# Person Attribute Search For Large-Area Video Surveillance

Jason Thornton, Jeanette Baran-Gale, Daniel Butler, Michael Chan, Heather Zwahlen
MIT Lincoln Laboratory, 244 Wood St., Lexington, Ma. 02420

***Abstract*—This paper describes novel video analytics technology which allows an operator to search through large volumes of surveillance video data to find persons that match a particular attribute profile. Since the proposed technique is geared for surveillance of large areas, this profile consists of attributes that are observable at a distance (including clothing information, hair color, gender, etc) rather than identifying information at the face level. The purpose of this tool is to allow security staff or investigators to quickly locate a person-of-interest in real time (e.g., based on witness descriptions) or to speed up the process of video-based forensic investigations. The proposed algorithm consists of two main components: a technique for detecting individual moving persons in large and potentially crowded scenes, and an algorithm for scoring how well each detection matches a given attribute profile based on a generative probabilistic model. The system described in this paper has been implemented as a proof-of-concept interactive software tool and has been applied to different test video datasets, including collections in an airport terminal and collections in an outdoor environment for law enforcement monitoring. This paper discusses performance statistics measured on these datasets, as well as key algorithmic challenges and useful extensions of this work based on end-user feedback.**[1]

## I. Introduction

Driven by recent advances in digital video camera design, video management systems, and efficient archiving, security at large publicly accessible facilities and urban sites increasingly deploy comprehensive video surveillance systems [1]. Such systems support both the ability to monitor wide areas in real time and the ability to conduct forensic reviews after an incident or a tip. Most analysis of surveillance video for security purposes requires the operators or investigators to search for particular types of video content (e.g., specific actions or behaviors, or vehicles, objects, or persons fitting a given description). While the fully attentive human visual system is very adept at interpreting video content, it is also typically limited in capacity to reviewing one camera view at a time at speeds near real-time, depending on the level of activity in the scene; as a result, searching over large amounts of surveillance video can be a slow and cumbersome process. In addition, the human attention span does not persist at full capacity over long time periods. This is the motivation for automatic video search technology, which can direct the attention of security personnel or investigators to potentially useful video content [2]. This paper considers a specific type of automated search that addresses a common operational need: the search for persons-of-interest across large surveilled areas.

Most research into video-based search for persons focuses on biometric recognition, using signatures that can be captured at a distance, such as gait [3] or especially face [4][5]. However, face recognition systems can only be employed when surveillance cameras capture face images at sufficient resolution and illumination. In fact, experimental studies indicate that face recognition performance begins to degrade at (compressed) image resolutions as high as 90 pixels between the eyes [6]; this is much greater resolution than typically provided by surveillance video, unless cameras are setup with narrow fields of view (e.g., monitoring specific doorways at close range). In addition, biometric recognition requires a database of prior enrollment records against which to perform identification. This does not apply when security personnel or investigators are working with an eyewitness description of a person of interest. An alternative approach is to perform automatic search for persons who match a basic appearance profile, including clothing, physical attributes (gender, hair color), and carried objects such as bags or backpacks. Note that many of these attributes are temporary in nature and taken together do not necessarily describe a unique individual in the area under surveillance, but they are observable through video at a distance. The ability to perform automatic searches based on these attributes has the potential to make many monitoring and forensic search tasks much more efficient.

There have been several recent attempts to incorporate appearance descriptions of persons into video search techniques, including so-called "soft biometrics." Both Vaquero et al. [7] and Demirkus et al. [8] localize persons and frontal faces in video using standard Haar feature cascade classifiers, and focus primarily on the characterization of facial attributes for images captured at relatively close range to the camera. For instance, the authors in [7] show promising performance on the detection of facial hair and glasses, while the authors in [8] demonstrate some ability to classify gender and ethnicity based on facial features. It is unclear how well these techniques would extend to longer ranges (on the order of tens of meters), where detailed face images are no longer available. In addition, both papers describe simple techniques for recognizing clothing color, based on the extraction of dominant colors in pre-set regions within the bounding box surrounding each person. We propose an extension of this sort of analysis, incorporating characteristics observable at a distance, based on a generative

[1]This work is sponsored by the Assistant Secretary of Defense for Research and Engineering and the Department of Homeland Security Science and Technology Directorate under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

probabilistic model that describes the expected variation in appearance given a particular attribute profile. The main advantages of this model are its ability to successfully recognize many different combinations of attributes (including multi-colored clothing and hand-carried objects) and its flexibility across different poses and body types.

Section II gives an overview of the main components of the proposed algorithm, as well as some details about its implementation. Section III describes an interactive software tool built to test the algorithm on several real surveillance datasets. Section IV provides some performance metrics on these datasets, and Section V discusses some of the challenges and useful extensions for this work.

