Chapter 14
Face Recognition at a Distance

Frederick W. Wheeler, Xiaoming Liu, and Peter H. Tu

14.1 Introduction

Face recognition, and biometric recognition in general, have made great advances in the past decade. Still, the vast majority of practical biometric recognition applications involve cooperative subjects at close range. Face Recognition at a Distance (FRAD) has grown out of the desire to automatically recognize people out in the open, and without their direct cooperation. The face is the most viable biometric for recognition at a distance. It is both openly visible and readily imaged from a distance. For security or covert applications, facial imaging can be achieved without the knowledge of the subject. There is great interest in iris at a distance, however it is doubtful that iris will outperform face with comparable system complexity and cost. Gait information can also be acquired over large distances, but face will likely continue to be a more discriminating identifier.

In this chapter, we will review the primary driving applications for FRAD and the challenges still faced. We will discuss potential solutions to these challenges and review relevant research literature. Finally, we will present a few specific activities to advance FRAD capabilities and discuss expected future trends. For the most part, we will focus our attention on issues that are unique to FRAD. Some of the main challenges of FRAD are shared by many other face recognition applications, and are thoroughly covered in other dedicated chapters of this book.

Distance itself is not really the fundamental motivating factor for FRAD. The real motivation is to work over large coverage areas without subject cooperation.

F.W. Wheeler (✉) · X. Liu · P.H. Tu
Visualization and Computer Vision Lab, GE Global Research, Niskayuna, NY 12309, USA
e-mail: wheeler@ge.com

X. Liu
e-mail: liux@ge.com

P.H. Tu
e-mail: tu@ge.com

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_14,

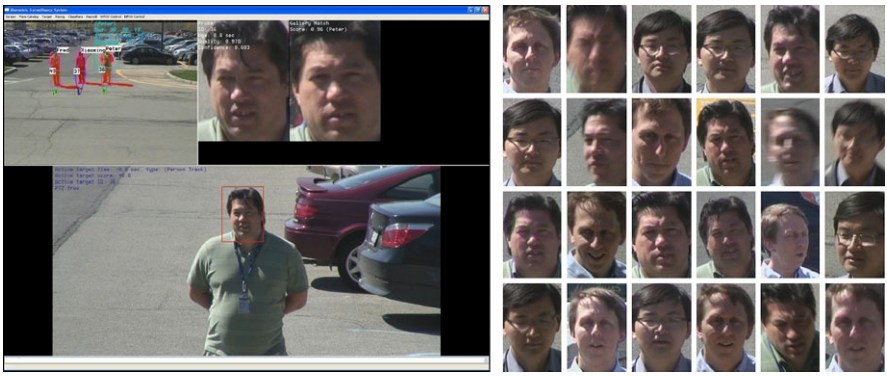

Fig. 14.1 On the left, a face recognition at a distance application showing tracked and identified subjects in wide field-of-view video (*upper left*), high-resolution narrow field-of-view video from an automatically controlled PTZ camera (*bottom*), and a detected and recognized facial image (*upper right*). On the right, some of the facial images captured by the system over a few minutes, selected to show the variation in facial image quality

The nature of the activity of subjects and the size of the coverage area can vary considerably with the application and this impacts the degree of difficulty. Subjects may be sparse and standing or walking along predictable trajectories, or they may be crowded, moving in a chaotic manner, and occluding each other. The coverage area may range from a few square meters at a doorway or choke point, to a transportation terminal, building perimeter, city block, or beyond. Practical solutions do involve image capture from a distance, but the field might be more accurately called *face recognition of noncooperative subjects over a wide area*. Figure 14.1 shows a FRAD system operating in a parking lot. There are two primary difficulties faced by FRAD. First, acquiring facial images from a distance. Second, recognizing the person in spite of imperfections in the captured data.

There are a wide variety of commercial, security, defense and marketing applications of FRAD. Some of the most important potential applications include:

- Access control: Unlock doors when cleared persons approach.
- Watch-list recognition: Raise an alert when a person of interest, such as a known terrorist, local offender or disgruntled ex-employee is detected in the vicinity.
- White-list recognition: Raise an alert whenever a person not cleared for the area is detected.
- Rerecognition: Recognize people recently imaged by a nearby camera for automatic surveillance with long-range persistent tracking.
- Event logging: For each person entering a region, catalog the best facial image.
- Marketing: Understand long-term store customer activities and behavior.

The *Handbook of Remote Biometrics* [53] also contains chapters on FRAD. The focus in that book is somewhat complementary, covering system issues and a more detailed look at illumination levels, optics and image sensors for face imaging at distances up to the 100–300 m range and beyond with both theoretical analysis and practical design advice.

14.1.1 Primary Challenges

In the ideal imaging conditions for 2D face recognition, the subject is illuminated in a uniform manner, is facing a color camera with a neutral expression and the image has a resolution with 200 or more pixels eye-to-eye. These conditions are easily achieved with a cooperative subject at close range.

With FRAD, the subject is by definition not at close range, but perhaps more importantly, the level of cooperation is reduced. Applications of FRAD for cooperative subjects are at best unusual and rare. In typical FRAD applications, subjects are not cooperative, and this is the scenario that is assumed in most research work on FRAD. Noncooperative subjects may be either unaware that facial images are being collected, or aware but unconcerned, perhaps due to acclimation. That is, they are neither actively cooperating with the system, nor trying to evade the system.

A much more challenging situation occurs when subjects are actively evasive. A subject may attempt to evade face capture and recognition by obscuring their face with a hat, glasses or other adornments, or by deliberately looking away from cameras or downward. In such situations it might still be beneficial for a system to automatically determine that the subject is evasive.

In a sense, FRAD is not a specific core technology or basic research problem. It can be viewed as an application and a system design problem. Some of the challenges in that design are specific to FRAD, but many are broader face recognition challenges that are discussed and addressed throughout this book. The main challenges of FRAD are concerned with the capture of facial images that have the best quality possibly, and with processing and face recognition that is robust to the remaining imperfections. These challenges can be organized into a few categories, which we discuss below.

The first challenge of FRAD is simply acquiring facial images for subjects who may be 10–30 m or more away from the sensor. Some of the optics issues to consider are lens parameters, exposure time, and the effect on the image when any compromise is made.

14.1.2 Optics and Light Intensity

As subject distance increases, a primary issue is the selection or adjustment of the camera lens to maintain field of view and image intensity. As the distance from the camera to the subject is increased the focal length of the camera lens must be increased proportionally if we are to maintain the same field of view, or image sampling resolution. That is, if the subject distance is doubled, then the focal length, F, must be doubled to maintain a facial image size of, say, 200 pixels eye-to-eye.

The light intensity a lens delivers to the image sensor is proportional to the f-number. The f-number, N, of a lens is the focal length divided by the diameter of the entrance pupil, D, or $N = F/D$. To maintain image brightness, and thus contrast and signal to noise ratio, the f-number must be maintained. So, if the subject

distance is doubled and the focal length is doubled, then the f-number of that particular lens must be maintained by doubling the pupil diameter. Of course, an image sensor with greater sensitivity may also be used to compensate for a reduction in light intensity from the lens.

If the pupil diameter for a lens is already restricted with an adjustable aperture stop, then increasing the pupil diameter to maintain the f-number is simply a matter of adjusting that setting. Adjustments to the pupil diameter are usually described in terms of the resulting f-number, and are typically called f-stops. The f-stops are defined as the f-number at the image when the lens is focused at infinity or very far away.

However, as subject distance is increased, eventually an adjustable aperture will be fully open and the pupil aperture will be limited by the size of the lens itself. So, imaging faces well at larger distances generally requires larger lenses, which have larger glass elements, are heavier and more expensive. Another drawback to increasing the diameter of the lens is a reduction in depth of field (DOF), the range over which objects are well focused. However, DOF is inherently larger at larger object distances, so this is often less of a concern for FRAD.

14.1.3 Exposure Time and Blur

There are a number of different types of image distortion that can be of concern when capturing faces at a distance. If an appropriate lens is selected, then a facial image captured at a distance can be as bright as an image captured at close range. However, this is not always the situation. Lenses with large diameters are expensive or simply may not be in place. When the lens does not have a low enough f-number (large enough aperture relative to the focal length), the amount of light reaching the image sensor will be too low and the image SNR will be reduced. If the image is amplified to compensate, sensor noise will be amplified as well and the resulting image will be noisy. Without amplification, the image will be dark.

If the light intensity at the sensor is too low, the exposure time for each image can be increased to compensate. However, this introduces a trade-off with motion blur. The subjects being imaged are generally in motion. If the exposure time is long enough that the motion of the subjects is significant during the exposure, then some degree of blurring will occur.

FRAD systems often utilize active pan-tilt camera control. Mechanical vibration of the camera can be significant in such systems and is another source of blur. At very long distances, atmospheric distortion and haze can also contribute to image distortion.

14.1.4 Image Resolution

In FRAD applications that lack an optimal optical system to provide an ideal image to the sensor, resolution of the resulting facial image can be low. In some cases,

it may be desired to recognize people in video from a stationary camera where facial image resolution is low due to subject distance. An active camera system with automatic pan, tilt and zoom may simply reach its capture distance limit, but one may still want to recognize people at greater distances. No matter how the optical system is designed, there is always some further desired subject distance and in these cases facial image resolution will be reduced. Facial recognition systems that deal with low-resolution facial images are certainly desirable.

14.1.5 Pose, Illumination and Expression

The Pose, Illumination and Expression (PIE) challenges are not unique to FRAD. There are many other face recognition applications that share these challenges. The PIE challenges are, however, somewhat customized and more pronounced in FRAD.

