An intelligent video framework for homeland protection

Peter H. Tu, Gianfranco Doretto, Nils O. Krahnstoever,
A. G. Amitha Perera, Frederick W. Wheeler, Xiaoming Liu,
Jens Rittscher, Thomas B. Sebastian, Ting Yu, and Kevin G. Harding

GE Global Research, One Research Circle, Niskayuna, NY, USA

ABSTRACT

This paper presents an overview of Intelligent Video work currently under development at the GE Global Research Center and other research institutes. The image formation process is discussed in terms of illumination, methods for automatic camera calibration and lessons learned from machine vision. A variety of approaches for person detection are presented. Crowd segmentation methods enabling the tracking of individuals through dense environments such as retail and mass transit sites are discussed. It is shown how signature generation based on gross appearance can be used to reacquire targets as they leave and enter disjoint fields of view. Camera calibration information is used to further constrain the detection of people and to synthesize a top-view, which fuses all camera views into a composite representation. It is shown how site-wide tracking can be performed in this unified framework. Human faces are an important feature as both a biometric identifier and as a method for determining the focus of attention via head pose estimation. It is shown how automatic pan-tilt-zoom control; active shape/appearance models and super-resolution methods can be used to enhance the face capture and analysis problem. A discussion of additional features that can be used for inferring intent is given. These include body-part motion cues and physiological phenomena such as thermal images of the face.

Keywords: intelligent video, surveillance, camera calibration, person detection, crowd segmentation, site-wide tracking, person reidentification, face modeling, face super-resolution, deception detection

1. INTRODUCTION

Increasingly, large networks of surveillance cameras are employed to monitor public and private facilities. This continuous collection of imagery has the potential for tremendous impact on public safety and security. Unfortunately, this potential is often unrealized since manual monitoring of growing numbers of video feeds is not feasible. As a consequence, surveillance video is mostly stored without being viewed and is only used for data-mining and forensic needs. However, the ability to perform computer-based video analytics is now becoming possible, enabling a proactive approach where security personnel can be continually appraised of who is on site, where they are, and what they are doing. Under this new paradigm, a significantly higher level of security can be achieved through the increased productivity of security officers.

The ultimate goal of intelligent video for security and surveillance is to automatically detect events and situations that require the attention of security personnel. Augmenting security staff with automatic processing will increase their efficiency and effectiveness. This is a difficult problem since events of interest are complicated and diverse.

In this paper we present an overview of work we have done for a variety of intelligent video applications. A comprehensive solution for automatic surveillance will use these and other components. Because of the difficulty of the problem, individual components, such as person detection, tracking, reacquisition, biometric identification, and behavior modeling each have strengths and weaknesses in different situations. To develop large scale solutions to automatic surveillance, our approach is to combine a complementary set of algorithms into a system that utilizes the strengths of each component and balances their weaknesses. This will ultimately provide security personnel with the high-level information that they require.

Automatic monitoring of people and interpretation their actions from surveillance imagery is a challenging task for many reasons. The range of appearance of objects such as people and vehicles is large. These objects can occlude

Further author information: (Send correspondence to P.H.T.)
P.H.T.: E-mail: tu@research.ge.com, Telephone: 1 518 387 5838
K.G.H.: E-mail: harding@research.ge.com, Telephone: 1 518 387 5827

one another and they exist in highly dynamic environments. However, before these issues can be addressed, the image formation process itself must be understood. In Section 2, a discussion of various aspects of illumination and its affects on site observations is given along with a discussion of lessons learned from the field of machine vision for automated inspection. The next step is to understand the geometric relationships between the camera and the physical site. This is defined in terms of calibration information, which is composed of camera position, orientation and focal length. In Section 3, an automatic approach to camera calibration is given.

Once the image formation process is understood, the task of detecting certain objects of interest can be considered. In this paper, we focus on the problem of detecting individuals. Recently there has been a wide variety of person detection mechanisms reported in the literature. We assert that a robust system cannot rely on a single approach to this problem. In Section 4, methods based on geometric constraints, machine learning, interest operators and assemblies of parts are presented. As site complexity increases, we cannot expect to observe individuals in isolation. In Section 5, a discussion of a global approach to crowd segmentation is provided.

Given effective person detection, the tracking of individuals over time can be achieved. Using camera calibration information, constraints based on person kinematics can be optimally employed and observations from different cameras can be fused into a single one-world-view enabling site-wide tracking of individuals (see Section 6). However, comprehensive camera coverage may not always be available, hence person reacquisition strategies based on general appearance must be developed (see Section 7).

Going beyond the tracking of individuals, an important aspect of site security is the determination of individual identity. Standard choke-point access control mechanisms are not always feasible. Hence, biometrics at a distance methods must be developed. One of the few biometrics suitable for this task is face recognition. However, it is well known that the quality of the captured facial imagery has a large impact on face recognition performance. In Section 8, methods for the optimal capture and processing of facial imagery at a distance are presented.

The ability to determine individual intent will have a major impact on proactive site surveillance. The observable cues that can be used to address this nascent task include gaze estimation, facial expression, articulated motion and physiological measurements such as thermal gradient measures of the face. Details regarding these types of measurements are presented in Section 9. This paper concludes with a discussion of how these and similar technologies can become the corner stone for our Homeland Protection strategies.

2. ENVIRONMENTAL CONDITIONS

2.1. Site Illumination

The desire today is to deploy video analytics in a wide range of situations and environments, ranging from crowded subways to open-air environments, but this diversity presents unique challenges to intelligent video on many levels. It is important to understand the image acquisition process and the lighting variation issues to address these challenges.

The light level in most offices is on the order of 5 times greater than what might be experienced in a more industrial setting or a subway station. But even the office environment is dark compared to outdoor lighting on even an overcast day, which can be more than an order of magnitude above indoor light levels. Direct sunlight coming in a window can be up to three orders of magnitude (1000) times brighter than typical indoor lighting alone.[1] Moreover, the very nature of how many public areas are lit, using skylights, large floor to ceiling windows, and indirect overhead lights poses many challenges.

Direct sunlight can create shadows that may hide or enhance certain features, whereas diffuse light from a cloudy sky can make skin that is wrinkled look smooth. Light reflected from a nearby object can put a feature on a face that is not there at all, perhaps a thin line of bright glare that appears to be a scar. People are adept at interpreting these phenomena; the same is not always true for machines.

2.2. Lessons from Machine Vision and Industrial Inspection

We have seen this difference in human versus video perception in the application of machine vision technology to industrial inspection problems.[2] In that situation, we attempt to control the lighting, using a low angle directional light to bring out surface pitting or machining marks, or a diffuse in-line light to make surface irregularities disappear so that we can locate a hole on the surface. We have developed a whole science around creating the right lighting in machine vision[3–7] to make the analysis of the image data as simple as possible. Unfortunately, such methods are rarely practical in security situations.

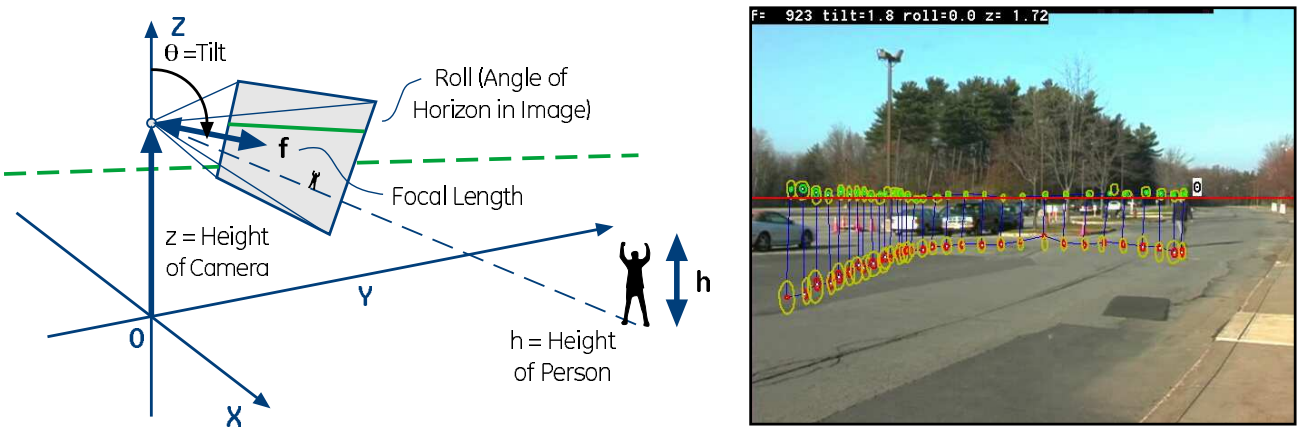

Figure 1. Autocalibration. A visualization of our camera autocalibration approach. Based on detecting and tracking observations of people, the system can calibrate itself automatically, which enables the estimation of 3D scene properties such as target size, target velocity, and distance between individuals. Furthermore it improves the performance of tracking and crowd segmentation algorithms.

