full video tracks of targets visible in a video surveillance system for the purpose of evidence retrieval. The current gold standard of video surveillance evidence retrieval remains manual tracking and export by one or more system operators. It is unknown whether applying face recognition to the evidence retrieval and video tracking problem is effective at reducing the time required for an operator to complete the task. The goal of this research is to create an evaluation methodology or protocol to determine the effectiveness of using an interactive video tracking system compared to current manual tracking methods.

1.3 Objectives

The main objective of this work is to provide clear performance measurements of the Search and Retrieve interactive video tracking system. The evaluation protocol should be operator agnostic and control for the variables of operator skill and knowledge. The goal is to measure the effectiveness of the face recognition and the map-interface tracking components of the S&R prototype individually and when used in conjunction with each other. By individually evaluating each component and when the components are combined, the advantages and disadvantages of the component combination should become evident.

Table 1.1: Breakdown of work done in this thesis and work done previously

	Previously Completed (CBSA)	This Work
S&R Prototype	All	None
Dataset	Footage Captured	Ground Truth Annotation Editing/Encoding
Evaluation	None	All

1.4 Work Breakdown

The work presented in this thesis was done in conjunction with the Canada Border Services Agency (CBSA). The Search and Retrieve prototype software was developed by the CBSA prior this work [1]. Dataset footage was captured by the CBSA, but its editing and annotation was completed as part of this thesis. The evaluation methodology and execution were completed as part of this thesis work. A work breakdown can be found in Table 1.1.

1.5 Contributions

The research process reported in this thesis has led to four primary contributions.

The first contribution was the formalization of the Airport Multi-Camera Video dataset. The dataset footage was provided by the CBSA and the annotations were manually created for all 83 subjects present. Footage was removed when unnecessary, and supporting documentation was created.

The second contribution was the creation of an evaluation protocol for interactive video trackers. The protocol made use of both machine vision video tracking, and human computer principals to capture the effect of as many potential variables to performance as possible.

The third contribution was the application of the created evaluation protocol to

the S&R prototype. The evaluation provides insight into the benefits of using the S&R prototype over manually tracking or relying on face recognition to track a target in the AMCV dataset.

The fourth contribution is a set of specific recommendations for a refined evaluation protocol based on the results of the application of the created evaluation protocol.

1.6 Thesis Outline

This thesis is split into seven chapters. Chapter 2 provides a summarized background into video tracking, face recognition, and human-computer interaction relevant to the evaluation of the Search and Retrieve prototype. The Search and Retrieve prototype is described in detail in Chapter 3. The new dataset, the Airport Multi-Camera Video (AMCV) dataset designed to simulate an airport customs environment is discussed in Chapter 4. The dataset features footage from an airport customs area and was developed to simulate a real operational airport environment as closely as possible. Chapter 5 details the general evaluation protocol and metrics used to measure the Search and Retrieve prototype's performance. The results of the evaluation are described in detail in Chapter 6 and include the precision recall of all the tracking methodologies used. In the concluding chapter, chapter 7, the findings and observations made throughout the thesis are summarized. Potential future work is included in this chapter.

Chapter 2

Background

The increased use of video surveillance in both commercial and government applications is presenting operators with unprecedented quantities of information to parse and interpret. Person recognition and tracking in video is performed regularly and without significant effort by these operators everyday. However, the speed a trained human operator can process and interpret video is severely limited and is unable to reasonably keep up with the drastic increase in available footage. Computers have been used to automate person recognition and tracking in video with varying degrees of success; computers are able to process video footage faster than a human operator, but the accuracy is diminished. To increase the effectiveness of such systems, integrating the success rate of human operators with the speed of automated programs would be ideal.

The aim of this thesis was to evaluate an interactive system designed to recognize and track an individual through an airport environment using footage from an internet protocol (IP) video surveillance system. The system can be thought of as three separate components: recognition accomplished through face recognition, video tracking, and a human-computer interface for interaction to be possible. What follows is a background discussion of each of these three components and datasets in

literature used to evaluate them.

2.1 Video Tracking

Video tracking is the ability to automatically follow an object in a camera's field of view (FOV). The problem of video tracking can be applied to a single camera or multiple cameras. This section describes the basics of video tracking.

2.1.1 Background

Video tracking consists of following an element or elements in a video sequence automatically [2]. Targets such as faces, individuals, or specific objects may be tracked by such a system. A video tracker generally includes two key components: matching and motion. The matching concept involves identifying the element to be tracking on a frame-by-frame basis; an example of which would be the Microsoft Kinect(R) finding the face or faces to be tracked in its single camera setup [3]. Once the object has been successfully found, it must be followed by the algorithm as it moves from its initial coordinates as time progresses - the motion aspect of video tracking. An ideal video tracker should have strong performance under the following conditions [4]:

Clutter: the tracker should maintain its tracking target in conditions where there are many similar objects in the field of view.

Occlusion: the tracker should maintain tracking in the event of temporary target occlusion (drop-out), and resumed correctly when the target reappears (drop-in). The tracker should also maintain tracking in the event of partial occlusion by another object in the field of view.

False positives/negatives: the tracker should be correctly identifying targets, and any other elements should be ignored. The number of incorrect and missed identifications should be limited.

Changes in Motion: the tracker should maintain tracking targets undergoing any change of speed and acceleration.

Changes in Pose: the tracker should maintain tracking targets that change profile such as rotation, deformation and translation.

Consistency: the tracking of the target should be maintained indefinitely over time.

Depending on the type of object to be tracked, video trackers may employ a wide variety of techniques to create both the matching and motion components and meet all of the parameters above. Other challenges to video tracking include limitations or changes in the background such as illumination and sensor noise. A video tracking system can be broken into the following primary components: [4]

1. Feature Extraction: the goal of feature extraction is to identify and segment relevant information in the video. Potential features include edges, corners, and blobs.

2. Target Representation: this step involves recreating a representation for the shape and appearance of the target to be tracked from the features extracted in the previous step. The target representation is also known as the state, the model of the object of interest to be tracked. The representation of the target will be a trade-off between detail and variance: the more detailed the representation the less able the tracker will be to handle changes in the target's appearance as a result of pose, rotation, occlusion, etc. The greater the variance

of the target, the higher chance the tracker will mistake a non-target object for the target.

3. Localisation: this involves the search of the state in video given its initial position.

4. Track Management: this step involves determining when to begin and terminate a given track for the target. A track can begin for example when the target enters the field of view of the camera from an edge, and can end when the target is occluded by another object in the field of view such as a vehicle. A track start and end are sometimes referred to as target birth and death respectively.

5. Trajectories: these are the path of the target from their appearance in the scene to their disappearance from the scene.

2.1.2 Interactive Tracking

Video trackers can be made interactive in an attempt to boost performance [4]. Tracking systems as a whole can be divided into three categories: manual, automatic, and interactive. Manual and automatic, as their names imply, are defined as when the tracking is done entirely by the operator or by the algorithm respectively. Interactive, or supervised, tracking systems will take elements from both a pure automatic and manual approach to complete the tracking problem.

Interactivity in a supervised tracking system can occur either before, after, or interspersed with the automated tracking being utilized. The operator can be assigned the task of selecting and labelling targets before automated tracking is applied. Verification of correct tracking can be done by the operator after automated tracking has been completed. Similar to face recognition, interactivity can take place in conjunction with the automated tracking's operation: the operator can assist in feature

extraction or assisting in track management by correcting when the track begins to lose a target. Significant challenges such as partial occlusion or track changes can be corrected by the operator.

2.2 Biometrics

A biometric is defined as the identification of an individual based on physiological or behavioural characteristics. A biometric system is commonly used to establish or authenticate the identity of an individual for a specific purpose, examples of which include surveillance and access control.

Modalities, or types of biometrics, include DNA, face, iris, fingerprint, gait, and voice. For a trait to be considered a strong and practical biometric it can be judged based on a number of key properties it should possess: universality, uniqueness, permanence, collectability, performance, acceptability, and circumvention [5]. Government applications of biometrics include ID cards, border control, passport control, forensics, and public surveillance.

- Universality: The chosen biometric characteristic should be common in the population at large. A strong biometric should be possessed by every individual in the population we want to identify.

- Uniqueness: The biometric is distinguishable amongst the population. The ideal biometric has significant differences between each individual in the population.

- Permanence: The biometric should remain unchanged over time. A biometric that changes drastically in a short period of time may not be useful as an identification or verification tool. The ideal biometric does not degrade over time and remains constant.

- Collectability: The biometric is ideally easily collected from an individual. Fingerprints are an example of a biometric that may require close proximity to the subject while a facial image can be taken at a greater distance.

- Performance: The chosen biometric modality meet the required accuracy and speed given the resources, operational, and environmental factors. A strong biometric would feature low error rates that can be obtained quickly.

- Acceptability: People should be willing to accept the use of their biometrics in the intended scenario. For example an individual may be more willing to accept use of face recognition to identify or verify their identity, but less willing to provide a DNA sample.

- Circumvention: The biometric should resist being copied or counterfeited thereby producing a false positive. A biometric that can be easily circumvented, copied, or counterfeited is less reliable.

2.2.1 Identification vs. Verification

A biometric system has two distinct operating modes: enrollment and authentication [6]. Enrollment or training is when an individual's biometric data is initially acquired and stored for future usage. Authentication mode can be further split one of two modes: identification and verification.

Verification, or 1:1 matching, involves the system attempting to confirm the identify of the individual [6]. Often verification involves an additional token that the user presents to begin the process. A current example of a verification system would be ePassport face recognition used by countries such as the UK, New Zealand and Australia [7, 8]. In this example, the user first presents the ePassport to the system which contains an individual's face image on a chip within the ePassport. The system

then compares the face image on the ePassport with the face image captured from the user at the time of presentation. The newly acquired face is compared against the face image on the ePassport to confirm the identity of the passport holder.

Identification, also referred to as one-to-many or 1:N matching, occurs when the system must search a database of biometric templates in an attempt to determine the identity of the system's user [6]. An example of an identification application would be checking for an individual's fingerprints in a criminal database such as during security background checks.

A biometric system's operation can be broken into four primary steps: acquisition, feature extraction, matching, and decision-making [6].

1. Acquisition: The process where the biometric data of an individual is captured and stored.

2. Feature Extraction: The processing of the data acquired from the individual to make matching possible. An example of feature extraction would be the extraction of minutia (ridge landmarks) in fingerprint recognition.

3. Matching: Features extracted in the previous step are compared against existing templates in either 1:1 or 1:N matching. Typically matching generates a score that can be used for decision making.

4. Decision-making: The individual's identity is either determined or verified in this stage based on the score or scores generated during the matching step.

2.2.2 Biometric Fusion

Biometric fusion involves the merging of two or more different biometrics [9]. The biometrics used in fusion may be of different modalities or of the same modality. For

example face and fingerprint recognition may be used in conjunction, or two different face recognition algorithms could be used simultaneously. The goal of biometric fusion is to decrease the system's overall false positive rate (FPR) and false negative rate (FNR).

Biometric fusion can be thought of as increasing the total dimensionality of the biometric system which ideally also increases the discrimination between biometric templates [9]. Each biometric collected has a number of distinguishing features that vary between individuals and can be measured [6]. Examples of such features include number, type and location of fingerprint ridge endings or in the case of face recognition the position, shape, and size of eyes. For each biometric feature, a dimension that can be used to compare two or more biometrics is introduced. By adding multiple biometric modalities or algorithms in a fusion scheme, the number of dimensions that can be used to differentiate the individual are increased [9].

Implementing a simple fusion scheme could produce modest improvements to a biometric system without significant development effort. Fusion may take place at the three of the four primary steps of a biometric system described in Section 2.2.1: feature extraction, matching and decision making.

1. Feature extraction fusion: Before comparison with a template, the feature extraction results in the creation of a multi-dimensional feature vector which can be compared against existing template feature vectors for identification or verification. In biometric fusion, multiple feature vectors can be combined into a single feature vector of a higher dimensionality which will ideally create greater discrimination.

2. Fusion at matching: After feature extraction, the matching step results in a match score for the biometric data when compared to each template. When

using multiple modalities, multiple scores can be created - combining these match scores can be used to reduce the FPR and FNR.

3. Fusion at decision making: A biometric system must generate a decision whether to accept or reject the biometric presented. Fusion at this stage involves combining multiple accept or reject decisions to create a single decision; the simplest example of which is a simple voting scheme where the majority of votes decides. For example, if face, fingerprint and iris biometrics were combined at the decision level if any two biometrics voted to accept the individual in verification, than the overall system would accept said individual regardless of the decision of the third biometric.

2.2.3 Interactive Biometric Systems

The ground truth is the data used as a training data for classifier's such as biometric systems [10]. For example, training and testing a face recognition system would utilize a dataset of face images. For the training data to be useful, it is important to know the truths, which face images are of the same individual and which are not. By knowing the truth allows for one to determine that system's error rates.

In many biometrics datasets and evaluations, the ground truth is created manually [11]. The creator of the ground truth, a human, is ultimately recognized as the gold standard for determining what actually happens during data collection [11]. For example in a face recognition dataset, a person creating the ground truth is tasked with correctly identifying each subject present in the dataset and linking the picture to the appropriate metadata. However, the creator of the ground truth can make mistakes, and variance will exist between ground truth annotators [10]. This limit in accuracy of the ground truth ultimately limits the accuracy of the test results that

are possible.

The idea of creating an interactive biometric system is to combine the user's accuracy with a computer's ability to process large amounts of data quickly. By allowing a user to classify low confidence biometric accepts or rejects the accuracy of a system can be theoretically improved given that the user's accuracy in classification is higher than that of an automated system [12].

Similar to how biometric fusion can take place at three different levels of a biometric system to improve performance, introducing a human's skill in recognition into a biometric system can boost its overall performance [12]. Human interaction can take place in at the same three levels as biometric fusion and in both identification and verification scenarios. Interactive biometric systems as used here is not to be confused with the concept of designing the acquisition component with user feedback to improve biometric sample quality.

A simple type of interaction is verification of selected information. For example, at feature extraction the user can be utilized to fine tune or correct the selection of features picked by the system. An example of a interactive classifier is the CAVIAR system [13]. CAVIAR was designed to assist users in classifying flowers: the system requires the user to assist and verify in annotating key features of the flower and provides several suggested matches after annotation is complete [13]. Specifically in the CAVIAR system the user is instructed to verify and modify the outline of flower pedals that will be used in its identification [13]. Verification can be extended to matching or decision making: the user can contribute to the system by verifying any matches or decisions made by the system. For example, in a 1:N match scenario the user may be presented with the top 5 potential matches by the system with which the user can select the correct match. Therefore in this manner, a system's performance can theoretically be improved from the top 1:N match rate to the top five 1:N match

rate.

Interactive biometric systems can be limited by two factors: the limitations of the human user, and the interface used. In most biometric cases, the human analysis can be considered the gold standard of recognition; this statement generally applies to trained users in the given biometric modality. For example, in face recognition humans are quite capable of identifying a person and comparing two images [14]. However, other biometric modalities such as fingerprint or iris do not perform well with untrained users; an untrained individual is unlikely to attempt fingerprint or iris recognition on a regular basis and may find that task difficult to accomplish [14]. A good interface may also be required to maximize the benefit of having a user involved in the process.

2.3 Face Recognition

Face recognition is the sole biometric modality used in this project. Face recognition is a non-intrusive method, and among the most common features used by humans to differentiate each other [15, 16]. Face recognition can be accomplished using still 'mug-shots' or from video creating a broad number of potential verification and identification applications. Face recognition performance can be significantly impacted by environmental factors such as illumination and viewing angle, and by user factors such as age, facial accessories worn, and pose.

Face recognition can also be broken into two separate scenarios: cooperative and uncooperative user scenarios [15]. As the name implies, cooperative applications consists of cases where the user is willing to present their face in a proper or specific way often in exchange for an access or privilege. Examples of cooperative applications include e-passports and access control scenarios. Uncooperative applications include

surveillance scenarios where the user is unaware face recognition is occurring. Typically, uncooperative applications feature an increased distance between the acquiring sensor and the user compared to cooperative applications. Ambient illumination in uncooperative applications may also be inconsistent.

Face as a biometric is attractive for a number of reasons [15, 17]. It has a high universality, and is easy to collect in an unobtrusive and convert manner. However, compared to other biometrics such as fingerprint and iris, its overall performance and permanence tends to lag behind. Face recognition performance is heavily influenced by image quality - algorithms have superior performance when used in the cooperative scenario. Illumination and pose can be more easily controlled in a cooperative scenario allowing for an overall higher sample quality with which face recognition can operate upon. Uncooperative scenarios such as in surveillance video will lag behind the cooperative scenario with regards to performance. Video as a source for face recognition introduces additional noise factors such as motion blur, and individual video frames are often of a lower resolution than a single still frame. Reduced control in illumination and pose reduce the overall sample quality.

The task of face recognition can be broken into four steps analogue to the typical biometric system: face detection, face normalization, feature extraction, and feature matching [15–17].

1. Face detection: Before a face recognition can begin, a face must first be segmented from the rest of the image or background. Any errors in this step would render face recognition impossible. Face detection often includes a coarse estimate of the location and scale of face and landmarking of features for future steps such as the eyes, nose, and mouth.

2. Face normalization: The process of changing the face geometrically and photometrically such that all captured images match. This includes factors such as changes in rotation.

3. Feature extraction: Performed on the normalized face, the goal is to extract useful information found in the image for identifying features. Face recognition features often fall into one of two categories: the location and shape of specific features such as eyes and nose, or a representation of the entire face such as principal component analysis or linear component analysis.

4. Feature matching: The extracted features from input faces are matched against one or many of the enrolled faces. The result is either a verification of identity in a 1:1 match scenario or the most likely candidate above a threshold in a 1:N match scenario.

Two common holistic feature extraction methods are principal component analysis (PCA) and linear discriminant analysis (LDA) [15, 18, 19]. These holistic methods involve simplifying the face data into lower dimensions for comparison while retaining the overall characteristics [18, 20]. Feature based approaches are also utilized which focuses and segments local features such as the individual's nose or eyes [17]. Examples of feature based approaches include Elastic Bunch Graph Matching or Gabor wavelets [17, 21].

2.3.1 Face Quality

A large variety of factors influence the ability for a face recognition algorithm to be successful. Table 2.1 contains factors affecting face recognition performance such as resolution, dynamic range, and the subject's distance from the capture device [22, 23]. Most of these qualities are reflected in standards such as ISO/IEC 19794-5 [22, 23].

It should also be noted that it is easier to identify an individual with motion than without [16].

2.4 Human-Computer Interaction

The interface is a crucial component of any system. To measure its impact on a system's performance, existing methods in human-computer interaction (HCI) can be used. Three key factors are considered in HCI: effectiveness, efficiency, and satisfaction. Each of these three factors must be considered when measuring any interactive system's performance and are defined below [24].

1. Effectiveness: Can the user accomplish the assigned task accurately?
2. Efficiency: How easily can the user accomplish the task using the interface?
3. Satisfaction: How enjoyable is the experience for the user?

When considering the performance of a system that relies on a human operator, the HCI, the accuracy of the system's internal processes and the skill of the human operator must all be considered [24]. Users may also have experience with similar systems which should be taken into account. It is important to note that as a user uses a system, their familiarity with the system increases. A familiarity curve can be used to model the increased productivity of the user with an unfamiliar system. There are numerous factors that can affect the familiarity or learnability curve including but not limited to previous experiences, and age.

Usability testing is a common method and is explored in section 2.4.1.

Table 2.1: List of Image Qualities affecting Face Recognition Performance

Quality	Description
Resolution	The overall resolution of the facial region often measured using the number of pixels between the eyes.
Pose Angle	Pitch and yaw angle of the head. Most algorithms can correct for the roll angle.
Brightness and Exposure	Dark or hotspots (bright, saturated) can mask out image features making recognition more difficult.
Dynamic Range, Contrast	Range of pixel values across the facial region which is an indicator of the overall entropy of the image.
Sharpness	Clearly visible features with no blurring are ideal for face recognition.
Face Shadows	Uneven lighting in either horizontal or vertical directions can cause shadows reducing contrast or create dark spots.
Eye Shadows	Dark spots created underneath the eyes usually caused by vertical ligthing.
Expression	Neutral expressions are ideal as extreme expressions can cause facial distortions making recognition more difficult.
Image Capture Distance	Distortion resultant from capturing an image at too close a distance. Also known as foreshortening.
Glasses	Thin-framed glasses are preferred. Thick-framed glasses, in particular when occluding eye features, are not acceptable. Dark sunglasses are not acceptable. Glasses with strong light reflections are not acceptable.
Occlusions	Any occlusion of eye or other facial features is unacceptable. These include hair over the eyes, low-sitting hats, extended head-coverings, and scarves over the mouth or nose.
Pixelation	Noise caused by poor image resizing techniques or low quality photo scanning.
Compression	Artifacts caused by too much image compression, such as JPG compression. The artifacts may mask features, effectively cause pixelation, and lower effective image resolution.