In many FRAD applications, it is desirable to mount cameras well above people's heads, as is done for most ordinary security cameras. This allows for the imaging of people's faces over a wider area with less occlusion. A disadvantage is that the viewing angle of faces has a slight downward tilt, often called the "surveillance perspective."

The pan angle (left-right) of faces in FRAD applications can in the worst cases be completely arbitrary. In open areas where there are no regular travel directions, this will be the case. Corridors and choke points are more favorable situations, generally limiting the directions in which people are facing. People tend to face the direction of their travel. When faces can be oriented in many directions, the use of a distributed set of active pan-tilt-zoom cameras can help [25]. Still, the variation of facial capture pan angle can be high.

There is some hope for this inherent pose problem with FRAD, and it is observation time. A wide-area active camera system may be able to observe and track a walking subject for 5–10 seconds or more. A stationary or loitering person may be observed for a much longer period of time. A persistent active face capture system, may eventually opportunistically capture a facial image of any particular subject with a nearly straight-on pose angle.

Most FRAD applications are deployed outdoors with illumination conditions that are perhaps the most challenging. Illumination is typically from sunlight or distributed light fixtures. The direction and intensity of the sunlight will change with the time of day, the weather and the seasons. Over the capture region there may be large objects such as trees and buildings that both block and reflect light, and alter the color of ambient light, increasing the variation of illumination over the area.

Subjects who are not trying to evade the system and who are not engaged in conversation will for the most part have a neutral expression. Of the PIE set of challenges, the expression issue is generally less of a concern for FRAD.

14.1.6 Approaches

There are two basic approaches to FRAD: high-definition stationary cameras, and active camera systems. FRAD generally means face recognition not simply at a distance, but over a wide area. Unfortunately, a wide camera viewing area that captures the entire coverage area results in low image resolution. Conversely, a highly zoomed camera that yields high-resolution facial images has a narrow field of view.

14.1.6.1 High-Definition Stationary Camera

If a FRAD capture sensor is to operate over a 20 m wide area with a single camera and we require 100 pixels across captured faces, then we would need a camera with about 15 000 pixels of horizontal resolution. Assuming an ordinary sensor aspect ratio, this is a 125 *mega*pixel sensor and not currently practical. If we were to use a high-definition 1080 by 1920 pixel camera to image the full 20 m wide area, a face would be imaged with a resolution of about 13 by 7 pixels.

If the coverage region is not too large, say 2 m across, then a single stationary high-definition 1080 by 1920 camera could image faces in the region with about 100 pixels eye-to-eye. This is not very high resolution for face recognition, but may be sufficient for verification or low-risk applications. If the desired coverage area grows, and stationary cameras are still used, then the camera resolution would have to increase, or multiple cameras would be required.

14.1.6.2 Active-Vision Systems

FRAD is more often addressed with a multi-camera system where one or more Wide field Of view (WFOV) cameras view a large area with low-resolution and one or more Narrow Field Of View (NFOV) cameras are actively controlled to image faces with high resolution using pan, tilt and zoom (PTZ) commands. Through the use of face detection or person detection, and possibly tracking, the location of people is determined from the WFOV video. The NFOV are targeted to detected or tracked people through pan and tilt control, and possibly also adaptive zoom control. The WFOV and NFOV cameras are sometimes called the master and slave cameras, and NFOV cameras are often simply called PTZ cameras.

This is also often described as a “foveated imaging” or “focus-of-attention” approach and it somewhat mimics the human visual system, where a wide angular range is monitored with relatively low resolution and the eyes are actively directed toward areas of interest for more detailed resolution. This is especially the case when the WFOV and NFOV cameras are co-located.

What we describe here is a prototypical approach. There are of course many possible modifications and improvements. A single camera may be used with a system that enables switching between WFOV and NFOV lenses, with the additional challenge that wide-field video coverage will not be continuous. A low-zoom WFOV

camera may not be stationary, but could instead pan and tilt with a high-zoom NFOV camera, more like the human eye or a finderscope. Instead of using a WFOV camera, a single NFOV camera could continuously scan a wide area by following a pan and tilt angle pattern. One could use many WFOV cameras and many NFOV cameras in a single cooperative network. Clearly, there are many options for improving upon the prototypical approach, with various advantages and disadvantages.

When multiple subjects are present a multi-camera system must decide somehow how to schedule its NFOV camera or cameras. It needs to determine when to point the camera at each subject. There has been considerable research effort put into this NFOV resource allocation and scheduling problem, which becomes more complicated as more NFOV cameras are utilized. Some of the factors that a scheduling algorithm may account for include: which subjects are facing one of the NFOV cameras, the number of times each subject's face has been captured thus far, the quality and resolution of those images, the direction of travel and speed of each subject, and perhaps the specific location of each subject. An NFOV target scheduling algorithm accounts for some desired set of factors such as these and determines when and where to direct the NFOV cameras using their pan, tilt and zoom controls.

14.1.7 Literature Review

14.1.7.1 Databases

Most test databases for face recognition contain images or video captured at close range with cooperative subjects. They are thus best suited for training and testing face recognition for access control applications. However, there are a few datasets that are more suitable for face recognition at a distance development and evaluation.

The database collected at the University of Texas at Dallas (UTD) for the DARPA Human ID program [40] includes close-up still images and video of subjects and also video of persons walking toward a still camera from distances of up to 13.6 m and video of persons talking and gesturing from approximately 8 m. The collection was performed indoors, but in a large open area with one wall made entirely of glass, approximating outdoor lighting conditions. A fairly low zoom factor was used in this collection.

Yao et al. [58] describe the University of Tennessee, Knoxville Long Range High Magnification (UTK-LRHM) face database of moderately cooperative subjects at distances between 10 m and 20 m indoors, and extremely long distances between 50 m and 300 m outdoors. Indoor zoom factors are between 3 and 20, and outdoor zoom factors range up to 284. Imaging at such extremes can result in distortion due to air temperature and pressure gradients, and the optical system used exhibits additional blur at such magnifications.

The NIST Multiple Biometric Grand Challenge (MBGC) is focused on face and iris recognition with both still images and video and has sponsored a series of challenge problems. In support of the unconstrained face recognition challenges, this

program has collected high-definition and standard definition outdoor video of subjects walking toward the camera and standing at ranges of up to about 10 m. Since subjects were walking toward the camera, frontal views of their faces were usually visible. MBGC is also making use of the DARPA Human ID data described above [39].

It is important to remember that as distance increases, people also have increased difficulty in recognizing faces. Face recognition results from Pittsburgh Pattern Recognition on MBGC uncontrolled video datasets (of the subjects walking toward the camera) are comparable to and in some cases superior to face recognition results by humans on the same data [39].

Each of these databases captures images or video with stationary cameras. FRAD sensors are generally real-time actively controlled camera systems. Such hardware systems are difficult to test offline. The evaluation of active camera systems with shared or standardized datasets is not feasible because real-time software and hardware integration aspects are not modeled. Components of these systems, such as face detection, person detection, tracking and face recognition itself can be tested in isolation on appropriate shared datasets. But interactions between the software and hardware components can only be fully tested on live action scenes. Virtual environments can also be used to test many aspects of an active-vision system [43–45] (Sect. 14.1.7.3).

14.1.7.2 Active-Vision Systems

There have been a great many innovations and systems developed for wide-area person detection and tracking to control NFOV cameras to capture facial images at a distance. We review a selected group of publications in this section, in approximate chronological order. A few of the systems described here couple face capture to face recognition. All are motivated by this possibility.

In some very early work in this area, Stillman et al. [52] developed an active camera system for person identification using two WFOV cameras and two NFOV cameras. This real-time system worked under some restricted conditions, but over a range of several meters, detected people based on skin color, triangulated 3D locations, and pointed NFOV cameras at faces. A commercial face recognition system then identified the individuals. Interestingly, the motivation for this effort, at its time in history, was to make an intelligent computing environment more aware of people present in order to improve the interaction. No mention is made of security.

Greiffenhagen et al. [21] describe a dual camera face capture system where the WFOV camera is an overhead omnidirectional camera and the NFOV has pan, tilt and zoom control. The authors use a systematic engineering methodology in the design of this real-time system, and perform a careful statistical characterization of components of the system so that measurement uncertainty can be carried through and a face capture probability of 0.99 can be guaranteed. While face recognition was not applied to the captured images, they are of sufficient quality for recognition. The system handles multiple subjects as long as the probability of occlusion is small.

Zhou et al. [64] have developed their Distant Human Identification (DHID) system to collect biometric information of humans at a distance, for face recognition and gait. A single WFOV camera has a 60° field of view and enables tracking of persons out to a distance of 50 m. A combination of background subtraction, temporal differencing, optical flow and color-based blob detection is used for person detection and tracking. The system aims to capture short zoomed-in video sequences for gait recognition and relatively high-resolution facial images. Person detections from the WFOV video are used to target the NFOV camera initially, and then the NFOV tracks the subject based only on the NFOV video data.

Marchesotti et al. [34] have also developed a two-camera face capture at a distance system. Persons are detected and tracked using a blob detector in the WFOV video, and an NFOV camera is panned and tilted to acquire short video clips of subject faces.