Some control over direct sunlight can (and should) be imposed by means of curtains, baffles (portable walls), and the like in some situations such as security check points. However, a person with a shiny plastic bag or raincoat might still direct glare in directions not expected. Early problems encountered in machine vision include the change in the angle of sunlight at different seasons as well as the presence of an operator or inspector wearing a white lab coat. Our eyes are so indifferent to changes in light levels and local hot spots so that we do not even see these variations. But a video camera can be overwhelmed by a four-fold increase in light, and processing algorithms may be deceived by a well defined shadow.

One method that has been explored by both the industrial inspection area as well as some security applications is the use of some degree of three-dimensional data. In industrial inspection, the prominent method to obtain 3D data is to use some form of structured light.[8–10] Light is projected on a part, often in the form of parallel lines of bright and dark areas. Viewing these lines on the part from some other angle allows the use of triangulation on the perceived shape of the lines to infer the shape of the surface. This structured light approach may be usable in some security situation working in the near IR region of the spectrum where people would not see it. But otherwise, highly controlled structured light may again be difficult to use in the wider application of intelligent video applications. Alternatively, the shape-from-*X* approaches, such as,[11] can be employed in a passive manner given sufficient geometric knowledge of the imaging conditions.

3. CAMERA CALIBRATION

In order to apply intelligent video processing over a large collection of cameras, the cameras must be calibrated. The process of camera calibration in general refers to the estimation of a camera's internal (e.g., focal length, aspect ratio, principal point) and external (e.g., location, orientation) parameters. In the context of visual surveillance and content extraction from video, knowledge about a camera's internal and external parameters is useful, as it allows a connection between image and world measurements to be established.[12] To enable convenient calibration, we recently developed a very powerful and robust autocalibration approach[13,14] that uses observations of people to perform the calibration.

Camera calibration in general is a challenging task to perform, as it either requires direct access to the sensor or extensive knowledge of the scene geometry.[15–17] In contrast, camera autocalibration approaches[18–20] exchange knowledge about scene geometry for knowledge of the camera motion or specific scene content. There is also an array of intermediate methods,[21–24] and it is in this category that our approach falls. The advantage of our method over existing work is that it is based on a Bayesian analysis of foot and head location measurements extracted from detections and tracks of people (see Figure 1). Through Bayesian analysis and the use of the recently discovered foot to head homology parameterization of the problem, our approach is able to tackle large amounts of noise and errors in the measurements while still being able to obtain accurate calibration estimates.

4. PERSON DETECTION

In surveillance applications we are primarily concerned with monitoring people as they move about a site, so detection of people in video streams is of primary importance for an intelligent video surveillance system. As discussed in Section 2, automated image analysis is a difficult problem, and detecting people in images is no exception. The appearance of people in unconstrained images varies greatly: people appear in different poses, they are often partially occluded, the lighting is different, etc. A number of approaches for person detection have been proposed in the literature, and we will outline some of them below.

4.1. Background subtraction

Background subtraction[25–27] is the most commonly used approach for object detection when the video camera is static. In this approach, the system builds a model (called the background model) of what the scene looks like, and whenever a pixel does not match this model, it is flagged as foreground (belonging to some object). The background model could be as simple as a static image of the scene when it is known that there are no people around. Practical solutions, however, use more complex models, and update them over time to allow for variations in the scene due to lighting, etc. The foreground pixels so detected form groups of "pixel blobs" corresponding to foreground objects, which may be people, other objects, or noise.[28] Further analysis using size, aspect ratio, etc., classifies the blobs into person or non-person.

When the camera is calibrated, the scene geometry can also be used in detecting people from the foreground pixels. In one approach,[29] we enumerate all possible person locations consistent with the geometry, and classify as a person those locations with sufficient support from the foreground pixels. Here, the scene geometry severely constrains the size and locations of putative windows containing a person, allowing the algorithm to run in real-time.

4.2. Motion segmentation

When the background pixels are constantly changing, it is difficult to maintain a good background model to apply background subtraction. This occurs, for example, when the camera is mounted on a moving platform, or when the background is extremely dynamic, such as in a traffic scene. The most common approach in this case is frame differencing.[30,31] Here, one takes two video frames that are separated by some fixed time (one second, for example), compensates for the overall camera motion using frame-to-frame registration, and subtracts the two frames. The idea is that the non-moving pixels in the two frame pixels will cancel each other out (because of the camera motion compensation), and thus large differences correspond to moving objects, which can be flagged as foreground pixels. However, the foreground pixels from frame differencing are far noisier than those from background subtraction. For example, with frame differencing often only the leading and trailing edge of an object is detected; and there are "ghost" pixels and objects. Compared to background subtraction, frame differencing requires far more complex analysis to obtain reasonable detections,[30] or requires support from higher-level processes.[32]

A different approach to motion segmentation is to detect interest points on a frame and track those points across frames, thus computing temporal trajectories for the points. The concept is that interest points that fall on a single object will have similar trajectories, and so that grouping the trajectories segments and groups the interest points into different objects.[33–35]

4.3. Single image person detection

Instead of relying on temporal cues, one can also attempt to detect people in a single image. This is far more difficult and computationally intensive, since it requires strong models describing how people appear in images. One advantage of such algorithms is that they tend to be easier to tune to a particular scene, and are generally more robust to the changes in the so-called nuisance parameters (lighting, color of clothing, and so on). The fact that this class of algorithms does not directly utilize temporal information can be an advantage: the algorithms can be directly applied to moving platforms, pan-tilt-zoom cameras, and highly dynamic scenes, without complex modeling of camera motion and pixel variation. Furthermore, it is relatively easy to add temporal consistency checks on top of the detection algorithms. For example, one could ensure that if there is a detection on one frame, a detection also occurs in roughly the same place on the next frame.

The single image detection approaches can be categorized into two broad groups: monolithic detection, where the entire person is detected; and parts-based detection, where the presence of a person is inferred by detecting constituent parts.

4.3.1. Monolithic detection

Broadly, the goal of monolithic detection is to analyze the pixels in a window occupied by a person to generate a feature vector which is different from that which would be generated had the pixels not been occupied by a person. To be invariant to many of the nuisance parameters, these features try to encode the overall shape of the object within a hypothesized detection window. The feature space is designed to be rich enough that the feature vectors generated by a detection window containing a person (in any acceptable pose) is different enough from those generated by a detection window not containing a person, that a classifier can then be trained to distinguish person from non-person. Detecting people is achieved by iterating over all possible detection windows, testing for the presence or absence of a person.

Some specific approaches to monolithic person detection include matching silhouette templates to image edges[36] and using dynamic point distribution models.[37] A recent, highly successful approach[38] has been to use histogram of gradients (HoG) over the detection window to generate the feature space, and use a Support Vector Machine[39] (SVM) to classify the features. HoG features and a cascade of classifiers has also been used to obtain real-time person detection with excellent detection performance.[40]

4.3.2. Parts-based detection

In contrast to monolithic detection, parts-based detection uses a set of detectors, with each detector being tuned to a particular body part.[41] This implies a detector for heads, for torsos, for arms, and so on. The idea is that although the body as a whole has a large number of poses, each individual part has far fewer valid poses. For example, the torso could be described by a rectangular blob regardless of whether the person is standing, running, jumping, bending over, and so on. Each part detector is simpler and more robust than a whole body detector because of the reduced variation.