2.4.1 Usability Testing

Usability is defined as how easily a system can be learned and used by its operator. Any interactive system or prototype can be usability tested. A usability test primarily focuses on tasks or goals that are or near to the tasks to be encountered by the system in its intended environment and purpose. Participants are usually drawn from end users or potential end users. Usability tests have three general components: a product or system to evaluate, participants who are either end users or representative of end users, and a task to perform using the product or system. Products or systems that can be evaluated are any technology that has meaningful interaction with the user to accomplish its purpose. These include, but are not limited to, software products, hardware products, and instruction manuals. Products or prototypes for usability testing can also vary in fidelity. Informal testing can make of use of low fidelity prototypes as simple as screen mockups printed on pieces of paper to high fidelity prototypes and finished products with all the expected functionality of the finalized version.

The measurement of usability is a key goal and challenge of usability testing [25]. Relating back to the three key HCI factors of effectiveness, efficiency, and satisfaction one can create general metrics for usability testing. Effectiveness refers to success and error rates of the given task. Efficiency refers to the time to complete the task, and satisfaction refers to the user's opinion of the product. All three of these factors can be measured quantitatively.

For a deeper understanding of a user's process in usability testing, the thinking aloud protocol can be used [24]. A commonly used technique, the thinking aloud protocol requires participants to constantly vocalize their thoughts and decision making during testing.

Usability testing can be broadly defined as one of four types [25]:

1. Exploratory: designed to explore high level design decisions early in the development process with a low fidelity prototype.

2. Assessment: used early or midway through development similar in goal to an exploratory test

3. Validation or verification: performed at the end of development to ensure that the product meets minimum usability requirements defined before development began. A high fidelity prototype or finished product should be used.

4. Comparison: performed at any point to compare two or more products. The goal is to identify strengths and weaknesses of each tested product. Compared products should be of a similar fidelity.

2.5 Datasets

There are numerous video tracking datasets currently available publicly. Examples of these include the i-LIDS multi-camera tracking scenario, the PETS 2009 Workshop dataset, and the Chokepoint dataset [26–28]. Other related datasets include CAVIAR (PETS04), EPFL, VIRAT, and TRECVID07 [29–32]. The intended operation environment for the Search and Retrieve prototype is an indoor airport customs area, which therefore the dataset selected should accurately represent for maximum test accuracy.

The Imagery Library for Intelligent Detection Systems (i-LIDS) dataset produced by the UK Home Office features six different scenarios including a multi-camera tracking scenario [27]. The multi-camera scenario features an airport environment, Gatwick Airport. It is limited to five cameras: two overlapping cameras and three non-overlapping cameras totalling about 50 hours of footage. i-LIDS features high

quality footage with a large number of individuals that may be tracked. An realistic modern surveillance system includes similarly high quality video with a greater number of cameras and overlapping cameras to represent a complete operational environment when compared to i-LIDS cross-section of an airport environment. The dataset includes a large number of individuals that may be tracked, but information outside of their silhouette coordinates is absent from the ground truth. The i-LIDS dataset falls short of representing an airport operational environment, because there are only a minimal number of cameras available.

A number of datasets have also been produced for the PETS series of workshop competitions of which the PETS 2009 dataset is the closest to a multi-camera system evaluation [26]. The dataset used in PETS 2009 featured 8 camera views for video tracking in an outdoor environment. The outdoor environment presents a large field of view [26].The PETS 2009 dataset falls short of representing an airport operational environment, because of its outdoor setting. By contrast the Chokepoint dataset features an indoor environment - specifically a large number of individuals moving through a narrow doorway or hallway [28]. The video quality and field of view closely resembles that of a narrow corridor or interview setting in the airport customs environment, but lacks overlapping cameras to evaluate multi-camera tracking algorithms [28]. The Chokepoint dataset does not represent an airport operational environment, because of the lack of overlapping cameras present.

TRECVID is features surveillance footage from an airport environment. TRECVID does not of represent an airport operational environment, because the field of view from the cameras is too limited and minimal cameras are used [32]. The CAVIAR dataset from 2004 has low quality surveillance footage, 384x288 resolution at 25fps. This video quality is below that of an average modern IP surveillance system offered by camera manufacturers, and therefore the CAVIAR dataset falls short of

representing an airport operational environment [29, 33, 34]. EPFL features a multi-camera pedestrian dataset with footage from cameras set at ground-level [30]. EPFL does not represent an airport customs environment because ground-level cameras are not often found in surveillance deployments due to high occlusion rates and the possibility of tampering or damage [30]. The VIRAT dataset consists of ground and aerial outdoor footage of various qualities [31]. VIRAT does not represent an airport environment as it is designed for event recognition, and is unsuitable for indoor multi-camera scenarios such as person tracking [31].

2.6 Chapter Summary

In this chapter we have introduced the concepts of video tracking, face recognition, and usability testing. The basic challenges of video tracking and a generalized process of how a typical algorithm works are covered. The concept of biometrics is introduced along with the properties required to make a good biometric. The generalized process for face recognition is discussed along with the image quality factors that affect face recognition. The definition and potential benefits of interactive biometric systems is discussed. A brief introduction to human-computer interaction is included along with the basics of usability testing. Finally, potential datasets that can be used in this thesis are examined.

Chapter 3

Search and Retrieve Prototype

Video retrieval is an important and challenging process for surveillance system operators [35]. Multi-camera surveillance systems may produce large amounts of video data for their users such as law enforcement that may be used as evidence in the event of an incident. After an incident where an investigation is warranted, footage must be pieced together to create an accurate and complete storyline of the individuals in question. Video retrieval is done manually by an officer; a painstaking and slow process that can take multiple working days depending on the number of cameras in the system.

One particular application of video tracking and retrieval is in the Canada Border Services Agency controlled area of an airport. An airport has two significant restrictive characteristics: an enclosed environment with restricted areas individuals are to traverse, and a specific flow between the areas individuals are to follow. The restrictions create specific chokepoints all individuals must pass through to move from area to area. Individuals may not move backwards through the chokepoints. To assist in video retrieval in an the airport environment, the Search and Retrieve prototype was created by CBSA.

3.1 System Overview

The goal of the Search and Retrieve prototype is to create a complete surveillance record of an individual's time in an airport environment known as the storyboard in the context of the prototype. The user adds footage to the storyboard using one of two distinct components listed below.

1. Face Recognition: Searching video using a commercial face recognition product with a simple recursive human-assisted fusion scheme.

2. Map Interface: Manually adding footage with the assistance of the interface consisting of a timeline and camera map. Existing footage added manually or by face recognition is included in the storyboard allowing for the search area to be narrowed.

The face recognition search process begins with a photo of a person to search for. Any image featuring a frontal image of the face can be used and may be extracted from video if a still image is unavailable. A higher quality photo will provide more reliable matches. It is possible to extract a facial image from video using the prototype if a suitable frontal image is otherwise unavailable. Using the initial search image, face recognition is applied to the archived video. Matches are presented to the user for verification with verified faces being used to search for new matches. Once all possible matches have been found, the user uses the map interface to fill in the remaining storyboard: this video will consist of cameras where face recognition has failed such as when the face is not visible or the camera is placed far away from the subject causing face recognition to fail. An example of when face recognition would fail would be a camera mounted on a high ceiling overlooking a wide area. A simplified flowchart of this process is presented below in Figure 3.1.

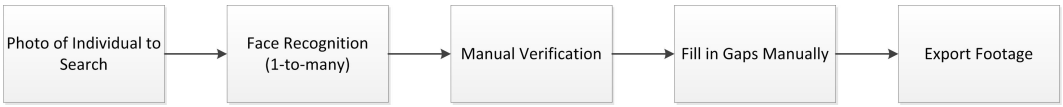


Figure 3.1: Full System Flowchart

Considering a video tracker addresses two problems, motion and matching, the Search and Retrieval prototype automatically exclusively deals with the matching problem leaving the motion side to the user. If the face is not visible, the onus is completely on the operator to track the individual through the scene. Problems such as occlusion where the face recognition will fail is left to the operator to handle. The Search and Retrieval prototype can be considered as face recognition software that has been adapted to video tracking by adding interactivity.

By making the Search and Retrieve prototype substantially interactive, the prototype has a few distinct differences over its automated counterparts. Relying on human interaction will invariably be slower than a process that requires the computer alone, but is ideally faster than a purely manual process. Requiring verification of face recognition matches increases the match rate precision. The recall of the system is unchanged by the verification process, but can be increased using the face recognition fusion process. The fusion process involves searching using multiple face images including any previously verified matches in the search query. Video tracks created through interaction should have a higher completion percentage.

3.2 Similar Systems

The Search and Retrieval prototype shares significant similarities with an image verification and classification system known as computer assisted visual interactive recognition (CAVIAR) [13]. The CAVIAR program was designed to assist in the classification of flowers or faces through an interactive mechanism. Previous attempts had been

made to use image based classifiers to classify wild flowers, and CAVIAR, produced by [36], had the goal of combining known techniques with an interactive interface to improve precision. The CAVIAR system works by superimposing the classification model upon the search image [36]. The user can then manipulate the model to align the recognition with visible features: in flowers the pedals outline can be manipulated and for face the position of the pupils can be manipulated [13]. The modified model is applied to the search index with the system presenting the user with the top three candidates [13]. The user has three choices: select one of the candidates as a true positive match, modify the model to improve the matches, or browse through the lower ranked matches in an attempt to find a correct match [13]. The CAVIAR system features a significantly different interaction than Search and Retrieve; the interaction takes place at the classifier level, and not the score level like in Search and Retrieve [13]. When changes take place at the classifier level, the user has a greater influence on the precision and recall of the system. Incorrect interactions can result in a reduced precision and recall potentially generating zero precision/recall [13]. By contrast, Search and Retrieval's interactivity takes place after match scores have been generated meaning the user has no influence on the recall of the system's initially query [13]. Subsequent queries utilize the initial search images and any verified matches which will ideally increase recall [13].

Interactive classifiers share many similarities with human computation project reCAPTCHA [37]. reCAPTCHA is based on CAPTCHA whose primary purpose is to differentiate a human user from a program or bot on the internet. reCAPTCHA differs from CAPTCHA by adding a secondary objective, to improve the optical character recognition (OCR) quality when digitizing books for the Internet Archive [38]. It does this by presenting the user with an image with two words and asks the user to type in the words. However, unlike a regular CAPTCHA where the word or words

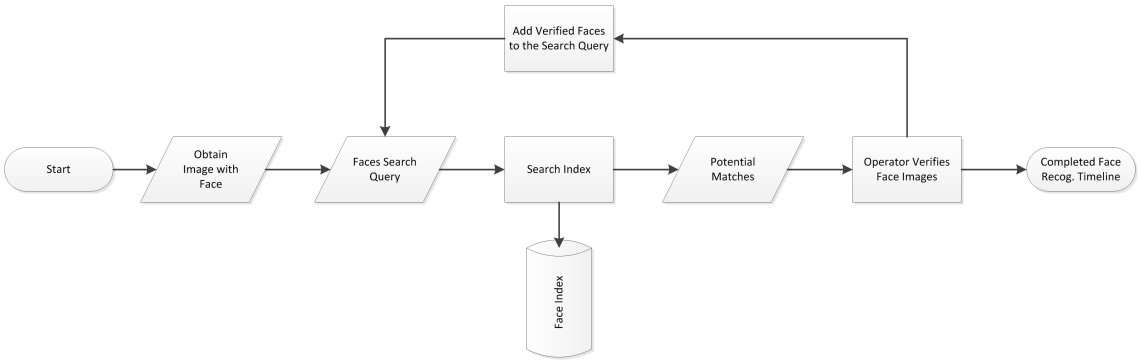


Figure 3.2: Face Recognition Flowchart

are known, at least one word in reCAPTCHA is a word that has failed OCR. Using reCAPTCHA allows for crowd-sourcing of words that fail OCR. The known word continues to be used for verification [38].

3.3 Program Components

The Search and Retrieve prototype consists of two components: a commercial face recognition with fusion, and the storyboard interface. This section describes these two components in greater detail in Sections 3.3.1 and 3.3.2.

3.3.1 Face Recognition and Fusion

A commercial off the shelf face recognition product was used in the search and retrieve prototype. Face recognition is achieved in two primary steps: face detection followed by matching. The complete process can be seen as a flowchart in Figure 3.2. Each step of the process is describes in Table 3.1 and a wireframe of the prototype is shown in 3.3.

Before face recognition is possible, an index of detected faces must be created. The face index includes all the extracted face images from the video stored in the

commercial face recognition product's proprietary format, and references to the original video frames from which the face was extracted. When queried with a face image, the face recognition algorithm searches the index in a 1:N search. The raw video is not used for face recognition due to performance concerns. This method of searching previously detected faces instead of raw video is necessary for performance – searching raw video directly is an intensive and time-consuming task. For the purposes of this study, the face index used by the Search and Retrieve prototype was generated offline from the test dataset before any user interaction occurs with the system. In a real operation environment, the face index would have to be generated as close to real time as possible. The commercial face recognition software included a face detection module that was run against all the test data.

The Search and Retrieval prototype can query the face index using one or multiple face images. During a single image query, every face in the index is assigned a match score to the query image. Scores that are above the user defined threshold are presented for validation by the user. The user validates images by viewing the associated video and confirming that the person being searched for is present in the video match. Validated facial images are then used by the prototype in future queries. When multiple face images are queried a simple fusion is used. Fusion is done at the score level. Each face image is used in a 1-to-many match producing multiple match scores for each potential image in the index. For an image to be a confirmed match, two conditions must be true: one match score must exceed a user defined threshold shown in equation 3.2, and the average of the scores must exceed a second lower threshold shown in equation 3.1.

$$\frac{\sum_{i=1}^{N} S_i}{N} > T_1 \tag{3.1}$$

Table 3.1: Face Recognition Stages with Descriptions

#	Step	Function
1	Obtain Image with Face	An image containing a visible face of the target individual is loaded into the program by the operator
2	Faces Search Query	The face in the search image is extracted
3	Search Index	The extracted face is compared against all existing faces (1:N comparison) in the face index
4	Potential Matches	Any scores during the previous step above the operator defined threshold are flagged as potential matches
5	Operator Verifies Face Images	Potential matches are shown the operator who can indicate through the user interface which are true positives
6	Add Verified Faces to Search Query	Facial images indicated as true positives are added to the initial face search query. The operator may now return to Step 2
7	Completed Face Recog. Timeline	If the operator believes no additional footage can be found or exists, the timeline is complete

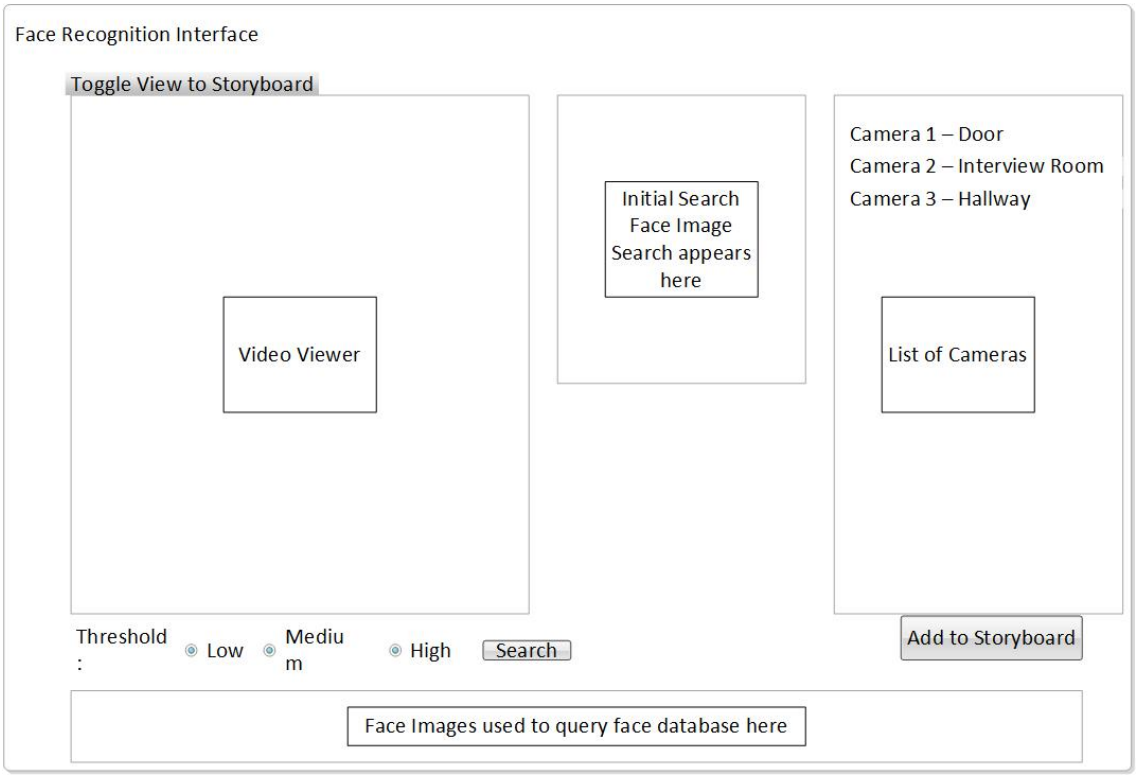


Figure 3.3: Face Recognition UI

$$S_i > T_2 \tag{3.2}$$

The goal of the fusion scheme was to reduce the number of false positives being detected. If all the validated images were used individually without fusion, it was found that the number of returned potential matches was overwhelming in number to the user with a high false positive rate. This was determined to be a product of poor image quality creating potential matches with only one of the validated query images. By requiring the average score to be above a threshold, the goal was to remove low scoring matches only matching with one of the many query images.

The fusion scheme has significant shortfalls. It becomes heavily influenced by image quality as the number of verified matches increase. Image quality is a factor in

the average match score; a lower image quality means lower overall scores resulting in an average that will not meet the threshold. As more images are added, fewer matches are found since average score will be pushed down by the poor image quality. A wireframe of the user interface used for face recognition and operator validation can be seen in Figure 3.3.

3.3.2 Storyboard Interface

The storyboard interface is designed to provide spatial and temporal context to the user when manually adding video segments to the system track. The interface consists of three components listed below.

1. Map of the airport's customs area including the location and direction of each camera
2. Timeline of video included in the system track
3. Window to view video footage from the selected camera

The map provides spatial context for the user showing the location and facing direction of each camera. Individual cameras are colored coded on the map: red cameras indicate the cameras in which the target was detected by face recognition, green cameras indicate cameras which were manually added to the storyboard by the user, and grey cameras indicate a camera currently not included in the storyboard.