A *face cataloger* system has been developed and described by Hampapur et al. [22]. This system uses two widely separated WFOV cameras with overlapping views of a 20 ft. by 19 ft. lab space. To detect persons, a 2D multi-blob tracker is applied to the video from each WFOV camera and these outputs are combined by a 3D multi-blob tracker to determine 3D head locations in a calibrated common coordinate system. An active camera manager then directs the two pan-tilt NFOV cameras to capture facial images. In this system, a more aggressive zoom factor is used when subjects are moving more slowly. The authors experimentally demonstrate a trade-off between NFOV zoom and the probability of successfully capturing a facial image. When the NFOV zoom factor is higher, the required pointing accuracy is greater, so there is a higher likelihood of missing the subject. This would be a general trend with all active camera systems. Later work described by Senior et al. [51] simplified and advanced the calibration procedure and was demonstrated outdoors.

Bagdanov et al. [2] have developed a method for capturing facial images over a wide area with a single active pan-tilt camera. In this approach, reinforcement learning is employed to discover the camera control actions that maximize the chance of acquiring a frontal face image given the current appearance of object motion as seen from the camera in its home position. Essentially, the system monitors object motion with the camera in its home position and over time learns *if*, *when* and *where* to zoom in for a facial image. A frontal face detector applied after each attempt guides the learning. Some benefits are that this approach uses a single camera and requires no camera calibration.

Prince [41, 42], Elder [19] et al. have developed a system that addresses collection and pose challenges. They make use of a foveated sensor using a stationary camera with a 135° field of view and a foveal camera with a 13° field of view. Faces are detected via the *stationary* camera video using motion detection, background modeling and skin detection. A pan and tilt controller directs the *foveal* camera to detected faces. Since the system detects people based on face detection it naturally handles partially occluded and stationary people.

Davis et al. [16, 17] have developed methods to automatically scan a wide area and detect persons with a single PTZ camera. Their approach detects human behavior and learns the frequency with which humans appear across the entire coverage

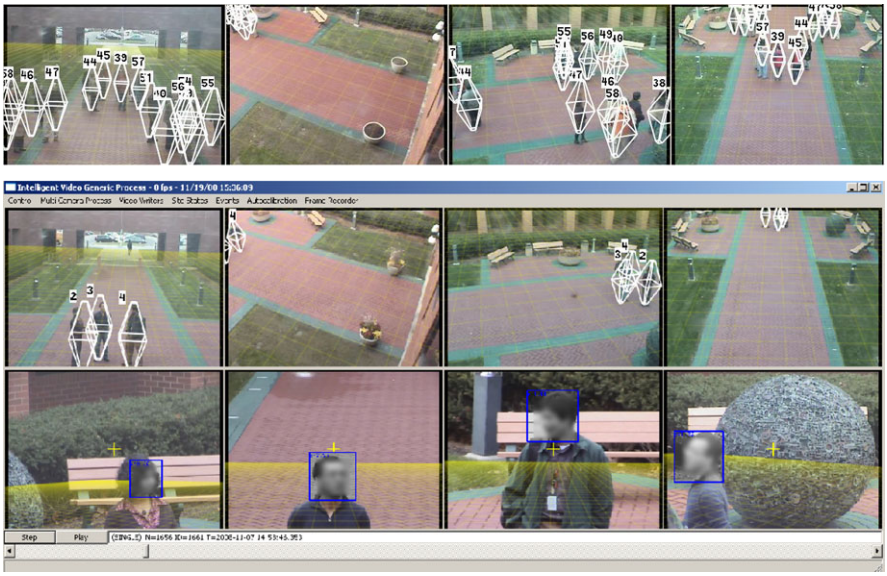

Fig. 14.2 Multi-camera person tracking with crowding (*above*) and person tracking and active NFOV face capture (*below*) (© 2009 IEEE, used with permission [61])

region so that scanning is done efficiently. Such a strategy might be used to create a one-camera active face capture system, or to greatly increase the coverage area if used as a WFOV camera.

Krahnstoever et al. [25] have developed a face capture at a distance framework and prototype system. Four fixed cameras with overlapping viewpoints are used to pervasively track multiple subjects in a 10 m by 30 m region. Tracking is done in a real-world coordinate frame, which drives the targeting and control of four separate PTZ cameras that surround the monitored region. The PTZ cameras are scheduled and controlled to capture high-resolution facial images, which are then associated with tracker IDs. An optimization procedure schedules target assignments for the PTZ cameras with the goal of maximizing the number of facial images captured while maximizing facial image quality. This calculation is based on the apparent subject pose angle and distance. This opportunistic strategy tends to capture facial images when subjects are facing one of the PTZ cameras.

Bellotto et al. [5] describe an architecture for active multi-camera surveillance and face capture where trackers associated with each camera, and high-level reasoning algorithms communicate via an SQL database. Information from persons detected by WFOV trackers can be used to assign actively controlled NFOV cameras to particular subjects. Once NFOV cameras are viewing a face, the face is tracked and the NFOV camera follows the face with a velocity control system.

In Yu et al. [61], the authors have used this system [25] to monitor groups of people over time, associate an identity with each tracked person and record the degree of close interaction between identified individuals. Figure 14.2 shows person tracking and face capture from this system. This allows for the construction of a social

network that captures the interactions, relationships and leadership structure of the subjects under surveillance.

14.1.7.3 NFOV Resource Allocation

Face capture with active cameras is faced with the problem of resource allocation. Given a limited number of NFOV cameras and a large number of potential targets, it becomes necessary to predict feasible periods of time in the future, during which a target could be captured by a NFOV camera at the desired resolution and pose, followed by scheduling the NFOV cameras based on these feasible temporal windows. Lim et al. [27, 28] address the former problem by constructing what is known as a "Task Visibility Interval" that encapsulates the required information. For the latter, these authors then utilize these "Task Visibility Intervals" to schedule NFOV camera assignments.

Bimbo and Pernici [6] have addressed the NFOV scheduling problem for capturing face images with an active camera network. They formulate the problem as a Kinetic Traveling Salesman Problem (KTSP) to determine how to acquire as many targets as possible.

A variety of NFOV scheduling policies have been developed and evaluated by Costello et al. [15] as well.

Qureshi and Terzopoulos [43–45] have developed an extensive virtual environment simulator for a large train station with behaviorally realistic autonomous pedestrians who move about without colliding, and carry out tasks such as waiting in line, buying tickets, purchasing food and drinks, waiting for trains and proceeding to the concourse area. The video-rendering engine handles occlusions, and models camera jitter and imperfect color response. The purpose is to develop and test active camera control and scheduling systems with many WFOV cameras and many NFOV cameras on a scale where real-world experiments would be prohibitively expensive. With such a simulator, visual appearance will never be perfect. However, this system allows the setup and evaluation of person tracking tasks and camera scheduling algorithms with dozens of subjects and cameras over a very large area with perfect ground truth. Then, the exact same scenario can be executed again with a change in any algorithm or aspect of the camera set-up.

14.1.7.4 Very Long Distances

Yao et al. [58, 60] have explored face recognition at considerable distances, using their UTK-LRHM face database. For indoor data, with a gallery of 55 persons and a commercial face recognition system, they show a decline in recognition rate from 65.5% to 47.3% as the zoom factor goes from 1 to 20 and the subject distance is increased to maintain an eye-to-eye image resolution of 60 pixels. It is also shown that the recognition rate at a zoom factor of 20 can be raised back up to 65.5% with wavelet-based deblurring.

Yao et al. [59, 60] have used a super-resolution approach based on frequency domain registration and cubic-spline interpolation on facial images from the UTK-LRHM face database [58] and found considerable benefit in some circumstances. Super-resolution seems most effective when facial images begin at a low-resolution. For facial images with about 35 pixels eye-to-eye, super-resolution increased the recognition rate from 10% to 30% with a 55-person gallery. Super-resolution followed by unsharp masking further increased recognition rate to 38% and yielded a cumulative match characteristic performance almost as high as when optical zoom alone was used to double the facial image resolution.

14.1.7.5 3D Imaging

Most 3D face capture systems use the stereo or structured light approach [9]. Stereo capture systems use two cameras with a known geometric relationship. The distance to feature points detected in each camera's image is then found via triangulation. Structured light systems use a light projector and a camera, also with a known geometric relationship. The light pattern is detected in the camera's image and 3D points are determined. Each system is characterized by a *baseline* distance, between the stereo cameras or between the light projector and camera. With either approach, the accuracy of the triangulated 3D data degrades with subject distance if the baseline distance is held constant. To maintain 3D reconstruction accuracy as subject distance increases, the baseline distance must be increased proportionally. This prohibits a physically compact system and is a fundamental challenge to 3D face capture at a distance with these methods. However, there are some newer systems under development that are overcoming the baseline challenge for 3D face capture at a distance. Below we review a few 3D face capture systems designed to operate at large distances.

Medioni et al. [35–37] have addressed FRAD for noncooperative individuals with a single camera approach and 3D face reconstruction. They propose a system where an ultra-high resolution 3048 by 4560 pixel camera is used by switching readout modes. Bandwidth limitations generally prevent the readout of the full resolution of such cameras at 30 Hz. However, full-frame low-resolution fast-frame-rate readouts can be used for person detection and tracking and partial-frame high-resolution readouts can be used to acquire a series of facial images of a detected and tracked person. Person detection is accomplished without background modeling, using an edgelet feature-based detector. This work emphasizes the 3D reconstruction of faces with 100 pixels eye-to-eye using shape from motion on data acquired with a prototype of the envisioned system. 3D reconstructions are performed at distances of up to 9 m. Though current experiments show 2D face recognition outperforming 3D, the information may be fused, or the 3D data may enable pose correction.