Given a set of body part detectors, a constellation model encodes the constraints between the locations of these parts. For example, the legs generally appear below the torso, but not too far below. The constellation model is often a probabilistic model to allow for noise in the image and errors in the part detectors. The model parameters are learned from a set of training images containing people in various poses. The person detection process is then based on running the part detectors over the whole image, and using the constellation model to evaluate which subset of detections are in a configuration that could reasonably arise from a single person. An advantage of the parts-based approach over the monolithic detection is that it can be more robust to partial occlusions, provided the high-level constellation model is able to reason about occlusions and missing parts.

Instead of developing specific detectors for the various constituent parts, an alternative is to also learn the parts.[42] The concept here is to run a series of interest operators, and learn where they tend to fire relative to the center of a detection window. This is used to estimate both the probability of a particular operator appearing at a particular location when there is a person in the detection window, and the probability of it appearing when a person is not in the detection window. By combining these probabilities for multiple interest points, one can derive a probability that a person is in the detection window and a probability that a person is not in the detection window. A thresholded likelihood ratio of these two probabilities is then a person detector.

4.3.3. Machine learning

Both monolithic and parts-based detection rely heavily on learning the parameters of the classifiers that provide the detections. There is a large body of work on machine learning for computer vision,[43,44] and much of it has been applied to person detection, and more generally, to object detection. Some of the more common approaches are: support vector machines, neural networks, cascaded classifiers, and AdaBoost.

Support vector machines are the basis for an approach to solve a non-linear classification problems by lifting the feature vectors into a higher-dimensional space such that the classification in the new space is linear. The approach is so-named because the feature vectors along the boundary are the “support vectors” that are used to define the boundary; the classifier is defined entirely by the support vectors and the lifting function.

Neural networks attempt to approximate the classification function through a series of layers, where the output of each node in a given layer is a non-linear function applied to a weighted sum of the nodes in the previous layer. Neural networks have been successfully applied to a variety of problems. The main difficulty in neural networks is in choosing the number of layers and the number of nodes in each layer.

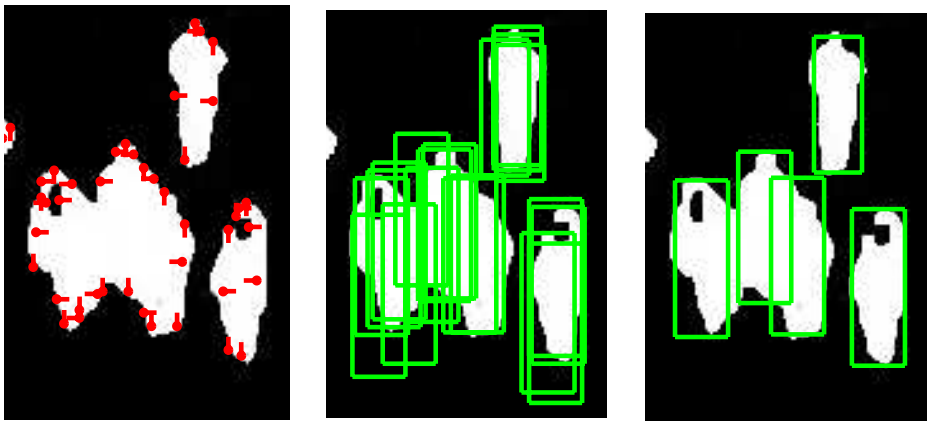

Figure 2. Image Features, Groupings, and Segmentation. The feature extraction, the set of possible groupings, and segmentation results are shown for one example image. A standard probabilistic background model was used to segment the image foreground. The set of image features illustrates that each feature is labeled as being at the top side or bottom of a person. The set of groupings illustrates that there may be a large set of possible groupings. The segmentation on the right shows the final segmentation.

A cascade of classifiers is a binary tree of classifiers such that each lower-level node improves upon the decision made on the path above it. The tree is often not balanced, giving the ability to quickly reject false hypothesis at the high-level nodes.

A cascade combines classifiers using only their decisions: the decision output of the classifier at each node is used to follow one of the two paths. In contrast, AdaBoost is a technique for combining directly the responses of multiple weak classifiers into a single, strong classifier. The strong classifier response is a weighted combination of the weak classifier responses that gives the best classification rate. The AdaBoost algorithm ("boosting") is an efficient way of determining these weights by considering the relative strengths of one weak classifier over another. Unlike a cascade, each weak classifier output is used every time to generate the strong classifier output.

5. CROWD SEGMENTATION

Although sophisticated person detectors such those previously described have been developed, they usually assume that people are well separated. A segmentation of the scene into individuals must be performed so that crowd activity can be characterized and disorderly events identified. Previous works such as[45] have used mechanisms such as head detectors to segment crowds, but these approaches tend to fail when the features cannot be directly observed. Another approach is to perform local feature grouping. Two examples of this method are Song et al.[46] and Mikolajczyk et al.[41] Since these methods do not consider all the data simultaneously, difficulties can occur when there is high ambiguity associated with the local context, such as in the center of dense crowds. For this reason, a global optimization may be desirable. Elgammal et al.[47] assumed prior knowledge regarding the number of people and their appearance, and used exhaustive local search to find the solution. Zhao et al.[48] made no assumptions about the people in the scene, and used Markov Chain Monte Carlo (MCMC), a form of random search, to perform their optimization, a process that is computationally expensive and sensitive to initialization. In our approach,[49,50] crowd segmentation is achieved using a variant of Expectation Maximization. The approach is efficient, insensitive to initialization, and requires no prior knowledge regarding the number of people in the scene. It operates in real-time in fixed camera CCTV surveillance systems.

Our crowd segmentation algorithm starts with a set of segmented foreground patches defined by the silhouette of the crowd. For each foreground patch, a set of local image features based on the outline of the silhouette are extracted. An accurate segmentation of individuals is achieved by grouping the image features corresponding to each individual. The first step in this process is to hypothesize all possible groupings, using knowledge of the camera geometry and the potential shape of people facilitates as constraints. Initially, a given image feature may be assigned to multiple groupings. An algorithm based on expectation maximization (EM) is used to find the optimal assignment of each image feature to at most

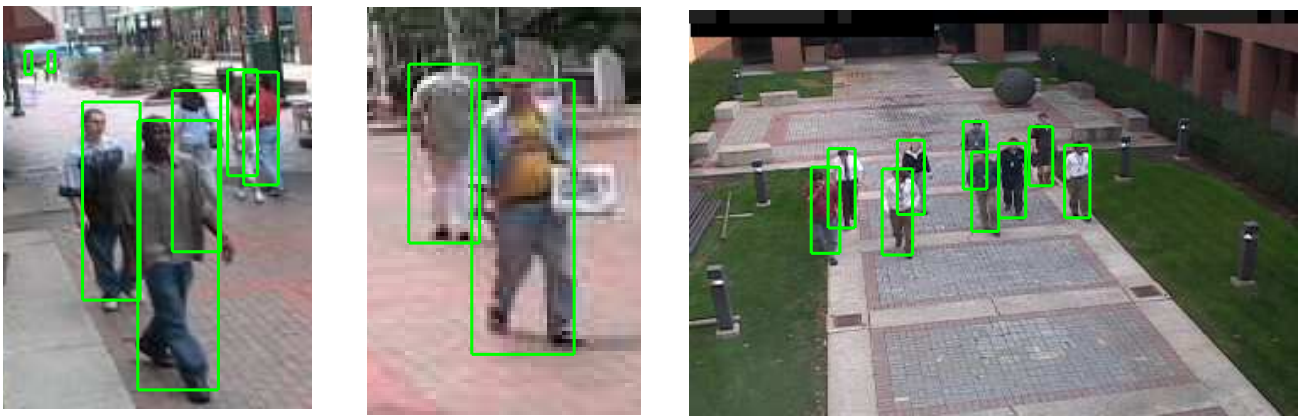

Figure 3. Challenges. Varying scale, clutter, shadows, as well as partial occlusion make these examples particularly challenging. Note that the algorithm generates plausible results.

one grouping. The final segmentation is based on the groupings that have a significant number of assignments. Figure 2 illustrates this process and Figure 3 shows a number of example results.

6. SITE-WIDE TRACKING

Now that we have established methods for jointly calibrating a set of cameras, and can detect persons in video, our next step is person tracking. In this section we will describe the basic problem of person tracking, and approaches that we and others have developed to solve this problem. For now, we will focus on tracking a person who is continuously visible through a set of networked cameras. In the next section we will discuss an approach to handle gaps in coverage.