The timeline provides temporal context to the user indicating the time and camera or cameras at which the tracked subject is visible. The video viewer allows the user to view video for the purposes of review or adding new footage to the timeline. The

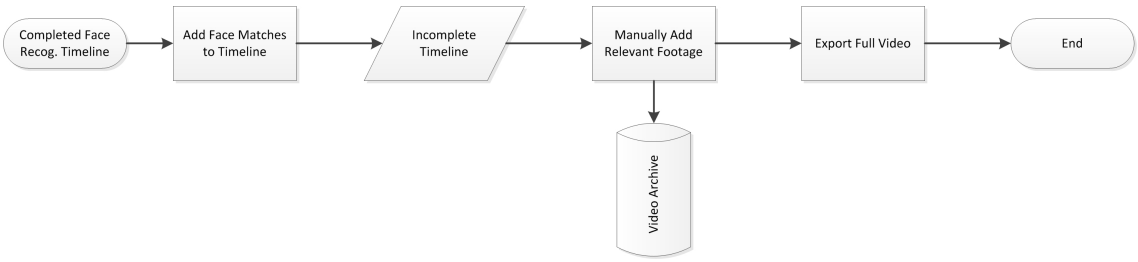


Figure 3.4: Map Interface Flowchart

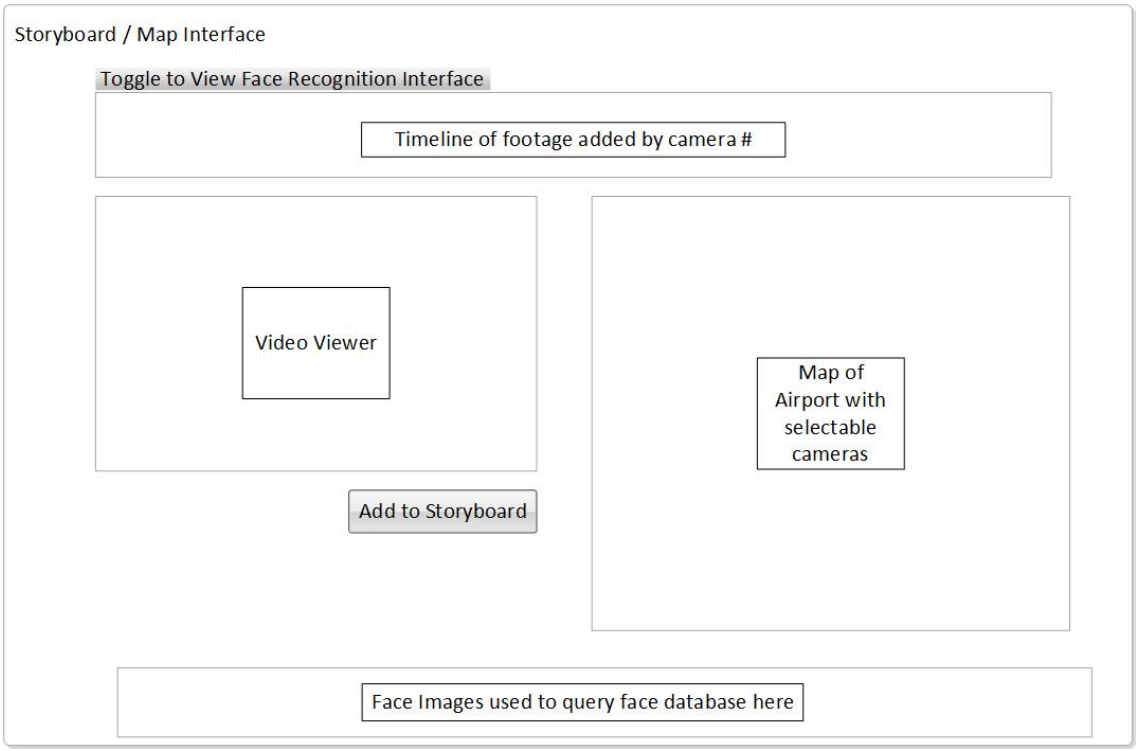


Figure 3.5: Storyboard UI

Table 3.2: Map Interface Stages with Descriptions

#	Step	Function
1	Incomplete Timeline	The operator may load an incomplete timeline, usually from the face recognition stage, into the prototype
2	Manually Add Relevant Footage	Footage is added to the timeline manually by the user. This is accomplished by tagging the footage in the viewfinder by time and camera number. The map of all cameras and the timeline view assist the user is finding the correct footage
3	Export Full Video	If desired, an AVI export of all the footage in the timeline can be created

process is shown as a flowchart in Figure 3.4 with a wireframe of the interface shown in Figure 3.5.

Manually adding footage using the map interface introduces an element of human error into the video tracking process. Areas of poor camera coverage or high occlusion significantly increase the likelihood of false positives and false negatives.

3.3.3 Future Work

The Search and Retrieve prototype has a number of potential improvements that could be implemented. Potential improvements to the prototype are proposed in this section. The current commercial face recognition algorithm can be replaced with different commercial algorithms or an in-house built algorithm in an attempt to improve performance. Using confirmed video tracks can be used to further narrow the search field. Additional video tracking may also be included. Single or multi-camera techniques can be used to improve the tracking process. Single camera techniques can be used to help fill the storyboard once an individual is detected by face recognition.

For example, a simple optical flow tracking could be implemented to better track a correctly identified face.

Time and spatial based searching is also a potential improvement for the Search and retrieve prototype. Confirmed matches included in the storyboard can be used to restrict the search parameters of either face recognition or video tracking algorithms implemented in the system. For example, if the person is correctly identified in two separate cameras, the search field can be restricted to footage in close proximity temporally and spatially. Deploying this type of feature temporally is relatively trivial compared to spatially which would require significant user fine-tuning based on the location of the cameras deployed.

Additional interaction similar to CAVIAR or reCAPTCHA may also benefit the Search and Retrieval prototype. Any additional interaction would require careful consideration of the interface to ensure new features do not significantly inconvenience the user. An example of an added interaction could be at the facial feature detection of the initial search image. The chosen algorithm would highlight facial features such as the eyes, and face shape. The user would be required to the discovered features in the event of an error by the algorithm such as an offset in eye coordinates. It is worth considering that additional interaction with intermediate steps in face recognition may not be possible while using a commercial product.

Improvements to the map interface may also be included. The map includes the direction and location of each camera included in the available videoset, but the inclusion of the complete field of view would assist the user in determining whether the target is visible in any single camera.

3.4 Chapter Summary

In this chapter we have introduced the Search and Retrieve prototype and its objective of finding and tracking an individual through archived footage from a surveillance system. The Search and Retrieve prototype has two primary interfaces: a face recognition based interface and map interface. The two interfaces, along with the operator, work in concert to create a full timeline of footage of the target.

Chapter 4

Airport Multi-Camera Video Dataset

In this chapter, the Airport Multi-Camera Video (AMCV) dataset is introduced. The AMCV dataset was developed with the CBSA to provide an accurate representation of an airport customs area. The dataset provides unique surveillance footage using a large number of cameras available to the CBSA. Due to privacy concerns, all of the individuals appearing in the dataset are volunteer actors. The dataset was developed with the purpose of being used to evaluate various video analytics such as multi-camera face recognition or video tracking. The dataset footage was captured by the Canada Border Services Agency with its editing and annotation being completed as part of this work.

4.1 Overview

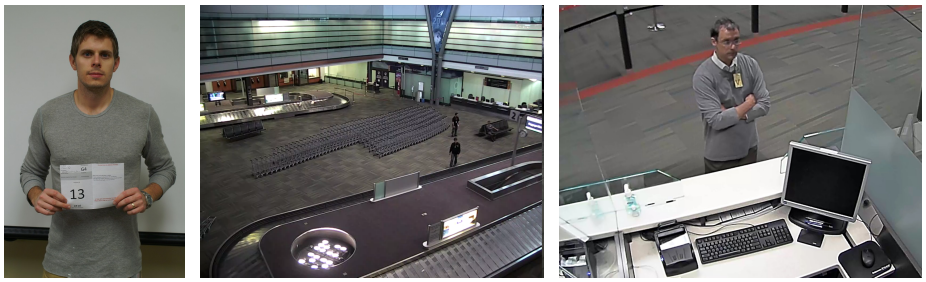
The analysis and extraction of event information from large scale multi-camera surveillance systems is of interest to the physical security community. Datasets are crucial to the testing and standardization of new potential computer vision algorithms. Many current datasets are often limited in the number of cameras,

space coverage, and pedestrian flow making them poor proxies for realistic scenarios [26–29, 31, 32]. The small number of cameras in a dataset often results in little to no overlap between cameras' field of views. A small number of cameras do not accurately represent the coverage available in a modern CCTV system. For multi-camera scenarios, datasets lack the sheer number of cameras typical of modern IP camera based surveillance deployments.

Many of these datasets are produced without a specific application in mind and do not accurately represent an operational environment. An operational environment can be defined as the real procedures and location conditions found by law enforcement or other users of a surveillance system. The metrics provided by a general video surveillance dataset therefore may not be reflective of the performance one might expect in a real operational context.

The Airport Multi-Camera video (AMCV) dataset was produced by the CBSA. The AMCV Dataset was created to provide an accurate operational representation of video surveillance in an airport border crossing environment. The video surveillance footage can be used to test a variety of video analytics algorithms such as face recognition or video tracking. The dataset includes approximately 34 hours of continuous surveillance footage representing roughly 30 minutes in real time from 76 cameras. The video was captured at the border crossing area of an international airport. The dataset features 83 CBSA employee volunteers acting as travellers traversing the border clearance process. Sample frames form the dataset and a volunteer photo are shown in Figure 4.1. We believe the dataset offers six primary advantages over its counterpart surveillance datasets:

1. Actual Airport Environment. The AMCV dataset was filmed in a currently operational airport border crossing area using the surveillance system setup



(a) (b) (c)

Figure 4.1: The AMCV Dataset features:(a) a sample mugshot of one volunteer actor; (b) overview camera with a wide field of view where identification is difficult; (c) camera viewing an interview where identification is possible, but has a small field of view.

deployed there. While the individuals filmed were not real travellers, the process they traverse is modelled after realistic border clearance procedures.

2. Extensive Camera Coverage. 76 different cameras were included in the AMCV dataset. Many datasets in operational contexts such as i-LIDS multi-camera indoor tracking dataset only include a small limited number of camera angles with minimal overlap between the cameras. The AMCV dataset was devised to provide as much information as a modern surveillance system would provide to its user resulting in 76 cameras' footage being included. Only cameras that were deemed privacy-invasive were excluded, such as cameras that captured non-volunteer individuals during the time of the dataset's creation. The result is a realistic amount of footage which must be parsed through quickly for a solution to be viable in a real-time operational context.

3. Participant Enrollment Image. Every actor in the dataset has two high quality photographs available. The mugshots can be used for applications such as to mimic a traveller watch-list scenario of interest to many public safety organizations or to begin post-event evidence retrieval after an incident.

4. Scripted Scenarios. Actors were scripted to follow the traveller protocol for international arrivals at an airport. A subset of the actors were instructed to carry out additional instructions including wearing different accessories, going to the bathroom, or using one's phone when possible.

5. Documented Capture Camera Characteristics. Included in the dataset are the characteristics and specifications of each camera used in its filming.

6. Detailed Annotations. The dataset includes an extensive annotations encoded in an XML format including timestamps and face coordinates for all 83 actors. The annotations were created and validated manually by a single individual and therefore cannot be considered on its own the ground truth without additional annotators. The XML format can be found in Appendix A.

The AMCV Dataset was not created for event recognition. There are no clearly defined events or unusual activity known to be present in the surveillance data or in the annotations. The dataset also does not include varying task difficulty – tracking or identifying any individual was not evaluated for difficulty relative to its peers.

4.2 Collection Methodology

The AMCV dataset was filmed at an airport international arrivals area with volunteers acting as travellers. Due to privacy concerns real footage of travellers could not be used and volunteer actors were used instead. The actor protocol was created to mimic the conditions and process real travellers would encounter at an airport international arrivals area. A currently used CCTV surveillance system was used to capture the resulting footage. The dataset is described in greater detail in the following three sections: the actor protocol, the camera fields of view, and the annotations

Table 4.1: Actor Protocol Steps

Step	Name	Description
1	Arrivals	Actors move from aircraft gate to PIL queue
2	Primary Inspection	Actors queue and are interviewed
3	Baggage Claim	Luggage can be claimed in this area
4	Exit Control	Individuals are either exit or sent to inspection
5	Secondary Inspection	Luggage is searched in this area.
6	Detention	Individual travellers that are detained are held here

provided.

4.2.1 Actor Protocol

Actors were instructed to follow the international arrivals process typically used in an airport environment as outlined in [39, 40]. The flow of individuals through an airport is strictly controlled: real world travellers who disembark an international flight follow a controlled path from the aircraft gate to the exit. The path's six steps are listed in Table 4.1. In the dataset, actors began at an aircraft gate and proceeded to the primary inspection (PIL) interview. Two aircraft gates serve as start sites in the dataset with the actors split between the two. Actors travelled in small groups of 4-5 from the start point to the PIL queue. Actors queued in front of the PIL which is a brief initial interview with a border officer. Actors proceeded from primary inspection to the baggage claim hall directly behind the interview booths. Actors were instructed to linger in the baggage claim hall for an indeterminate amount of time until being verbally instructed to proceed to the exit. At the exit, certain actors were directed to the secondary inspection area by a border officer. Once directed to the exit, actors will not re-appear in the dataset. If directed to secondary inspection, the actor's luggage is searched and interviewed by a border officer at which point they

Table 4.2: Actor Roles in AMCV Dataset

Protocol	Number
No special conditions	48
Adding or removing outerwear in the bathroom	6
Spending time in a bathroom	4
Looking down when walking	4
Wearing sunglasses	11
Loitering outside the bathroom	2
Walking and texting	3
Wearing a hat	16

may be detained or allowed to exit. All of these instructions were relayed to actors and border officers in a small mock passport printout they were instructed to quickly read and memorize. Actors were asked to refrain from reading the instructions during the exercise and to follow the other actors if a step was forgotten. The mock passport printouts were also used to indicate to the officers when to send an actor to secondary inspection. A complete flowchart of the international arrivals process in an airport can be seen in Figure 4.2.

In addition to following the protocol outlined above, actors in the dataset were

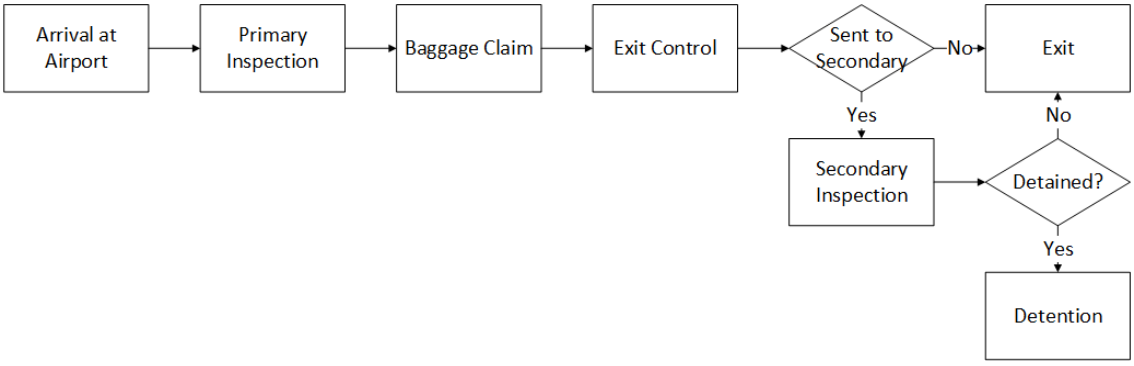


Figure 4.2: International Arrivals Flowchart

given specific actions to carry out or accessories to wear. Some actors were given multiple roles at once such as wearing both a hat and sunglasses. In total, 35 of the 83 actors were given one of the listed roles. The eight different roles are listed below:

1. Adding or removing outerwear in the bathroom: the actor was instructed to enter the bathroom and to add or remove a winter jacket before approaching the first PIL.

2. Spending time in a bathroom: the actor enters the bathroom and waits for several minutes before proceeding.

3. Looking down when walking: the actor looks at the floor as much as possible when moving from location to location.

4. Wearing sunglasses: the actor wears sunglasses, but is required to remove them whenever interacting with an officer.

5. Loitering outside the bathroom: the actor is to stand outside of the bathroom for several minutes before proceeding.

6. Walking and texting: the actor is to be using their mobile device when walking.

7. Wearing a hat: the actor was instructed to wear a hat. Unlike wearing sunglasses, removing the hat is not required when interacting with an officer.

8. Detention: the actor was sent to the detention cell for a brief period of time.

The number of actors assigned to each role is listed Table 4.2.

4.2.2 Cameras Field of View

The camera fields of view present in a customs area can be generally categorized as one of three view types: overview, corridor and interview. An overview camera views

Figure 4.3: (a) Baggage claim overview FOV; (b) corridor camera FOV; (c) upper interview camera FOV; (d) lower interview camera FOV;

wide areas and may be used to provide scene context. Corridor views show subjects moving through a bottleneck, usually a narrow hallway or queuing line. Interview cameras are placed to show travellers in situations where one must stop to interact with an officer.

An example frame from an overview camera can be seen in Figure 4.3a. Identification of an individual's features in an overview camera such as this is a difficult task due to the distance of the camera from the scene and the corresponding large field of view. The particular camera in Figure 4.3a is mounted on a high ceiling and overlooks a baggage claim area in the airport. As traffic increases in an overview camera, high levels of occlusion are probable.

In Figure 4.3b, an example frame from a corridor camera view is shown. Similar to the Chokepoint dataset, travellers must pass through a doorway, queuing line or other narrow area to proceed. Cameras are strategically placed in these areas to capture a frontal image of individuals as they pass through this area. Faces, for example, are typically identifiable from this view. Occlusion may still be a challenge depending on

how high the camera is mounted relative to the travellers in this circumstance.

The final camera view presents travellers at an interview booth in the airport. Example frames are shown in Figure 4.3c & 4.3d. Identification is easiest in these views as the camera is in close proximity to the individual and no occlusion is present. However, due to the proximity of the camera to the individual, it is not uncommon for features such as a face to be presented at a large angle to the camera offering a more significant video processing challenge.

4.3 Annotations

The dataset's annotations includes four separate file types which are listed below:

1. Actors Protocol: The actor protocol is stored in a comma separated file containing each actor's protocol, unique ID number, and their corresponding annotation XML files

2. Enrollment Photographs: Two high quality color photographs of each actor is included in the dataset as JPG files to make identification easier. One image was taken with flash; one image was taken without flash.

3. Actor Track: Each actor has a corresponding file containing the face coordinates along with the camera IDs and times the actor appears in the dataset. The actor track XML consists of a timeline with each item representing times of the appearance of the actor in a camera's field of view. The start time and end time are included in the timeline item. A single camera may have multiple timeline items as an actor may leave and re-enter a scene. Full occlusion of an actor in a scene is treated as the actor leaving the camera's field of view, and is subsequently broken into multiple timeline items. Non-travellers, such as

border officers' mock-interviewing the actors, are not included in the dataset's annotations. The XML includes face coordinates, face width and height, unique camera IDs, and the start and end time of the actor's appearances.

4. Camera Metadata: Each camera has a corresponding file containing the specifications of each camera's video file. This includes frame rate, frame width and height, angle, rotation, filename, filepath, unique camera ID, and the camera position on a 2D plane in XY coordinates.

The annotation tracks were generated and validated manually by a single individual. In total over 2GB of XML metadata is included in 163 files: 76 camera metadata files and 87 actor track XML files. Each actor in the dataset has an XML annotation file with one actor having five XML annotation files which can be used to estimate the annotation variance. Each file includes timestamps for all camera appearances for the actor. Each timestamp includes a start and end time indicating when the actor first appears and subsequently leaves the camera's field of view. Only complete occlusion of an actor was considered as an actor leaving a scene. Other forms of occlusion, such as an actor being partially visible behind another actor or only standing partially within the field of view, are included within a timestamp.

The included annotations do not include either the face or silhouette of actors in the dataset. Due to the size of the dataset, adding detailed frame by frame annotations is an ongoing process and has not yet been completed.

4.4 Chapter Summary

In this chapter we have introduced the Airport Multi-Camera Video (AMCV) dataset: a new dataset that features footage from a real surveillance system in an airport customs environment. The dataset provides realistic operational footage and camera

angles that law enforcement agencies currently utilize. 76 cameras and 83 individual actors are featured with approximately 30 minutes of footage available for most cameras. Cameras in the dataset observe a variety of scene types including wide areas, corridors and close-up actor interactions with a law enforcement officer. Individual actors were given different scripts to follow such as going to the bathroom or texting while walking through the environment. Commercial face recognition was applied to the dataset to provide a preliminary evaluation of the dataset's performance.

Chapter 5

Evaluation Framework

The success of the Search and Retrieval system in tracking an individual is based not only on the performance of the face recognition, but the face recognition skill of human operator, the ability for the operator to utilize the interface, and the difficulty of the task being completed. Therefore evaluation of the interactive system involves considering the automated algorithm, human computation and human-computer interface.

However, there is also the additional complication that the face recognition fusion involves both automated and human elements. Under this consideration, the evaluation of the system was designed to attempt to separate these two elements and associate an independent performance to the automated and human elements.

The tracking of an object or individual varies in difficulty depending on the amount of track fragmentation from occlusion, colouration, and amount of movement. It is assumed in this situation given the same series of cameras that overall image quality including illumination and noise remains relatively constant from individual to individual tracked.

5.1 Evaluation Protocol

When considering the interactive video tracking system, Search and Retrieve, it was determined that automatic video tracking metrics are insufficient to completely describe a system's overall performance. The inclusion of a human operator necessitates the use of expanded metrics to capture three critical criteria:

1. Accuracy or Effectiveness: the degree to which the video track of the subject corresponds to the ground truth video track.

2. Efficiency: the amount of time or effort needed to create the video track.

3. Satisfaction: Attitude of the user when performing the task. High satisfaction is important to maintain consistent performance over long periods of time.

To evaluate the combined human-computer system, our primary goal is to isolate and evaluate each part of the system independently to ascertain how each component performs and the performance benefits of a combined solution. In the case of the Search and Retrieve problem, this methodology translates into separating the face recognition and manual search components. Accuracy is the easiest to measure: if the user is able to correctly identify and track the target in the dataset then the system can be considered accurate. Efficiency would be measured simply by how long it took the user to achieve an accurate result. Satisfaction is more difficult and relies on the responses of the participants to a survey. A fourth measure, repeatability should also be considered, but requires a sample size larger than available [41].