Rara et al. [46, 47] acquire 3D facial shape information at distances up to 33 m using a stereo camera pair with a baseline of 1.76 m. An Active Appearance Model localizes facial landmarks from each view and triangulation yields 3D landmark positions. The authors can achieve a 100% recognition rate at 15 m, though the gallery

size is 30 subjects and the collection environment is cooperative and controlled. It is noted that depth information at such long distances with this modest baseline can be quite noisy and may not significantly contribute to recognition accuracy.

Redman et al. [48] and colleagues at Lockheed Martin Coherent Technologies have developed a 3D face imaging system for biometrics using Fourier Transform Profilometry. This involves projecting a sinusoidal fringe pattern onto the subject's face using short eye-safe laser bursts and imaging the illuminated subject with a camera that is offset laterally from the light source. Fourier domain processing of the image can recover a detailed 3D image. In a sense, this falls into the class of structured light approaches, but with a small baseline requirement. A current test system is reported to capture a 3D facial image at 20 m subject distance with a range error standard deviation of about 0.5 mm and a baseline distance of only 1.1 m.

Redman et al. [49] have also developed 3D face imaging systems based on digital holography, with both multiple-source and multiple wavelength configurations. With multiple-wavelength holography, a subject is imaged two or more times, each time illuminated with a laser tuned to a different wavelength, in the vicinity of 1617 nm for this system. The laser illumination is split to create a reference beam, which is mixed with the received beam. The interference between these beams is the hologram that is imaged by the sensor. The holograms at each wavelength are processed to generate the 3D image. The multi-wavelength holographic system has been shown to capture a 3D facial image at a 100 m subject distance with a range error of about 1–2 mm, though this has been performed in a lab setting and not with live subjects. With this approach there is zero baseline distance. The only dependence of accuracy on subject distance is atmospheric transmission loss.

Andersen et al. [1] have also developed a 3D laser radar and applied it to 3D facial image capture. This approach uses a time-of-flight strategy to range measurement, with a rapidly pulsed (32.4 kHz) green nD:YAG laser and precisely timed camera shutter. 50–100 reflectivity images are captured and processed to produce a 3D image. This system has been used to capture 3D facial images at distances up to 485 m. Each 3D image capture takes a few seconds. Though at this stage not many samples have been collected, the RMS range error is about 2 mm at 100 m subject distance and about 5 mm at 485 m subject distance. Atmospheric turbulence, vibrations and system errors are the factors that limit the range of this system.

The Fourier Transform Profilometry and Digital Holography approaches operate at large distance, but do not naturally handle a large capture region. Coupled with a WFOV video camera and person detection and tracking system, these systems could be used to capture 3D facial images over a wide area.

14.1.7.6 Face and Gait Fusion

For recognition at a distance, face and gait are a natural pair. In most situations, a sensor used for FRAD will also be acquiring video suitable for gait analysis. Liu et al. [31] exploit this and develop fusion algorithms to show a significant improvement in verification performance by using multi-modal fusion gait and face recognition with facial images collected outdoors at a modest standoff distance.

Fig. 14.3 The Biometric Surveillance System, a portable test and demonstration system on a wheeled cart with two raised camera nodes (*left*), and a close-up view of one node (*right*)

Zhou et al. [62, 63] have recognized that the best gait information comes from profile views, and have thus focused on fusing gait information with profile facial images. In initial work [62], face recognition was done using curvature features of just the face profile. Later efforts [63] use the whole side-view of the face and also enhanced the resolution of the side-view using multi-frame super-resolution, motivated to use all available information. For 45 subjects imaged at a distance of 10 ft, the best recognition rates are 73.3% for single-frame face, 91.1% for multi-frame enhanced face, 93.3% for gait, and 97.8% for fused face and gait. Facial profile super-resolution gives a considerable improvement, as does fusion.

14.2 Face Capture at a Distance

GE Global Research and Lockheed Martin have developed a FRAD system called the Biometric Surveillance System [56]. The system features reliable ground-plane tracking of subjects, predictive targeting, a target priority scoring system, interfaces to multiple commercial face recognition systems, many configurable operating modes, an auto-enrollment mechanism, and network-based sharing of auto-enrollment data for re-identification. Information about tracking, target scoring, target selection, target status, attempted recognition, successful recognition, and enrollments are displayed in a highly animated user interface (Fig. 14.1 on page 354).

The system uses one or more networked nodes where each node has a co-located WFOV and NFOV camera (Fig. 14.3). Each camera is a Sony EVI-HD1, which features several video resolution and format modes and integrated pan, tilt, zoom and focus, all controllable via a VISCA™ serial interface. The WFOV camera is operated in NTSC video mode, producing 640 by 480 video frames at 30 Hz. The pan, tilt and zoom settings of the WFOV camera are held fixed. The NFOV camera is configured for 1280 by 720 resolution video at 30 Hz, and its pan, tilt and zoom setting are actively controlled. Matrox® frame grabbers are used to transfer video streams to a high-end but standard workstation.

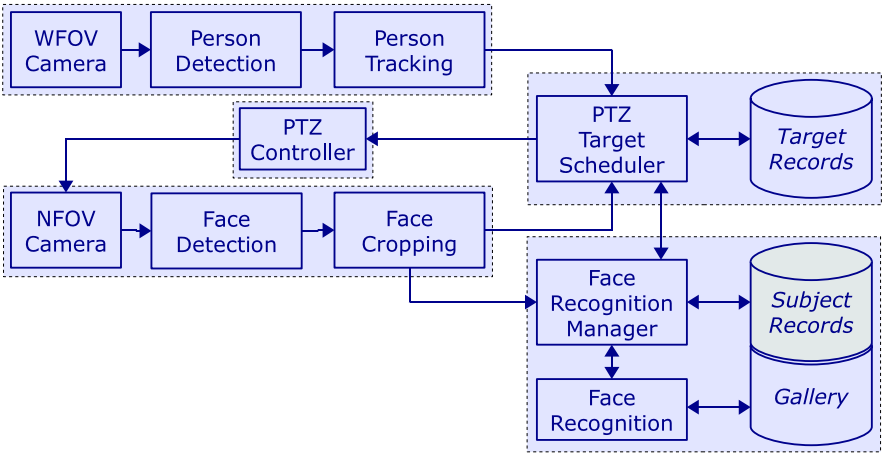

Fig. 14.4 System diagram showing main computational components of the Biometric Surveillance System

A system diagram is shown in Fig. 14.4. The stationary WFOV camera is used to detect and track people in its field of view. The WFOV camera is calibrated to determine its internal and external parameters, which include the focal length, principal point, location, and orientation. This defines a mapping between real-world metric coordinates and the WFOV camera image. Since the camera is stationary, a background subtraction approach is used for moving object detection. The variation of each color component of each pixel is learned and adapted using a non-parametric distribution. Grayscale imagery may be used as well, but color increases detection rate. Whenever a pixel does not match this model, it is declared a foreground pixel. From the camera calibration information and the assumption that people are walking upright on the ground plane, feasible sizes and locations of persons within the image plane are established. Blobs of foreground pixels that conform to these feasible sizes are detected persons. A ground-plane tracker based on an extended Kalman filter is applied to detected persons [7, 24]. The use of the Kalman filter makes the tracker robust to intermittent occlusions and provides the velocity, travel direction and predicted locations of subjects [54].

The automatically controlled NFOV camera is also calibrated with respect to the real-world coordinate system, when it is in its home position with its pan and tilt angles at 0° and its zoom factor set to 1. Further calibration of the NFOV camera determines how pan, tilt and zoom settings affect its field of view. An important part of this calibration is the camera's *zoom point*, or the pixel location that always points to the same real-world point as the zoom factor is changed. The zoom point is not necessarily the exact center of the image, and even a small offset can affect targeting accuracy when a high zoom is used for distant subjects. Effectively, this collective calibration data allows for the specification of a region in the WFOV image, and determines the pan, tilt and zoom settings to make that region the full image for the NFOV camera.

Table 14.1 The factor and clipping range used for each parameter to score targets

Parameter	Factor	Clipping Range
Direction cosine	10	$[-8, 8]$
Speed (m/s)	10	$[0, 20]$
Capture attempts	-2	$[-5, 0]$
Face captures	-1	$[-5, 0]$
Times recognized	-5	$[-15, 0]$

14.2.1 Target Selection

When multiple persons are present, the system must determine which subject to target for high-resolution face capture. From the WFOV person tracker, it is straightforward to determine the distance to a subject, the degree to which a subject is facing (or at least moving toward) the cameras, and the speed of the subject. Further, because the person tracker is generally quite reliable, a record can be kept for each tracked subject. This subject record includes the number of times we have targeted the subject, the number of times we have successfully captured a facial image and the number of times the subject has been successfully identified by the face recognition algorithm. All of this information is used by the target selection mechanism.

Detected and tracked persons are selected for high-resolution facial capture based on a priority scoring mechanism. A score is produced for each tracked subject, and the subject with the highest score is selected as the next target. Several parameters are used in the scoring process, and for each parameter, a multiplicative factor is applied and the result is clipped to a certain range. For example, the subject's speed in m/s is multiplied by the factor 10.0, clipped to the range $[0, 20]$ and added to the score. Table 14.1 shows the complete set of parameters and factors currently in use, though not yet optimized. The direction cosine parameter is the cosine of the angle between the subject's direction of travel and the line from the subject to the NFOV camera. This parameter indicates the degree to which the subject is facing the NFOV camera. The net overall effect of this process is to favor subjects moving more quickly toward the cameras who have not yet been satisfactorily imaged. In practice, a target selection strategy like this causes the system to move from subject to subject, with a tendency to target subjects from which we are most likely to get new and useful facial images.