## II. Search Algorithm

The proposed search technique has two main components. The first is a detection process that ingests raw video and finds instances of moving persons within each camera view. This process is designed to run in real-time, analyzing one frame per camera per second and storing the detections as time-stamped and location-stamped records in a database. The second component matches detections to a specified attribute profile. Given a person description, this process scores the likelihood that each detection in the database fits that profile, by evaluating a probabilistic generative appearance model. Both components must be accurate and computationally efficient in order to support useful automatic queries.

### *A. Moving Person Detection*

Because of the high redundancy of video content from one frame to the next, the detection processing chain first samples the video stream at one frame per second. The goal of this processing is to find persons in motion, since there is no need to store stationary subjects repeatedly to the detection database. As is typically done, we use a sliding window approach to evaluate candidate detection locations within each frame, where a candidate location is specified by the position of a bounding box $(x, y)$ and its dimensions $(w, h)$; since we assume a fixed aspect ratio for the bounding box dimensions, there are three parameter values to sweep through during search. Broadly, each detection must satisfy three criteria: it must exhibit a human shape, it must fit the ground plane constraints of the scene, and it must contain a high ratio of foreground pixels (as opposed to static background). Figure 1 depicts sample detections that meet these three criteria.

There has been significant progress over the last decade on image processing techniques that recognize human shapes or contours. Most notably, Dalal and Triggs [9] proposed the use of histograms of oriented gradients (HOG) features as the basis for a human contour detector. Histograms of oriented gradients provide a characterization of object contours which is somewhat invariant to lighting conditions and color contrast, and can be used to train a support vector machine (SVM) classifier that detects the human form with impressive accuracy. However, even state-of-the-art contour-based detectors tend to produce multiple false positives per frame for the

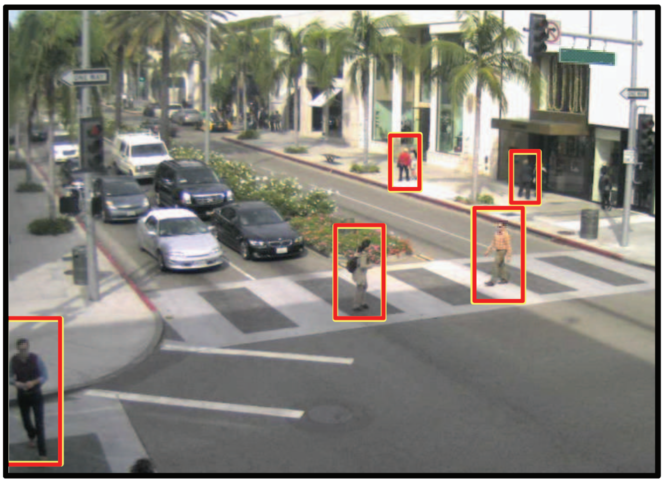

Fig. 1. Detections of moving persons in a sample video frame.

types of large and visually complex scenes common in video surveillance applications. In order to reduce the frequency of false positives, we also take advantage of prior information about the scene and the dynamic nature of video.

Most surveillance cameras capture scenes in which the activity of interest occurs along an approximate ground plane that remains fixed (notable exceptions include active pan-tilt-zoom cameras and views that depict multiple levels, stairs, etc.). When this assumption is reasonable, we can use information about the ground plane (learned during an offline training step) to eliminate false positives based on the size and position of the detection. The ground plane information is encoded in a perspective model similar to the one used in [10]. More specifically, this model learns a planar homography which maps foot position to expected head position. If all persons in the scene are standing upright on a planar surface, this ensures that the foot-to-head mapping is an affine map between planes in 3D space (specif cally, translation by the height of the person), and therefore induces a plane homography when projected into image coordinates. For a candidate bounding box, we measure the normalized "perspective error" of bounding box position $\vec{p} = [x, y, w, h]$ as:

$$\epsilon_{per}(\vec{p}) \;=\; \frac{|h - h_{est}(\vec{p})|}{h} \qquad (1)$$

where $h_{est}$ is the estimated height for position $\vec{p}$, computed by mapping the foot position to a head position according to the scene homography, then taking the difference between projected head position and actual foot position.

Finally, since our objective is to detect moving persons, and a significant ratio of human contour false positives are generated by static background patterns, we also learn a dynamic background model of the video scene. Typical "background subtraction" techniques learn adaptive statistical models of static background at the pixel level, then compare new frames of video to these models in order to estimate which pixels depict moving foreground. We use the background subtraction technique proposed by Stauffer and Grimson [11], which maintains an adaptive Gaussian mixture model in 3D color space for every pixel. This method offers some robustness to shadow effects because of the multimodal form of its background model. We measure a foreground ratio for each candidate bounding box, computed as the ratio of all pixels within the box that are labeled as foreground by this model.