Many tracking systems, most notably template based approaches, separate the tracking problem into two tasks: track initialization followed by a track update step that involve local search in the prediction gate of the tracks. In contrast, many traditional tracking systems outside of the computer vision community, due to the nature of the sensors used, separate the tracking problem into three steps: detection, data association and track state update. In such approaches, tracks only utilize the data that is associated to them to update their internal state. The robust detection of targets is necessary for such traditional tracking approaches to succeed. The recent progress in the real-time detection of faces[51] and people[38] as well as the robust detection of people in crowded scenes[42,48,50] have facilitated the use of such traditional tracking approaches. One such system is outlined in,[29] where a fast person and object detection algorithm supplies a tracking module with a list of detections at every frame. The detections contain information about the class of the target (e.g., 'person', 'object' etc.), its location and location uncertainty in the image. In addition, the detector provides sufficient information to (i) project the location information onto the ground plane and (ii) to recover information about the physical height and width of targets. An approach to do this is outlined in,[13] where a person detector supplies bounding boxes for people in the image, based on which foot and head location estimates are obtained. This information, in conjunction with the metric projection matrix of the camera can be used to project the location onto the ground plane as well as to obtain the physical dimensions for each detection. This approach works well when people occur in isolation, but breaks down in situations where people occur in groups or are only partially visible in the image (due to the image borders). This is a major challenge for many practical applications. Hence, in the approach outlined in[29] detections are projected into the ground plane and supplied to a centralized tracker that processes the locations of these detections from one or more camera views. At this stage the tracking of extended targets in the imagery has been reduced to tracking 2D point locations in the ground plane, which can be performed very efficiently. Detections are processed by the following stages:

Track Prediction - The location for each track is predicted forward in time according to its current state and its dynamical model. The time stamp of the currently processed detection batch determines how far forward in time the prediction is performed.

Data Association - Each track in the set of currently active tracks is assigned to at most one detection using the Munkres algorithm.[52] The distance between tracks and detections is measured using the Mahalanobis distance where the covariance given by the sum of the current track gate, the uncertainty of the detection and a base uncertainty. The Munkres algorithm

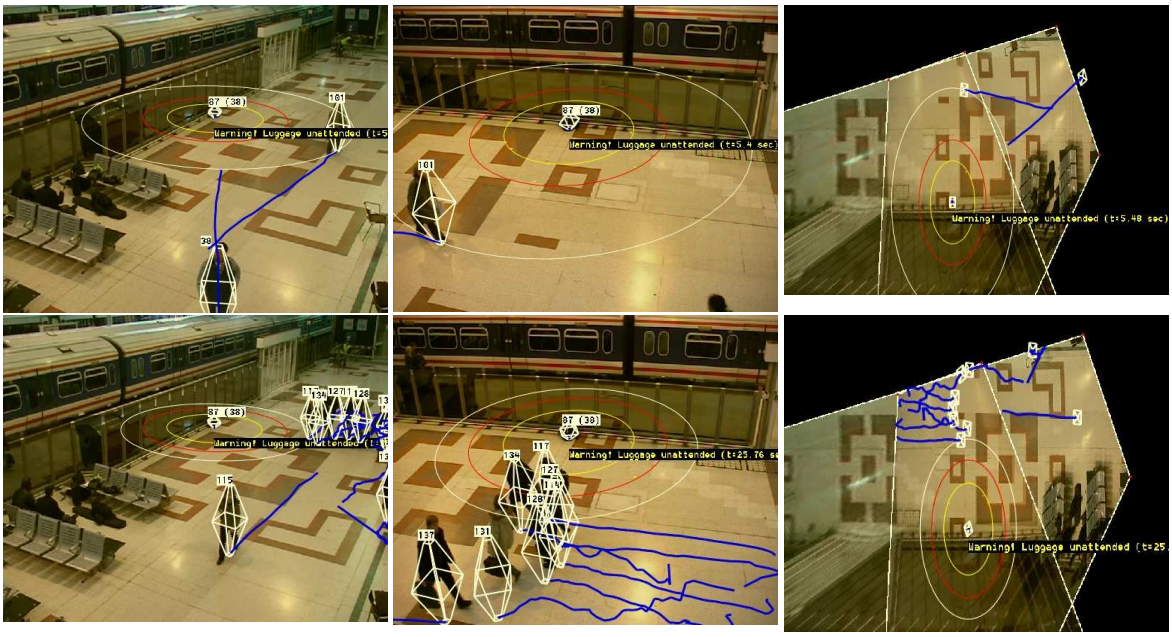

Figure 4. Tracker. Example of a tracking system tracking multiple targets from multiple views (from[29]). The right images show virtual top-down views of the site activity.

obtains the optimal assignment between tracks and detections under this distance measure. Tracks that are too far away from their assigned detections are considered to be non-associated. The success of this general nearest neighbor approach to tracking[53] depends on the performance of the detector and the clutter probability. More sophisticated approaches such JPDAF or MHT can also be utilized.

Track Update - After association, tracks are updated according to their assigned observations. If a track was assigned to any observation, the update step is performed with a virtual observation that has infinite uncertainty, amounting to an update that does not correct the predicted location but increases the uncertainty of the state estimate. Track states are maintained using extended Kalman filters with suitable dynamical models (e.g., a constant velocity turn model described in[53]).

Track Maintenance - Tracks are marked for deletion if the state uncertainty becomes too large, if the track goes out of view (of the entire camera network) or if it has not been associated with a detection within a certain time window. Upon deletion, a determination is made as to whether the track was a false alarm, based on several criteria involving the lifetime of the track and its motion pattern.

Track Formation - Each detection that is not associated with an existing track leads to the formation of a new track if its spatial dimensions (height, width) passes a number of tests designed to limit the number of spurious tracks that are created. Steps are in place for detecting and patching 'ghosts' in the foreground image, created by targets that have been assimilated or initialized into the background.

The above described tracking system constitutes a generalized nearest neighbor tracker.[53] It is computationally very efficient and hence suited for tracking a large number of targets in many camera views simultaneously. If accurate and persistent target tracking (even in dense groups and crowds) is desired, more sophisticated and computationally more expensive approaches such as JPDAF,[54] MHT[55] or Bayesian multi-target trackers[56] can be employed.

Figure 4 shows the tracker in operation on sequence S1 of the PETS 2006 dataset. The tracker deals well with isolated targets as well as with crowds. It should be stressed that the availability of multiple calibrated camera views helps greatly in constraining the target tracking process in this work.

7. PERSON REIDENTIFICATION

In the previous section we considered the problem of person tracking when the subject is continuously within the coverage region of a network of cameras. When subjects move in and out of the coverage region, or move between camera coverage regions we will generate disconnected tracks. In this section we consider the person reacquisition problem, which is to link those disconnected tracks based on the appearance of the subject.

Many applications require the ability to reidentify an individual across multiple disjoint fields of view. Among the approaches that use passive biometrics such as face[57] and gait,[58] here we focus on reidentification algorithms that rely on the overall appearance of the individual. An appearance-based algorithm must deal with several challenges such as: different camera angles and illumination conditions, variation in pose and the rapidly changing appearance of loose or wrinkled clothing. However, we make the standard assumption that individuals do not change their clothing between sightings. This is reasonable for many applications such as airport and subway surveillance. In such scenarios, temporal reasoning and spatial layout of the different cameras can be used for pruning the set of candidate matches.

Several approaches have been proposed where invariant signatures based on the global appearance of an individual are compared. In[59] a color histogram of the region below the face (found by a face detector) serves as the signature for comparison. See[60] for a related approach using clothing color descriptors. Recently, the brightness transfer functions between different cameras have been used to track individuals over multiple non-overlapping cameras.[61,62] It has been shown that the brightness transfer functions lie in a low-dimensional subspace, and can be learned using a set of corresponding calibration objects.[62] Reidentification is then achieved by comparing the adjusted color histograms.

In contrast to the global appearance based methods previously discussed, recent advances in object recognition have demonstrated that comparing multiple local signatures can be effective in exploiting spatial relationships and achieving some robustness with respect to variations in appearance.[63,64] The key to this methodology is the ability to establish correspondences between objects. Two approaches that are successful in this regard are interest point operators[64,65] and model fitting.[66]

There are two aspects to the person reidentification problem. First, one needs to establish correspondences, i.e., determine which parts of one image should be compared to which parts in the second image. Second, one needs to generate invariant signatures for comparing the corresponding parts. In[67] it is shown that dealing with this two issues leads to improved person reidentification, also because the appearance variation due to articulation is inherently addressed.