The goal of the Search and Retrieve program was to improve on the performance of automated video tracking solutions using human intervention. The program may not be as efficient as an automated video tracking system, but can ideally achieve greater accuracy thanks to manual verification of results. Manual video tracking would have

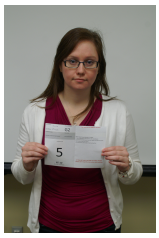


Figure 5.1: Sample Mugshot

high accuracy, but with lengthy processing times are inefficient and unsatisfactory to the user as it is a tedious process. Therefore we hypothesize that the Search and Retrieve program should have a higher accuracy than automated tracking, and a higher efficiency and satisfaction than manual tracking.

The evaluation of the Search and Retrieve interaction tracking system was modelled after usability testing and split into three separate scenarios: using solely face recognition tracking (automated tracking), using the Search and Retrieval system (interactive tracking), and using no face recognition assistance (manual tracking). In all cases, the system input is a mugshot photo of the same subject the system is attempting to track. For the purposes of manual tracking, the mugshot was provided as a printout, while for interactive tracking the mugshot was provided to the participant as a JPEG image that could be imported into the system. Participant interactions were recorded using screen capture and mouse capture software. All users were given a specified amount of time, 20 minutes, to complete the task to the best of their abilities. It was accepted that the users would not be able to complete the task in the given 20 minutes, but due to the length of time the task could take a fixed time was determined to be necessary. Contributing factors to the 20 minute time limit included that the participants in the study were not compensated, and the testing location required a minimum 15 minute one way commute for most participants.

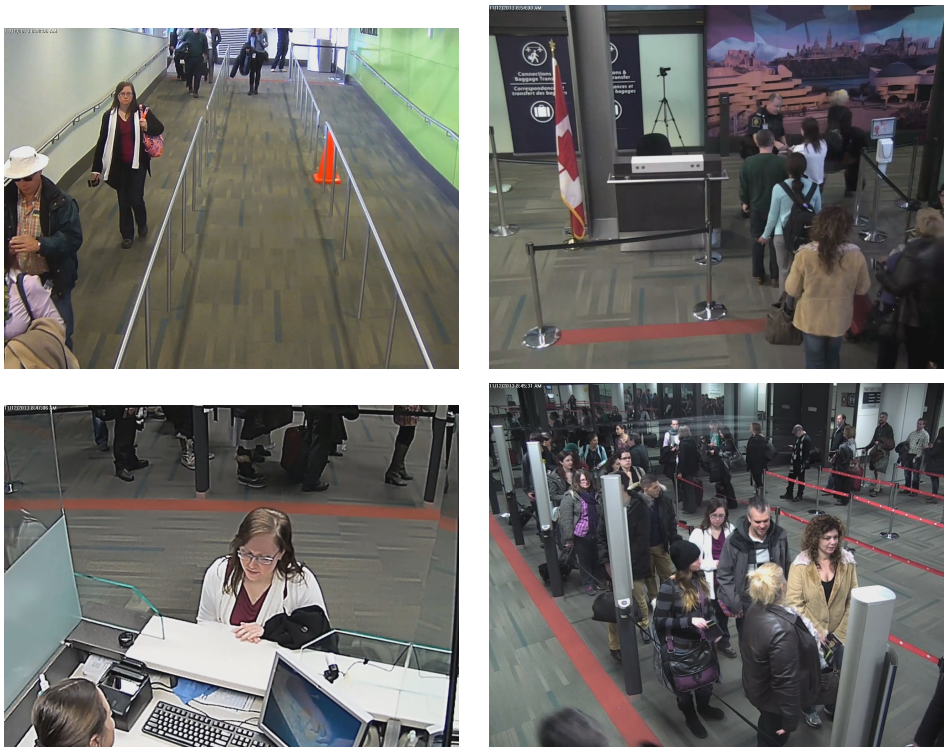


Figure 5.2: Example frames from AMCV dataset of the target to be found by participants in the evaluation

For our evaluation, the AMCV dataset described in Chapter 4 was used to simulate the task of finding an individual in an airport surveillance system and tracking their progress through customs. The target individual in this circumstance to be found was female and is shown in Figure 5.1. The target in one instance changes clothing in the bathroom. Manually tracking the target is made easier by the light clothing worn by the target providing high contrast to the more typical dark clothing worn by other individuals in the dataset, and by a brightly colored handbag carried by the target. It should also be noted that the individual wears glasses in the dataset reducing somewhat the effectiveness of face recognition. Four example frames of target at various points in the AMCV dataset are shown in Figure 5.2.

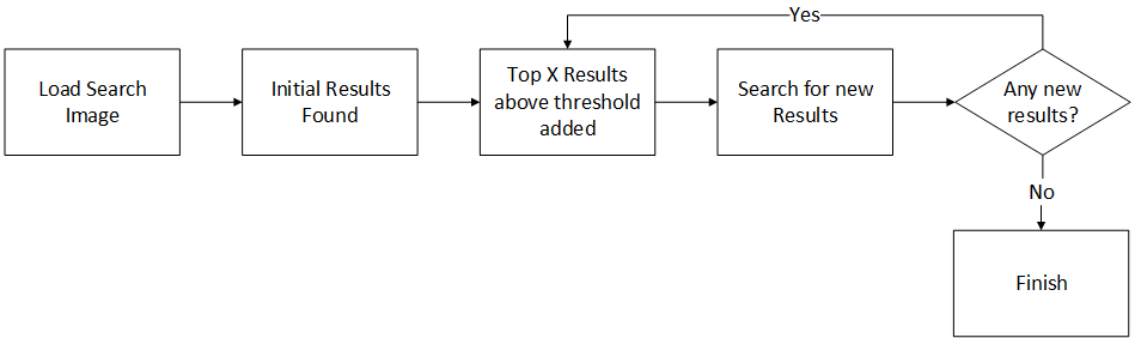


Figure 5.3: Automatic Evaluation Flowchart

5.1.1 Automatic Face Recognition Tracking

The face tracking component of the system was evaluated to simulate if it was a separate automatic face recognition system. The system was tested repeatedly using multiple thresholds, using a different number of top scores above the single match threshold to perform the recursive fusion. Once no additional matches are found during recursion, the system has completed its search. Given the nature of face recognition, a large number of tracks are expected to be missing and a high track fragmentation/lost-track ratio is expected. Of particular interest however is the false alarm detection rate. For ease of execution, the steps above were conducted manually using the Search and Retrieval software instead of reprogramming the software. The process is shown as a flowchart in Figure 5.3.

A sliding threshold was also used to generate an automatic face recognition tracking result. In this case, the threshold started low and matches were found in the same manner as above. However, when no new matches are found, the threshold is lowered and a new search is initiated. This new thresholding is repeated twice with a total of three total thresholds used in a sliding threshold evaluation. The sliding threshold evaluation was designed to simulate how a user might act if no new matches are found, but in such a way as to be result agnostic. Since the computer does not know

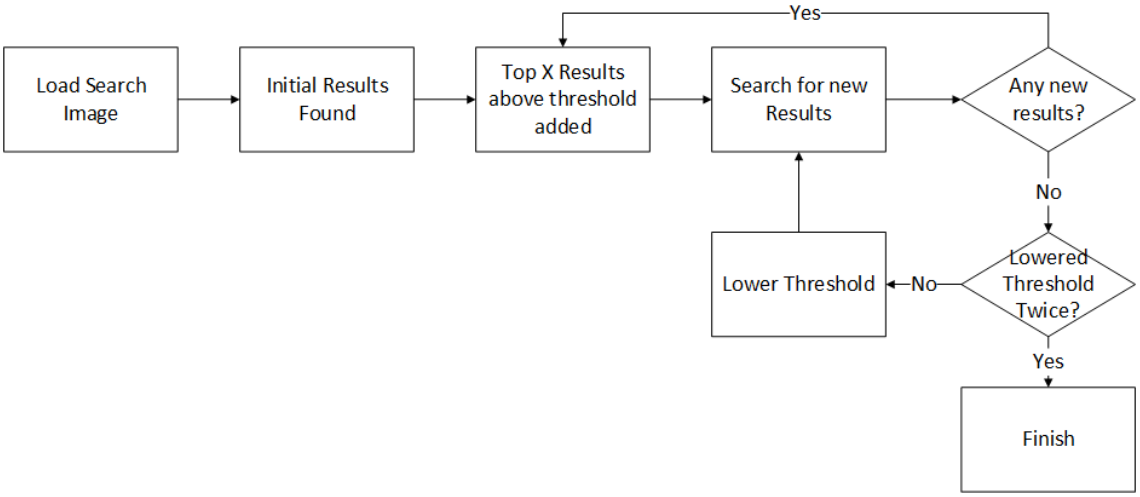


Figure 5.4: Sliding Automatic Evaluation Flowchart

if initial results added to the face recognition fusion were correct, any subsequent potential true matches may have lower scores. The converse is also true wherein if the computer adds true matches to the fusion and lowers the threshold, additional matches that otherwise would not have been found at the higher threshold can now be found. The goal of sliding the threshold was to determine if moving the threshold in this manner has any appreciable change in the overall automated face recognition result. The process is shown as a flowchart in Figure 5.4.

5.1.2 Manual Tracking

The evaluation protocol for manual tracking involved recruiting volunteers to perform manual tracking without the use of any automated algorithm, face recognition or otherwise. Participants were given an individual mugshot photos and asked to track them using the system without use of the face recognition. The participants were not given any additional information about the individual except for the photo. To perform the task, participants were given the use of the Storyboard UI as shown in

Chapter 3. Without the user interface, the task was decided to be too difficult and frustrating to attempt as it would involve individually opening the video files using a native video player and tracking the start-end times manually either using a text editor or pen and paper.

Unlike the automated face recognition tracking system, supervised evaluation was required. A total of 20 minutes was provided to each participant to complete as much of the task as possible. Instructions were provided primarily by a small booklet containing screen captures of the program demonstrating the functionality. Written instructions were preferred over verbal instruction for consistency although prompts and some verbal instruction was used when users were stuck. The instructions given can be seen in Appendix B. The same metrics can be used to evaluate the final output such as track fragmentation, correctly detected tracks, and lost-track ratios. A pre-evaluation and post-evaluation survey was used to gauge the participant's perceptions of the efficiency, and satisfaction of the task. Since the total time allotted to participants was significantly less than required to fully complete the task, the choice of when a video track is complete is removed. Standardizing the time therefore removes the potential of certain users spending more time with the program to achieve a higher overall accuracy which then must be somehow accounted for in the results. The time given to the participant should be sufficient to prevent any sort of user fatigue or boredom that may set in at different rates for each participant, but long enough to become proficient in the use of the prototype. It was determined that 20 minutes was sufficient for testing.

5.1.3 Interactive Face Recognition Tracking

Interactive face recognition tracking is accomplished by combining the methods in the manual and automated tracking. In this scenario, participants can make use of the

full Search and Retrieval program covered in Chapter 3. Both the face recognition and storyboard UI components described in Chapter 3 are utilized. Participants were again given an instruction booklet consisting of frame captures of the program and comments highlighting individual elements and potential actions that may be taken to accomplish the task. No additional verbal instruction was provided to the participants unless it was clear that no progress was being made. Similar to the manual tracking protocol, participants were given 20 minutes to attempt to complete tracking the target. A pre-evaluation and post-evaluation survey was again used to gauge the participant's perceptions of the efficiency, and satisfaction of the task.

5.1.4 Participants

Participants were students recruited from Carleton University. A major logistical challenge in recruitment was presented by the fact that any interaction with the system had to take place at the CBSA offices away from the university limiting the number of recruitable participants. All users chosen were untrained and had no exposure to the test program before participation in the study. Recruitment of CBSA employees was attempted with either current Border Officers or Border Officer Trainees being ideal participants as they are the proposed end-users of the Search and Retreive program. However, these participants were ultimately unavailable to complete the study.

Ideally an interactive system would provide similar or improved accuracy in results while taking less time to complete then performing the same task manually. By having untrained users with limited to no exposure with video tracking and surveillance systems, the goal was to reduce the impact of user skill on the interactive system's accuracy. Users of the system were restricted by time. Limiting the task to 20 minutes was chosen to make participation easier, but to complete a video tracking task in the

AMCV dataset can take multiple hours depending on the user.

Each participant performed either interactive or manual tracking, and never both. This was decided given the limited time the participants would have to complete the task. Part of the interface and program operation are also shared by both methods and to ensure uniformity in how familiar the participants were the program they would be exposed to only one of the two tasks. It was believed that comparing the two methodologies would be valid if the user skill level could be accounted for and the same target was tracked controlling for task difficulty.

5.2 Video Tracking Metrics

This section provides specific definitions for the tracking metrics used to evaluate the ability of the system to correctly identify and track a target. Tracking metrics have similarities to common biometric confusion matrix metrics that has been applied to a video tracking system. In all the metrics, the system track refers to the video track created by the system being tested. The system track is compared against the ground truth track: the track created entirely manually and is considered the gold standard for the dataset. The ground truth track has an expected error or variance associated with it that must also be considered. For the context of this research, the only ground truth available are the AMCV dataset annotations. The metrics that follow could be used with any annotations or ground truth.

The following metrics were used in this thesis: correctly detected track, false alarm track, track detection failure, track fragmentation, and the lost-track ratio.

5.2.1 Correctly Detected Track / True Positive

A track is considered correctly detected if two conditions are met: the temporal overlap and the spatial overlap between the ground truth and system track exceeds given thresholds [42]. Spatial overlap in a camera is defined in literature as the number of shared pixels between the system track and the ground truth track. If the precise pixels of the subject are not known as is the case of Search and Retrieval, it can be more loosely defined as equivalent to temporal overlap. Temporal overlap is defined as number of shared video frames between the ground truth and system tracks. Correct track detection is expressed mathematically in equation 5.1 with an arbitrary threshold, T_{OV}. The threshold can be expressed as a percentage of the total number of frames in the sequence, or in other terms the minimum amount of time the target must be visible in the field of view to be considered a true positive.

$$\frac{Frames_{GT} \cap Frames_{ST}}{Frames_{ST}} \geq T_{OV} \tag{5.1}$$

where $Frames_{GT}$ is the number of frames of the ground truth track, $Frames_{ST}$ is the number of frames of the system track and T_{OV} is the overlap threshold [42]. In the case of manual or interactive tracking, the threshold, T_{OV}, has to be defined by the user and may not be consistent between multiple users. It is possible to define a threshold for the user and provide tools to assist the user in making the determination, but these were not included as part of this trial.

An example of the correctly detected track equation being used is shown below. In this example, the ground truth track includes frames 20-30, where the system track is 17-26. The threshold for the example is 0.5.

$$\begin{gathered}\frac{Frames_{GT} \cap Frames_{ST}}{Frames_{ST}} \geq T_{OV} \\ \frac{(20-30) \cap (17-26)}{10} \geq 0.5 \\ \frac{7}{10} \geq 0.5 \\ 0.7 \geq 0.5\end{gathered} \tag{5.2}$$

5.2.2 False Alarm Track / False Positive

A system track that fails to exceed the true positive condition can be considered a false positive or false alarm track [42]. The false positive condition is defined in equation 5.3. A false positive is produced when the system has indicated footage contains the target individual, but it is not included in the ground truth.

$$\frac{Frames_{GT} \cap Frames_{ST}}{Frames_{ST}} < T_{OV} \tag{5.3}$$

where $Frames_{GT}$ is the number of frames of the ground truth track, $Frames_{ST}$ is the number of frames of the system track and T_{OV} is the overlap threshold [42]. An example of the false alarm track equation being used is shown below. In this example, the ground truth track includes frames 20-30, where the system track is 29-38. The threshold for the example is 0.5.

$$\begin{gathered}\frac{Frames_{GT} \cap Frames_{ST}}{Frames_{ST}} < T_{OV} \\ \frac{(20-30) \cap (29-38)}{10} < 0.5 \\ \frac{2}{10} < 0.5 \\ 0.2 < 0.5\end{gathered} \tag{5.4}$$

5.2.3 Track Detection Failure / False Negative

Track detection failure is when the system fails to detect a track that is included in the ground truth [42]. A track is considered a detection failure or false negative if one of the two conditions described in equation 5.5 is met [42].

$$\frac{Frames_{GT} \cap Frames_{ST}}{Frames_{GT}} < T_{OV} \tag{5.5}$$

where $Frames_{GT}$ is the number of frames of the ground truth track, $Frames_{ST}$ is the number of frames of the system track and T_{OV} is the overlap threshold [42]. An example of the track detection failure equation being used is shown below. In this example, the ground truth track includes frames 20-30, where the system track is 22-23. The threshold for the example is 0.5.

$$\begin{gathered} \frac{Frames_{GT} \cap Frames_{ST}}{Frames_{GT}} < T_{OV} \\ \frac{(20-30) \cap (22-23)}{11} < 0.5 \\ \frac{2}{11} < 0.5 \\ 0.18 < 0.5 \end{gathered} \tag{5.6}$$

5.2.4 Track Fragmentation

Track fragmentation is used to describe the lack of continuity in the system's tracking of an object [42]. Fragmentation is measured as the number of system tracks created per ground truth track as shown below in Figure 5.5. An ideal tracking system should have a track fragmentation of 0 indicating a stable and continuous tracking of the subject. Track fragmentation is easily expressed as one of two situations. Either there are more than one system tracks overlapping with a single ground truth track

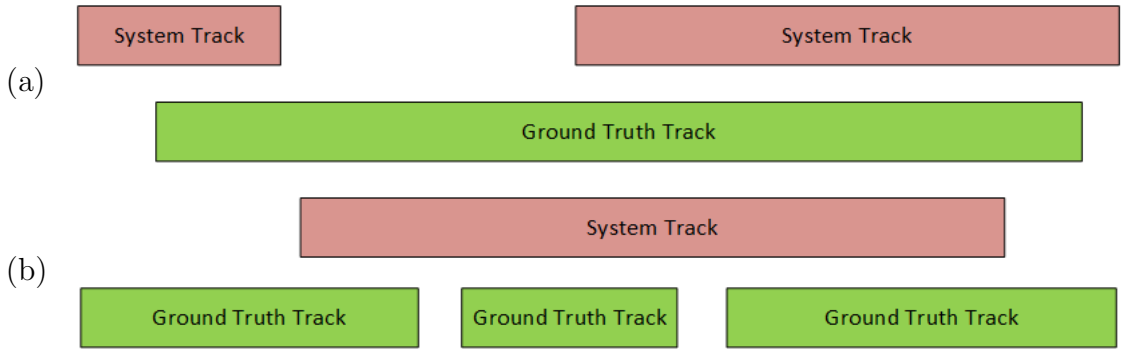


Figure 5.5: Track Fragmentation with (a) multiple system tracks per ground truth or (b) multiple ground truth tracks per system track

Table 5.1: Example Data for Sample Calculations

Ground Truth Frames	21 - 100
System Track Frames	15 - 24
	51 -70
	81 - 100

or there are multiple ground truth tracks overlapping with only one system track. One or more of these system tracks may also be classified as true positives if they exceed the threshold, T_{OV}. Track fragmentation is shown graphically in Figure 5.5.

Track fragmentation can be calculated as a ratio of the total number of system tracks to the number of true positive tracks in the sample as shown in equation 5.7. When the track fragmentation ratio approaches one, all of the true positive ground truth tracks in the sample are either fragments of the ground truth or contain multiple ground truth tracks.

$$\text{Track Fragmentation Ratio} = \frac{\text{Total Track Fragments}}{\text{Total TP Tracks}} \tag{5.7}$$

where Total Track Fragments is the total number of track fragments that exist

and Total TP Tracks is the total number true positive tracks. An example of track fragmentation is shown below in Table 5.1. In this sample data, there are 3 total track fragments for one true positive track. Therefore the track fragmentation ratio in this instance would be 3. In otherwords, there are three system tracks where there is only one ground truth track.