Each of the three detection criteria produces a real-valued score that indicates the degree to which that criterion is met for detection position $\vec{p}$: the output of the SVM human contour detector $s_{cont}(\vec{p})$, the perspective error $\epsilon_{per}(\vec{p})$, and the foreground ratio $r_{fore}(\vec{p})$. The fused detections result from a function $f(s_{cont}(\vec{p}), \epsilon_{per}(\vec{p}), r_{fore}(\vec{p}))$ mapping all three values to either 1 (detection) or 0 (no detection). For reasons of computational efficiency, we decompose this function into a combination of individual thresholds applied to each value. This way, we can use a cascaded detection strategy, where the computationally simpler criteria (foreground ratio and perspective error) are evaluated first in order to rule out most candidate bounding box positions. As a result, the computationally expensive HOG features must only be generated at relatively sparse locations and scales. To achieve additional speedup, we use Dalley's GPU implementation of parallelized HOG feature computation [12]. All three thresholds for this method were selected empirically from training data.

In addition to person detection, we also attempt to classify each detection by gender. This classif er is obtained by retraining a SVM discriminant in HOG feature space. As a result, classification is based on contour characteristics, which appear to capture information about hair style, clothing style, and body frame. Since HOG features are already computed during detection, performing this additional classification is computationally inexpensive.

### *B. Generative Appearance Model*

In our search framework, the operator specifies an attribute profile including some subset of the following characteristics: gender, hair or hat color, clothing color, and bag descriptions (type and color). Given a particular attribute profile, however, there is quite a bit of variation in the way an image chip manifests as a set of pixel values, due to changes in view angle, lighting conditions, body type, clothing and bag style, etc. In order to explain this expected variation, we formulate a generative probabilistic model that gives the likelihood of observing a set of pixel values given an attribute profile. This model has a hierarchical structure, with key factors of the image formation encoded as latent variables in the hierarchy. The model, as visualized in Figure 2, has the following structure.

Let $\vec{A}$ represent the set of attributes in the attribute profile, comprised of a collection of flags and real-valued color specifications. The f rst level of the model hierarchy partitions the image chip into its component parts as depicted in Figure 2: head, torso, lower body, and (optionally) one or multiple bags. Let vectors $\vec{z}_{head}$, $\vec{z}_{torso}$, $\vec{z}_{lower}$, and $\vec{z}_{bag}$ encode the 2D positions of their respective components as rectangles within the image chip, and let vector $\vec{z}_{body} = [\vec{z}_{head}, \vec{z}_{torso}, \vec{z}_{lower}]$

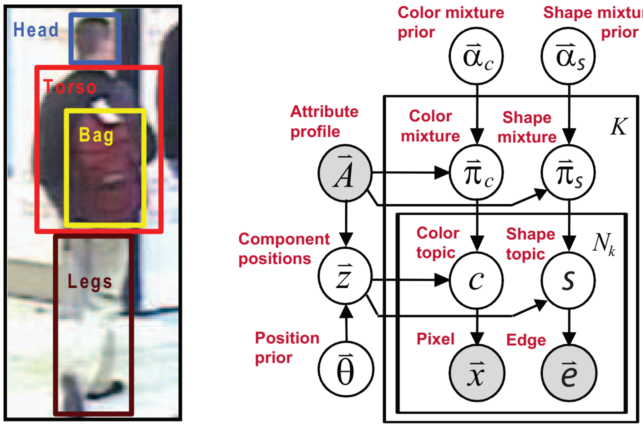


Fig. 2. Left: Illustration of image chip decomposition into its component parts. Right: Graphical representation of the hierarchical generative model. Nodes represent variables, either directly observed (shaded), latent, or set during training, while arrows represent dependencies. Each image chip contains $K$ component parts and $N_k$ pixels within each component.

contain the first three position vectors appended together. The generative model jointly selects the basic position values from a multivariate Gaussian distribution:

$$\vec{z}_{body} \sim N(\vec{\mu}, \Sigma) \tag{2}$$

where parameter set $\theta = \{\vec{\mu}, \Sigma\}$ contains the mean expected position set and the covariance matrix of position changes due to body, clothing, and view angle variation. In addition, if attribute set $\vec{A}$ contains bag specifications, each bag position is drawn from one of three non-parametric distributions, depending on the specified bag type (backpack, hand-carried bag, or rolled luggage). Note that bag position is defined to be relative to the position of the body components, so that the generation of $\vec{z}_{body}$ and $\vec{z}_{bag}$ vectors are not independent.