7.1. Interest Operator and Model Fitting Reidentification

In this section we give an overview of,[67] where two person reidentification approaches are presented. The first one uses interest operators, whereas the other one exploits model fitting, for establishing spatial correspondences between individuals.

The first approach has to deal with the problem that the responses of many interest point operators will not persist over extended periods of time due to the dynamic nature of the appearance of a person.[68] This issue is addressed by using an operator that generates a large number of responses in regions with high information content, thus increasing the probability of establishing true correspondences between images of the same individual. The Hessian affine invariant operator[68] is used for this purpose. Signature matching is used to establish correspondences between two sets of interest points. A match score is computed based on the cardinality of the final set of correspondences.

In contrast to the interest point operator approach which generates a large number of potential correspondences, model-based algorithms establish a mapping from one individual to another. In[67] a decomposable triangulated graph[69,70] is used to model the articulated shape of a person. A dynamic-programming algorithm is used to fit the model to the image of the person.[69,70] Model fitting localizes different body parts such as arms, torso, legs and head, thus facilitating the comparison of appearance and structure between corresponding body parts.

We now describe how the invariant signatures for comparing different regions are generated by combining color and structural information. The color information is captured by histograms based on hue and saturation. Some invariance to differences in ambient illumination is achieved via normalization. Unlike most rigid objects, the structural appearance of loose fitting or wrinkled clothing on perambulating individuals is highly dynamic. Hence, the application of a traditional edge operator[71] will produce many spurious edges corresponding to wrinkles and folds in clothing. To address this issue, a spatio-temporal segmentation algorithm that generates salient edgel information is applied to the imagery. The watershed

Figure 5. Person reidentification. The left and right groups show the top ten matches to six query images using the model-based algorithm. The query image is shown in the left column, and the remaining columns are the top matches ordered from left to right. A box is used to highlight when a match corresponds to query. Third row of the left group shows an example where the correct match is not present in the top ten matches.

algorithm is used to generate an over-segmentation of each frame. A spatio-temporal graph is then generated by treating each region as a node and placing edges between spatially and temporally adjacent regions. A graph partitioning algorithm that models each cluster as a minimum spanning tree is then used to generate salient edgels corresponding to the boundaries of each type of clothing. Finally, the region signatures are then augmented with local histograms of these salient edgels. Figure 7.1 shows some matching results for the model-based approach.

8. FACE SHAPE MODELING AND SUPER-RESOLUTION

For many real-world law enforcement surveillance applications, automatic face recognition at a significant standoff distance is highly desirable. Certainly, in the type of comprehensive automatic surveillance system we are considering here, it would be beneficial to determine the real identity of tracked individuals, and to compare them against a given watch-lists. However, the performance of existing face recognition systems is often inadequate at surveillance distances, primarily due to the low-resolution of the subject probe images.[72] This section describes methods that we have developed to apply multi-frame image super-resolution to video of faces to improve the accuracy and extend the range of commercial face recognition systems. In our initial work in this area, we have shown that even combining multiple facial images through registered averaging followed by restoration with a Wiener filter can improve the performance of face recognition from low-resolution video.[73]

Image super-resolution is the process of using multiple images or video frames of the same object or scene to estimate one image of superior resolution.[74–78] Quality improvement can come from noise reduction through averaging, deblurring, and de-aliasing. The image formation process, including face motion, camera Point Spread Function (PSF), and sampling, is modeled for each frame. Finding the super-resolved image that is consistent with each of the input video frames is then a matter of solving a constrained optimization problem.

A prerequisite to super-resolution is accurate image registration. In general it is best to use a registration formulation that can accurately model the actual frame-to-frame motion, with no additional freedom. With this in mind we use an Active Appearance Model[79,80] (AAM) for face registration. The AAM is designed specifically for the shape and appearance of the face. Previous work in face super-resolution has used optical flow for registration,[81] or shift-only registration.[82] Optical flow can certainly model facial motion, but it is computationally complex and its generality brings the risk of over-fitting. Other face super-resolution results by Baker[82] are quite impressive, but much of power of this approach comes from the prior model of facial appearance. This brings the risk of hallucinating, i.e., reconstructing visible facial features not justified by the actual data.

An important feature of our overall approach is the face-specific methods used for frame registration, and the data-driven methods used for super-resolution, to avoid reconstructing features not justified by the data. Since our intended application is forensic analysis, we must be careful to not introduce additional information through aggressive use of a prior model of facial appearance during super-resolution processing. In the remainder of this section, we will go into more detail on the face registration and super-resolution process and show some example results.

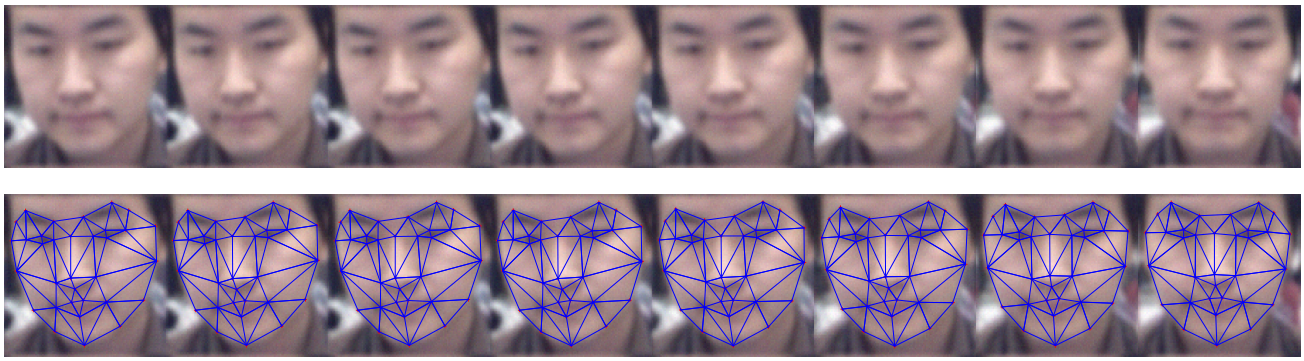

Figure 6. Face shape fitting. Faces from 8 consecutive video frames and the fitted AAM shape model. The appearance part of the model is not shown.

8.1. Active Appearance Model

An Active Appearance Model applied to faces is a two-stage model of both facial shape and appearance designed to fit the faces of different persons at different orientations. The shape model describes the distribution of the locations of a set of landmark points. The 33 feature points used in this work can be seen as the triangle vertices in the fitting example of Figure 6. The shape model is trained using a set of about 500 images from the Notre Dame Biometrics database Collection D[83,84] on which the feature point locations were found manually. Principle Components Analysis (PCA) and training data are used to reduce the dimensionality of the shape space while capturing the major modes of variation across the training set population.

The AAM shape model includes a mean face shape that is the average of all face shapes in the training set and a set of eigenvectors. The mean face shape is the canonical shape and is used as the frame of reference for the AAM appearance model. Each training set image is warped to this frame of reference, so that all faces are normalized as if they had the same shape. With shape variation now removed, the variation in appearance of the faces is modeled in this second stage, again using PCA to select a set of appearance eigenvectors for dimensionality reduction.

The complete trained AAM can produce face images that vary continuously over appearance and shape. For our frame-to-frame registration purposes, the AAM is fit to a new face as it appears in a video frame. This is accomplished by solving for the face shape and appearance parameters (eigen-coefficients) such that the model-generated face matches the face in the video frame using the Simultaneous Inverse Compositional (SIC) algorithm.[80] Figure 6 shows an example of AAM fitting results for video frames. The AAM used in this work[85] has two significant additional features: it is multi-resolution so the AAM appearance model resolution is kept close to the actual video frame resolution; and the model is iteratively refined during training, significantly reducing fitting time and making fitting more robust to initialization.

The AAM provides the registration needed to align the face across the video frames. The landmark positions are the vertices of 49 triangles over the face as seen in Figure 6. 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 triangle vertices.

