A Video Analysis Framework for Soft Biometry Security Surveillance

Yuan-Fang Wang
Department of Computer Science
University of California
Santa Barbara, CA 93106
yfwang@cs.ucsb.edu

Edward Y. Chang
Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106
echang@ece.ucsb.edu

Ken P. Cheng
Proximex Corporation
6 Results Way
Cupertino, CA
ken.cheng@proximex.com

ABSTRACT

We propose a distributed, multi-camera video analysis paradigm for aiport security surveillance. We propose to use a new class of biometry signatures, which are called soft biometry including a person's height, built, skin tone, color of shirts and trousers, motion pattern, trajectory history, etc., to ID and track errant passengers and suspicious events without having to shut down a whole terminal building and cancel multiple flights. The proposed research is to enable the reliable acquisition, maintenance, and correspondence of soft biometry signatures in a coordinated manner from a large number of video streams for security surveillance. The intellectual merit of the proposed research is to address three important video analysis problems in a distributed, multi-camera surveillance network: sensor network calibration, peer-to-peer sensor data fusion, and stationary-dynamic cooperative camera sensing.

1. INTRODUCTION

Objectives. Our project is aimed at developing a robust and intelligent video analysis paradigm for large distributed camera networks, such as the surveillance camera networks that are routinely deployed at all major airports around the world these days. The proposed paradigm is to enable the reliable acquisition, maintenance, and correspondence of a new class of biometry signatures (*soft biometry*, to be defined later) in a coordinated manner from a large number of video streams. Theoretically, we propose to research novel algorithms in sensor calibration, data fusion, and cooperative sensing in distributed, multi-camera networks. In practice, we are working with our industrial partner to perform "rubber-meets-road" validation by deploying our soft biometry video-analysis system in major US airports for security monitoring.

Motivation. On September 4th 2004, three terminals at Los Angeles International Airport were shut down for more than three hours. Apparently, a passenger bypassed a security checkpoint without being properly searched. Three connected terminal buildings were evacuated with all passengers inside re-screened. This incident had happened at a most inopportune time—the busy Labor Day travel weekend, and delayed about a hundred flights, inconvenienced thousands of passengers, and caused major traffic tie up on the surrounding San Diego and Santa Monica Freeways.

What is troubling is that this was not an isolated incident, and similar incidents had occurred many times before: At Miami International Airport on November 14th, 2002, two passengers bypassed security when they strolled through an exit lane. Five concourses were evacuated and some 40 flights were delayed for more than three hours. At Chicago O'Hare International Airport on October 16th, 2002, the United Airlines terminal was evacuated after a man avoided a security checkpoint by entering the terminal through an exit. At San Louis Lambert International Airport on October 4th, 2002, a passenger managed to walk away from a checkpoint with a suspicious bag before the security agents could act. All passengers in the East Terminal were forced to be re-screened. At Dallas-Fort Worth International Airport on January 8th, 2003, a man bolted through a passenger checkpoint and disappeared into the crowd. The action forced the evacuation of thousands of people from three terminals. At Seattle-Tacoma International Airport on January 5th, 2003, TSA security personnel stationed at an exit lane fell asleep and left the exit unguarded. Hundreds of people who had already passed through security checkpoints, including those who had boarded planes, were brought back and re-screened. The perpetrators of these incidents were never caught.

While fixing such security lapses appears unglamorous and mundane in the grand scheme of national security and combating terrorism, it is important for at least two reasons: (1) The above incidents might turn out to be false alarms; however, the resulted emotional stress, scheduling chaos, traffic tie up, and financial toll are very real and compelling, and (2) as we learned from our industrial partners at Proximex Corp., who attended a gathering of airport security personnel of major US airports last fall in Detroit, that dispatching security agents to track down false alarms consumes a significant portion of their budgets and ties down agents who could have performed other more useful tasks in securing an airport. Hence, an automated, or even semi-automated, surveillance and identification system—that quickly narrows down the search area of the errant passengers, avoids forced shutdown and evacuation, and improves public safety—is solely needed.

So what kind of technologies can be employed to deal with these types of incidents? While significant strides have been made in biometry, such as fingerprint, iris scan, voice identification, and face recognition, these techniques are not suitable for many reasons. Finger print, iris scan, and voice identification are more applicable to cooperative subjects. They are applied under a controlled environment, with a slow speed, and for a relatively small passenger volume. Hence, these definitely are the wrong approaches for identifying an errant passenger in a busy terminal building. Even the vaunted face recognition techniques are of little practical use

VSSN Singapore '05

here. The perpetrators in the above cases very likely were ordinary, law-abiding citizens who might have made an honest mistake or were in a real hurry to catch their flights. They were not criminals or terrorists with existing biometric profiles that can be searched for and compared. Even if they were, no face detection and recognition technique that we are aware of is capable of reliably correlating FBI mug shots with airport surveillance videos that might be available in these cases. It is extremely difficult, if not impossible, even for humans to ID a face image, taken with a low resolution surveillance camera covering a large area, of a moving subject appearing at a great distance, with varying lighting and body pose, and with significant occlusion and ingenious disguise of facial features. Hence, it is not reasonable to expect a face recognition system to achieve such an impossible feat in the foreseeable future.