Once the component positions have been determined, the attribute description governs the color and shape information expressed by the pixel values of each sub-region. The model employs a technique similar to latent Dirichlet allocation (LDA) [13] to generate the mixture of observed color or shape information. As an example, consider the generation of color information. The model defines a discrete number of color "topics," which are selected to be twelve common perceptual colors (black, white, gray, tan, brown, yellow, orange, pink, red, green, blue, purple). Each local region $k$, within a component, is characterized by a distribution over the color topics $\vec{\pi}_k$ (a 12-dimensional real-valued vector that sums to one). First, the model draws the color topic distribution from a Dirichlet prior:

$$\vec{\pi}_k \sim Dir(\vec{\alpha}_k) \tag{3}$$

The Dirichlet prior distribution accounts for variation in the color mixture due to factors such as style, material composition, shadow effects, or object occlusion. The Dirichlet parameters $\vec{\alpha}_k$ are set based on the attribute profile (e.g., a "dark red torso" specif cation may have a bias toward selection of mixtures dominated by red and black topics).[2]

Once the topic distribution has been selected, each pixel within the component region independently selects a color topic $c$ according to the distribution given by $\vec{\pi}_k$. Finally, the pixel value is drawn from the corresponding color topic density within the three-dimensional color space. If $\vec{x}$ represents a single three-dimensional point defined in hue-saturation-value (HSV) color space, the probability that this point was generated by a color density characterized by mean $\vec{\mu}_c$ and covariance matrix $\Sigma_c$ is given by the quasi-Gaussian expression:

$$p(\vec{x}) = \phi \cdot \exp\left(-0.5 \cdot \vec{d}(\vec{x}, \vec{\mu}_c)^T \Sigma^{-1} \vec{d}(\vec{x}, \vec{\mu}_c)\right) \tag{4}$$

where constant $\phi$ normalizes the distribution and $\vec{d}$ is the difference vector between $\vec{x}$ and $\vec{\mu}_c$, with the first element $d_1$ computed to account for the cyclical nature of the angular hue axis, which spans from 0 to 1:

$$d_1(x_1, \mu_{c1}) \;=\; \mathrm{mod}(x_1 - \mu_{c1} \;+\; 0.5,\; 1) \;-\; 0.5 \tag{5}$$

Forming color densities in HSV space, as opposed to RGB space, provides more tolerance to inconsistent lighting conditions.

A similar generative model structure applies for shape information within each component region. Instead of using color topics and observed pixel values, the model replaces these with "shape topics" (e.g., multi-scale gradients) and edge-based primitives centered around each pixel location. In this case, the Dirichlet distribution that drives the model is governed by the label of the component (head, torso, etc.).

Finally, the real-valued gender score $g$ produced during the detection stage is assumed to be drawn independently from the rest of the model, according to an exponential distribution:

$$s(g) \;\sim\; \exp(\lambda) \tag{6}$$

where $s(\cdot)$ adjusts the sign of the gender score according to the gender label in the attribute profile (or sets it to zero if gender is unspecified), and parameter $\lambda$ controls the importance of the gender specification relative to all other attribute specifications.

The model structure described above has a set of observed variables $O$ (the pixel values, edge primitives, and attribute profile), a set of hidden, or latent, variables $H$ (the component positions, topic mixtures, and topic labels), and a set of learned model parameters $\Theta$ (describing various Gaussian, Dirichlet, exponential, and non-parametric distributions). When a given image chip is evaluated for its degree of match to a given attribute profile, the match score $m$ is computed by maximizing the joint distribution of the model with respect to its hidden variables:

$$m \;=\; \max_{H} \; P\left(O, H, \Theta\right) \tag{7}$$

This is equivalent to estimating the most likely hidden states to explain the observed image; if a high-probability explanation exists that is compatible with the attribute description and the observed values, this will lead to a high match score. Parts of this maximum likelihood optimization problem can be solved in closed-form. For the variables which cannot be estimated in closed form (i.e., the component positions), we employ a greedy iterated conditional modes (ICM) algorithm [14], which is initialized at the mean position vector and converges quickly to a local optimum.

[2]Note that in contrast to typical LDA models, the topics in this model are pre-defined, and the parameters of the Dirichlet distribution are set according to the attribute description, not learned in an unsupervised manner.

### C. Model Implementation

After experimentation with multiple versions of the model framework described above, we selected a simplif ed implementation for use in all experiments discussed in this paper. This implementation has the following details. The component position variables represent boundary locations along the vertical axis only, while positions along the horizontal axis are pre-determined and fixed. All location variables are expressed as ratios of image chip height in order to make their distributions invariant to resolution. In addition, this version of the model is limited in scope to color topic mixture representations (not shape mixture representations) in order to improve computational efficiency.

In order to learn the model parameters, we labeled ground truth for training purposes from several hundred images sampled from multiple video datasets. The labeled ground truth included component descriptions, positions, and perceptual color labels. The mean vectors and covariance matrices of all multivariate Gaussian distributions were set using maximum likelihood estimation, while each of the three bag position distributions was learned directly from labeled examples using kernel density estimation. Finally, the method for setting the Dirichlet parameters according to the attribute profile was selected through empirical testing.