8.2. Multi-Frame Super-Resolution

To super-resolve faces, we adapt the robust method of Farsiu et al.[86] which models the image formation process and does not rely on a facial image prior, thus avoiding hallucination.[82] As is typically done for super-resolution methods, we will describe the algorithm using standard notation from linear algebra, assuming each image has all of its pixel values in a vector. In the actual implementation, the solution process is carried out with more practical operations on 2D pixel arrays.

The super-resolution process uses an image formation model relating each of the input frames to an unknown super-resolved image, which has about twice the pixel resolution of the input frames. The image formation process accounts for the face motion using the AAM, camera blur, and detector sampling. This model maps the unknown super-resolved image to generate images that match each of the input images. The difference between any input image and the corresponding

(a) Original Video Frame (b) Wiener Filter (c) Super-Resolution

Figure 7. Super-resolution. Example original video frames (a), Wiener filter results (b), and super-resolution results (c) with enlarged views of the left eye. Only the facial region is enhanced in (c); the background is taken from a single input frame. The Wiener filter results in (b) show considerable amplification of interlacing artifacts.

generated image indicates how consistent the super-resolved image is with that image. A steepest descent optimization then finds the super-resolved image that is simultaneously consistent with all of the input images.

When the observed video is color, super-resolution processing is applied to the luminance component only. The initial image is converted to the NTSC color space (YIQ), and the luminance (Y) component is computed for all input frames. The super-resolved luminance result is combined with the chrominance components from the initial image. In practice, we have found this to yield good results, without color distortion, and feel it is justified considering the eye's limited sensitivity to resolution in the chrominance components.

To solve for the super-resolution image, it is first set to an initial image generated by warping and averaging each frame. Then, as is done in[86] for ordinary non-facial images, a steepest descent search using the analytic gradient of the cost function with a fixed number of iterations is used. Only the face region is registered by the AAM, so the background region of the reconstructed image has no data constraints. Since it is initialized to a reasonable starting point, this causes no problems. After optimization, the background region is replaced by blending the super-resolved face with the background from a single input frame. Figure 7 shows example super-resolved faces from a PTZ surveillance video camera.

9. DETERMINATION OF INTENT

Every day, hundreds of thousands of people pass through airport security checkpoints, border crossing stations, or other security checks. Through countless interactions, security professionals must ferret out high-risk individuals who represent

a danger to other citizens. During each interaction, the security professional must decide whether the individual is being forthright or deceptive. This task is difficult because of the limits of human vigilance and perception and the small percentage of individuals who actually have hostile intent. Security personnel cannot halt the flow of people and material to extensively gauge the truthfulness of every interaction, so homeland protection would greatly benefit from automatic techniques to identify deception and ill intent.

An ideal homeland protection system should be able to detect ill intent as early as possible in order to enable proactive action by security professionals. This means that we should seek approaches that are effective in unconstrained settings, require no subject cooperation, and work at a standoff distance, which is a tall order. On the other hand, one could think of integrating several capabilities, and determine the intent of a person by fusing the information coming from a pool of sensors/subsystems. For instance, millimeter wave sensors have been developed to detect person-borne weapons, or explosives,[87] and other sensors have been developed to detect chemical, biological, and radiological/nuclear weapons. Among all the sensors, if there was one that could analyze the temporal variation of the emotional status of individuals, one would be able to tell whether the analyzed subject is deceptive, and therefore dangerous.

The classical approach to deception detection is achieved through the temporal analysis of vital signs acquired during a confrontation session. This is the polygraph test, which aims at performing a micro behavior analysis of the subject. It has the main disadvantage of being an invasive technique, and requiring a controlled environment. The former disadvantage could be removed if it was possible to acquire the vital signs with a wireless sensor. It turns out that this is possible by means of a thermal camera, and in Section 9.1 we are going to describe this promising approach.

Automated ill intent detection can also be achieved by means of a variety of potential behavioral indicators of deception. For example, it has been shown that the analysis of micro-momentary facial expressions can reveal emotions, and in particular deception.[88] Although this is a non-invasive technique which requires a less constrained environment, it is not the most flexible in that it requires the analysis of unobstructed, high-quality video of the face. Rather then examining micro behavior, from video it is easier to analyze macro behavior, which also requires less subject cooperation. For instance, by analyzing the video of the interview of a criminal suspect for movement patterns, and comparing it to known deceptive and truthful subjects, one could potentially gain insight into the honesty of the person. We are going to articulate more about this approach in Section 9.2.

9.1. Thermal Imaging for Deception Detection

It has been shown by Pavlidis[89] that the analysis of thermal video of the face of an individual allows extraction of a physiological signature directly associated with his/her stress levels. This physiological response could be considered part of the “fight or flight” syndrome triggered by the autonomic nervous system, whereby blood redistributes peripherally towards musculoskeletal tissue. Experimentation has demonstrated that during stress produced by startle stimuli in the lab, subjects exhibited elevated blood perfusion in the orbital muscle area (called periorbital region), which resulted in localized elevated temperature. Such a heat signature can be captured by a highly sensitive thermal imaging system and analyzed using pattern recognition methods. Based on this principle, Pavlidis has developed an imaging system for quantifying stress during polygraph examinations. A comparison between this and the traditional polygraph techniques, performed by the Department of Defense Polygraph Institute, has later revealed that the accuracy of the two modalities is equivalent (around 80%).

Recent developments by Pavlidis have shown that through the use of sophisticated computer vision algorithms, it is possible to maintain good performance while releasing many of the environmental assumptions previously applied. More precisely, the improved system takes advantage of an accurate tracker to detect, follow and extract the periorbital region of the face of the individual under examination. This important feature allows the subject to move naturally, provided that he stays in front of the thermal camera. Also, an ad-hoc designed segmentation algorithm allows to extract the right subpart of the periorbital region that is then used to compute the thermal signature. The main advantage of this technique is that it is non-invasive, and coupled with sophisticated computer vision and pattern recognition algorithms, holds the promise of performing ill intent detection at a distance in uncontrolled scenarios.

Figure 8 shows the thermal pattern change in the periorbital region of two subjects undergoing a sustained stress test, where the subjects were asked to compute a series of subtractions. The temperature increase is consistently visible regardless of the subject, and is caused by the positive gradient of superficial blood perfusion, that in the periorbital region is very noticeable due to the high concentration of superficial blood vessels.

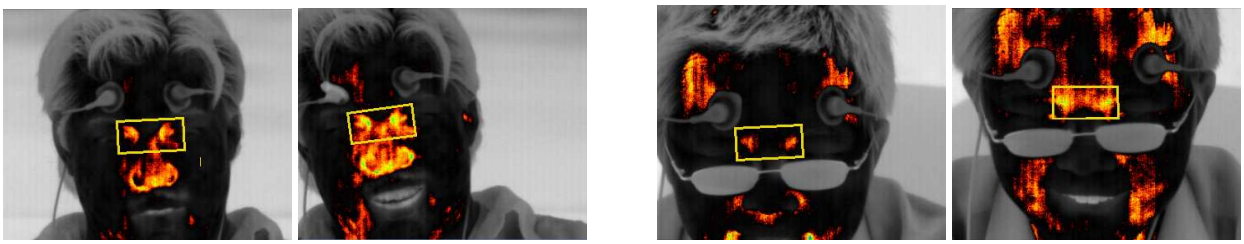

Figure 8. Thermal patterns. The left and right groups of thermal images show the facial thermal patterns of two subjects performing a sustained stress test. For each group the left image was taken before starting the test, whereas the right image was taken towards the end of the test. The rectangle highlights the periorbital region. The images were acquired by the ATHEMOS system.[90]

9.2. Movements and Behavioral Patterns

In[91] the authors describe a behavioral approach to deception detection, which is appealing because it can be used unobtrusively without the cooperation of the subject. The idea is to extract a set of features from head and hands movements in video, and infer deception or truthfulness from them. Such a system could potentially have great impact in augmenting human abilities to assess credibility.