Specific Tasks. We argue that a new class of biometry, which we term *soft biometry* including a passenger's height, built, skin tone, color of shirts and trousers, motion pattern, trajectory history, etc., can be inferred from airport surveillance videos and used to track and ID errant passengers without having to shut down a whole terminal building and cancel multiple flights. We call these signatures "*soft*" biometry because they change over time, are not unique traits of a person, and are not legally accepted ID signatures like fingerprint or DNA. Nonetheless, soft biometry does offer significant merits in the video-analysis scenarios identified above: that many such signatures are not expensive to compute, do not require the cooperation of the surveillance subjects, can be sensed at a distance in a crowded environment, serve as a good screening tool to narrow down the search for suspects, and have wide applications beyond airport security surveillance. For example, in a crowded amusement park, missing children can be searched for by their height, cloth colors, and motion history. In an industrial factory or a college campus, soft biometry can augment paging for localizing people and equipment.

One might suspect, and we concur, that it is not very difficult to compute some of these soft biometric signatures from individual, properly-segmented image frames. *The real challenge is in designing a robust and intelligent video-analysis system to support the reliable acquisition, maintenance, and correspondence of soft biometry signatures in a coordinated manner from a large number of video streams gathered in a large camera network*. Our research thus aims at developing novel video-analysis methods to support *data fusion* and *event analysis* tasks [1, 3] in distributed, multi-camera surveillance networks. We identify below three research tasks to advance fundamental theories and develop algorithms that can significantly improve the operation of large camera networks, quality of data fusion, and accuracy of event analysis.

• *Task A*: *Sensor network calibration*. In order to correctly correlate and fuse information from multiple cameras, calibration is of paramount importance. Cameras deployed in a large network have different physical characteristics, such as location, field-of-view (FOV), spatial resolution, color sensitivity, and notion of time. The difference makes answering even simple queries exceedingly difficult. For example, if a subject moves from the FOV of one camera to another, which has different color sensitivity and operates under dissimilar lighting conditions, drastic changes in color signatures do occur. To reliably compute soft biometry to assist the identification of subjects across the FOVs of multiple cameras therefore requires careful color calibration. We have developed and integrated a suite of algorithms for *spatial, temporal*, and *color* calibration for cameras with both overlapped and non-overlapped FOVs.

• *Task B*: *Peer-to-peer sensor data fusion*. As cameras have limited FOVs, multiple cameras are often stationed to monitor an extended surveillance area, such as an indoor arrival/departure lounge or an outdoor parking lot. Collectively, these cameras provide complete spatial coverage of the surveillance area. (A small amount of occlusion by architectural fixtures, decoration, and plantation is often unavoidable.) Individually, the event description inferred from a single camera is likely to be incomplete. (E.g., the trajectory of a vehicle entering a parking lot is only partially observed from a certain vantage point.) We have developed algorithms to fuse video data from multiple cameras for reliable event detection.

• *Task C*: *Stationary-dynamic cooperative camera sensing*. To achieve effective wide-area surveillance with limited hardware, a surveillance camera is often configured to have a large FOV. However, once suspicious persons/activities have been identified through video analysis, selected cameras ought to obtain close-up views of these suspicious subjects for further scrutiny and identification (e.g., to obtain a close-up view of the license plate of a car or the face of a person). Our solution is to employ stationary-dynamic camera assemblies to enable wide-area coverage and selective focus-of-attention through cooperative sensing. That is, the stationary cameras perform a global, wide FOV analysis of the motion patterns in the surveillance zone. Based on some pre-specified criteria, the stationary cameras identify suspicious behaviors or subjects that need further attention. The dynamic camera, mounted on a mobile platform and equipped with a zoom lens, is then used to obtain close-up view of the subject to reliably compute soft biometry signatures. We have studied research issues to enable cooperative camera sensing, including dynamic camera calibration and stationary-dynamic camera sensing using a visual feedback paradigm.

Significance and Impact. Significant progress has been made in video surveillance. Mature technology is increasingly being applied to real-world problems and spawns new commercial opportunities. The momentum started back in 1997, when DARPA began a three-year program to develop video surveillance and monitoring (VSAM) technology. The pace accelerated after the September 11 Attack. While extensive research has been conducted on many component technologies, our research makes contribution in three specific aspects of *robustness, integration, and validation*.

2. TASK A: SENSOR NETWORK CALIBRATION

To fuse data in a network of multiple cameras, it is important that a consistent notion of space, time, and color is established to facilitate the exchange of sensor data. These correspond to spatial, temporal, and color calibration. Spatial calibration is a problem that has been thoroughly researched in computer vision. We have also developed temporal registration techniques to determine the time skew between cameras' clocks by matching the trajectories of the same object observed in multiple video streams. In this paper, we present our color registration algorithm.

Color registration is difficult because the "sensed" color and the "true" color of an object can be drastically different. Three important factors, the physical content of the scene, the illumination of the incident light, and the characteristics of the camera, affect color sensing. The ability of a vision system to diminish, or in the ideal case, remove color variation from fluctuation in source illumination and receiver characteristics, and therefore "see" the physical scene precisely, is called color constancy. Many color constancy algorithms exist dating back a couple of decades, including grey-world algorithms, retinex methods, linear decomposition, gamut mapping, Bayesian correlation, and many others.