## III. Interactive Search Tool

The system described in this paper processes surveillance video in order to extract instances of moving persons and then stores these instances (as time and location tagged descriptors) to a database. In order to allow an operator to quickly and easily search over these records, we constructed an interactive search tool that operates directly on the database. The tool works by taking in a set of search criteria through a graphical user interface, retrieving all records from the detection database that fall within the specified time period and location, applying the probabilistic model described in Section II-B to score how well each record matches the specified attribute profile, ranking them in order of best match, and then displaying the top set of matches back to the operator for further investigation of the relevant video.

Figure 3 shows example screenshots of the search tool. Launching an attribute-based person search brings up a search criteria input menu, divided into an attribute profile section (upper half) and a search localization section (lower half). Within the attribute profile section, the operator may specify any subset of attribute options represented in the menus; any inputs not provided by the operator are left as unspecified by

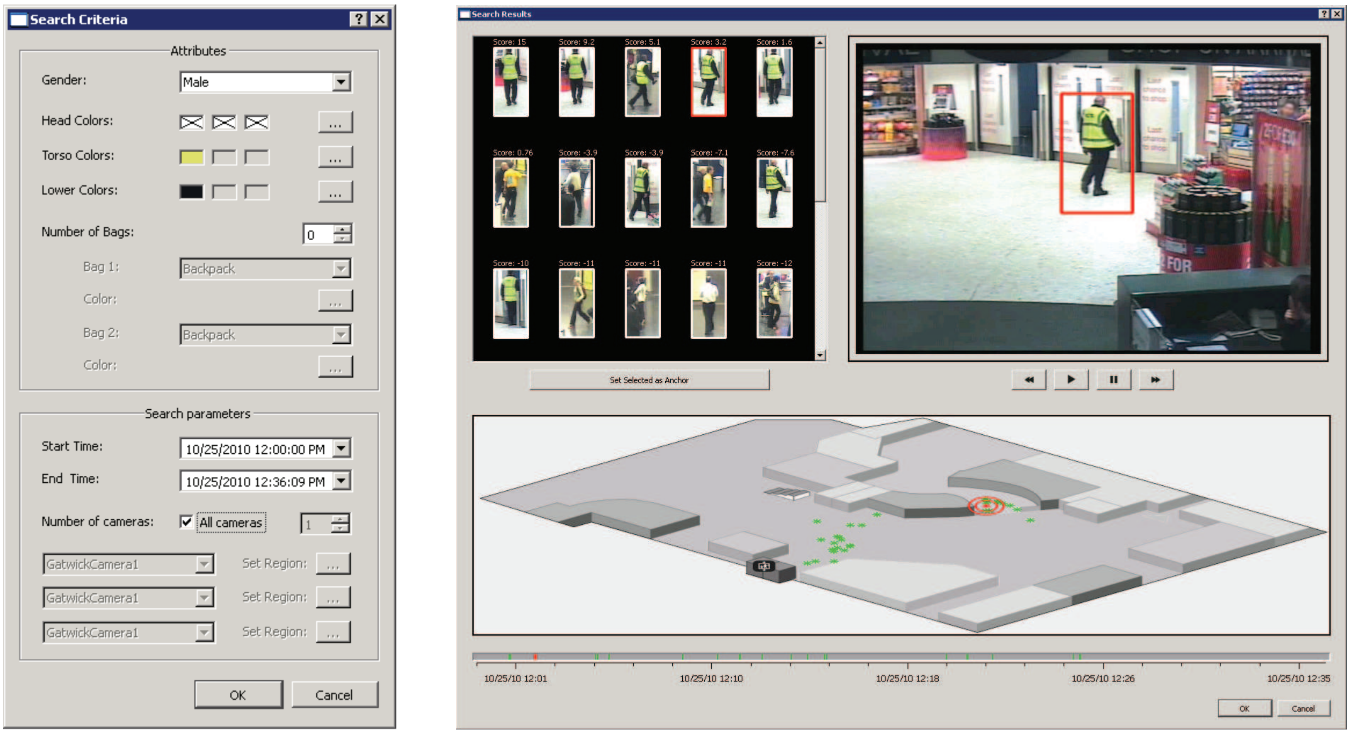


Fig. 3. Sample screenshots of the interactive search tool. Left: Search criteria input interface, in which the operator has specified "male" gender, "yellow" shirt color, and "black" pants. Right: Search results exploration interface, displaying top matches for this search criteria. Image chips are arranged in order of descending match score, from left to right and top to bottom. Note that some image chips farther down the progression contain inexact but close matches, such as females with the specified clothing, or subjects with off-white rather than yellow shirts. This search was performed over approximately 30 minutes of data from each of three camera views, obtained from the i-LIDS dataset [15] collected at Gatwick Airport in London.