Similar to track fragmentation, the lost-track ratio, λ, can also be used to estimate number of frames per track that are lost when comparing the system and ground truth tracks. The lost-track ratio is defined below in Equation 5.8 [43].

$$\lambda = \frac{N_{ST}}{N_{GT}} \tag{5.8}$$

where N_{ST} is the number of correct frames in the system track and N_{GT} is the number of frames in the ground truth. Using the previous example in Table 5.1, we can apply the lost-track ratio as follows:

$$\begin{aligned} \lambda &= \tfrac{N_{ST}}{N_{GT}} \\ \lambda &= \tfrac{4}{80} \\ \lambda &= 0.05 \end{aligned} \tag{5.9}$$

5.2.5 Precision and Recall

In the context of video tracking, precision is a method of measuring the rate of success within the system track. Similarly, recall indicates how much of the ground truth footage was captured by the system being evaluated. From the true positive and false positives rates generated above, the precision and recall can be calculated as shown in equation 5.10 and 5.11 [44]. Precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.10}$$

where TP is the number of true positive frames and FP is the number of false positive frames. Recall is subsequently defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.11}$$

where TP is the number of true positive frames and FN is the number of false negative frames. This definition of recall differs from lost-track ratio as it includes multiple system tracks and ignores fragmentation whereas lost-track ratio only incorporates one system track. Using example data shown in Table 5.1 we can calculate recall and precision as follows:

$$\begin{aligned} \text{Precision} &= \tfrac{TP}{TP+FP} \\ \text{Precision} &= \tfrac{44}{44+6} \\ \text{Precision} &= 0.88 \end{aligned} \tag{5.12}$$

$$\begin{aligned} \text{Recall} &= \tfrac{TP}{TP+FN} \\ \text{Recall} &= \tfrac{44}{44+36} \\ \text{Recall} &= 0.55 \end{aligned} \tag{5.13}$$

5.3 User Metrics

This section describes some of the metrics used to measure the user skill and satisfaction with the test system. Two primary methods were used to capture and measure the human interaction with the system: desktop capture, and user surveys. These

two methods are described in greater detail in the following sections.

5.3.1 Desktop Capture

Two programs were used on the participant's PC machine to capture their interactions. CamStudio, a free streaming video software, was used to record the desktop screen of the participants during the entire test [45]. CamStudio was chosen over alternatives such as the VLC media player for desktop capture because of its cost and low memory usage [46]. The Search and Retrieve program requires significant RAM to function and having a low impact recording software was a major priority so that the recording did not impact the performance of the Search and Retrieve program and therefore the overall evaluation. For mouse capture, JitBit's Macro Recorder was used [47]. A macro recorder was useful as if data was lost, the recorded macro could be used to hypothetically recreate the user's interaction.

The desktop capture was evaluated qualitatively, but the mouse capture data can be analyzed quantitatively. A long pause in between actions or multiple long pauses in a short span of time are of interest. These spans of inaction by the user can be the product of multiple things such as the user being confused on how to progress or the user is considering the information being presented to them by the program. When a delay between actions is substantially longer than other delays, it is an anomaly that can be captured through the mouse capture data. However small actions such as movement of the mouse are captured producing significant noise in the data. To reduce the number of data points and noise, a rolling or moving average graph was created such as that shown below in Figure 5.6. In the example figure, the rolling average was created using a window of 100 samples at a time. Each of the spikes are points of interest that can be further investigated through the desktop video capture for further analysis. The horizontal line on the graph represents the mean delay for

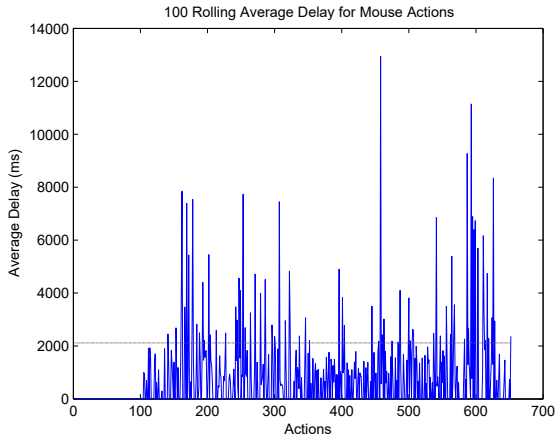


Figure 5.6: An example of a Mouse Capture Delays as a Rolling Average Graph. Abnormally long delays of no user action appear as large spikes above the overall average delay between mouse actions (horizontal line)

the subject and was used as a threshold to filter random delays with what delays were of interest. Since all interactions with the program are mouse based, keyboard capture was deemed unnecessary.

5.3.2 User Surveys

Two user surveys were provided to the participants: one before any interaction with the Search and Retrieve program and one after the interaction with the program. The pre-use survey was designed to determine whether the user had any prior experience that might influence their proficiency with the program including any prior experience with surveillance software such as a video management software (VMS) or face recognition. Knowledge of the customs process or the airport in which the AMCV dataset was filmed was also asked of the participants as familiarity with the scene captured in the AMCV dataset would presumably allow for an increase in proficiency when tracking the target.

The post-use survey was designed to measure satisfaction with the Search and Retrieve program. The questionnaire was based on the IBM Computer System Usability Questionnaire for determining subjective usability in a non-laboratory setting [48]. The rating system for the survey used was reversed from that in the literature (1 is low, 5 is high) [48]. A number of questions were removed that had no contextual basis in the Search and Retrieve program such as error messages being useful as the program does not provide any error messages. These oversights are a result of the Search and Retrieve program being a prototype and not a full commercialized product at the time of evaluation. Both surveys are included in Appendix C.

5.4 Chapter Summary

In this chapter we have introduced and described the evaluation protocols and metrics to be used in evaluating the Search and Retrieve program. These include specific metrics for measuring how close the program produced video track is to the ground truth, user impressions of the program, and how often the user is idle when using the program. The AMCV dataset will be used in conjunction with three separate test cases: automated tracking, manual tracking, and interactive tracking.

Chapter 6

Evaluation Results

6.1 Data Analysis

The test case scenarios produced an XML file containing metadata for all footage added to the video track. Of interest to the analysis is the start and end times of the video track alongside the camera identification number. All of the segments together create the completed video track.

Five separate case scenarios were considered when comparing the ground truth to the resultant XML metadata. A true match is found whenever the camera number and track time in the XML coincides with the same camera number and time found in the ground truth. When compared to the ground truth, each segment of the sample video track falls into one of five cases below:

1. Sample track is completely within by the ground truth track (Correctly Detected Track (CDT))

2. Ground truth track is completely within the sample track (CDT)

3. Sample track begins after ground truth track (partial CDT)

4. Sample track begins before ground truth track (partial CDT)

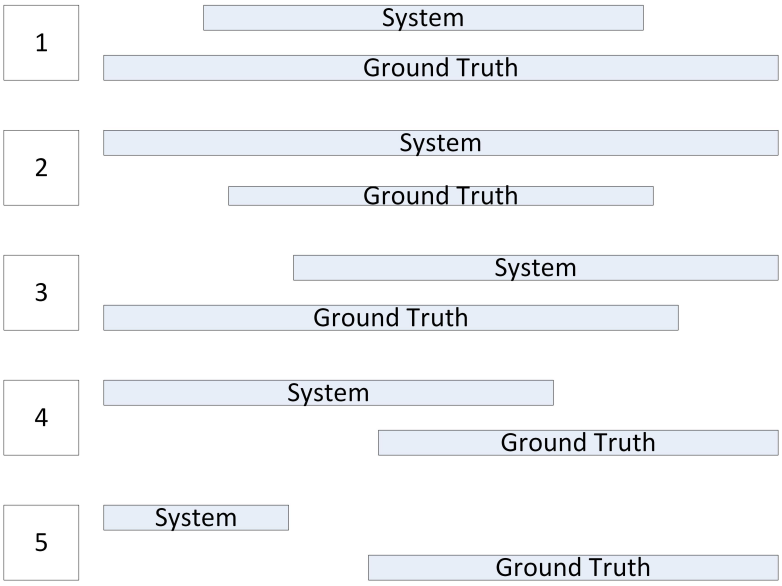


Figure 6.1: Five cases for a sample video track when compared to the ground truth track

5. Sample track and ground truth track have no overlapping footage (False Positive(FP))

All five cases are shown graphically in Figure 6.1. Analysis was completed using a custom MATLAB script file written to compare the XML metadata of each test case against the manually created ground truth XML. For each video track in the XML metadata, the MATLAB script assigns it to one of the above five cases. The lost-track ratio is then calculated for each ground truth video track with one or more system tracks associated to it. Any ground truth track with two or more system tracks associated to it, or a system track with two or more ground truth video track associated to it, are considered track fragments by the script.

Table 6.1: Ground Truth Track Stats

Statistic	Number
GT Tracks	56
GT Length (s)	1322.117
Dataset Length (s)	122400

6.2 Results

The following sections describe the results collected from the evaluation protocol. The first section describes a few key statistics of the ground truth used to compare against the different evaluation cases. Initially ten participants were planned with 5 assigned to manual tracking and 5 assigned to interactive tracking. However, due to extenuating circumstances, only 8 individuals participated in the study. Each participant was assigned a unique identification (ID) number from 1 to 8 with IDs 1-5 performing interactive tracking and IDs 6-8 performing manual tracking.

6.2.1 Ground Truth

Key ground truth statistics used throughout this section can be seen in Table 6.1. With approximately 122400 seconds of total footage in the dataset, 1.1% of the footage in the dataset has the target visible and is therefore included in the ground truth. In this situation, the only ground truth available are the AMCV dataset annotations. Since the annotations were done by a single individual, the results that follow are highly susceptible to the noise in the annotations. For example, the exact frame where the target enters or exits the frame may differ based on the annotator. Without a suitable substitute, the analysis was carried out with the AMCV dataset annotations acting as the ground truth.

Table 6.2: Face Recognition Only Results

Condition	Threshold (Exp. FPR)	Precision	Recall	Total Tracks
Top 1	0.0001	0.39	0.010	7
	0.001	0.83	0.024	8
	0.005	0.061	0.0064	14
Top 3	0.0001	0.65	0.016	8
	0.001	0.79	0.023	9
	0.005	0.061	0.0064	22
Top 5	0.0001	0.76	0.016	7
	0.001	0.90	0.030	13
	0.005	0.65	0.025	11

6.2.2 Automated Tracking

Based on the procedure presented in Section 5.1.1, the precision and recall were calculated for a face recognition only solution. Three specific cases were tested where the recursive feature used the top 1, 3, and 5 results to generate additional potential matches which are shown in Table 6.2. Thresholds are presented by the algorithm as expected false positive rate (FPR) with three being selected for testing.

From the results a few things are noteworthy. As one might expect, the recall for all of the cases is small with a maximum recall of 3%. This aligns with the assumption that solely face recognition makes a poor video tracking algorithm as faces tend to not be visible in video for extended periods of time. Reducing the threshold, or in this case raising the expected FPR, has the desired effect of increasing the total number of tracks captured by the algorithm. With a greater number of tracks and a higher expected FPR, the precision and recall correspondingly decrease.

In Table 6.3 the results of sliding the threshold down once no new results are generated are presented. The same threshold values were used as found in Table 6.2.

Table 6.3: Face Recognition Only with Sliding Threshold

Condition	Precision	Recall	Total Tracks
Top 1	0.030	0.010	12
Top 3	0.63	0.019	10
Top 5	0.73	0.022	10

No significant benefit in recall or precision is particularly notable compared to the previous method.

Using a medium threshold of 0.001 FPR resulted in the best performance of the four cases tested. This would support the expectation that the system has an optimal operating point that maximizes the precision and recall of the system. Sliding the threshold from high to low, and changing the threshold each time the system ran out of potential matches achieved increased precision and recall as the number of matches held per iteration increased. This result supports the idea that the S&R fusion scheme, while simple, can improve the face recognition results in at least some circumstances.

The lost-track ratio, λ as defined in equation 5.8 and track fragmentation, defined in equation 5.7, were compared to the ground truth and calculated for all automated tracking results and are shown in Table 6.4. The average lost-track ratio was 0.29. Track fragmentation is expressed as the ratio of track fragments over the number of true positive tracks. A track fragmentation true positive ratio of 1 indicates that all of the true positive tracks in the system track were fragments of a single ground truth track. Neither the lost-track ratio nor the track fragmentation showed a specific pattern or correlation.

Table 6.4: Lost-Track Ratio, λ, and Track Fragmentation for Automated Tracking

Condition	Threshold (Exp. FPR)	λ	TF/TP
Top 1	0.0001	0.24	1.0
	0.001	0.033	0.50
	0.005	0.43	0
Top 3	0.0001	0.32	0.50
	0.001	0.40	0.56
	0.005	0.12	0.50
Top 5	0.0001	0.38	0.43
	0.001	0.37	0.46
	0.005	0.39	0.22
Top 1	Sliding	0.080	1.0
Top 3	Sliding	0.38	0.38
Top 5	Sliding	0.33	0.4
Avg.		0.29	0.50

6.2.3 Manual Tracking

Based on the procedure presented in Section 5.1.2, the precision and recall were calculated for the manual tracking only solution. The results are presented below in Table 6.5. In the given 20 minutes provided to participants, each was able to complete nearly the same number of video tracks: 11 to 12. Subject ID 8 in Table 6.5 was able to obtain a substantially higher recall due to video tracks being of a substantially greater length than any of the other video tracks captured by any of the other participants. Three system tracks created by subject ID 8 accounted for 84% of the total footage that subject captured and accounts for the discrepancy in recalls. A difference in ability to successfully identify the target qualitatively observed during testing is a likely cause for the discrepancy in precision.

The lost-track ratio and track fragmentation compared to the ground truth were

Table 6.5: Manual Tracking Results

ID #	Precision	Recall	Total Tracks
6	0.59	0.011	12
7	0.78	0.014	12
8	0.54	0.29	11
Avg.	0.63	0.10	12

Table 6.6: Manual Lost-Track Ratio, λ, and Track Fragmentation

ID #	λ	TF/TP
6	0.081	0
7	0.13	0.091
8	0.79	0.45
Avg.	0.33	0.18

also calculated for all manual tracking results and are shown in Table 6.6. Again we can see a much larger lost-track ratio for one subject, number 3, than the others. The large amount of footage covered by the three system tracks discussed before is represented here as an increase in track fragmentation. In this situation, track fragmentation was caused by one system track existing where multiple ground truth tracks exist.

6.2.4 Interactive Face Recognition Tracking

Based on the procedure presented in Section 5.1.3, the precision and recall were calculated for the interactive face recognition solution. The results are presented below in Table 6.7. The average number of added video tracks to the solution is 25 with two outliers at 9 and 55. Subject 4 added a large number of tracks, but had a small recall indicating that each individual video track is short. Conversely, subject 3 had the smallest number of tracks, but an above average recall. This is an indication

Table 6.7: Interactive Tracking Results

ID #	Precision	Recall	Total Tracks
1	0.89	0.19	19
2	0.51	0.049	20
3	0.85	0.19	9
4	0.78	0.070	55
5	0.74	0.24	22
Avg.	0.75	0.15	25

that the average track length for subject 3 was above average which is corroborated by the raw data. Subject 2 had a significantly lower precision than the other four subjects in the evaluation. It should be noted that while the interactivity allows for an increase in identification precision, the overall track overlap may be lower if the user is careless or inattentive to when the target precisely enters and leaves a camera view.

The lost-track ratio, λ, and track fragmentation results are shown in Table 6.8. There appears to be a weak positive correlation between lost-track ratio and track fragmentation. This appears to be counter intuitive as one would expect the track fragmentation to shrink as the lost-track ratio increases. Namely, less track fragmentation would correlate to more complete coverage by the system track corresponding to a lost-track ratio closer to one. However, if time gaps between track fragments are extremely small, the lost-track ratio can be large without any change or even an increase in track fragmentation. Small time gaps between track fragments could be the result of user error or differences in the user's judgment on when the target has left the field of view long enough that the tracker should no longer consider the target as part of the scene.

Table 6.8: Interactive Lost-Track Ratio, λ, and Track Fragmentation

ID #	λ	TF/TP
1	0.39	0.77
2	0.49	0.35
3	0.78	0.89
4	0.21	0.64
5	0.49	0.81
Avg.	0.47	0.69

Table 6.9: Pre-use Survey Results

Subject ID	1	2	3	4	5	6	7	8
Age	21	32	20	18	21	21	19	22
Interactive/Manual	Inter.	Inter.	Inter.	Inter.	Inter.	Man.	Man.	Man.
Used Video Surveillance Before?	N	N	N	N	N	N	N	N
Used FR before?	N	N	N	N	Y	N	Y	Y
Last at AMCV Airport	N	N	N	N	1 month	3 years	N	1 year
Last at Customs	10 years	3 years	1 year	3 years	1 month	3 years	6 years	1 year
Edited or Annotated Video?	Y	N	Y	Y	Y	Y	N	N

6.2.5 User Feedback

The results of the short pre-use survey and post-use survey are presented here. None of the participants had any substantive experience with face recognition, video surveillance systems, or the customs procedure. A few participants had editing video experience. The results of the pre-use survey are presented below in Table 6.9. Only one participant had been at the airport where the AMCV dataset was collected within the past year, and the same participant was the only who had gone through a customs process in the past year. Generally, the previous experiences of all the participants is mostly uniform and is not expected to substantially impact the participant's performance relative to each other.

Below in Table 6.10 are shown the summary results of the post-use questionnaire.

Table 6.10: Post Survey Results

ID	1	2	3	4	5	6	7	8
Interactive/Manual	Inter.	Inter.	Inter.	Inter.	Inter.	Man.	Man.	Man.
Overall	4.35	4.21	4.69	3.30	3.00	3.64	4.28	3.00
System Use	4.50	4.25	4.62	3.37	3.00	3.62	4.37	2.62
Information Quality	4.50	4.00	4.50	2.5	2.50	3.50	4.00	4.00
Interaction Quality	4.50	4.50	5.00	3.5	3.00	3.50	4.00	3.50

The overall score is the average score for all questions asked. The System Use category combined questions that ask the user to consider how easy it was to operate the S&R prototype. The Information Quality category combines questions related to how the users felt about the software's presentation of information. The Interaction Quality combines questions related to how the users' felt about the interface and whether it was 'liked' and if it was easy to learn.

Four of the users were generally content with the system overall with a score above 4 in most or all of the four categories of questions. The split does not seem to be between users who utilized the face recognition interface or not. Without a clear correlation between the satisfaction of users who used the face recognition or not, no conclusions over the benefits of using either the face recognition or not using it to complete the task can really be drawn. Ideally one would ask the same user to compare using each half of the program, but this was not included in the evaluation protocol due to time constraints and to prevent improved accuracy in the data for whichever task, manual or interactive tracking, was completed second by the participant. It should also be noted that the interfaces for manual and interactive recognition are not substantially different, and many of the same complaints such as an inability to view surveillance footage in a larger window are shared between the two interfaces. Overall the surveys indicate a lack of comfort with the system by some of the users.

6.2.6 User Mouse Capture

During testing, all the user's mouse actions were captured using JitBit's Macro Recorder [47]. The user does not make use of the keyboard so all of the user's interactions with the S&R prototype are captured by mouse capture. Each action can be a mouse click or mouse movement. Included in the mouse capture metadata is the delay in milliseconds between each action. While a small amount of delay is expected, a user with a clear goal and proficiency with the software being used is hypothesized to not have significant large delays between actions. Long delays above the average can be considered abnormal. The aberrations may be due to user unfamiliarity with the program and how to proceed, or the user pausing to make a decision. In either case, a user with significantly less overall long delays or pauses is potentially more skilled in the program's operation than others with many long delays.

Table 6.11 below provides a summary of the number of above average delays. The average delays were calculated by first removing all of the mouse capture delays below 20 milliseconds which were considered noise. The remaining data was graphed using a rolling average of 100 samples. Mouse action delays are then calculated and presented in two ways: the number of delays over a baseline, and the number of delays over a baseline thresholded to remove data points that are close together. Unfortunately two of the mouse capture files were corrupted and were not available for analysis so the conclusions that can be drawn are limited. Overall one manual user, subject ID 2, appears to be far more proficient than the other users given substantially less major delays between actions. Apart from the one outlier, the number of delays appears to be relatively consistent. As most the users are without any training or prior experience with a video tracking problem or surveillance system, this result seemingly confirms the initial goal of all the users being of relatively equal skill.