Color perception is an extremely complicated and nonlinear science. To simplify the analysis, many color constancy models as-

sume a single camera; a fixed, frontal surface orientation; and often times, a point light source or spatially-invariant illumination. Or color-constancy research is often confined to the Mondrian world—a world of flat, frontally presented collages of color papers. In contrast, in order to color register spatially-distributed surveillance cameras operating under different lighting conditions and with varying color sensitivity, our scheme needs to take into consideration variation in surface orientation, extended light, secondary reflection, and limited spatial resolution and varying color sensitivity of multiple cameras.

We have developed a robust color calibration procedure as follow: We quantize the entire color space into 11 bins (black, white, red, yellow, green, blue, brown, purple, pink, orange, and gray). These colors are usually referred to as culture colors. Representing the entire color space using a small number of primitives is advantageous for at least two reasons: (1) In most surveillance applications, surveillance subjects occupy small screen areas, and hence, pixels available to construct the color signature of an object are usually quite limited. Coarse quantization of the color space (into 11 bins in the case of culture colors) avoids random fluctuation of color signatures due to insufficient pixel samples, and (2) culture colors facilitate posing query as these colors are universally perceived and widely used across multiple cultures.

For each sensor, we collect images of calibration markers that are known to be of certain culture colors. (These can be as simple as people wearing certain colored shirts walking around in the FOV of the camera.) The sensed red, green, blue pixel values are recorded in a table $CC_{i,k}^{r,g,b}$, where $1 \leq i \leq 11$, and $1 \leq k \leq n_i$, and n_i is the number of color samples collected for the i-th culture color. We form the discrimination function of the i-th culture color for a sensor as (this function can be different for different image regions of a single sensor, for the same sensor operating under different lighting conditions, and for different sensors)

$$f_i(C^{r,g,b}) = \sum_{k=1}^{n_i} \Phi(||C^{r,g,b} - CC_{i,k}^{r,g,b}||) \qquad (1)$$

where Φ is a suitable kernel function (e.g., Gaussian). (In real implementation, we do not use all color samples in Eq. 1 as it is highly inefficient. Instead we use kernel methods to locate support vectors in classification.) A query color sample $C^{r,g,b}$ is then assigned to the culture color with the highest discrimination score. While this scheme seems naive, we show elsewhere that its operation is sensible and corresponds closely to the notion of color similarity in the real-world.

3. TASK B: PEER-TO-PEER MULTI-CAMERA DATA FUSION

In a distributed camera network, the server receives video streams from distributed cameras that each has limited spatial and temporal coverage, is potentially noisy, and is susceptible to occlusion and scene clutter. We propose here a hierarchical peer-to-peer fusion scheme to deal with these problems.

Sensor data fusion refers to the task of combining multiple sensor data in a complementary and synergistic way to improve data availability, reduce noise, and improve robustness in the analysis. Sensor data fusion can be for multiple sensors of the same or different types and can occur at data, feature, and decision levels. Data and feature fusion strategies are often used for combining heterogeneous sensor data, e.g., in fusing inertia, ultrasonic, and vision sensors for mobile robotics applications, and in fusing multi-image modalities (e.g., infrared and vision sensors) for target recognition and scene interpretation. IBR (image-based-rendering) techniques

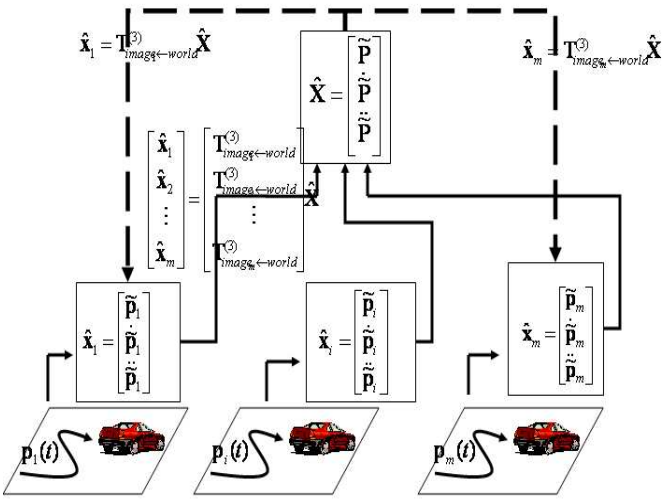

Figure 1: Two-level hierarchical Kalman Filter configuration.

can also be considered a data fusion strategy where a single sensor or multiple sensors, often of the same kind, are used to to construct an environment map. Decision fusion strategies have the root in pattern recognition with many well-established algorithms that are readily applicable. Our unique contribution is in using two-level hierarchical Kalman Filters with both bottom-up and top-down analysis for data fusion and information dissemination from and to multiple sensors, thus improving tracking reliability.

We used the Kalman Filter as the tool for fusing information spatially and temporally from multiple cameras for event detection. Suppose that a vehicle (or a person) is moving in a surveillance zone. Its trajectory in the global reference system is $\mathbf{P}(t) = [X(t), Y(t), Z(t)]^T$. The trajectory may be observed in camera i, as $\mathbf{p}_i(t) = [x_i(t), y_i(t)]^T$, where $i = 1, \cdots, m$ (the number of cameras used). The goal is to optimally track, correlate, and fuse individual camera trajectories into a consistent, global description.