When a subject is selected, the system uses the Kalman filter tracker to predict the location of the subject's face at a specific target time about 0.5–1.0 s in the future. The NFOV camera will point to this location and hold until the target time has passed. This gives the camera time to complete the pan and tilt change, and time for vibration to settle. Facial images are captured when the subject moves through the narrow field-of-view as predicted. We have already discussed the trade-off between zoom factor and probability of successfully capturing a facial image. This system uses an adaptive approach. If there have been no face captures for the subject, then the initial face resolution goal will be a modest 30 pixels eye-to-eye. However, each

time a facial image is successfully captured at a particular resolution, the resolution goal is increased by 20%. Each subject tends to be targeted and imaged many times by the system, so the facial image resolution goal rapidly increases. For a particular target, this resolution goal and the subject distance determines the zoom factor of the NFOV camera. The subject distance is also used to set the focus distance of the NFOV camera.

14.2.2 Recognition

The NFOV video is processed on a per-frame basis. In each frame, the Pittsburgh Pattern Recognition FT SDK is utilized to detect faces. If there is more than one detection, we use only the most central face in the image, since it is more likely to be the face of the targeted subject. A detected face is cropped from the full frame image and passed to the face recognition manager. The target scheduler is also informed of the face capture, so the subject record can be updated.

When the face recognition manager receives a new facial image, a facial image capture record is created and the image is stored. Facial images can be captured by the system at up to about 20 Hz, but face recognition generally takes 0.5–2 s per image, depending on the algorithm. Recognition cannot keep up with capture, so the face recognition algorithm is operated asynchronously. The system can be interfaced to Cognitec FaceVACS®, Identix FaceIt®, Pittsburgh Pattern Recognition FTR, or an internal research face recognition system. In a processing loop, the face recognizer is repeatedly applied to the most recently captured facial image not yet processed, and results are stored in the facial image capture record. Face recognition can use a stored gallery of images, manual enrollments, automatic enrollments or any combination.

The face recognition manager queries the target scheduler to determine which tracker subject ID a facial image came from, based on the capture time of the image. With this information, subject records are created, keyed by the tracker ID number, and the facial image capture records are associated with them.

The auto-enrollment feature of this system makes use of these subject records. This is a highly configurable rule-based process. A typical rule is that a subject is an auto-enroll candidate if one face capture has a quality score exceeding a threshold, one face capture has a face detection threshold exceeding a threshold, recognition has been attempted at least 4 times and has never succeeded, and the most recent face capture was at least 4 seconds ago. If a subject is an auto-enroll candidate, the facial image with the highest quality score is selected and enrolled in the face recognition gallery, possibly after an optional user confirmation.

In indoor and outdoor trials, the capabilities of the system have been evaluated [56]. Test subjects walked in the vicinity of the system in an area where up to about 8 other nonsubjects were also walking in view. An operator recorded the subject distance at the first person detection, first face capture and first successful

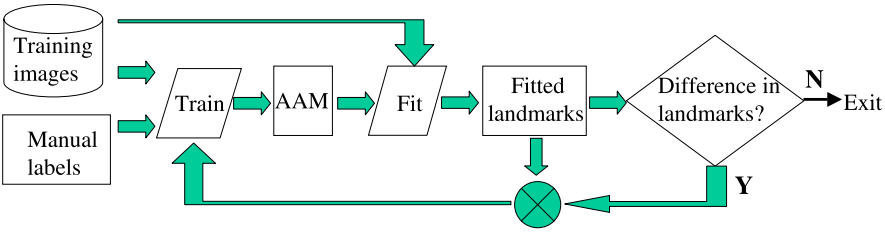

Fig. 14.5 Diagram of the AAM enhancement scheme (© 2006 Xiaoming Liu, et al., used with permission [33])

face recognition with a gallery of 262 subjects. In this experiment, the mean distance to initial person detection was 37 m, the mean distance to initial facial image capture was 34 m and the mean distance to recognition was 17 m.

14.3 Low-Resolution Facial Model Fitting

Face alignment is a process of overlaying a deformable template on a face image to obtain the locations of facial features. Being an active research topic for over two decades [12], face alignment has many applications, such as face recognition, expression analysis, face tracking and animation, etc. Within the considerable prior work on face alignment, *Active Appearance Models (AAMs)* [14] have been one of the most popular approaches. However, the majority of existing work focuses on fitting the AAM to facial images with moderate to high quality [26, 29, 30, 57]. With the popularity of surveillance cameras and greater needs for FRAD, methods to effectively fit an AAM to low-resolution facial images are of increasing importance. This section addresses this particular problem and presents our solutions for it.

Little work has been done in fitting AAMs to low-resolution images. Cootes et al. [13] proposed a multi-resolution Active Shape Model. Dedeoglu et al. [18] proposed integrating the image formulation process into the AAM fitting scheme. During the fitting, both image formulation parameters and model parameters are estimated in a united framework. The authors also showed the improvement of their method compared to fitting with a single high-resolution AAM. We will show that as an alternative fitting strategy, a multi-resolution AAM has far better fitting performance than a high-resolution AAM.

14.3.1 Face Model Enhancement

One requirement for AAM training is to manually position the facial landmarks for all training images. This is a time-consuming and error-prone operation, which certainly affects face modeling. To tackle the problem of labeling error, we develop an AAM enhancement scheme (see Fig. 14.5). Starting with a set of training images

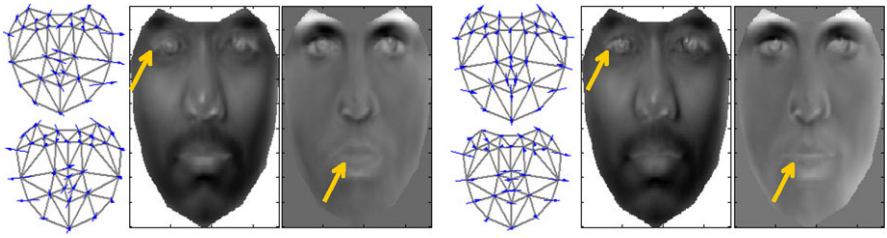

Fig. 14.6 The 6th and 7th shape basis and the 1st and 4th appearance basis before (*left*) and after enhancement (*right*). After enhancement, more symmetric shape variation is observed, and certain facial areas appear sharper (© 2006 Xiaoming Liu, et al., used with permission [33])

and manual labels, an AAM is trained using the above method. Then the AAM is fit to the same training images using the Simultaneous Inverse Compositional (SIC) algorithm, where the manual labels are used as the initial location for fitting. This fitting yields new landmark positions for the training images. This process is iterated. This new landmark set is used for face modeling again, followed by model fitting using the new AAM. The iteration continues until there is no significant difference between the landmark locations of the consecutive iterations. In the face modeling of each iteration, the basis vectors for both the appearance and shape models are chosen such that 98% and 99% of the energy are preserved, respectively.

With the refined landmark locations, the resulting AAM is improved as well. As shown in Fig. 14.6, the variation of landmarks around the outer boundary of the cheek becomes more symmetric after enhancement. Also, certain facial areas, such as the left eye boundary of the 1st appearance basis and the lips of 4th appearance basis, are visually sharper after enhancement, because the training images are better aligned thanks to improved landmark location accuracy.

Another benefit of this enhancement is improved compactness of the face model. In our experiments, the numbers of appearance and shape basis vectors reduce from 220 and 50 to 173 and 14, respectively. There are at least two benefits of a more compact AAM. One is that fewer shape and appearance parameters need to be estimated during model fitting. Thus the minimization process is less likely to become trapped in a local minimum, and fitting robustness is improved. The other is that model fitting can be performed faster because the computation cost directly depends on the dimensionality of the shape and appearance models.

14.3.2 Multi-Resolution AAM

The traditional AAM algorithm makes no distinction with respect to the resolution of the test images being fit. Normally the AAM is trained using the full resolution of the training images, which is called a *high-resolution AAM*. When fitting a high-resolution AAM to a low-resolution image, an up-sampling step is involved in interpolating the observed image and generating a warped input image, $I(\mathbf{W}(\mathbf{x};\mathbf{P}))$. This

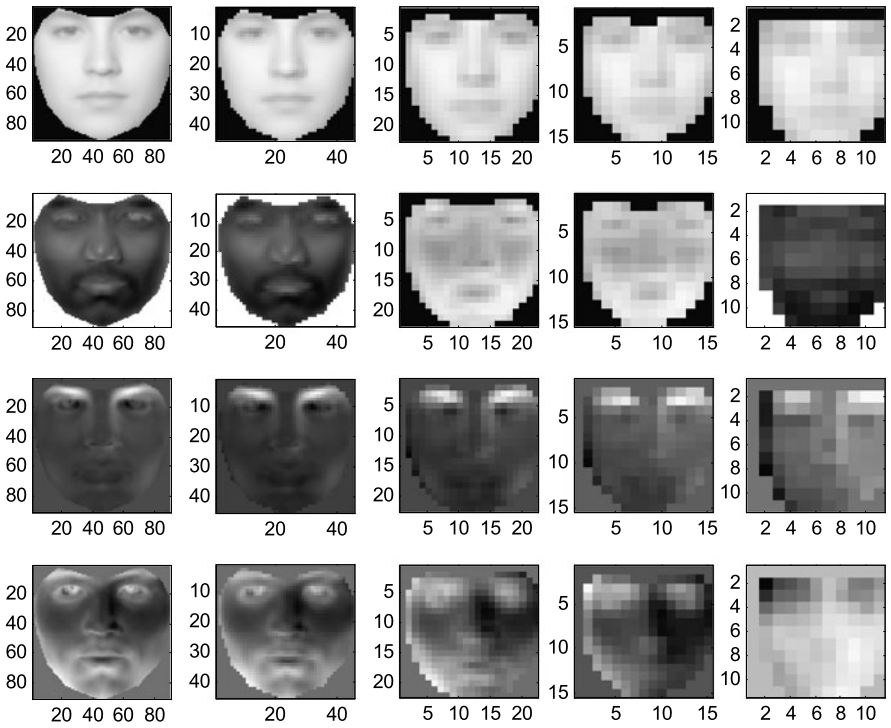

Fig. 14.7 The appearance models of a multi-res AAM: Each column shows the mean and first 3 basis vectors at relative resolutions 1/2, 1/4, 1/8, 1/12 and 1/16 respectively (© 2006 Xiaoming Liu, et al., used with permission [33])

can cause problems because a high-resolution AAM has high frequency components that a low-resolution image does not contain. Thus, even with perfect estimation of the model parameters, the warped image will always have high frequency residual with respect to the high-resolution model instance, which, at a certain point, will overwhelm the residual due to the model parameter errors. Hence, fitting becomes problematic.