default, meaning they will not factor into the search results. A gender menu allows for selection of male or female. A head color menu allows for selection of color information corresponding to either hat or hair. A torso clothing color menu allows specification of upper body clothing color(s), which typically correspond to shirt, coat, or blouse color descriptions. A lower body clothing color menu takes in information about pants, shorts, skirts, and/or prominent shoes. Finally, the bag options allow the user to specify one or more bags in the possession of the person-of-interest. Each bag takes a type description (backpack, hand-carried bag, or rolled luggage) in addition to color patterns. All color selections are made through a palette interface that allows selection from among a range of shades within multiple common perceptual color categories.

In addition to the attribute profile, the operator selects a start time and end time for conducting the search, which may span up to multiple days of archived video. The operator may also choose the subset of camera views over which to search, focusing in on particular regions within individual camera scenes (e.g., doorways or pedestrian passageways) if desired. Once all attribute and search localization criteria has been entered, the search process launches and the software accesses the detection database to computes all relevant match scores. The time required to finish this processing is typically at least a few seconds, and scales in rough proportion to the area and time window of the search, depending on variations in pedestrian frequency.

The search results interface has three panels, as depicted on the right side of Figure 3. The first panel contains a set of image chips representing the top matches to the attribute search criteria. These chips are filtered to avoid redundant displays of the same person in consecutive frames of video. The operator may scroll through these image chip results and click on any of them to retrieve the corresponding video content. The second panel displays the video frame from which the match was detected, and allows the operator to view the video using a standard set of playback controls. The third panel displays both a timeline and a map of the facility, containing estimated times and locations for each of the top matches. The operator has the option to retrieve video by selecting an individual detection directly from the map or from the timeline, rather than from the image chip set.

As it is set up, the search tool provides both a clear way for users to enter structured search queries and a mechanism to explore potential sightings of a person-of-interest throughout all available surveillance video. It is important to note that if there is a useful video detection, it will not always appear as the first (most highly scored) image chip, due to inaccuracies in witness descriptions, limitations in the type of attribute inputs accepted by the software, or imperfections within the generative appearance model. However, if the useful content appears somewhere among the top match results, it is typically a much easier and quicker process for an operator to browse through a set of image chips and focus in on the relevant content than it is for that operator to manually scan large amounts of raw video.

## IV. Experimental Results

We have tried the search tool described above on several surveillance video datasets, including the i-LIDS dataset [15] collected at Gatwick Airport, sample video collected at a major US airport, and sample video collected at street intersections by outdoor law enforcement surveillance cameras. However, our ability to measure performance quantitatively is limited by the availability of detailed ground truth for these datasets. In order to compute the results discussed in this section, we selected a particular area surveilled during the US airport video data collection and labeled ground truth for all pedestrian activity occurring within that area over a defined period of time. The surveillance video of this (approximately 50 square meter) region has image resolution along the height of pedestrians ranging from about 80 pixels to 120 pixels. We collected ground truth corresponding to the location and attribute-based appearance of each person who passed through this region. Analyzing sampled frames of video at a rate of 1 Hz, there was a total of approximately 1,000 instances (i.e., frame appearances) of 150 unique pedestrians observed during ground truth labeling.

In addition, we also labeled the location of every pedestrian passing through a typical outdoor street scene captured by law enforcement surveillance video. This video contains examples of pedestrians (at about 50 to 200 pixels of height resolution) as well as moving vehicles. We again generated ground truth at a rate of 1 Hz, covering approximately 2,200 instances of 100 unique pedestrians.

### A. Detection

The output of the moving person detection algorithm depends upon three threshold values, corresponding to the three criteria for detection described in Section II-A. Varying these threshold values affects both the probability of detection (PD) and the false positive rate (FPR). Therefore, we conducted a parameter sweeps for the indoor and outdoor surveillance video to find threshold values that yielded the best detection performance, using the ground truth data described above to evaluate PD and FPR metrics. For each person passing through the region of interest, we counted a correct detection if the algorithm flagged that person in at least one analyzed frame of video. On the other hand, we counted any detection that did not correspond to an entire person as a false positive.

Table I shows PD and FPR for the selected detection thresholds of each dataset. The algorithm achieved the best results on the indoor airport video, detecting an instance of almost every pedestrian while pulling in false positives at a relatively low rate (so that they constitute a small percentage of records in the detection database). The outdoor scene proved more difficult for several reasons. Some of the missed detections were from pedestrians who never completely enter the scene before exiting, or who appear partially occluded behind vehicles. In addition, most of the false positives were generated by parts of moving vehicles or by trees blowing in the wind (clutter which is not present in the indoor scene). Overall, we found that the detection process supports good search capability, especially when the video captures a complete view of the subject, but that the performance depends heavily on the characteristics of the scene.