We formulate the solution as a two-level hierarchy of the Kalman Filters. Referring to Fig. 1, at the bottom level of the hierarchy, we employ for each camera a Kalman Filter to estimate, independently, the position $\mathbf{p}_i(t)$, velocity $\dot{\mathbf{p}}_i(t)$, and acceleration $\ddot{\mathbf{p}}_i(t)$ of the object, based on the tracked image trajectory in *the local camera reference frame* ("$\hat{}$" denote estimated quantities, and "$\tilde{}$" denote quantities in homogeneous coordinates in Fig. 1). Or in the Kalman Filter jargon, the position, velocity, and acceleration vectors establish the "state" of the system while the image trajectory serves as the "observation" of the system state. At the top level of the hierarchy, we use a single Kalman Filter to estimate the object's position $\mathbf{P}(t)$, velocity $\dot{\mathbf{P}}(t)$, and acceleration $\ddot{\mathbf{P}}(t)$ in *the global world reference frame*—this time, using the estimated positions, velocities, and accelerations from multiple cameras ($\mathbf{p}_i(t)$, $\dot{\mathbf{p}}_i(t)$, $\ddot{\mathbf{p}}_i(t)$) as observations (the solid feed-upward lines in Fig. 1). This is possible because camera calibration and registration are used for deriving the transform matrices ($\mathbf{T}_{image \leftarrow world}$ and $\mathbf{T}_{world \leftarrow image}$ in Fig. 1). These matrices allows $\mathbf{p}_i$, measured in the reference frame of a camera, to be related to $\mathbf{P}$ in the global world system.

An interesting scenario occurs when one (or more) cameras in the sensor network loses track of an object. This can happen because of scene clutter, self- and mutual-occlusion, or the tracked objects exiting the FOV of a camera, among many other possibilities. The camera could switch from a "track" mode into a "reacquire" mode by searching the whole image for telltale signs of the

object. However, doing so inevitably slows down event-processing and introduces a high degree of uncertainty in the resulted event description. Instead, we allow the dissemination of fused information to individual cameras (the dashed feed-downward lines in Fig. 1) to help guide the reacquisition process. The Kalman Filter, being a flexible information-fusion algorithm, can readily use the fused information (instead of sensor data) for maintaining and updating state vectors. This hierarchical feed-upward (for sensor data fusion) and feed-downward (for information dissemination) filter structure thus provides a powerful and flexible mechanism for joining sensor data spatially.

4. TASK C: STATIONARY-DYNAMIC CO-OPERATIVE CAMERA SENSING

To achieve effective wide-area surveillance and selective focus-of-attention places conflicting constraints on the system configurations and camera parameters. For instance, a large surveillance FOV is achieved using a lens with a short focal length, whereas selective focus-of-attention requires a lens with a long focal length, and the ability to dynamically adjust the aim of the camera.

We propose cooperative sensing using a stationary-dynamic camera assembly to achieve these two conflicting goals. In our design, an extended surveillance area is covered by multiple stationary (or master) cameras with wide FOVs to perform a global analysis of the motion patterns in the surveillance zone. Based on some pre-specified criteria, the stationary cameras identify suspicious behaviors or subjects that need further attention. These behaviors may include loitering around sensitive or restricted areas, entering through an exit, leaving packages behind unattended, driving in a zigzag or intoxicated manner, circling an empty parking lot or a building in a suspicious and reconnoitering manner, among many others. A dynamic (or slave) camera is mounted on a mobile platform and equipped with a zoom lens; the aim and zoom of a dynamic camera are both put under program control (or a PTZ camera). Once a suspicious event/subject has been identified, the stationary cameras will guide the dynamic cameras to focus on the region of interest (e.g., the license plate of a car or the face of a person) for selective attention and analysis.

A large number of R&D issues need to be addressed related to the configuration, calibration, and operation of a stationary-dynamic camera assembly. While many important research questions—ranging from low-level image processing to high-level intelligent event analysis—will be of interest to the CV community, we address two specific problems that present unique challenges in using stationary-dynamic cameras for video surveillance: (1) off-line calibration of both stationary and dynamic cameras, and (2) on-line selective focus-of-attention by cooperative stationary-dynamic camera sensing.

While using dynamic PTZ cameras to augment stationary cameras for surveillance is not new, our contribution is in making some fundamental algorithmic improvement in calibration and operation to make the idea practical, robust, and efficient. We contrast our approaches with the state-of-the-art methods in off-line calibration and on-line selective focus-of-attention. Davis and Chen [2] presented a technique for calibrating a pan-tilt camera off-line. The technique adopted a general camera model that did not assume that the rotational axes were orthogonal or that they were aligned with the camera's imaging optics. Furthermore, [2] argued that the traditional methods of calibrating stationary cameras using a fixed calibration stand were impractical for calibrating dynamic cameras, because a dynamic camera had a much larger working volume. Instead, a novel technique was adopted to generate virtual calibration landmarks using a moving LED. The 3D positions of the LED were inferred, via stereo triangulation, from multiple stationary cameras