The basic idea of applying multi-resolution modeling to AAM is straightforward. Given a set of facial images, we down-sample them into low-resolution images at multiple scales. We then train an AAM using the down-sampled images at each resolution. We call the pyramid of AAMs a *multi-res AAM*. For example, Fig. 14.7 shows the appearance models of a multi-res AAM at relative resolutions 1/2, 1/4, 1/8, 1/12 and 1/16. Comparing the AAMs at different resolutions, we can see that the AAMs at lower resolutions have more blurring than the AAMs at higher resolutions. Also, the AAMs at lower resolutions have fewer appearance basis vectors compared to the AAMs at higher resolutions, which will benefit the fitting. The landmarks used for training the AAM for the highest resolution are obtained using the enhancement scheme above. The mean shapes of a multi-res AAM differ

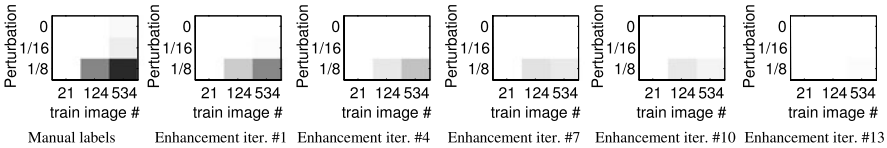

Fig. 14.8 The convergence rate of fitting using an AAM trained from manual labels, and AAM after enhancement iteration number 1, 4, 7, 10 and 13. The brightness of the block is proportional to the convergence rate. Continuing improvement of fitting performance is observed as the enhancement process progresses (© 2006 Xiaoming Liu, et al., used with permission [33])

only by a scaling factor, while the shape basis vectors from different scales of the multiple-resolution AAM are exactly the same.

14.3.3 Experiments

Our experiments are conducted on a subset of the ND1 face database [10], which contains 534 images from 200 subjects. Regarding the fitting performance measurement, we use the convergence rate (CR) with respect to different levels of perturbation on the initial landmark locations. The fitting is converged if the average mean squared error between the estimated landmarks and the ground-truth is less than a threshold. Given the true landmarks of one image, we randomly deviate each landmark within a rectangular area up to a certain range, and the projection of the perturbed landmarks in the shape model is used as the initial shape parameters. Three different perturbation ranges, R, are used: 0, $1/16$, and $1/8$ of the facial height.

Another varying factor is the number of images/subjects in the training set. When multiple images of one subject are used for training an AAM, the resulting AAM is considered as a person-specific AAM. When the number of subjects in the training set is large, the resultant AAM is a generic AAM. The more subjects used, the more generic the AAM is. Using the ND1 database, we test the modeling with three different population sizes, where the numbers of images are 21, 124, 534, and the corresponding numbers of subjects are 5, 25, 200, respectively.

Figure 14.8 shows the CR of AAM fitting after a varying number of model enhancement iterations. The leftmost plot shows the CR using an AAM trained from manual labels only, with varying population size and perturbation window size. Each element represents the CR, which is computed using the same training set as test images. There are some non-converged cases when more generic models are used with a larger perturbation window size. The rest of the plots show the CR using the AAM trained after 1, 4, 7, 10 and 13 iterations of the enhancement algorithm. Continuing improvement of fitting performance is observed with additional enhancement iterations. After the model enhancement is completed, the fitting process converges for all testing cases, no matter how generic the model or how large the perturbation of the initialization.

The second experiment is to test the fitting performance of a multi-res AAM on images with different resolutions. The same dataset and test scheme are used as in

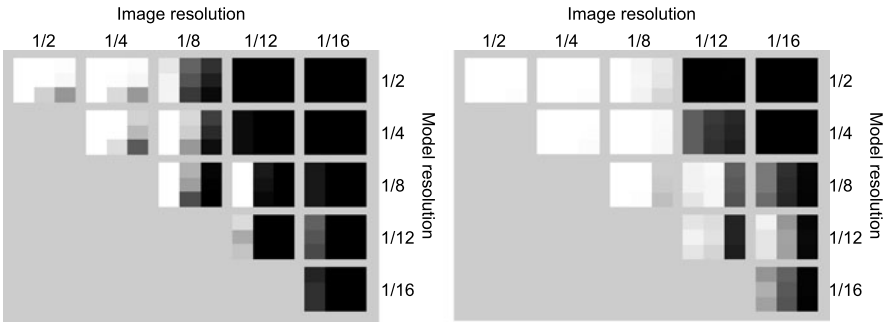

Fig. 14.9 The convergence rate of fitting a multi-res AAM trained with manual labels (*left*) and enhanced landmarks (*right*) to images with different resolution. Each 3 by 3 block has the same axes as in Fig. 14.8 (© 2006 Xiaoming Liu, et al., used with permission [33])

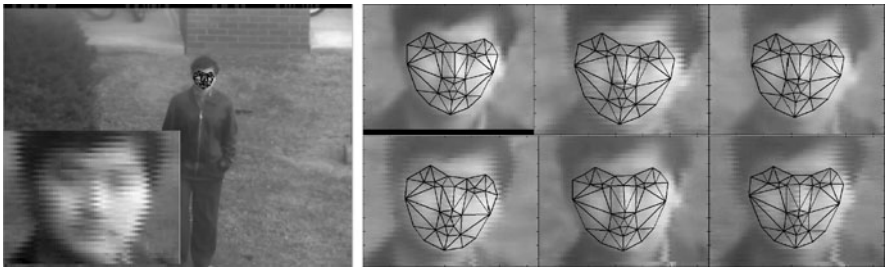

Fig. 14.10 Fitting a multi-res AAM to an outdoor surveillance video: One sample frame with zoom in facial area (*left*) and six zoom in frames overlaid with fitting results (*right*), (© 2006 Xiaoming Liu, et al., used with permission [33])

the previous experiment, except that the different resolutions of down-sampled training images are also used as test images for fitting. Model fitting is conducted with all combinations of AAM resolution and image resolution, where the model resolution no less than the image resolution. As shown in Fig. 14.9, the AAM trained with enhanced landmarks performs much better than the AAM trained from manual labels. Also, for low-resolution images, the best fitting performance is obtained when the model resolution is slightly higher than the facial image resolution, which is far better than fitting using the AAM with the highest model resolution. This shows that the additional appearance detail in the higher resolution AAM seems to confuse the minimization process and results in degraded fitting performance.

The last experiment is model fitting on a surveillance video captured using a PTZ camera positioned about 20 m from the subject. Sample fitting results using a multi-res AAM are shown in Fig. 14.10. Although the frame size is 480 by 640 pixels, the facial area is not only at a low resolution, but also suffers from strong blurring, specular, and interlacing effects, which makes fitting a very challenging task. Our multi-res AAM continuously fits around 100 frames and provides reasonable results.

However, the high-resolution AAM only successfully fits the first 4 frames in this sequence.

14.4 Facial Image Super-Resolution

In situations where lack of image resolution is an impediment to recognition, multi-frame image super-resolution can be a differentiator that makes FRAD possible. In typical FRAD systems facial images are captured with a video camera, and many images of the face are recorded over a short period of time. While conventional face recognition systems operate on a single image, it is desirable to improve recognition performance by using all available image data. There are a few approaches that may be taken. In this section we will describe our super-resolution approach [55], Chap. 13 covers the direct use of video for recognition, and multi-sample fusion [50] is another viable approach.

Super-resolution is the process of producing one high-resolution image from multiple low-resolution images of the same object or scene [8, 11, 32]. Resolution improvement can come from dealiasing, deblurring and noise reduction. A key aspect of super-resolution processing is the exploitation of the fact that the object or camera moves between video frames, so that image pixels in different frames have sub-pixel offsets and thus contain new information.

In this section, we describe a method for the super-resolution of faces from video. Super-resolution algorithms generally have two parts: frame-to-frame registration, and the super-resolution image processing itself. In our method, registration is accomplished with an Active Appearance Model (AAM) designed specifically for the shape of the face and its motion [33]. To solve for the super-resolved image, we define and optimize a cost function with an L_1 data fidelity component and a Bilateral Total Variation (BTV) regularization term, as described by Farsiu [20].

Most image super-resolution algorithms use a parameterized whole-image transformation for registration, such as a homography or rigid translation [23]. This is suitable when the scene is planar or the perspective distortion due to depth and camera motion is insignificant. Facial images have been super-resolved quite dramatically by Baker and Kanade [4], but the motion model used is translation-only and does not account for 3D face shape, effectively assuming that the subject is always facing the camera. To deal with the nonrigid motion of moving faces, optical flow has been used for the registration step [3]. While optical flow certainly can track facial motion, it is computationally complex and its generality brings the risk of overfitting. The super-resolution approach by Mortazavian et al. [38], like that described here, also uses a facial model fitting approach for registration.