TABLE I
Moving person detection performance.

| Video data | PD | FPR |
|---|---|---|
| Airport indoor | 97% | 0.0045 per sec |
| Law enforcement outdoor | 89% | 0.26 per sec |

### B. Appearance Model

The probabilistic appearance model is a mechanism to score the likelihood that each image chip depicts a person with the specified attribute set. When the model functions correctly, all examples of persons matching the provided description will appear at the top of the match list, which is ranked in order of descending likelihood scores. To test the performance of the model, we ran multiple sample searches over the portion of video for which we had ground truth labels. The attribute profiles for each of eleven test searches are listed in the legend of Figure 4; these were selected arbitrarily by taking descriptions from the ground truth labels that matched one or more persons who appeared within the video (excluding gender, which was tested separately).

For each test search, we generated a performance curve by varying the number of top search results returned by the search. Adding to the number of results increases the chance of finding all true matches, but it also increases the occurrence of false positives. The y-axis in Figure 4 plots the recall, or the percentage of all true matches returned by the search, while the x-axis plots the number of returned false positives, normalized by the total number of false positive individuals in the database. Note that by these metrics, an algorithm that assigned random match scores to image chips would have an expected tradeoff represented by the dotted line in Figure 4.

As expected, all search results using the proposed model perform significantly better than the random-scoring baseline. However, there is noticeable variation in error rates depending on the particular attribute set. Five of the eleven sample searches (represented by the red line in Figure 4) found all true matches before returning any false positives. Other sample searches returned multiple false positives before recovering all matches. The more specif c search queries (especially those with only one true match among all pedestrians) tend to show better results because these profiles contain the most discriminating information. On the other hand, more generic descriptions seem to pull in some false positives along with all of the true positives.

Finally, we also evaluated the accuracy of the gender classifier by testing it on multiple surveillance datasets. We

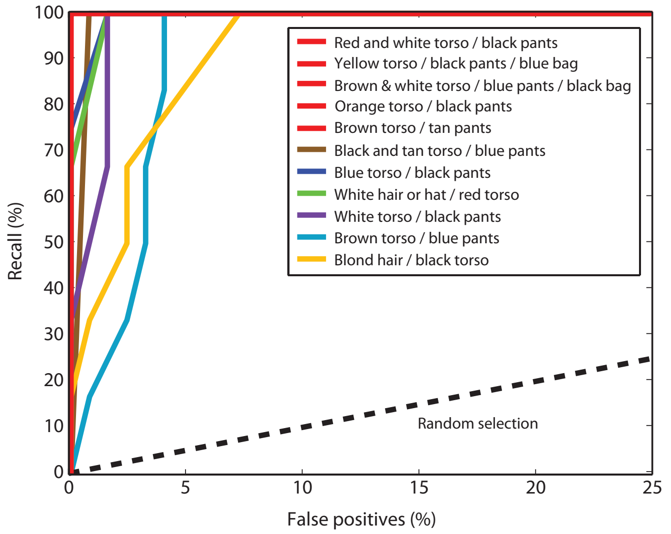


Fig. 4. Performance curves for eleven sample attribute-based searches. Curves plot recall (y-axis) vs. normalized false positives (x-axis). The dotted line represents expected performance of a random selection search algorithm. The first five searches in the legend have completely accurate search results.

measured correct classification rates ranging from 75% to 85%. While these results are far from perfect, when used in combination with stronger attribute cues, the classifier does tend to move more matches with the specified gender toward the top of the match list. This indicates that it may also be useful to include other "weak" classifiers based on contour analysis, such as height or body frame information.

## V. Discussion

The search capabilities described in this paper provide a useful way for security personnel or investigators to sift through large quantities of video data to find a particular person-of-interest, given a recent observation or witness report describing that person. This capability may be applied in two different modes: to perform real-time monitoring for any persons that match an attribute prof le (e.g., at all entryways to a facility), or to perform forensic searches over archived video data. In this way, the search tool provides the operator with a number of leads to potentially valuable and easily accessible video content.

However, the performance of the search algorithm certainly depends on both the quality of the recorded video (especially the resolution and look angle) and the characteristics of the environment under surveillance. For instance, outdoor views of street intersections seem to present more challenges than relatively controlled indoor settings; there are more sources of motion, such as vehicles and trees blowing in the wind, in addition to more variation in illumination conditions due to natural light changes and building shadows. Another important factor is crowd density. A typical subject, in any given frame of video evaluated in this work, has a high likelihood of at least some type of partial occlusion from other movers or static objects; however, the video dataset does not contain solid masses of people, which would make successful detection and analysis much more difficult. In addition, different locations may exhibit sufficiently different clothing styles to degrade the accuracy of the probabilistic appearance model. When this is the case, the model may be re-trained directly on new (and more representative) sample data.