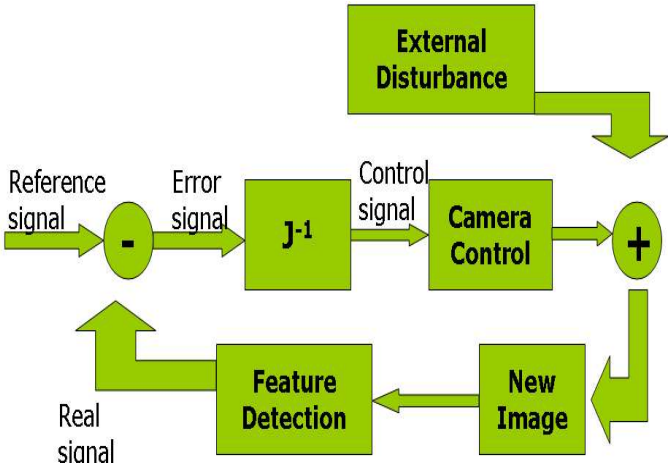

Figure 2: Selective focus-of-attention as visual servo.

in the environment. To solve for the camera parameters, an iterative minimization technique was proposed.

Zhou et al. [4] presented a technique to achieve selective focus-of-attention on-line using a stationary-dynamic camera pair. The procedure involved identifying, off-line, a collection of pixel locations in the stationary camera where a surveillance subject could later appear. The dynamic camera was then *manually* moved to center on the subject. The pan and tilt angles of the dynamic camera were recorded in a look-up table indexed by the pixel coordinates in the stationary camera. At run time, the centering maneuver of the dynamic camera was accomplished by a simple table-look-up process, based on the locations of the subject in the stationary camera and the pre-recorded pan-and-tilt maneuvers.

Compared to the state-of-the-art, our contributions are twofold. In terms of off-line camera calibration:

1. Three pieces of information are needed to uniquely define pan and tilt: position of the rotation axis, orientation of the axis, and rotation angle. Although [2] assumes this general model, it explicitly calibrates only the position and orientation of the axis. Our technique calibrates all these degrees-of-freedom (DOF).

2. Our results show that the iterative minimization technique of [2] is computationally expensive and does not guarantee convergence. Our technique solves for all intrinsic and extrinsic camera parameters for both stationary and dynamic cameras using a closed-form solution that is both efficient and accurate.

3. While the virtual landmark approach in [2] is interesting, we will show that such a technique is less accurate than the traditional techniques using a small calibration pattern (e.g., a checkerboard). We argue that traditional techniques can also provide large angular ranges for calibrating pan and tilt DOFs effectively.

In terms of on-line selective focus-of-attention:

1. In order for the procedure proposed in [4] to work, surveillance subjects must appear at the same depth each time they appear at a particular pixel location in the stationary camera. This assumption is unrealistic in real-world applications. Our technique does not impose this constraint, but allows surveillance subjects to appear freely in the environment with varying depths.

2. [4] manually builds a table of pan and tilt angles, which is time consuming. Furthermore, the process needs to be repeated at each surveillance location, and it will fail if the environmental layout changes later. Our technique does not use such a static look-up table, but adapts automatically to different locales.

We formulate selective, purposeful focus-of-attention as a visual

servo problem. The framework is modeled as a feedback control loop shown in Fig. 2. As mentioned, the stationary cameras perform visual analysis to extract the soft biometry signatures (color, texture, position, and velocity) of the suspicious persons/vehicles. A similar analysis is performed by the dynamic cameras under the guidance of the stationary cameras. Soft biometry features of the subjects (e.g., position and size of a car license plate or the face of a person) are computed and then serve as the input to the servo algorithm (the real signals). The real signals are compared with the reference signals, which specify the desired position (e.g., at the center of the image plane) and size (e.g., covering 80% of the image plane) of the image features. Deviation between the real and reference signals generates an error signal that is used to compute a camera control signal (i.e., desired changes in the pan, tilt, and zoom DOFs). Executing these recommended changes to the camera's DOFs will train and zoom the camera to minimize the discrepancy between the reference and real signals (i.e., to center the subject with a good size). Finally, as we have no control over the movements of the surveillance subjects, such movements are considered external disturbance (noise). This loop of video analysis, feature extraction and comparison, and camera control (servo) is then repeated over time.

5. EXPERIMENTAL RESULTS

We present sample video analysis results of using soft biometry signatures for video surveillance. These examples demonstrate our existing capabilities in camera registration, background modeling, video tracking, multi-camera data fusion, sequence data (i.e., motion trajectories) analysis, and identification using skin tone and clothing colors. These results show that (1) it is possible to automatically analyze video footages to extract soft biometry signatures and use the signatures to assist tracking and identification, and (2) the analysis can be performed on real airport surveillance videos, and real-world problems in airport access control can be facilitated using video analysis and soft biometry. We have used real airport surveillance video footages in analysis tasks that are particular to airport secruity surveillance. We have also used generic video footages to demonstrate our tracking and foreground/background modeling abilities.

Fig. 3 shows the results of using skin tone (one of the soft biometry signatures) for detecting unauthorized "piggy-backing" access patterns. In airports, access to sensitive areas requires authorization and many doors are equipped with an access control system. To unlock a door, a person must produce a special access card for the card reader to scan. In rush hours many airport employees may pass through a secured access door in a short time. Often, one employee might swipe the card to allow multiple employees to enter (piggy-backing); a practice is now disallowed due to tightened airport security. To catch unauthorized piggy-backing access, a security camera is used to monitor the access door. Once the card reader registers a scan, it is desirable, by examining the ensuing video clips, to count how many people have gone through the door before it closes. By correlating how many times access cards are read and how many people are detected in the video, one is then able to assert if piggy-backing access has occurred.