14.4.1 Registration and Super-Resolution

Given video of a subject we fit an AAM [33] to the face in each frame. The AAM defines 33 landmark positions that are the vertices of 49 triangles over the face as

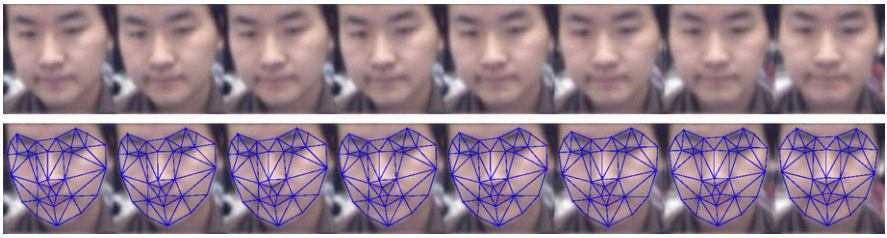

Fig. 14.11 Faces from 8 video frames showing the fitted AAM shape model. The fitted AAM will allow frame-to-frame registration even as the face rotates (© 2009 IEEE, used with permission [55])

seen in Fig. 14.11. The registration of the face between any two frames is then a piecewise affine transformation, with an affine transformation for each triangle defined by the corresponding vertices. A set of about $N = 10$ consecutive frames are combined to produce the super-resolved image.

We will describe the super-resolution algorithm using linear algebra notation, as if each image has its pixel values in a single vector. The computation is actually carried out with operations on 2D pixel arrays. The super-resolution process uses an image formation model relating each of the input frames Y_i, to an unknown super-resolution image, X, with twice the pixel resolution. The image formation process accounts for the face motion, camera blur and detector sampling. For each input frame, F_i is the registration operator that warps X to be aligned with Y_i, but at twice the resolution. The camera blur operator, H, applies the Point Spread Function (PSF). For most installed surveillance cameras it is difficult to determine the true PSF, so we assume a Gaussian shaped PSF with hand selected width, σ. Finally, the sampling operation of the detector is represented by the sparse matrix D that extracts every other pixel in each dimension, yielding an image that should match our real observed image. If we let V_i represent additive pixel intensity noise, the complete linear image formation process is then,

$$Y_i = DHF_iX + V_i.$$

This is the forward model of our observed low-resolution images Y_i given the unknown high-resolution image X. Our goal is to solve the inverse problem to recover X.

The super-resolved image X is produced by optimizing a cost function that is the L_1 norm of the difference between the model of the observations and the actual observations, plus a regularization term, $\Psi(X)$,

$$\hat{X} = \text{argmin}_X \left[\sum_{i=1}^{N} \| DHF_iX - Y_i \|_1 + \lambda \Psi(X) \right].$$

The L_1 norm is used in the data fidelity portion of the cost function for robustness against incorrect modeling assumptions and registration errors. For the regulariza-

tion term, we use Bilateral Total Variation (BTV) [20],

$$\Psi(X) = \sum_{l=-P}^{P} \sum_{m=-P}^{P} \alpha^{|m|+|l|} \left\| X - S_x^l S_y^m X \right\|_1.$$

Here S_x^l and S_y^m are operators that shift the image in the x and y direction by l and m pixels. With BTV, the neighborhood over which absolute pixel difference constraints are applied can be larger (with $P > 1$) than for Total Variation (TV). The size of the neighborhood is controlled by parameter P and the constraint strength decay is controlled by $\alpha \in (0, 1)$. L_1-based regularization methods such as BTV or TV are selected to preserve edges. By contrast, L_2-based Tikhonov regularization is effectively a smoothness constraint, which is counter to our goal of increased resolution.

To solve for the super-resolution image, X is first initialized using straightforward warping and averaging. A steepest descent search using the analytic gradient of the cost function is used [55]. With the original frames normalized to a pixel range of [0, 1], we have found that setting the regularization strength parameter λ to 0.025 gives the best visual results and we use that value for the experiments presented here.

14.4.2 Results

Figure 14.12 shows sample super-resolution results, including: (a) the face from the original video frame; (b) that single frame restored with a Wiener filter; and (c) the result of multi-frame super-resolution using $N = 10$ consecutive frames. The increase in sharpness and clarity is visually apparent. In another experiment with a gallery of 700 images from unique subjects, we tested 138 facial video clips collected from three individuals with surveillance cameras that had an eye-to-eye distance ranging from 17 to 48 pixels. With this challenging data collected at about 10 m, super-resolution processing instead of using just single frames increased the rank-1 recognition rate from 50% to 56%.

14.5 Conclusions

Face recognition at a distance is a challenging problem with a large number of beneficial applications. We have reviewed the primary challenges, approaches and research literature on this topic, and we have described some specific work that we have carried out to create a prototype FRAD system, fit alignment models to faces at very low resolutions and super-resolve facial images. Still, there are a great many open issues that may lead to enhanced functionality or new applications. We conclude this chapter by highlighting a number of these potential future avenues of research.

Fig. 14.12 Example original video frames, Wiener filter results, and super-resolution results with enlarged views of the left eye. In the Wiener filter results, artifacts in the face region are primarily due to enhancement of interlacing artifacts. Some ringing due to circular convolution is present, but only near the image edges. The increased resolution and clarity in the super-resolution results is clearly visible (© 2009 IEEE, used with permission [55])

Commercially available face recognition systems are typically designed and trained for access control applications with high-quality facial images. The need for facial recognition algorithms that work well under FRAD imaging conditions is well understood, and this is an active research area. It will be key to understand which facial features are present and absent in facial images collected at a distance so that recognition algorithms can focus on those that remain. If face recognition can be performed fast enough, then immediate recognition results, or even face quality analysis can be utilized more actively for NFOV resource allocation and active capture control loops. The use of incremental fusion of face recognition results during rapid video capture of faces may also make active capture systems more efficient.

One of the challenges of FRAD is subject pose. Strategies to attract the attention of subjects to certain locations near cameras may help this issue in some situations. In all of the active vision systems, we have discussed both the WFOV and NFOV cameras are stationary. The use of cameras on movable platforms, such as guide wires or robots could enable much more effective facial image collection and open a new surveillance and biometric identification paradigm.

Acknowledgements Section 14.2 of this report was prepared by GE Global Research as an account of work sponsored by Lockheed Martin Corporation. Information contained in this report constitutes technical information which is the property of Lockheed Martin Corporation. Neither

GE nor Lockheed Martin Corporation, nor any person acting on behalf of either; a. Makes any warranty or representation, expressed or implied, with respect to the use of any information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or b. Assume any liabilities with respect to the use of, or for damages resulting from the use of, any information, apparatus, method, or process disclosed in this report. Sections 14.3 and 14.4 were supported in part by award #2005-IJ-CX-K060 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

References

1. Andersen, J.F., Busck, J., Heiselberg, H.: Long distance high accuracy 3-D laser radar and person identification. In: Kamerman, G.W. (ed.) Laser Radar Technology and Applications X, vol. 5791, pp. 9–16. SPIE, Bellingham (2005)
2. Bagdanov, A., Bimbo, A., Nunziati, W., Pernici, F.: Learning foveal sensing strategies in unconstrained surveillance environments. In: AVSS (2006)
3. Baker, S., Kanade, T.: Super resolution optical flow. Tech. Rep. CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (1999)
4. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. Pattern Anal. Mach. Intell. **24**(9), 1167–1183 (2002)
5. Bellotto, N., Sommerlade, E., Benfold, B., Bibby, C., Reid, I., Roth, D., Fernández, C., Gool, L.V., Gonzàlez, J.: A distributed camera system for multi-resolution surveillance. In: Proc. of the ACM/IEEE Intl. Conf. on Distributed Smart Cameras (ICDSC) (2009)
6. Bimbo, A.D., Pernici, F.: Towards on-line saccade planning for high-resolution image sensing. Pattern Recognit. Lett. **27**(15), 1826–1834 (2006)
7. Blackman, S., Popoli, R.: Design and Analysis of Modern Tracking Systems. Artech House, Norwood (1999)
8. Borman, S.: Topics in multiframe superresolution restoration. Ph.D. thesis, University of Notre Dame, Notre Dame, IN (2004)
9. Bowyer, K.W., Chang, K., Flynn, P.: A survey of approaches and challenges in 3D and multimodal 3D + 2D face recognition. Comput. Vis. Image Underst. **101**(1), 1–15 (2006)
10. Chang, K., Bowyer, K., Flynn, P.: Face recognition using 2D and 3D facial data. In: Proc. ACM Workshop on Multimodal User Authentication, pp. 25–32 (2003)
11. Chaudhuri, S. (ed.): Super-Resolution Imaging, 3rd edn. Kluwer Academic, Dordrecht (2001)
12. Cootes, T., Cooper, D., Tylor, C., Graham, J.: A trainable method of parametric shape description. In: BMVC, pp. 54–61 (1991)
13. Cootes, T., Taylor, C., Lanitis, A.: Active shape models: Evaluation of a multi-resolution method for improving image search. In: BMVC, vol. 1, pp. 327–336 (1994)
14. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
15. Costello, C.J., Diehl, C.P., Banerjee, A., Fisher, H.: Scheduling an active camera to observe people. In: Proc. of the ACM Intl. Workshop on Video Surveillance and Sensor Networks, pp. 39–45 (2004)
16. Davis, J., Morison, A., Woods, D.: An adaptive focus-of-attention model for video surveillance and monitoring. Mach. Vis. Appl. **18**(1), 41–64 (2007)
17. Davis, J., Morison, A., Woods, D.: Building adaptive camera models for video surveillance. In: WACV (2007)
18. Dedeoglu, G., Baker, S., Kanade, T.: Resolution-aware fitting of active appearance models to low-resolution images. In: ECCV (2006)
19. Elder, J.H., Prince, S., Hou, Y., Sizintsev, M., Olevskiy, Y.: Pre-attentive and attentive detection of humans in wide-field scenes. Int. J.Comput. Vis. **72**, 47–66 (2007)

20. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super-resolution. IEEE Trans. Image Process. **13**(10), 1327–1344 (2004)
21. Greiffenhagen, M., Ramesh, V., Comaniciu, D., Niemann, H.: Statistical modeling and performance characterization of a real-time dual camera surveillance system. In: CVPR (2000)
22. Hampapur, A., Pankanti, S., Senior, A., Tian, Y.L., Brown, L., Bolle, R.: Face cataloger: multi-scale imaging for relating identity to location. In: AVSS, pp. 13–20 (2003)
23. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2000)
24. Krahnstoever, N., Tu, P., Sebastian, T., Perera, A., Collins, R.: Multi-view detection and tracking of travelers and luggage in mass transit environments. In: Proc. Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS) (2006)
25. Krahnstoever, N., Yu, T., Lim, S.N., Patwardhan, K., Tu, P.: Collaborative real-time control of active cameras in large scale surveillance systems. In: Proc. Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2) (2008)
26. Liang, L., Wen, F., Xu, Y., Tang, X., Shum, H.: Accurate face alignment using shape constrained Markov network. In: CVPR (2006)
27. Lim, S.N., Davis, L.S., Mittal, A.: Constructing task visibility intervals for a surveillance system. ACM Multimedia Systems Journal **12**(3) (2006)
28. Lim, S.N., Davis, L., Mittal, A.: Task scheduling in large camera network. In: ACCV (2007)
29. Liu, X.: Discriminative face alignment. IEEE Trans. Pattern Anal. Mach. Intell. **31**(11), 1941–1954 (2009)
30. Liu, X.: Video-based face model fitting using adaptive active appearance model. Image Vis. Comput. **28**(7), 1162–1172 (2010)
31. Liu, Z., Sarkar, S.: Outdoor recognition at a distance by fusing gait and face. Image Vis. Comput. **25**(6), 817–832 (2007)
32. Liu, K.R., Kang, M.G., Chaudhuri, S. (eds.): IEEE Signal Processing Magazine, Special edition: Super-Resolution Image Reconstruction, vol. 20, no. 3. IEEE (2003)
33. Liu, X., Tu, P.H., Wheeler, F.W.: Face model fitting on low resolution images. In: BMVC (2006)
34. Marchesotti, L., Marcenaro, L., Regazzoni, C.: Dual camera system for face detection in unconstrained environments. In: ICIP (2003)
35. Medioni, G., Choi, J., Kuo, C.H., Choudhury, A., Zhang, L., Fidaleo, D.: Non-cooperative persons identification at a distance with 3D face modeling. In: BTAS (2007)
36. Medioni, G., Fidaleo, D., Choi, J., Zhang, L., Kuo, C.H., Kim, K.: Recognition of non-cooperative individuals at a distance with 3D face modeling. In: 2007 IEEE Workshop on Automatic Identification Advanced Technologies, pp. 112–117 (2007)
37. Medioni, G., Choi, J., Kuo, C.H., Fidaleo, D.: Identifying noncooperative subjects at a distance using face images and inferred three-dimensional face models. IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. **39**(1), 12–24 (2009)
38. Mortazavian, P., Kittler, J., Christmas, W.: A 3-D assisted generative model for facial texture super-resolution. In: BTAS, pp. 1–7 (2009)
39. NIST Multiple Biometric Grand Challenge. http://face.nist.gov/mbgc
40. O'Toole, A., Harms, J., Snow, S., Hurst, D., Pappas, M., Ayyad, J., Abdi, H.: A video database of moving faces and people. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 812–816 (2005)
41. Prince, S., Elder, J., Hou, Y., Sizinstev, M., Olevsky, E.: Towards face recognition at a distance. In: Proc. of the IET Conf. on Crime and Security, pp. 570–575 (2006)
42. Prince, S., Elder, J., Warrell, J., Felisberti, F.: Tied factor analysis for face recognition across large pose differences. IEEE Trans. Pattern Anal. Mach. Intell. **30**(6), 970–984 (2008)
43. Qureshi, F., Terzopoulos, D.: Surveillance in virtual reality: System design and multi-camera control. In: CVPR, pp. 1–8 (2007)
44. Qureshi, F., Terzopoulos, D.: Multi-camera control through constraint satisfaction for persistent surveillance. In: AVSS, pp. 211–218 (2008)
45. Qureshi, F., Terzopoulos, D.: Smart camera networks in virtual reality. Proc. IEEE **96**(10), 1640–1656 (2008)

46. Rara, H., Elhabian, S., Ali, A., Miller, M., Starr, T., Farag, A.: Distant face recognition based on sparse-stereo reconstruction. In: ICIP, pp. 4141–4144 (2009)
47. Rara, H., Elhabian, S., Ali, A., Miller, M., Starr, T., Farag, A.: Face recognition at-a-distance based on sparse-stereo reconstruction. In: CVPR Workshop on Biometrics, pp. 27–32 (2009)
48. Redman, B., Höft, T., Grow, T., Novotny, J., McCumber, P., Rogers, N., Hoening, M., Kubala, K., Sibell, R., Shald, S., Uberna, R., Havermann, R., Sandalphon, D.: Low-cost, stand-off, 2D+3D face imaging for biometric identification using Fourier transform profilometry. In: 2009 Military Sensing Symposia (MSS) Specialty Group on Active E-O Systems, vol. 1. Las Vegas, NV (2009)
49. Redman, B., Marron, J., Seldomridge, N., Grow, T., Höft, T., Novotny, J., Thurman, S.T., Embry, C., Bratcher, A., Kendrick, R.: Stand-off 3D face imaging and vibrometry for biometric identification using digital holography. In: 2009 Military Sensing Symposia (MSS) Specialty Group on Active E-O Systems, vol. 1. Las Vegas, NV (2009)
50. Ross, A.A., Nandakumar, K., Jain, A.K. (eds.): Handbook of Multibiometrics. Springer, Berlin (2006)
51. Senior, A., Hampapur, A., Lu, M.: Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration. In: WACV, vol. 1, pp. 433–438 (2005)
52. Stillman, S., Tanawongsuwan, R., Essa, I.: A system for tracking and recognizing multiple people with multiple cameras. In: Proc. of 2nd Intl. Conf. on Audio-Vision-based Person Authentication, pp. 96–101 (1998)
53. Tistarelli, M., Li, S.Z., Chellappa, R. (eds.): Handbook of Remote Biometrics for Surveillance and Security. Springer, Berlin (2009)
54. Tu, P.H., Doretto, G., Krahnstoever, N.O., Perera, A.G.A., Wheeler, F.W., Liu, X., Rittscher, J., Sebastian, T.B., Yu, T., Harding, K.G.: An intelligent video framework for homeland protection. In: Proc. of SPIE Defense & Security Symposium, Conference on Unattended Ground, Sea, and Air Sensor Technologies and Applications IX. Orlando, FL (2007)
55. Wheeler, F.W., Liu, X., Tu, P.H.: Multi-frame super-resolution for face recognition. In: BTAS (2007)
56. Wheeler, F.W., Weiss, R.L., Tu, P.H.: Face recognition at a distance system for surveillance applications. In: BTAS (2010)
57. Yan, S., Liu, C., Li, S.Z., Zhang, H., Shum, H.Y., Cheng, Q.: Face alignment using texture-constrained active shape models. Image Vis. Comput. **21**(1), 69–75 (2003)
58. Yao, Y., Abidi, B., Kalka, N., Schmid, N., Abidi, M.: High magnification and long distance face recognition: Database acquisition, evaluation, and enhancement. In: Proc. Biometrics Symposium (2006)
59. Yao, Y., Abidi, B., Kalka, N.D., Schmid, N., Abidi, M.: Super-resolution for high magnification face images. In: Prabhakar, S., Ross, A.A. (eds.) Proceedings of the SPIE, Biometric Technology for Human Identification IV, vol. 6539. Orlando, FL (2007)
60. Yao, Y., Abidi, B.R., Kalka, N.D., Schmid, N.A., Abidi, M.A.: Improving long range and high magnification face recognition: Database acquisition, evaluation, and enhancement. Comput. Vis. Image Underst. **111**(2), 111–125 (2008)
61. Yu, T., Lim, S.N., Patwardhan, K., Krahnstoever, N.: Monitoring, recognizing and discovering social networks. In: CVPR (2009)
62. Zhou, X., Bhanu, B.: Feature fusion of face and gait for human recognition at a distance in video. In: ICPR, vol. 4, pp. 529–532 (2006)
63. Zhou, X., Bhanu, B.: Integrating face and gait for human recognition at a distance in video. IEEE Trans. Syst. Man Cybern., Part B, Cybern. **37**(5), 1119–1137 (2007)
64. Zhou, X., Collins, R., Kanade, T., Metes, P.: A master-slave system to acquire biometric imagery of humans at distance. In: ACM International Workshop on Video Surveillance (2003)