We have considered several extensions to this work. Feedback from end users indicates that it would be useful to extend the search capabilities to vehicle descriptions, at the level of color and generic type (sedan, truck, SUV, etc.). It would also be advantageous to search over motion or action primitives, such as a person running in a specif ed direction, or a person entering or exiting a vehicle. Finally, we have been working on integrating the search capabilities into a more general situational awareness tool for security and investigation, which will allow multiple users to access video and perform and share content searches over a web-based interface.

## Acknowledgments

This work was partially sponsored by Jalal Mapar at the Department of Homeland Security Science and Technology Directorate.

## References

[1] J. Vlahos, "Surveillance society: New high-tech cameras are watching you," *Popular Mechanics*, January 2008.

[2] N. Gagvani, "Challenges in video analytics," in *Advances in Pattern Recognition: Embedded Computer Vision*, 2008, ch. 12, pp. 237–256.

[3] J. Boyd and J. Little, "Biometric gait recognition," in *Advanced Studies in Biometrics: Revised Selected Lectures and Papers*, ser. Lecture Notes in Computer Science, M. Tistarelli, J. Bigun, and E. Grosso, Eds., 2005, vol. 3161, pp. 19–42.

[4] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[5] P. Phillips, "Overview of the multiple biometric grand challenge version 2," in *MBGC Workshop*, December 2009. [Online]. Available: http://www.nist.gov/itl/iad/ig/mbgc-presentations.cfm

[6] ——, "Mbgc still face challenge version 2 preliminary results," in *MBGC Workshop*, December 2009. [Online]. Available: http://www.nist.gov/itl/iad/ig/mbgc-presentations.cfm

[7] D. Vaquero, R. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *Proceedings of IEEE Workshop on Applications of Computer Vision*, 2009.

[8] M. Demirkus, K. Garg, and S. Guler, "Automated person categorization for video surveillance using soft biometrics," in *Proceedings of SPIE, Biometric Technology for Human Identif cation VII*, 2010.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[10] X. Liu, P. Tu, J. Rittscher, A. Perera, and N. Krahnstoever, "Detecting and counting people in surveillance applications," in *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2005.

[11] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1999.

[12] G. Dalley, "Improved robustness and efficiency for automatic visual site monitoring," Ph.D. dissertation, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 2009.

[13] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, January 2003.

[14] J. Huang, Q. Morris, and B. Frey, "Iterated conditional modes for cross-hybridization compensation in DNA microarray data," University of Toronto, Tech. Rep., 2004.

[15] K. Sage, A. Nilski, and I. Sillett, "Latest developments in the ilids performance standard: New imaging modalities," *IEEE Aerospace and Electronic systems Magazine*, vol. 25, July 2010.

## ABOUT THE AUTHORS

***Dr. Jason Thornton*** *is a member of the technical staff in the Homeland Protection Division at MIT Lincoln Laboratory. His work at the Laboratory includes research related to sensor fusion, threat assessment and decision support, and image and video processing. He holds a PhD degree in electrical and computer engineering from Carnegie Mellon University. Dr. Thornton may be contacted at Jason.thornton@ll.mit.edu.*

***Daniel Butler*** *is a graduate student in computer science and engineering at the University of Washington in Seattle. He received his ScB in applied mathematics and computer science from Brown University in 2009. From 2009 to 2010 he was assistant technical staff at MIT Lincoln Laboratory, and from 2010 to 2011 he was a visiting researcher at the Academia Sinica Institute for Mathematics, where he was supported by a Fulbright scholarship.*

***Jeanette Baran-Gale*** *received a BS degree in Computer Engineering from The Pennsylvania State University in 2001 and a MS in Mathematics from the University of Massachusetts at Lowell in 2007. Jeanette was an associate staff member of MIT Lincoln Laboratory from 2004 to 2011; she is currently working on her PhD in bioinformatics and computational biology at the University of North Carolina at Chapel Hill.*

***Dr. Michael Chan*** *has been a Research Staff Member at MIT Lincoln Laboratory since 2006. His recent work focuses on developing advanced data analytics algorithms and net-centric systems to improve situation awareness and reduce workload for human operators of large surveillance systems. Dr. Chan received his bachelor's degree in electrical engineering from Oxford University, and his MS and PhD degrees in computer science from the University of Pennsylvania.*

***Dr. Heather Zwahlen*** *is a technical staff member at MIT Lincoln Laboratory in the Advanced System Concepts Group. She received a PhD in physics from Michigan State University (MSU) in 2005. She is currently leading a program focusing on technology solutions for homeland protection.*

---