This approach, which is based on the automatic analysis of gesture from video, is motivated by the fact that researchers have found that deceivers, in an attempt to retain credibility and deflect suspicion, express patterns of atypical behavior.[91] It has been observed that deceivers often suppress the normal gestures that accompany interaction and appear over-controlled. Moreover, when they do move, the movement tends to be abrupt. Truthful subjects, on the other hand, maintain more smooth and congruent presentations. The discrepancies between the signatures of features extracted from movement on video can be quite telling.[91]

Similarly to gesture analysis, gaze and gait analysis[92] could be used to augment the behavioral pattern description. From the computer vision point of view, the most difficult problem would be to extract particular configurations of gaze direction, gait, and gesture. In order to extract all this information in a robust manner, an articulated person model could be fit to image measurements, similar to what has been described in Section 7 to extract person descriptors. The temporal variation of such model would then enable the extraction of proper signatures that could subsequently be analyzed using pattern recognition techniques.[43]

10. CONCLUSIONS

The previous sections are evidence of the progress that has been made in the field of intelligent video. However, from a homeland protection point of view, significant adoption of these technologies has yet to occur. The question that must be addressed is how and when will truly intelligent surveillance systems arrive? Some predict that intelligent video will soon be reduced to a commodity technology, with algorithms such as person detection and face recognition directly embedded on many surveillance cameras and that this will be the trigger needed for wide scale acceptance of intelligent video. From a network bandwidth and cost perspective, embedding is certainly a very attractive option. However these capabilities are just tools and in isolation may not have the power needed to address the complex demands of homeland protection.

For homeland protection, and for many other domains, it is desirable to be able to easily deploy intelligent video systems that do not require adaptation to specific sites or scenarios. This approach may be successful for a number of constrained applications, however we assert that it is still too early for universal adoption of this strategy. A major challenge is to achieve a state where algorithms degrade gracefully and do not become overwhelmed by circumstances not anticipated by system developers. Systems must be able to capitalize on what can be done and compensate for what can't be done. This requires a comprehensive evolutionary system of systems approach which can adapt to challenging, unanticipated site conditions and ever changing user needs. In this paper we have outlined a number of intelligent video sub-systems that we believe will contribute to the foundation of such a system.

ACKNOWLEDGMENTS

Face restoration work described in Section 8 was supported 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. RCA, Harrison, NY, *RCA Electro Optics Handbook*, 1974.
2. K. Harding, "Optical considerations in machine vision," in *SME Machine Vision Workshop series, 1985-. Included Machine Vision Capabilities for Industry*, N. Zuech, ed., pp. 115–151, SME, 1986.
3. M. P. Coletta and K. Harding, "Lighting science - tools to guide machine vision applications," in *SME Vision*, (Chicago), 1989.
4. K. Harding, "Light source models for machine vision," in *SPIE Conf. Optics, Illumination and Sensing for Machine Vision*, Svetkoff, ed., (Philadelphia), Nov. 1989.
5. G. T. Uber and K. G. Harding, "Illumination and viewing methods for machine vision," in *Opcon 90*, (Boston), November 4–9 1990.
6. R. F. Rauchmiller Jr., K. G. Harding, M. A. Michniewicz, and E. A. Kaltenbacher, "Design and application of a lighting test bed," in *SME Vision '90*, pp. 1–14, (Detroit), November 12–15 1990.
7. K. Harding, "Machine vision—lighting," in *Encyclopedia of Optical Engineering*, R. G. Driggers, ed., pp. 1227–1336, Marcel Dekker, 2003.
8. K. G. Harding, "Sensors for the '90s," *Manufacturing Engineering* **106**, pp. 57–61, April 1991.
9. K. Harding, "Three-dimensional noncontact sensing," in *Encyclopedia of Optical Engineering*, R. G. Driggers, ed., pp. 2818–2827, Marcel Dekker, 2003.
10. K. G. Harding, "The promise and payoff of 2D and 3D machine vision: Where are we today?," in *Two- and Three-Dimensional Vision Systems for Inspection, Control, and Metrology*, B. G. Batchelor and H. Hugli, eds., **5265**, pp. 1–15, February 2004.
11. B. K. P. Horn, "Obtaining shape from shading information," in *The Psychology of Computer Vision*, P. H. Winston, ed., pp. 115–155, McGraw-Hill, New York, 1975.
12. O. D. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, Artificial Intelligence Series, MIT Press, Cambridge, USA, 1993.
13. N. Krahnstoever and P. Mendonça, "Bayesian autocalibration for surveillance," in *Proc. of IEEE International Conference on Computer Vision (ICCV'05), Beijing, China*, October 2005.
14. N. Krahnstoever and P. Mendonça, "Autocalibration from tracks of walking people.," in *Proc. British Machine Vision Conference (BMVC), Edinburgh, UK, 4-7 September*, 2006.
15. D. C. Brown, "Close-range camera calibration," *Photogrammetric Eng. and Remote Sensing* **37**, pp. 855–866, Aug. 1971.
16. R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE Journal of Robotics and Automation* **RA-3**(4), pp. 323–344, 1987.
17. Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Analysis and Machine Intel.* **22**, pp. 1330–1334, Nov. 2000.
18. S. Maybank and O. D. Faugeras, "A theory of self-calibration of a moving camera," *Int. Journal of Computer Vision* **8**, pp. 123–151, Aug. 1992.
19. B. Triggs, "Autocalibration and the absolute quadric," in *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 609–614, (San Juan, Puerto Rico), June 1997.
20. M. Pollefeys, R. Koch, and L. Van Gool, "Self-calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters," *Int. Journal of Computer Vision* **32**, pp. 7–25, August 1999.
21. B. Caprile and V. Torre, "Vanishing points for camera calibration," *Int. Journal of Computer Vision* **4**, pp. 127–139, 1990.

22. M. Armstrong, A. Zisserman, and R. Hartley, "Self-calibration from image triplets," in *Proc. 4th European Conf. on Computer Vision*, B. Buxton and R. Cipolla, eds., *Lecture Notes in Computer Science 1064* **I**, pp. 3–16, Springer-Verlag, (Cambridge, UK), Apr. 1996.
23. L. de Agapito, "Self-calibration of rotating and zooming cameras," *Int. Journal of Computer Vision* **45**, pp. 107–127, Nov. 2001.
24. K.-Y. K. Wong, P. R. S. Mendonça, and R. Cipolla, "Camera calibration from surfaces of revolution," *IEEE Trans. Pattern Analysis and Machine Intel.* **25**, pp. 147–161, Feb. 2003.
25. C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. 12th IEEE Computer Vision and Pattern Recognition, Santa Barbara, CA*, **2**, pp. 246–252, 1998.
26. K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th Int. Conf. on Computer Vision, Corfu, Greece*, pp. 255–261, 1999.
27. J. Rittscher, J. Kato, S. Joga, and A. Blake, "A probabilistic background model for tracking," in *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, **2**, pp. 336–350, 1999.
28. S. Intille, J. Davis, and A. Bobick, "Real time closed world tracking," in *Proc. 11th IEEE Computer Vision and Pattern Recognition, San Jaun, PR*, pp. 697–703, 1997.
29. N. Krahnstoever, P. Tu, T. Sebastian, A. Perera, and R. Collins, "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), New York*, 2006.
30. A. Perera, G. Brooksby, A. Hoogs, and G. Doretto, "Moving object segmentation using scene understanding," in *Proceedings of the IEEE Workshop on Perceptual Organization in Computer Vision*, 2006.
31. M. Irani, P. Anandan, and D. Weinshall, "From reference frames to reference planes: Multi-view parallax geometry and applications," in *Proc. 5th European Conf. on Computer Vision*, H. Burkhardt and B. Neumann, eds., *Lecture Notes in Computer Science 1407* **II**, pp. 829–845, Springer-Verlag, (Freiburg, Germany), June 1998.
32. R. Kaucic, A. G. A. Perera, G. Brooksby, J. Kaufhold, and A. Hoogs, "A unified framework for tracking through occlusions and across sensor gaps," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 990–997, 2005.
33. R. Vidal and R. Hartley, "Motion segmentation with missing data using PowerFactorization and GPCA," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, **II**, pp. 310–316, 2004.
34. G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, **I**, pp. 594–601, 2006.
35. S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. J. Kriegman, and S. Belongie, "Beyond pairwise clustering.," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, **II**, pp. 838–845, 2005.
36. D. Gavrila and V. Philomin, "Real-time object detection for smart vehicles," in *Proc. 7th Int. Conf. on Computer Vision, Corfu, Greece*, pp. 87–93, 1999.
37. J. Giebel, D. Gavrila, and C. Schnörr, "A Bayesian framework for multi-cue 3D object tracking," in *Proc. 8th European Conf. Computer Vision, Prague, Czech Republic*, pp. 241–252, 2004.
38. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Intl. Conference on Computer Vision and Pattern Recognition*, 2005.
39. M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates.," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 193–, 1997.
40. Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients.," in *CVPR (2)*, pp. 1491–1498, IEEE Computer Society, 2006.
41. K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. 8th European Conf. Computer Vision, Prague, Czech Republic*, **1**, pp. 69–82, 2004.
42. B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA*, 2005.
43. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, John Wiley and Sons, Inc., 2nd ed., 2001.
44. D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, Upper Saddle River, USA, 2003.
45. I. Haritaoglu, D. Harwood, and L. S. Davis, "HYDRA: Multiple people detection and tracking using silhouettes," in *IEEE International Workshop on Visual Surveillance*, pp. 6–13, 1999.