Table 6.11: Mouse Action Delays Results

	Interactive			Manual		
ID	3	4	5	1	2	3
Delay over baseline.	78	78	52	69	9	49
Thresholded	17	20	11	12	4	11

Table 6.12: All Collected Results

ID	1	2	3	4	5	6	7	8
Interactive/Manual	Inter.	Inter.	Inter.	Inter.	Inter.	Man.	Man.	Man.
Precision	0.89	0.51	0.85	0.78	0.74	0.59	0.78	0.54
Recall	0.19	0.049	0.19	0.07	0.24	0.011	0.014	0.29
Number of Tracks	19	20	9	55	22	12	12	11
0 Lost-Track Ratio	0.39	0.49	0.78	0.21	0.49	0.081	0.13	0.79
Track Fragmentation	0.77	0.35	0.89	0.64	0.81	0	0.091	0.45
Survey Results								
Overall Score	4.35	4.21	4.69	3.3	3	3.64	4.28	3
System Use	4.5	4.25	4.62	3.37	3	3.62	4.37	2.62
Information Quality	4.5	4	4.5	2.5	2.5	3.5	4	4
Interaction Quality	4.5	4.5	5	3.5	3	3.5	4	3.5

6.3 Discussion

6.3.1 Data Comparison

A full table of all collected results can be found in Table 6.12. The greatest disparity in performance is between automated FR tracking performance and the other two tracking methods. Note that the recall rates for interactive and manual tracking are on a per 20 minute basis. On average, the automated tracking achieved a recall of only 0.018. Based on this average recall, an automated tracker would only be able to extract about 24 seconds of relevant footage from the over 1300 seconds of

Table 6.13: Average Precisions and Recalls for all three tracking methodologies

Method	Avg. Precision	Avg. Recall
Interactive	0.75	0.15 per 20min
Manual	0.64	0.11 per 20min
Automated	0.57	0.018

ground truth footage in the dataset. At best, the automated tracking method was able to achieve a recall of 0.030 at a precision of 0.90. Compared to manual tracking for 20 minutes, the average recall was 0.11 or over 6 times that of the automated tracking method. The interactive tracking for 20 minutes saw the greatest average recall at 0.15 and average precision at 0.75. A summary of the average precision and recalls for the three methods is shown in Table 6.13. The recall rate for manual and interactive tracking is likely not a linear function. It is probable that as familiarity of the users increased over time, the recall rate per minute of interactivity would increase as familiarity increased.

A precision-recall plot with all three methodologies presented is visible below in Figure 6.2. The graph clearly shows two groups: the automated tracking results to the left of the graph with low recalls and the interactive tracking results to their immediate right with higher recalls. The overall average precision of the interactive results is higher than that of the automated results. The manual tracking results have more variation with one outlier data point with a significantly higher recall than the others as described earlier in Section 6.2.3.

From the precision-recall plot, it can be shown that a rather substantial increase in recall can be achieved using the interactive methodology as opposed to either an automated or manual methodologies. The manual outlier however has a highest recall despite operating under the same time limit perhaps indicating that a skilled

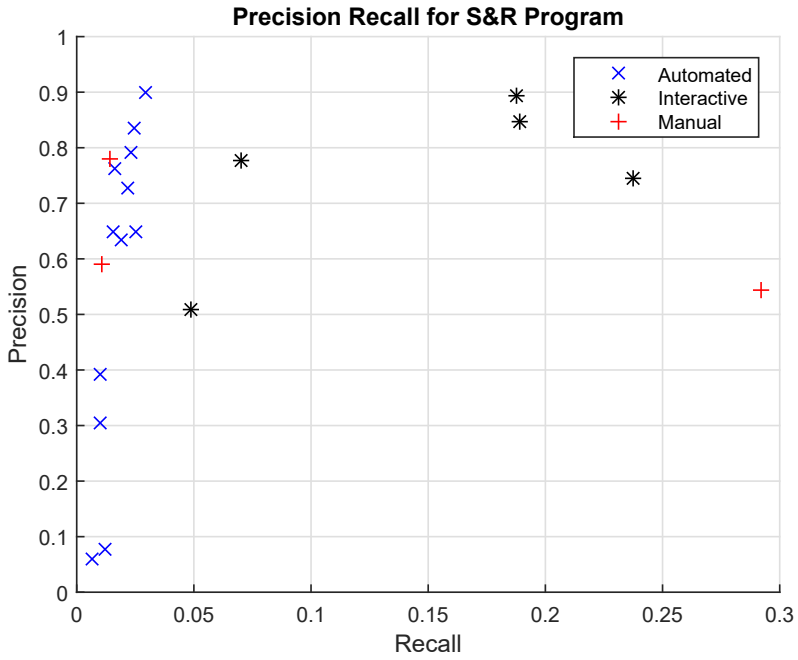
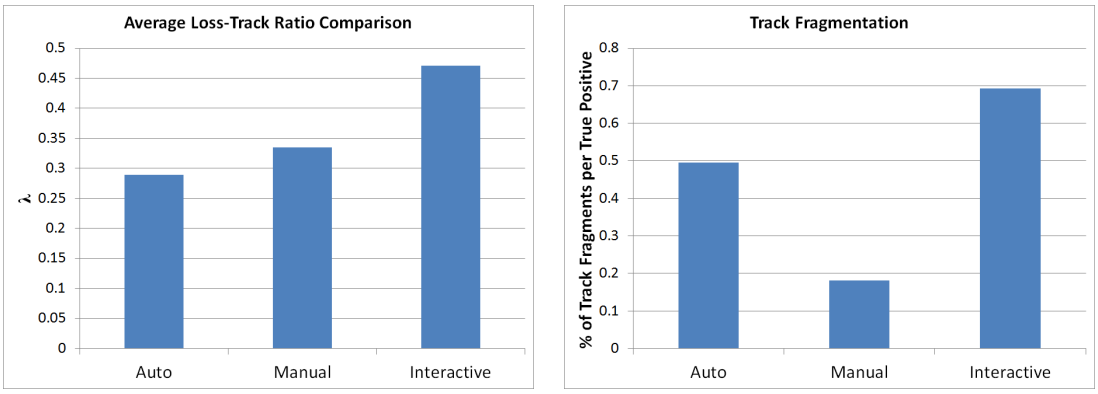


Figure 6.2: Precision Recall Curve for S&R Program with automated face recognition only (FR), Interactive Face Recognition (Mixed) and Manual methodologies present

user can achieve equal or greater recall than the interactive methodology. The precision between the three methods does not seem to vary significantly, with only the average automated method precision lagging behind in certain circumstances. The performance gap of the automated method would appear to be correctable with performance optimization. With the time limited for participants to complete the system track to the best of their abilities, it is highly probable that the recall of both the manual and interactive methodologies would increase as the time given to complete the task was increased.

In Figure 6.3 are the average lost-track ratios and track fragmentations for each of the three protocols followed in the evaluation. The interactive method had the highest lost-track ratio which further supports the precision-recall conclusion that the interactive method had the higher track completeness or recall than the other



(a) (b)

Figure 6.3: Lost-Track Ratio (a) and Track Fragmentation Comparison (b)

methods. However, the interactive method also had the highest track fragmentation of any method. This could be because of the user adding footage on top of that created automatically by the face recognition methodology. If done without properly consolidating the results of the automated process and manual process of adding video tracks, the overall track fragmentation would become the sum of the automated and manual track fragmentations. Proper training and additional time given to complete the task should reduce both track fragmentation for both the manual and interactive methods.

6.3.2 Limitations

The evaluation methodology has significant limitations that should be considered. The limitations of the evaluation extend from limitations of the AMCV dataset, the participants included, the sample size, the Search and Retrieve program itself, and length of time given to complete the video tracking task.

The AMCV dataset is limited with each camera view only providing approximately 30 minutes of footage. The 30 minutes of footage can be viewed quickly when sped

up by a user and is not representative of true surveillance systems that may include upwards of a month of searchable footage. The limitation of the footage makes longer scenarios impossible to be tested with this dataset: examples such as long loitering, disappearing and reappearing in cameras after long absences difficult to test. These scenarios would likely see an increased benefit from the use of face recognition in search and retrieval when compared to the more limited scope of the AMCV dataset. The AMCV dataset only includes small crowds consistent with the real world scenario of a single small flight making it unrepresentative of larger crowds and airport configurations. The Pan-Tilt-Zoom (PTZ) cameras present in the dataset are fixed in position; a situation that may not be the case in real world scenarios. Having a single individual be responsible for what we defined as the ground truth annotations make the results highly susceptible to noise.

The evaluation only included untrained users as participants. This does not provide any representation of the performance of trained users which would make up the bulk of the proposed end users. Untrained users were chosen for the ease of recruitment. All participants had little to no prior knowledge of surveillance systems, and a limited exposure to the airport shown in the dataset. Prior experience with face recognition was limited to commonly available consumer technologies such as those built into software products like Facebook and Google+ Photos. Having only untrained users does provide a more uniform sample than attempting to accurately measure the varied skill of trained users. Untrained users would also not be biased by experience with other surveillance system products. Using untrained users as participants meant repeat trials with the same participants was not possible further limiting potential recruits. If participants underwent repeated trials, they would become more proficient with the software likely skewing later trials recall.

There are limitations in having only one individual as the target. Each search

target in the dataset has a track difficulty associated to it based on a variety of factors some of which are difficult to quantify. Examples of these factors include the colour contrast of the target's clothing or belongings, height of the target affecting the amount of target occlusion, and the path of the target takes through the airport. Some factors also affect face recognition non-uniformly compared to human face recognition such as whether the target is wearing glasses or not. The evaluation methodology only included the ability for the participants to search for one individual target. The task length was the major factor in this decision: more than one track could take too long for participants to complete and too long a task could result in user fatigue resulting in decreased performance.

Task length is likely a major factor in the low recall rates seen by the interactive and manual methodologies. The maximum recall achieved in the test results was under 0.3. In other words less than 30% of relevant footage was captured by any of the participants. This is not entirely unexpected: anecdotally CBSA surveillance operators have indicated that a fully complete video track for a single individual can take several hours. While it would have been ideal to create a test scenario wherein the participants created the full track to the best of their abilities, the amount of time required for the trial was determined to be unreasonable. One could conjecture that if 20-30% of footage can be captured in 20 minutes, the remaining 70-80% could be captured in an additional 80 minutes, but one must also consider user fatigue. Untrained participants may not have the attention span or willingness to conduct the video tracking task for a long period of time. After 20 minutes users reported being 'bored' and 'tired' of the task.

The prototype of the Search and Retrieval program has several features either missing or incomplete that inhibit overall theoretical performance that must be taken into account. The video streaming playback feature of the prototype was unstable

and disabled for testing. In order to view the video, individual frames are extracted and displayed. While effective, this technique causes slow loading times when viewing footage quickly or switching camera viewpoints frequently. It was also not possible to increase the video viewfinder size in the UI. The machine running the prototype software also had a hardware malfunction affecting one of the trials.

While the mouse capture metrics support the idea that all the users were of relatively equal skill, more robust methods of measuring skill and user interaction would improve the certainty in that result. One would assume that user skill and familiarity would influence overall system performance in the case of the S&R prototype, but without more robust metrics and a diverse population of skills and experiences to measure the two variables it is difficult to gauge. Similarly the user surveys provide interesting insight into the comfort level of the participants with the S&R prototype. However without either a much larger sample size where variance can be reasonably determined, or an additional baseline metric, it is impossible to determine what effect the user satisfaction had, if any, on overall performance.

6.4 Recommendations for Future Evaluations

The evaluation methodology can be improved in a number of ways to increase its validity and improve subsequent results. These potential improvements are outline in this section as recommendations for future evaluations based on the presented methodology. The small sample size of our evaluation is a large weakness and would be easily rectified in the future by increasing the number of participants. The primary changes described include improving the dataset annotations, measuring participant skill, participant training, and modifying the participant activities. The recommendations are summarized by a step-by-step itemized process.

The AMCV dataset annotations were completed by a single individual. This presents a challenge when conducting the evaluation as we considered these annotations as the ground truth. If at least one tracking task was completed by a multitude of individuals, one could measure the variation between annotations. Estimating the variation is important to provide statistical validity to the results produced by the evaluation: without it any result could be due to noise in the annotations being used as the ground truth. A tracking difficulty metric should be added to the dataset to give at minimum a relative measure of how difficult each target in the dataset is to track. This difficulty metric would make it easier for participants to multiple targets and compare the results.

A consistent measure of user skill is also required to improve the evaluation. Instead of estimating user skill during the evaluation, it is recommended that an additional test is used before the participant's use the program to be evaluated. For face recognition, the Cambridge Face Memory Test (CFMT) could be used to estimate the participant's ability to recognize faces [49]. The selected test should reflect the task given to the user so in our case an additional test for tracking an individual in video without a face would be a good addition.

User training was an important consideration in the evaluation. Training was not given to users in this work and instead a short instruction manual was provided. More extensive training could be considered in future evaluations. Instead of providing a manual, the participants could be given time with the program with training instructions. The instructions could be written such as an interactive software tutorial, or provided verbally by the researcher or another individual familiar with the software being evaluated. If training is given verbally, it should be scripted to try to maintain uniformity between participants and their training. An interactive tutorial that is heavily scripted would be ideal, but could take significant fine tuning to be useful

and therefore training with an experienced trainer may be preferred. Participants in this evaluation were students and not the target demographic of the software being evaluated. If possible, the target user group should be used in the evaluation.

The activities each participant also conducted could be modified. In this evaluation, a participant performed either the manual or interactive tracking on a single target and was limited to 20 minutes with the program. Future evaluations may achieve superior results if participants performed manual and interactive tracking. This could be done either in sequence immediately after the first was completed, or, perhaps more ideally, on separate days if circumstances permitting. Multiple targets could be tracked by the participants, and the time restriction could be increased or removed. As demonstrated by the significant variation in the manual tracking performance, it was also determined that where the user chose to start highly influenced the speed at which footage was potentially captured. To control for this in the future, a starting point for the participant to begin the task should be considered.

By implementing some of these recommendations, the evaluation methodology should be applicable to other interactive systems and datasets. The recommendations would change the structure of the evaluation protocol presented as follows:

1. The dataset to be used has at least one track annotated by multiple individuals, as many as possible, to estimate variance in the ground truth.

2. Participants for the evaluation are recruited from the video tracking task's user base. In our case it would be CBSA border services officers who are responsible for managing and exporting video footage from surveillance systems.

3. Participants would take a standardized test, like CFMT, to measure user skill. A survey to gauge previous experiences can also be applied here.

4. A standardized training of the software to be evaluated would be presented to

the participants. Training would include hands on time with the software.

5. The primary task, in this case video tracking with or without face recognition assistance, would begin. A common starting point for all participants would be selected beforehand. Participants would perform both the manual and interactive tasks. At minimum each participant would complete one manual and one interactive tracking task. The tracking targets can be randomized if a measure of tracking difficulty is available, or otherwise should be restricted to a narrow pool so participants are tracking the same sample of subjects in the dataset. If possible, a time limit would not be imposed, but if necessary would be maximized as much as possible. Mouse capture, desktop capture, and eye tracking should be employed here to record participant activities.

6. Following the completion of all tasks, a survey is given to participants to measure impressions on usability of the software. Given a large enough sample size the results may be significant.

6.5 Revised Evaluation Protocol

Based on the recommendations made in Section 6.4, a revised evaluation protocol for the S&R prototype was developed. This section describes the new step-by-step evaluation protocol with all the modifications to be made. The AMCV dataset is assumed to be used, but these instructions could be applied to another dataset.

1. Select two targets for participants to track from the dataset for the evaluation. The two targets should share the same actor script: for example if one goes to the bathroom, the other should do the same. For this reason, AMCV dataset subject identifications (ID) 1 and 12 are recommended as they do not take any

special actions, but a number of other suitable combinations exist.

2. Create the ground truth for the two selected targets. At least five different individuals that are not participants in the evaluation should manually track the two targets in the entire dataset. The frame or time the targets enter and exit each camera's field of view should be included in this ground truth. For each target, the variance in annotations should be calculated. The more individuals that can help create the ground truth the better the variance measurements will be.

3. Run automated tracking (face recognition only) on the two selected targets. This process will be automated by modifying the S&R prototype to include more thresholds than included here in this thesis.

4. Prepare a training plan to introduce and train participants in the S&R prototype's operation. The plan will include both the face recognition component, and the storyboard or map component. The training should be scripted to ensure every participant is shown and told the same thing, but participants may ask questions. Repeated questions should be noted and added to future trainings. A third target from the dataset, not one of the two previously selected targets for evaluation, should be chosen for participants to practice tracking during the training.

5. Select a user skill test to determine and quantify the user's baseline skillset. For face recognition the Cambridge Face Memory Test appears to be suitable, but should be tested before it is utilized [49].

6. Prepare test setup area. A single workstation should include a desktop computer with a wide-screen monitor, mouse and keyboard. The computer should have

the S&R prototype installed along with appropriate desktop capture and mouse capture software. CamStudio and JitBit Macro Recorder were used for these functions respectively. For performance reasons, the AMCV dataset should be stored locally on a local SATA hard drive, or a solid state drive. Multiple workstations may be set up provided each is staffed by a member of the research team. The Cambridge Face Memory Test, which will be used as the user skill test, should be accessible from the workstation computer.

7. Participants are recruited for the evaluation. Participants would be recruited from CBSA Border Officers as they would be the user base of the S&R prototype if it were in production. At least 20 participants are ideal with more being better.

8. Create the pre-evaluation and post-evaluation surveys for the participants to fill out. The pre-evaluation survey should, for Border officers, include questions about length of service, familiarity with the airport where AMCV dataset was captured, and exposure to surveillance system operation. The post-evaluation survey used in this thesis can be used unchanged as it may provide useful information with an increased sample size.

9. Conduct the evaluation with participants. Participants that are not conducting the evaluation should be kept in a separate room while waiting.

   (a) Explain to participants the entire evaluation process and the purpose of the evaluation. Any necessary consent forms should be completed at this step.

   (b) Participants should now complete the user skill test selected in Step 5.

   (c) The S&R prototype training should now be provided based on the plan

produced above. Any saved data from training should be cleared from the S&R prototype before proceeding to the next step.

(d) Half of the participants will now begin manual tracking of the first target (ID 1) using the same instructions as in Section 5.1.2. Upon completing the manual tracking, this half of the participants would proceed to interactively track the second target (ID 12) based on the procedure in Section 5.1.3. The other half of the participants would start by interactively tracking the first target (ID 1) followed by manually tracking the second target (ID 12). In between the two the tracking tasks, the data collected should be saved, and the program cleared. All tracking tasks should start at the same camera location. Camera 77 at time 0 in the AMCV dataset is recommended as the starting point since every subject in the dataset walks towards and past this camera with a visible face. A time limit for any of the tracking tasks should not be imposed if possible, but if required, an hour would be given to complete the part. Both tracking tasks should be recorded by desktop capture and mouse tracking.

(e) After completing both tracking tasks, participants should be given the post-evaluation survey to complete.

(f) Participants should now be thanked and any compensation promised can be provided at this stage.

10. With the evaluation complete, the analysis of the data can be conducted. Based on the variance in the ground truth, the statistical validity of the results should be easier to test.

6.6 Chapter Summary

In this chapter we have presented and discussed the results of the evaluation protocol described in Section 5. First we presented the individual results for automated, manual and interactive tracking. The mouse capture and user survey data was also presented followed by a comparison and discussion of all the data captured. Strong conclusions cannot be drawn from the evaluation data, but overall the recall was found to be significantly higher for interactive tracking over manual and automated tracking. Limitations of the evaluation and recommendations for future evaluations were also discussed.

Chapter 7

Conclusion

7.1 Thesis Conclusions

The main objective of this thesis was to produce and demonstrate an evaluation protocol for interactive video tracking. The developed evaluation protocol split evaluation into three separate components: automated tracking, manual tracking, and interactive tracking. The metrics to be tracked were designed to encapsulate the three major human-computer interaction principals: effectiveness, efficiency and satisfaction. Effectiveness was captured by video tracking metrics, efficiency was captured by restricting all users to the same time length of interaction with the system, and satisfaction was captured by user surveys.

The created evaluation protocol was applied to the Search and Retrieve prototype. By using the S&R prototype's interactive tracking, the limited results showed that the precision and recall could be increased when compared to solely using either the manual tracking or automated tracking components. The face recognition only automated tracking method yielded the poorest average precision and recall with manual tracking performing between automated and interactive tracking. Interactive tracking yielded an average 33% increase in precision over automated tracking and

an 18% increase in precision over manual tracking. Recall saw greater increases with an 8× increase in recall over automated tracking and a 39% increase in recall over manual tracking in a 20 minute span. In the limited sample, interactive tracking achieved the greatest lost-track ratio of the three methods at the expense of increased track fragmentation. Mouse capture was used in an attempt to measure average user delays with minimal success and no discernible differences between the user dependent methods: manual and interactive tracking.

7.2 Summary of Contributions

This thesis has led to three primary contributions. The Airport Multi-Camera Video dataset footage was formalized into a full dataset with annotation XML metadata, and documentation. The dataset features 83 different trackable targets featured in 76 cameras and over 34 hours of total footage. An evaluation protocol was created for interactive video tracking that made use of both machine vision video tracking, and human computer principals. The primary driver of the new protocol was to capture the effectiveness of interactivity in a video tracker. The developed evaluation protocol was then applied to the Search and Retrieve prototype

The third contribution was the application of the created evaluation protocol to the S&R prototype. The evaluation provides insight into the benefits of using the S&R prototype over manually tracking or relying on face recognition to track a target in the AMCV dataset. The fourth contribution was a set of specific recommendations for a refined evaluation protocol which are summarized in a revised evalaution protocol for the S&R prototype.