To accomplish automated piggy-backing detection, it is often not enough to just analyze movements in the scene. Our experience with real airport surveillance videos indicated that multiple people can pass through the door in quick succession, resulted in overlapped silhouettes that are hard to separate. Advanced face detection algorithms have also not fared well because the video resolution is low, and hence, a face often occupies too small an image region to be reliably detected. Instead, we have used skin tone detection, coupled with shape (a face region should not be too elongated like an arm region) and location (a face region is often close to the top of a moving region) cues. These results (Fig. 3) demonstrate that soft biometry can be a good compromise between sophisticated face detection and recognition techniques and naive motion detection algorithms (e.g., frame differencing or background subtraction) to achieve robust video analysis at a reasonable computational cost.

Fig. 4 shows the video analysis results of acquiring soft biometry signatures of passengers passing through metal detectors to establish a browsable departure record. Here, background clutter, moving shadows, and the security agents' abrupt and unpredictable movements complicate the analysis. We have observed many times in a 5-minute airport surveillance video clip that the security agents entered the metal detector from both sides, to provide instructions to the passengers waiting in line and to retreat back to his station at the front of the detector. Hence, it is important that the visual tracking algorithm isolates the movements of the security agents and records only when a passenger passes through the metal detector, as shown in Fig. 4.

Fig. 5 shows a sample result of our tracking algorithm that uses soft biometry signatures of clothing colors for tracking multiple people. Sample image frames are displayed from left to right and then from top to bottom and numbered from 1 to 8. Tracked targets are identified by bounding boxes of different colors (yellow color is used to represent unidentified objects due to entering, exiting, and merging). Frames 2 and 4 show temporary occlusion of silhouettes. The occlusion was quickly resolved in frames 3 and 5, when silhouettes no longer overlapped. In frame 6, two of the targets had exited the field of view of the camera, but were correctly reacquired and recognized in frames 7 and 8.

Fig. 6 shows another examples of tracking multiple targets (both vehicles and people) using soft biometry signatures. We have collected hours of video using multiple video cameras in a parking lot. The video frames depicted both human and vehicular motion. The motion patterns for vehicles included entering, exiting, turning, backing up, circling, zigzag driving, and many more. For human motion, we recorded actions involving both individuals and groups, with patterns such as following, following-and-gaining, stalking, congregating, splitting, and loitering, among many others. Some of these patterns (like zigzag driving and stalking) were acted out by our group members, while others represented behaviors commonly observed in the parking lot. Sample results for tracking the movements of people in a parking lot are shown in Fig. 6(a) and (b). Of the three cameras we used, the views of two were partially occluded by parked cars[1]. The individual camera trajectories could therefore be broken. However, by using our data fusion algorithm, we were able to fill in the gap, smooth out sensor noise, and fuse individual trajectories into a complete, global description. Fig. 6(c) and (d) show the analysis of a vehicle's driving pattern when two cameras were used. Note that even with a very small overlap in the fields-of-view of the two cameras and a circling motion covering a large spatial area (hence, each camera observed only a part of the motion trajectory), we were able to fuse the individual camera trajectories to arrive at a complete description.

Fig. 7 shows sample foreground background identification results. Our experience indicated that correct foreground background identification is one of the most critical elements in object tracking. It is customary to model the colors of background pixels as

[1]The camera positions in these figures indicate only the general directions of camera placement. The actual cameras were placed much far away from the scene and always pointed to the parking lot.

Figure 3: Sample results of detecting unauthorized piggy-backing access patterns. In the top row, three legal single person entrance scenarios are shown. In the bottom row, two piggy-backing access patterns are shown. The image on the bottom left shows two face regions detected in a single frame. The two images on the bottom right show two face regions detected in different video frames within a short time interval to trigger piggybacking alarm (the two face regions are marked as 0 and 1, respectively.)

mixture-of-Gaussian distributions—if the background is stationary and the lighting condition is stable. This is often the case for indoor surveillance. Then a foreground pixel (occupied by moving objects) is identified as one with a pixel color that deviates significantly from the pre-established background color clusters. However, in outdoor surveillance scenarios, it is often difficult to distinguish foreground pixels from background pixels based purely on color information. As background pixels often experience large change in color attributes, just like their foreground counterpart. This is the case shown in Fig. 7 that background pixels, such as those depicting (a) ocean, (b) water foundation, and (c) vegetation and shadow, all experienced significant changes in color. Hence, a more sophisticated algorithm is needed to distinguish purposeful object motion from random or periodical movements that often characterize outdoor background. Our algorithm, which uses an image graph model (or random Markov field model) with belief propagation and Bayesian learning, achieves satisfactory results shown in Fig. 7.

For calibrating PTZ cameras, while we use the same governing equation as [2], we have developed a closed-form solution that is much more accurate and efficient. We will illustrate two points:
1. *Theoretically*, under the same simulation conditions, our method produces more accurate calibration results without failure, while convergence cannot be guaranteed in [2].
2. *Practically,* our set up using a traditional calibration mark placed near the camera produces more reliable results than the virtual landmark approach of [2], regardless of the calibration procedure used.

We verify the first claim as follow: We conducted 100 synthetic experiments. In each experiment, we generated 50 3D calibration marks randomly in an $8m \times 8m \times 8m$ volume (similar to the one used in [2]). We projected these 50 landmarks using a synthetic camera that closely mimicked the real-world Sony EVI-D30 PTZ camera. We then applied both calibration procedures, [2] and ours, to estimate the pan and tilt camera parameters using these 50 2D and 3D coordinates. In all simulation runs, we had chosen the initial guess of $\mathbf{T}_p$ and $\mathbf{T}_t$ to be zero, and $\mathbf{n}_p$ and $\mathbf{n}_t$ to be parallel to the CCD's y and x axes. We report the errors in calculating both the axis position and orientation in Table 1 averaged over these 100 runs. For [2], we also recorded the percentage of times the algorithm failed to converge, and if it did converge, the number of iterations needed. As can be easily seen in Table 1 that under the same experimental conditions, our algorithm obtained more accurate results and did not suffer from convergence problem.

The second claim above deserves some explanation. We adopt

Table 1: Comparison of calibration accuracy (For Davis and Chen, 51% simulation runs failed to converge. If the simulation did converge, 85 iterations were needed in average.)

	[2]	ours
Average % error in axis position	51.76%	35.48%
Average error in axis orientation	1.54 (rad)	.22 (rad)

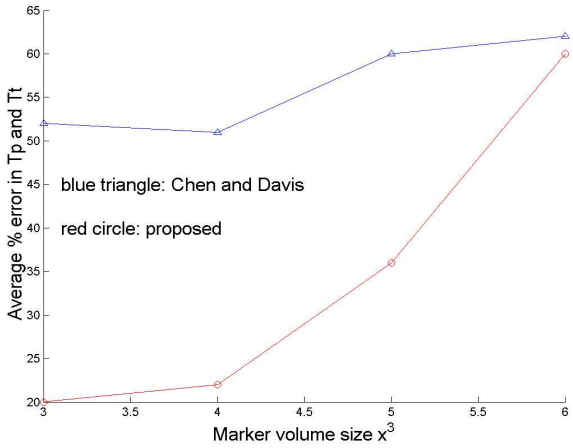

Figure 8: Comparison of calibration accuracy as a function of experimental setup (using a CCD of 300×300).

Figure 4: Sample results of acquiring soft biometry signatures of passengers passing through metal detectors. Snapshots of the passengers were taken and soft biometry information on height and clothing colors was recorded. The left column shows that three passengers passed through the metal detectors and their soft biometry signatures were correctly captured. The right column shows that the analysis program did not confuse the security agents with the passengers (the agents' snapshots and soft biometry were not recorded).

the traditional method of constructing a planar checkerboard pattern and then placing it at different depths before the camera to supply 3D calibration landmarks. While [2] advocates a different method of generating virtual 3D landmarks by moving an LED around in the environment. The argument used in [2] to support the virtual landmark approach is the need of a large working space to fully calibrate the pan and tilt DOFs. While this is true, there are different ways to obtain large angular ranges. Because $\theta \approx r/d$, a large angular range can be achieved by either (1) placing a small calibration stand (small r) nearby (small d) or (2) using dispersed landmarks (large r) placed far away (large d). While [2] advocates the latter, we adopt the former approach.

Our reason is that to calibrate $\mathbf{T}_p$ and $\mathbf{T}_t$ accurately, we want their effects to be as pronounced as possible and easily observable in image coordinates. This makes a near-field approach better than a far-field approach. Another reason is that to provide the same angular calibration range, using the same focal length and CCD, would imply that the CCD's fixed and limited spatial resolution is used to cover either a small spatial range (r) in a near field or a large spatial range in a far field. Hence, the spatial resolution power necessarily becomes poorer when the calibration markers are placed afar. Fig. 8 verifies the calibration error as a function of the volume occupied by the 3D calibration marks. As we shrank down the volume in front of the camera, the error in calibrating $\mathbf{T}_p$ and $\mathbf{T}_t$ dropped for both techniques as expected; however, our techniques outperformed [2] in all cases.

6. REFERENCES

[1] Robert T. Collins, Alan J. Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, N. Enomoto, Osamu Hasegawa, Peter Burt, and Lambert Wixson. A system for video surveillance and monitoring (VSAM project final report). *CMU Technical Report CMU-RI-TR-00-12*, 2000.

[2] J. Davis and X. Chen. Calibrating Pan-Tilt Cameras in Wide-area Surveillance Networks. In *Proc. Int. Conf. Comput. Vision*, Nice, France, 2003.

[3] G. Wu, Yi Wu, Long Jiao, Yuan-Fang Wang, and Edward Y. Chang. Multi-camera spatio-temporal fusion biased sequence-data learning for video surveillance. *ACM International Conference on Multimedia*, November 2003.

[4] X. Zhou, R. T. Collins, T. Kanade, and P. Metes. A Master-Slave System to Acquire Biometric Imagery of Humans at Distance. In *Proceedings of 1st ACM Workshop on Video Surveillance*, Berkeley, CA, 2003.