46. Y. Song, L. Goncalves, and P. Perona, "Monocuolar perception of biological motion - clutter and partial occlusion," in *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, **2**, pp. 719–733, 2000.
47. A. Elgammal and L. Davis, "Probabilistic framework for segmenting people under occlusion," in *Proceedings Eighth IEEE International Conference on Computer Vision, Vancouver, BC, Canada*, **2**, pp. 145–152, 2001. Endnote.
48. T. Zhao and R. R. Nevatia, "Bayesian human segmentation in crowded situations," in *IEEE Computer Vision and Pattern Recognition, Madison, Wisconsin*, **2**, pp. 459–466, 2003.
49. P. H. Tu and J. Rittscher, "Crowd segmentation through emergent labeling," in *Statistical Methods in Video Processing: ECCV 2004 Workshop SMVP2004*, pp. 187–198, May 2004.
50. J. Rittscher, P. Tu, and N. Krahnstoever, "Simultaneous estimation of segmentation and shape," in *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
51. P. Viola and M. Jones, "Robust real-time face detection," in *International Conference on Computer Vision*, **2**, p. 747, (Vancouver, Canada), 2001.
52. F. Burgeois and J. C. Lassalle, "An extension of the Munkres algorithm for the assignment problem to rectangular matrices," *Communications of the ACM* **14**, pp. 802–806, December 1971.
53. S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*, Artech House Publishers, 1999.
54. C. Rasmussen and G. Hager, "Joint probabilistic techniques for tracking multi-part objects," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16–21, 1998.
55. I. J. Cox and S. L. Hingorani, "An efficient implementation and evaluation of Reid's multiple hypothesis tracking algorithm for visual tracking," in *Intl. Conference on Pattern Recognition*, 1994.
56. M. Isard and J. MacCormick, "BraMBLe: A Bayesian multiple-blob tracker," in *IEEE Proc. Int. Conf. Computer Vision*, **2**, pp. 34–41, 2001.
57. R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Trans. on PAMI* **24**, pp. 696–706, May 2002.
58. L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. on PAMI* **25**, pp. 1505–1518, Dec. 2003.
59. G. Jaffré and P. Joly, "Costume: A new feature for automatic video content indexing," in *Proceedings of RIAO*, pp. 314–325, 2004.
60. J. M. Seigneur, D. Solis, and F. Shevlin, "Ambient intelligence through image retrieval," in *International Conference on Image and Video Retrieval*, pp. 526–534, Springer, 2004.
61. F. Porikli, "Inter-camera color calibration by correlation model function," in *Proc. of IEEE ICIP*, **2**, pp. 133–6, Sept. 14–17, 2003.
62. O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Proc. CVPR*, 2005.
63. S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. on PAMI* **24**, pp. 509–522, 2002.
64. D. Lowe, "Distinctive image features from scale-invariant key points," *IJCV* **60**, pp. 91–110, 2004.
65. J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proc. CVPR*, pp. 1470–1477, Oct. 13–16, 2003.
66. T. Zhao and R. Nevatia, "Car detection in low resolution aerial image," in *Proc. ICCV*, **1**, pp. 710–717, (Vancouver, BC, Canada), 2001.
67. N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. CVPR*, **2**, pp. 1528–1535, 2006.
68. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *IJCV* **65**(1-2), pp. 43–72, 2006.
69. Y. Amit and A. Kong, "Graphical templates for model registration," *IEEE Trans. on PAMI* **18**, pp. 225–236, Mar. 1996.
70. P. F. Felzenszwalb, "Representation and detection of deformable shapes," *IEEE Trans. on PAMI* **27**, pp. 208–220, Feb. 2005.
71. J. Canny, "A computational approach to edge detection," *IEEE Trans. on PAMI* **8**, pp. 679–698, Nov. 1986.
72. D. M. Blackburn, J. M. Bone, and P. J. Phillips, *FRVT 2000 Evaluation Report*, February 2001.

73. F. W. Wheeler, X. Liu, P. H. Tu, and R. Hoctor, "Multi-frame image restoration for face recognition," in *to appear in SAFE 2007: IEEE Signal Processing Society Workshop on Signal Processing Applications for Public Security and Forensics*, 2007.
74. S. Chaudhuri, ed., *Super-Resolution Imaging*, Kluwer Academic Publishers, 3rd ed., 2001.
75. K. R. Liu, M. G. Kang, and S. Chaudhuri, eds., *IEEE Signal Processing Magazine, Special edition: Super-Resolution Image Reconstruction*, vol. 20, no. 3, IEEE, May 2003.
76. M. Ng, T. Chan, M. G. Kang, and P. Milanfar, eds., *EURASIP Journal on Applied Signal Processing (JASP) Special Issue on Super-Resolution Enhancement of Digital Video*, Hindawi Publishing Corporation, 2006.
77. S. Borman, *Topics in Multiframe Superresolution Restoration.* PhD thesis, University of Notre Dame, Notre Dame, IN, May 2004.
78. F. W. Wheeler, R. T. Hoctor, and E. B. Barrett, "Super-resolution image synthesis using projections onto convex sets in the frequency domain," in *Proc. of the IS&T/SPIE Symposium on Electronic Imaging, Conference on Computational Imaging*, (San Jose, CA), January 2005.
79. T. Cootes, D. Cooper, C. Tylor, and J. Graham, "A trainable method of parametric shape description," in *Proc. 2nd British Machine Vision Conference*, pp. 54–61, Springer, September 1991.
80. S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision* **56**, pp. 221–255, March 2004.
81. S. Baker and T. Kanade, "Super resolution optical flow," Tech. Rep. CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, October 1999.
82. S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, pp. 1167–1183, September 2002.
83. K. Chang, K. W. Bowyer, and P. J. Flynn, "Face recognition using 2D and 3D facial data," in *ACM Workshop on Multimodal User Authentication*, pp. 25–32, December 2003.
84. P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *Audio and Video-Based Biometric Person Authentication*, pp. 44–51, 2003.
85. X. Liu, P. H. Tu, and F. W. Wheeler, "Face model fitting on low resolution images," in *Proc. of the British Machine Vision Conference (BMVC)*, (Edinburgh, UK), 2006.
86. S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super-resolution," *IEEE Transactions on Image Processing* **13**, pp. 1327–1344, October 2004.
87. P. J. Costianes, "An overview of concealed weapons detection for homeland security," in *Proceedings of the Applied Imagery and Pattern Recognition Workshop*, Oct. 19–21, 2005.
88. M. G. Frank and P. Ekman, "The ability to detect deceit generalizes across different types of high-stake lies," *Journal of Personality and Social Psychology* **72**(6), pp. 1429–1439, 1997.
89. P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. T. Pavlidis, M. G. Frank, and P. Ekman, "Imaging facial physiology for the detection of deceit," *IJCV* **71**, pp. 197–214, Feb. 2007.
90. P. Buddharaju, J. Dowdall, P. Tsiamyrtzis, D. Shastri, I. Pavlidis, and M. G. Frank, "Automatic thermal monitoring system (ATHEMOS) for deception detection," in *Video Proc. CVPR*, **2**, June 20–25, 2005.
91. T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, J. F. J. Nunamaker, D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intelligent Systems* **20**, pp. 36–43, Sept./Oct. 2005.
92. Y. Matsumoto, T. Ogasawara, and A. Zelinsky, "Behavior recognition based on head pose and gaze direction measurement," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, **3**, pp. 2127–2132, (Takamatsu), Oct. 31–Nov. 5, 2000.