7.3 Future Work

The Search and Retrieve prototype has significant room to be expanded. One such expansion includes utilizing different face recognition algorithms, either separately or in conjunction with the current COTS product. Adding other video tracking methods such as blob tracking or optical flow can be implemented to further assist the operator. As discovered in the evaluation process, a number of improvements to the user interface would likely improve operator effectiveness and efficiency. These include the ability to resize or full screen the video being watched, watching multiple video streams simultaneously, improved tool-tips, and overall improved performance optimization with respect to processor, hard drive, and graphical load on the host machine. Improvements to the S&R prototype have the potential to impact the performance metrics in our evaluation, and could therefore be repeated after the program is adjusted.

The AMCV dataset currently can be used to conduct experiments to evaluate many machine vision or video analytics algorithms designed to work on a multi-camera surveillance system such as video tracking or face recognition. To provide more in-depth metrics, and expanded annotations are required which would include frame-by-frame coordinates for both the face and silhouette of individuals in the dataset. Having multiple individuals perform annotations on at least one video track would bring the dataset annotations closer to a formalized ground truth.

The AMCV dataset is limited in scope despite its overall size. While it has a large number of available camera angles to be viewed, the total footage per camera is less than an hour. A real tracking scenario could include footage across up to 30 days. While a dataset that large would be difficult to manage, it would be necessary to get a true representation of how the S&R prototype would function in a real life

scenario. Alternatively, it would be ideal to measure performance in a pilot study should the S&R prototype or a similar program were implemented in a real operational environment.

Only one target was tracked in the AMCV dataset during the evaluation. One target was chosen for simplicity and to shorten the task length. With one target, certain characteristics are static between trials for the target such as the target's gender, actions taken, and clothing worn. Every target to be tracked has an innate difficulty. The tracking difficulty is influenced by a variety of factors; examples of factors include wearing bright or dark colours affecting contrast, time spent partially or fully occluded and overall distinctiveness of the face to the operator. There is unfortunately no easy unified metric for determining how difficult a video tracking task is with a multi-camera setup. Future work could include repeating the evaluation using a multitude of targets and varying the target tracking difficulty.

The evaluation protocol's implementation was successful, but can be redone with a more relevant user group and a larger sample size. Using untrained users were a uniform group and more accessible, but performing further evaluation on future prototypes or a finished commercialized product should be done with potential end users, CBSA officers. The number of metrics included can be expanded to include coordinate based tracking if the dataset being tested included these values. Implementing eye tracking would improve the overall usability test results and could provide greater insight how users can meaningfully interact with the S&R prototype. Varying the time given to complete the task could be insightful. If users were allowed to spend as much time as they deemed necessary to complete the video tracking task, one could determine what the maximum recall level would be for interactive and manual tracking. By closely following and measuring progress over the course of the task, we could measure how quickly the user becomes proficient and get a reasonable recall rate over

time. The evaluation could also be used to determine how much footage is not found based on the user's judgement of when the task is complete. A user may decide the video track is complete when not all the possible footage has been found; how much footage the typical user does not find could be important performance concern depending on how much footage is lost. Additional metrics from other classification and tracking fields could be considered and employed in future work such as radar tracking, and Kalman filtering.

List of References

[1] J.-P. Bergeron and D. Bissessar, "Accelerated evidence search report," tech. rep., Defence Research and Development Canada, Centre for Security Science, Ottawa ON (CAN);Canada Border Services Agency, Science and Engineering Directorate, Ottawa ON (CAN), July 2014.

[2] E. Trucco and K. Plakas, "Video tracking: a concise survey," *Oceanic Engineering, IEEE Journal of*, vol. 31, no. 2, pp. 520–529, 2006.

[3] Z. Zhang, "Microsoft kinect sensor and its effect," *MultiMedia, IEEE*, vol. 19, no. 2, pp. 4–10, 2012.

[4] E. Maggio and A. Cavallaro, *Video tracking: theory and practice*. John Wiley & Sons, 2011.

[5] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 1, pp. 4–20, 2004.

[6] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of biometrics*. Springer Science & Business Media, 2007.

[7] UK Government Digital Service, "Registered traveller service." https://www.gov.uk/registered-traveller, 2015.

[8] Australian Department of Immigration and Border Protection, "Arrivals SmartGate." https://www.border.gov.au/Trav/Ente/Goin/Arrival/Smartgateor-ePassport, 2015.

[9] A. A. Ross, A. K. Jain, and K. Nandakumar, "Information fusion in biometrics," *Handbook of Multibiometrics*, pp. 37–58, 2006.

[10] P. Smyth, U. Fayyad, M. Bhowpublished, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," 1995.

[11] T. Dunstone and N. Yager, *Biometric system and data analysis: Design, evaluation, and data mining*. Springer Science & Business Media, 2008.

[12] A. Evans, J. Sikorski, P. Thomas, S.-H. Cha, C. Tappert, J. Zou, A. Gattani, and G. Nagy, "Computer assisted visual interactive recognition (caviar) technology," in *Electro Information Technology, 2005 IEEE International Conference on*, pp. 6–pp, IEEE, 2005.

[13] J. Zou and G. Nagy, "Visible models for interactive pattern recognition," *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2335–2342, 2007.

[14] A. Adler and M. E. Schuckers, "Comparing human and automatic face recognition performance," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 37, no. 5, pp. 1248–1255, 2007.

[15] A. K. Jain and S. Z. Li, *Handbook of face recognition*, vol. 1. Springer, 2005.

[16] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[17] K. Meethongjan and D. Mohamad, "A summary of literature review: Face recognition," 2007.

[18] M. Turk, A. P. Pentland, *et al.*, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pp. 586–591, IEEE, 1991.

[19] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional datawith application to face recognition," *Pattern recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.

[20] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[21] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg, "The bochum/usc face recognition system and how it fared in the feret phase iii test," in *Face Recognition*, pp. 186–205, Springer, 1998.

[22] J. Sang, Z. Lei, and S. Z. Li, "Face image quality evaluation for iso/iec standards 19794-5 and 29794-5," in *Advances in Biometrics*, pp. 229–238, Springer, 2009.

[23] X. Gao, S. Z. Li, R. Liu, and P. Zhang, "Standardization of face image sample quality," in *Advances in Biometrics*, pp. 242–251, Springer, 2007.

[24] M. G. Helander, *Handbook of human-computer interaction*. Elsevier, 2014.

[25] J. A. Jacko, *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. CRC press, 2012.

[26] J. Ferryman and A. Ellis, "PETS2010: Dataset and challenge," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pp. 143–150, IEEE, 2010.

[27] Home Office Scientific Development Branch, "Imagery library for intelligent detection systems (i-LIDS)." https://www.gov.uk/guidance/imagery-library-for-intelligent-detection-systems, 2007.

[28] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 81–88, IEEE, June 2011.

[29] EC's Information Society Technology's programme project, "Context aware vision using image-based active recognition (CAVIAR) dataset."

[30] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple Object Tracking using K-Shortest Paths Optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[31] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3153–3160, IEEE, 2011.

[32] M. Campbell, A. Haubold, M. Liu, A. Natsev, J. R. Smith, J. Tesic, L. Xie, R. Yan, and J. Yang, "Ibm research trecvid-2007 video retrieval system," in *TRECVID*, 2007.

[33] Axis Communications, "Axis Product Guide: Network video solutions." http://classic.www.axis.com/files/brochure/pg_video_en_60161_1410_lo.pdf, 2014.

[34] Panasonic Corporation, "Network camera comparison chart." http://ssbu-t.psn-web.net/Catalogue/Comparison_Chart_2B-007LA.pdf, 2013.

[35] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *Image Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 1168–1181, 2007.

[36] T.-H. Hsu, C.-H. Lee, and L.-H. Chen, "An interactive flower image recognition system," *Multimedia Tools and Applications*, vol. 53, no. 1, pp. 53–73, 2011.

[37] A. J. Quinn and B. B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1403–1412, ACM, 2011.

[38] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "re-CAPTCHA: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[39] Canada Border Services Agency, "Arriving by air." http://www.cbsa-asfc.gc.ca/travel-voyage/aba-apa-eng.html, 2013.

[40] US Government Publishing Office, "CFR 235 - Inspection of Persons Applying for Admission, Part 235." http://www.gpo.gov/fdsys/pkg/CFR-2012-title8-vol1/pdf/CFR-2012-title8-vol1-part235.pdf, 2012.

[41] A. Dix, *Human-computer interaction*. Springer, 2009.

[42] F. Yin, D. Makris, and S. A. Velastin, "Performance evaluation of object tracking algorithms," in *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2007), Rio de Janeiro, Brazil*, 2007.

[43] T. Nawaz and A. Cavallaro, "Pft: A protocol for evaluating video trackers," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 2325–2328, IEEE, 2011.

[44] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[45] CamStudio, "CamStudio(TM) Open Source." http://camstudio.org, 2015.

[46] VideoLAN Organization, "VLC media player." https://www.videolan.org/vlc, 2015.

[47] JitBit, "Macro recorder." https://www.jitbit.com/macro-recorder, 2015.

[48] J. R. Lewis, "Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995.

[49] B. Duchaine and K. Nakayama, "The cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants," *Neuropsychologia*, vol. 44, no. 4, pp. 576–585, 2006.

Appendix A

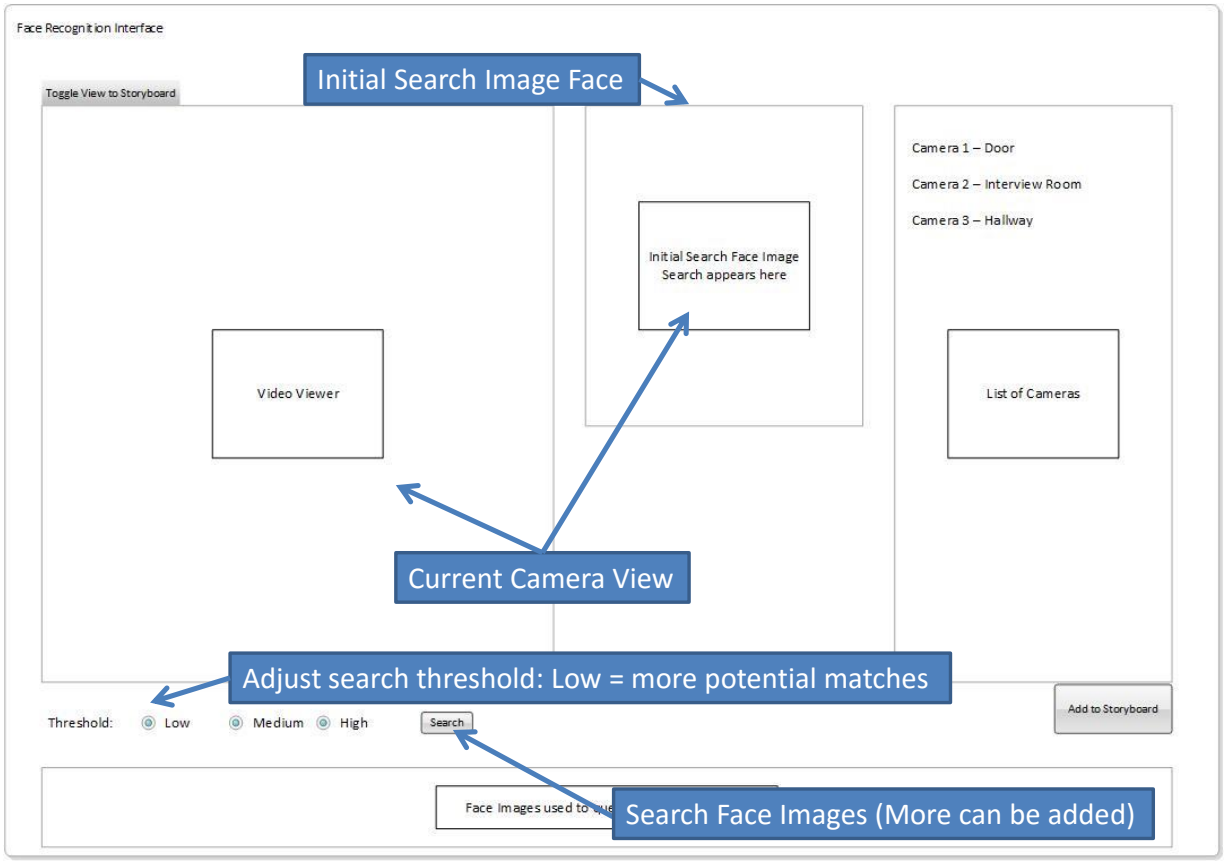
AMCV Dataset Ground Truth XML Syntax

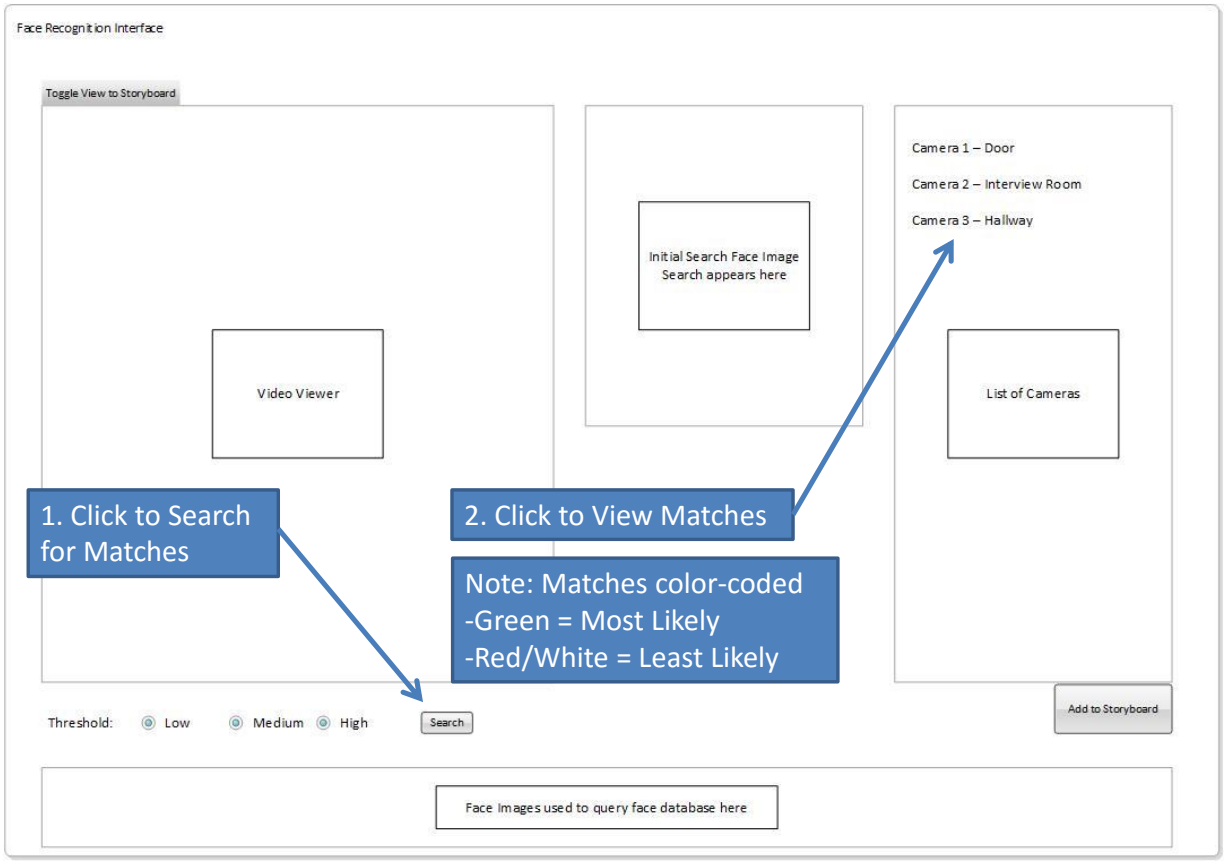
```
<?xml version="1.0" ?>
<DataModel xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

<passenger>
        <id>  </id>                                     //Unique ID for Passenger
        <role>  </role>                                 //Scenario for this Passenger
        <mugshotPath> </mugshotPath>                    //File path for mugshot
        <item>
                <videoFile>
                        <angle> </angle>                //Angle of the camera (usu. 90)
                        <rotation> </rotation>          //Rotation of video file (usu. 0)
                        <cameraID> </cameraID>          //Camera's unique ID
                        <_FPS> </_FPS>                  //FPS of video file
                        <_width> </_width>              //width of video file
                        <_height> </_height>            //height of video file
                        <file> </file>
                        <fileName> </fileName>
                        <InitialTime> </InitialTime>    //Start time of video
                        <cameraPosition>                //Coordinates of camera within MCIA
                                <X> </X>
                                <Y> </Y>
                        </cameraPosition>
                </videoFile>
                <startTime> </startTime>                //start time of frames of interest
                <endTime> </endTime>                    //end time of frames of interest
                <frames>
                        <frame>                         //For each frame in the video
                                <face_coordinates>      //Coordinates of face in the frame
                                        <X> </X>
                                        <Y> </Y>
                                        <height> </height>
                                        <Width> </width>

                                </face_coordinates>
                        </frame>
                        ...
                </frames>
        </item>
        ...
</passenger>
...
```