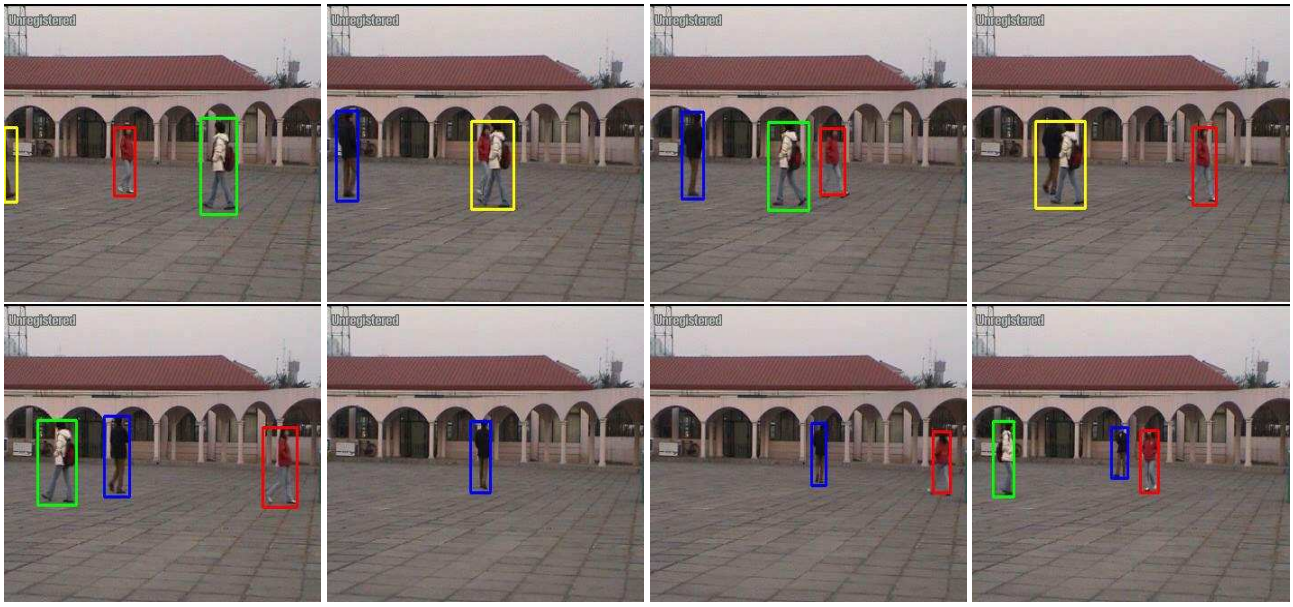

Figure 5: Sample people tracking results using soft biometry signatures of cloth colors. Identified moving objects are enclosed in colored bounding boxes. Yellow bounding boxes represent unidentified objects.

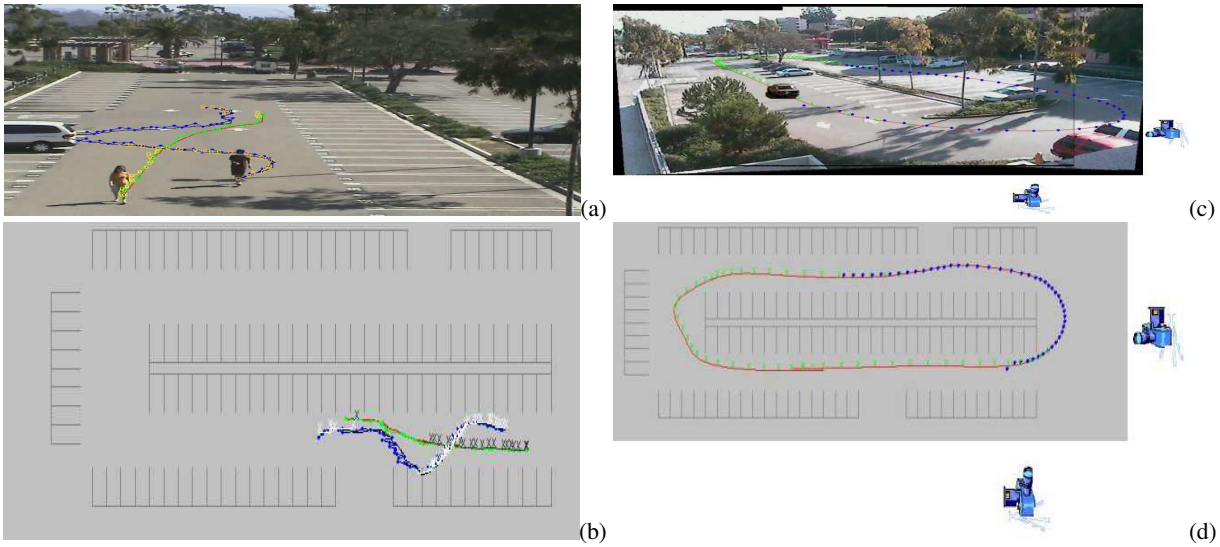

Figure 6: (a) A simulated stalking behavior in a parking lot and (b) trajectories of the sample stalking behavior. (c) and (d): similar data fusion results for vehicular motion. In these figures, the "-" is the fused trajectory; "." is the tracked trajectory from camera 1; "x" is the tracked trajectory from camera 2; and "o" is the tracked trajectory from camera 3.

Figure 7: Sample foreground/background identification results. Pixels marked red are identified as foreground pixels. Even though large color changes due to wave, water, and vegetation movements are observed in these pictures, the algorithm correctly eliminates background motion from consideration.