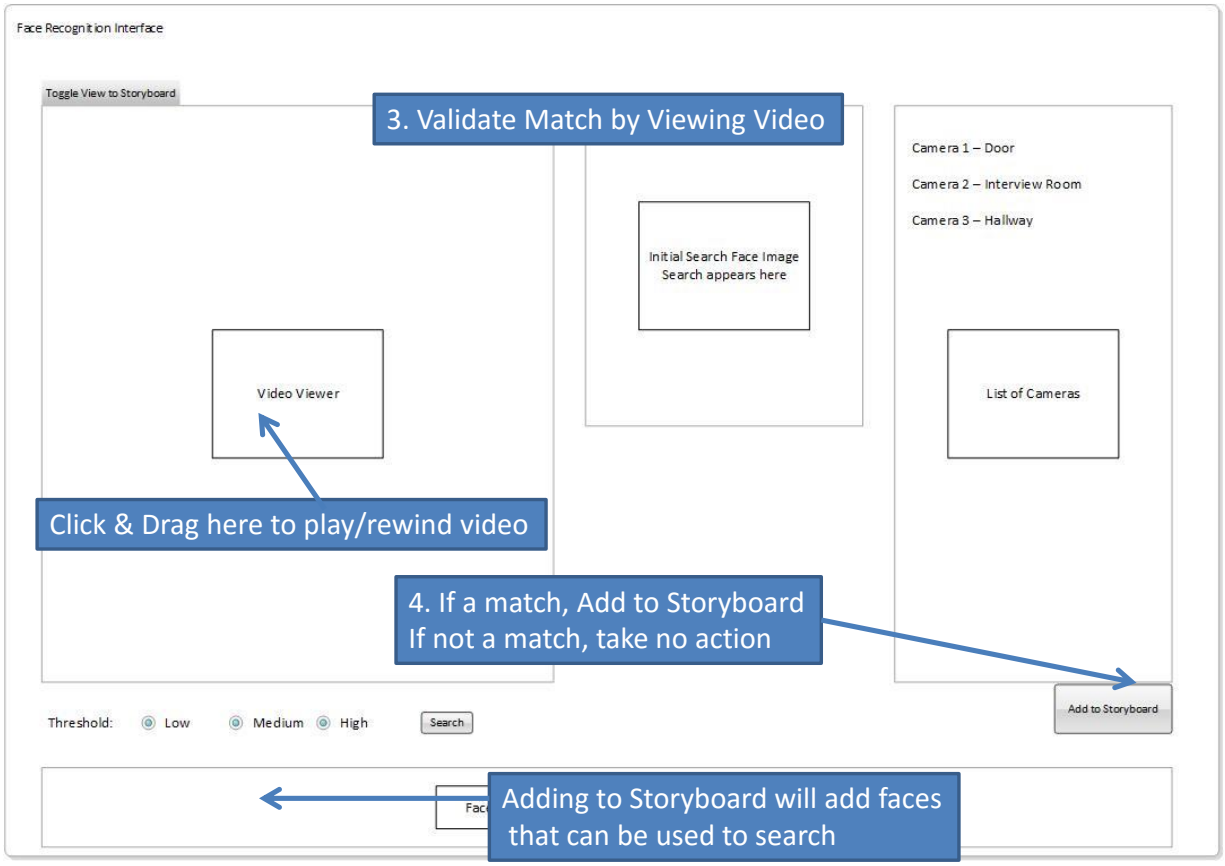
Appendix B

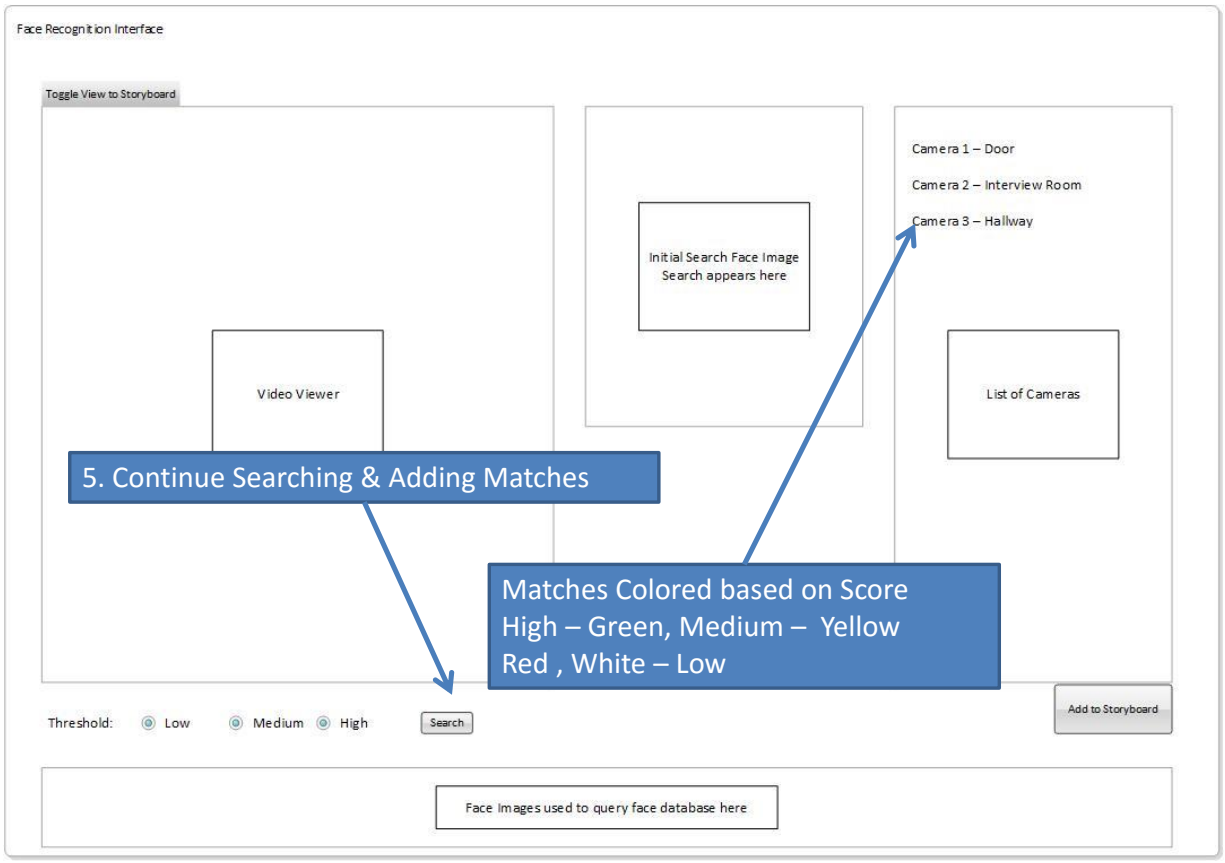
Evaluation Instructions given to Participants

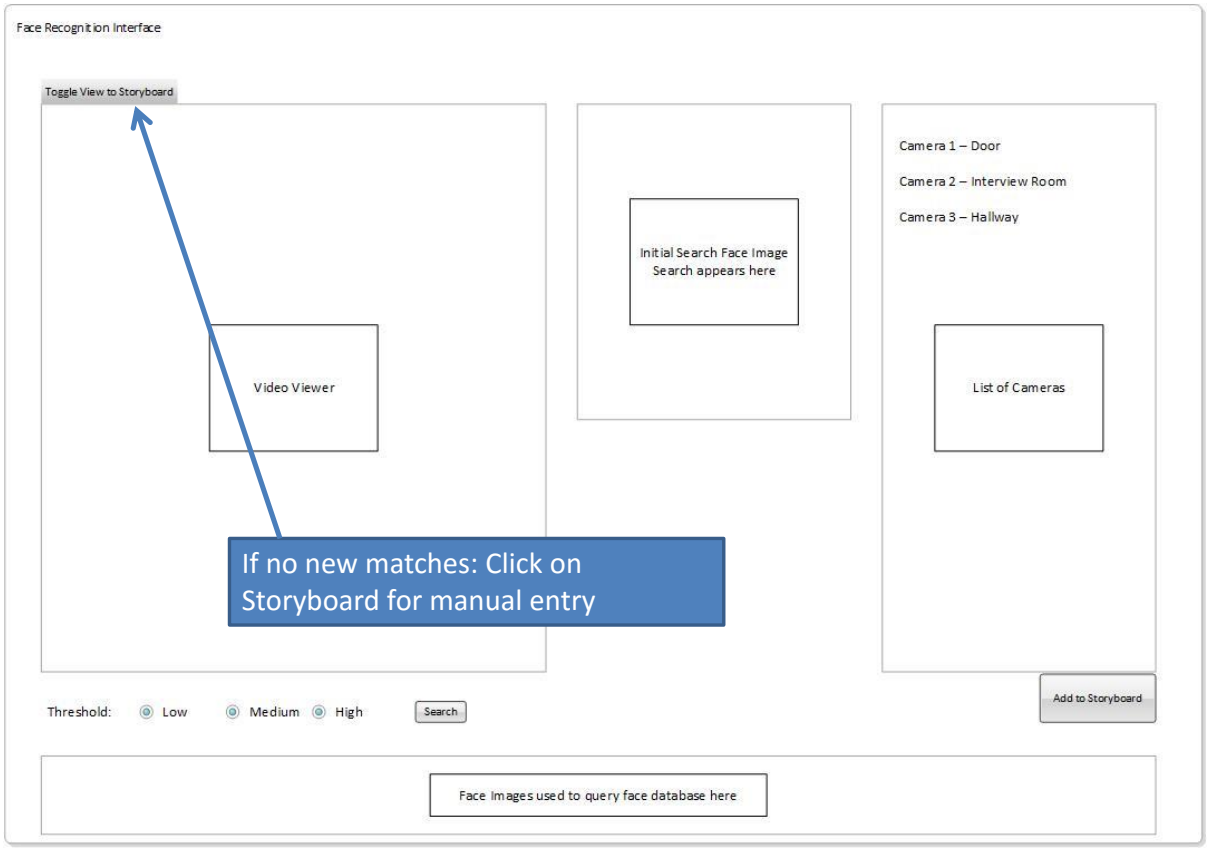
What follows are the instructions given to the participants in the evaluation study of the Search and Retrieve prototype. In the real instructions, actual screenshots taken from the program were used and not the wireframes presented. CBSA has asked that the actual program user interface be kept confidential.

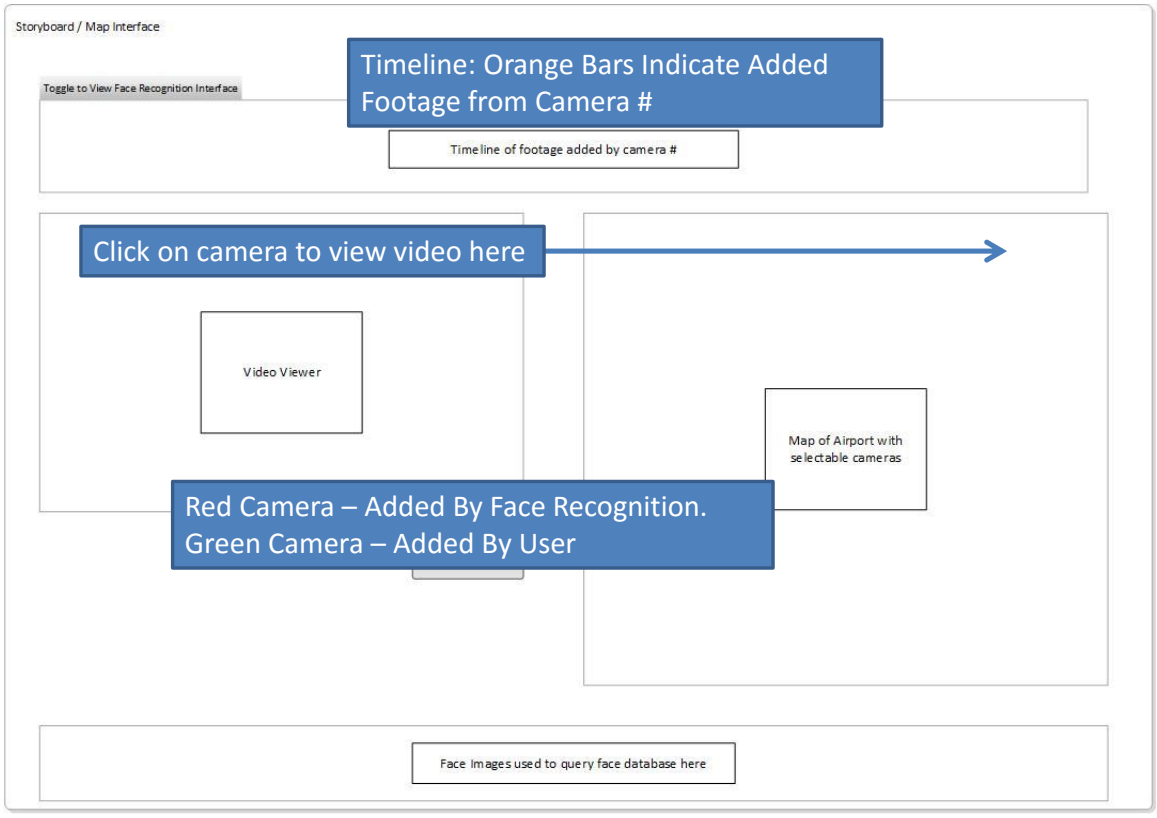

Face Recognition Interface
Initial Search Image Face
Toggle View to Storyboard
Camera 1 – Door
Camera 2 – Interview Room
Camera 3 – Hallway
Initial Search Face Image Search appears here
Video Viewer
List of Cameras
Current Camera View
Adjust search threshold: Low = more potential matches
Threshold:
Low
Medium
High
Search
Add to Storyboard
Face Images used to
Search Face Images (More can be added)

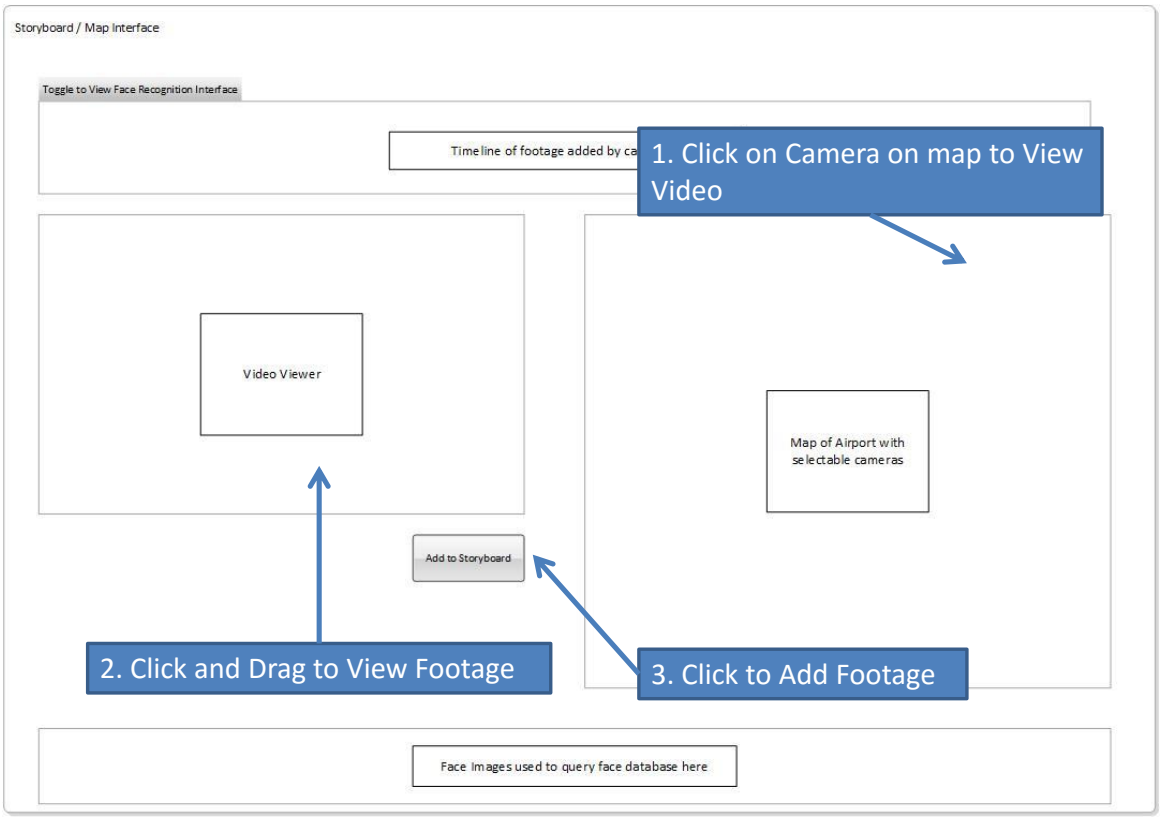

Face Recognition Interface
Toggle View to Storyboard
Video Viewer
Initial Search Face Image Search appears here
Camera 1 – Door
Camera 2 – Interview Room
Camera 3 – Hallway
List of Cameras
1. Click to Search for Matches
2. Click to View Matches
Note: Matches color-coded
-Green = Most Likely
-Red/White = Least Likely
Threshold:
Low
Medium
High
Search
Add to Storyboard
Face Images used to query face database here

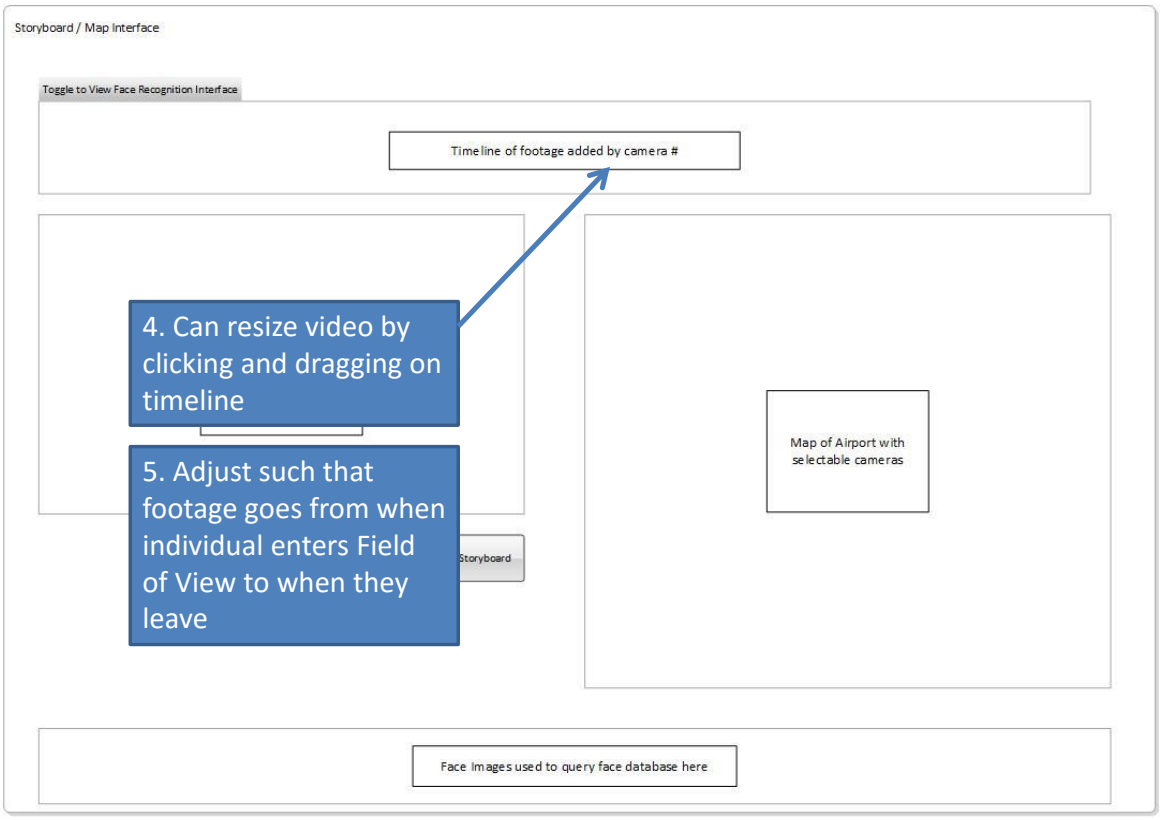

Face Recognition Interface
Toggle View to Storyboard
3. Validate Match by Viewing Video
Initial Search Face Image
Search appears here
Camera 1 – Door
Camera 2 – Interview Room
Camera 3 – Hallway
Video Viewer
List of Cameras
Click & Drag here to play/rewind video
4. If a match, Add to Storyboard
If not a match, take no action
Add to Storyboard
Threshold:
Low
Medium
High
Search
Adding to Storyboard will add faces
that can be used to search


Face Recognition Interface
Toggle View to Storyboard
Initial Search Face Image Search appears here
Camera 1 – Door
Camera 2 – Interview Room
Camera 3 – Hallway
Video Viewer
List of Cameras
5. Continue Searching & Adding Matches
Matches Colored based on Score
High – Green, Medium – Yellow
Red , White – Low
Threshold:
Low
Medium
High
Search
Add to Storyboard
Face Images used to query face database here


Face Recognition Interface
Toggle View to Storyboard
Video Viewer
Initial Search Face Image
Search appears here
Camera 1 – Door
Camera 2 – Interview Room
Camera 3 – Hallway
List of Cameras
If no new matches: Click on
Storyboard for manual entry
Threshold:
Low
Medium
High
Search
Add to Storyboard
Face Images used to query face database here


Storyboard / Map Interface
Toggle to View Face Recognition Interface
Timeline: Orange Bars Indicate Added Footage from Camera #
Timeline of footage added by camera #
Click on camera to view video here
Video Viewer
Map of Airport with selectable cameras
Red Camera – Added By Face Recognition.
Green Camera – Added By User
Face Images used to query face database here


Storyboard / Map Interface
Toggle to View Face Recognition Interface
Timeline of footage added by ca
1. Click on Camera on map to View Video
Video Viewer
Map of Airport with selectable cameras
Add to Storyboard
2. Click and Drag to View Footage
3. Click to Add Footage
Face Images used to query face database here


Storyboard / Map Interface
Toggle to View Face Recognition Interface
Timeline of footage added by camera #
4. Can resize video by clicking and dragging on timeline
5. Adjust such that footage goes from when individual enters Field of View to when they leave
Storyboard
Map of Airport with selectable cameras
Face Images used to query face database here

Appendix C

User Surveys

PRE-STUDY SURVEY

Have you ever used a video surveillance system before?	YES	NO
Have you ever used face recognition before (e.g. Facebook)	YES	NO
If yes, please estimate how long:		
Have you ever been to the International Arrivals Hall of the Ottawa Airport? (Macdonald-Cartier International Airport)	YES	NO
If yes, when was the last time you were there?		
Have you ever gone through Customs at an Airport in Canada	YES	NO
If yes, when was the last time you have done so?		
Have you ever edited or annotated video?	YES	NO

SEARCH AND RETRIEVE POST-STUDY SURVEY

Age:		
Gender:		
Did you watch someone else operate the program prior to use?(Circle Response)	YES	NO
If yes, please estimate how long:		

For each item identified below, circle the number
to the right that best fits your judgment of its quality.
Use the rating scale to select the quality number.

Item	Scale				
	Poor	Good			Excellent
1. Overall, I am satisfied with how easy it is to use this system	1	2	3	4	5
2. It was simple to use this system	1	2	3	4	5
3. I can effectively complete the video tracking task using this system	1	2	3	4	5
4. I am able to complete the task quickly using this system	1	2	3	4	5
5. I am able to efficiently complete the task using this system	1	2	3	4	5
6. I feel comfortable using this system	1	2	3	4	5
7. It was easy to learn to use this system	1	2	3	4	5
8. I believe I became productive quickly using this system	1	2	3	4	5
9. Whenever I make a mistake using the system, I recover easily and quickly	1	2	3	4	5
10. The organization of information on the system screens is clear	1	2	3	4	5
11. The interface of this system is pleasant	1	2	3	4	5
12. I like using the interface of this system	1	2	3	4	5
13. This system has all the functions and capabilities I expect it to have	1	2	3	4	5
14. Overall, I am satisfied with this system	1	2	3	4	5

Comments: