Attribute-Based People Search in Surveillance Environments

Daniel A. Vaquero[1], Rogerio S. Feris[2], Duan Tran[3], Lisa Brown[2], Arun Hampapur[2], Matthew Turk[1]

[1]U. of California, Santa Barbara
{daniel,mturk}@cs.ucsb.edu

[2]IBM Research
{rsferis,lisabr,arunh}@us.ibm.com

[3]U. of Illinois at Urbana-Champaign
ddtran2@uiuc.edu

Abstract

We propose a novel framework for searching for people in surveillance environments. Rather than relying on face recognition technology, which is known to be sensitive to typical surveillance conditions such as lighting changes, face pose variation, and low-resolution imagery, we approach the problem in a different way: we search for people based on a parsing of human parts and their attributes, including facial hair, eyewear, clothing color, etc. These attributes can be extracted using detectors learned from large amounts of training data. A complete system that implements our framework is presented. At the interface, the user can specify a set of personal characteristics, and the system then retrieves events that match the provided description. For example, a possible query is "show me the bald people who entered a given building last Saturday wearing a red shirt and sunglasses." This capability is useful in several applications, such as finding suspects or missing people. To evaluate the performance of our approach, we present extensive experiments on a set of images collected from the Internet, on infrared imagery, and on two-and-a-half months of video from a real surveillance environment. We are not aware of any similar surveillance system capable of automatically finding people in video based on their fine-grained body parts and attributes.

1. Introduction

In traditional surveillance scenarios, users are required to watch video footage corresponding to extended periods of time in order to find events of interest. However, this process is resource-consuming, and suffers from high costs of employing security personnel. The field of intelligent visual surveillance [6] seeks to address these issues by applying computer vision techniques to automatically detect specific events in long video streams. The events can then be presented to the user or be indexed into a database to allow queries such as "show me the red cars that entered a given parking lot from 7pm to 9pm on Monday" or "show me the faces of people who left a given train station last week."

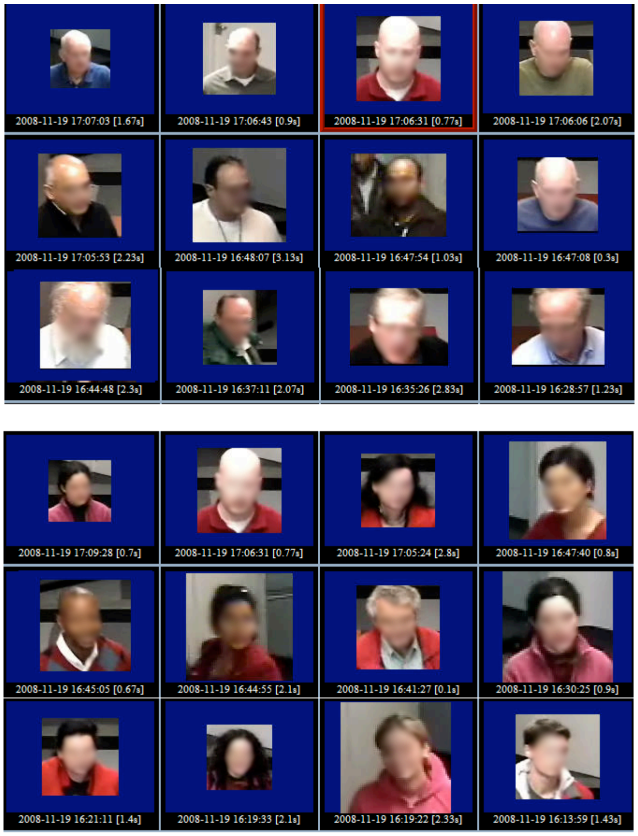

2008-11-19 17:07:03 [1.67s]
2008-11-19 17:06:43 [0.9s]
2008-11-19 17:06:31 [0.77s]
2008-11-19 17:06:06 [2.07s]
2008-11-19 17:05:53 [2.23s]
2008-11-19 16:48:07 [3.13s]
2008-11-19 16:47:54 [1.03s]
2008-11-19 16:47:08 [0.3s]
2008-11-19 16:44:48 [2.3s]
2008-11-19 16:37:11 [2.07s]
2008-11-19 16:35:26 [2.83s]
2008-11-19 16:28:57 [1.23s]
2008-11-19 17:09:28 [0.7s]
2008-11-19 17:06:31 [0.77s]
2008-11-19 17:05:24 [2.8s]
2008-11-19 16:47:40 [0.8s]
2008-11-19 16:45:05 [0.67s]
2008-11-19 16:44:55 [2.1s]
2008-11-19 16:41:27 [0.1s]
2008-11-19 16:30:25 [0.9s]
2008-11-19 16:21:11 [1.4s]
2008-11-19 16:19:33 [2.1s]
2008-11-19 16:19:22 [2.33s]
2008-11-19 16:13:59 [1.43s]

Figure 1. First search results obtained from queries for "bald people" (top 3 rows) and "red shirts" (bottom 3 rows, best seen in color). The facial details have been intentionally blurred for privacy.

In this work, we are interested in the analysis of people and in how we can use the extracted information to search for them in surveillance videos. Current research on this topic focuses on approaches based on face recognition [18, 2], where the goal is to establish the identity of a person given an image of a face. However, face recognition is still a very challenging problem, especially in low resolution images with variations in pose and lighting, which is often the case in surveillance data. State-of-the-art face recognition systems [4] require a fair amount of resolution in order to produce reliable results, but in many cases this level of detail is not available in surveillance applications.

We approach the problem in an alternative way, by avoiding face recognition and proposing a framework for finding people based on parsing the human body and exploiting part attributes. Those include visual attributes such as facial hair type (beards, mustaches, absence of facial hair), type of eyewear (sunglasses, eyeglasses, absence of glasses), hair type (baldness, hair, wearing a hat), and clothing color. While face recognition is still a difficult problem, accurate and efficient face detectors based on learning approaches [24, 7] are available. Those have been demonstrated to work well on challenging low-resolution images, with variations in pose and lighting. In our method, we employ this technology to design detectors for facial attributes from large sets of training data. We have an extensible architecture which could include non-visual attributes as well, such as temperature and voice obtained from other sensors. This enables efficient search on situations where face recognition does not perform well; in scenarios suited for the application of face recognition algorithms, the detected attributes could be integrated with face recognition approaches, making them more powerful.

Our technique falls into the category of *short term recognition* methods [1], taking advantage of features present in brief intervals in time, such as clothing color, hairstyle, and makeup, which are generally considered an annoyance in face recognition methods. There are several applications that naturally fit within a short term recognition framework. An example is in criminal investigation, when the police are interested in locating a suspect. In those cases, eyewitnesses typically fill out a *suspect description form*, where they indicate personal traits of the suspect as seen at the moment when the crime was committed. Those include facial hair type, hair color, clothing type, etc. Based on that description, the police manually scan the entire video archive looking for a person with similar characteristics. This process is tedious and time consuming, and could be drastically accelerated by the use of our technique. Another application is on finding missing people. Parents looking for children in an amusement park could provide a description including clothing and eyewear type, and videos from multiple cameras in the park would then be automatically searched for children matching those characteristics. Finally, the personal attributes could be used for tracking and correspondence across multiple related cameras, as it is very unlikely that people will change clothes or shave as they walk between two different places during a short time period. Our main contributions in this work are:

- We propose a novel framework based on human parts and their attributes to search for people in surveillance data;
- We implement the first video-based visual surveillance system with retrieval capabilities based on people's fine-grained parts and attributes. Figure 1 shows two queries, one for "bald people," and another for "red shirts," where keyframes for each video segment are displayed [1]. The performance of our approach is evaluated by conducting extensive experiments on Internet data and two-and-a-half months of surveillance video from a real environment;
- We show promising initial results on facial attribute detection in the thermal infrared domain. Attributes such as eyeglasses have traditionally been considered a nuisance in face recognition methods; we demonstrate how they can be used as a positive source of information in attribute-based search applications.

This paper is organized as follows. We begin by reviewing related work in Section 2. We then present our search framework in Section 3 and describe our implementation. Section 4 shows the architecture of our surveillance system, and Section 5 lists quantitative performance evaluation results. We present experiments on images collected from the Internet, on surveillance video from a real environment, and on infrared imagery.

2. Related Work

Our framework shares many aspects with short-term person recognition techniques, which use features present in short time periods, but that are not sufficient to establish identity of an individual in a longer term. Soft biometrics techniques [9] are examples of methods in this category. Clothing features [1, 5] have also been used to improve the accuracy of face recognition, and to cluster face images together when labeling large face sets [15]. Wu *et al.* [28] suggested the usefulness of clothing color in surveillance scenarios, and [26] mentions its application to tracking across multiple cameras.

Although many content-based retrieval methods have been proposed to search for objects in large image and video databases, our method is unique in the search based on fine-grained body parts and attributes. The MIT Photobook [14] is an interactive search tool that allows browsing of face images and other objects based on image content. Other works [19, 12] proposed the use of facial attributes to search for faces. However, they only deal with static face images.

Human parsing techniques [16] aim to segment the regions corresponding to individual body parts in images of people. Recently, some attention has been devoted to detecting specific facial features (such as hair [29], eyeglasses

[1] See the supplementary video for a demonstration of our system.

[27] and beards [13]). Our framework for searching based on body parts introduces the use of the semantic information (attributes) from the multiple parts in order to search for people in videos.

3. A Parts-and-Attributes Search Framework

In this section, we introduce a general framework for searching for people in surveillance data. The framework is composed of three main elements: sensors, body parts in multiple scales, and their attributes.

Sensors. These are responsible for collecting data from the environment where people may appear. The most obvious example is a video camera, but the concept can be generalized to include other sensors, such as temperature, audio, odor, and heartbeat rate. This would allow composite search across multiple modalities.

Body parts. Given a set of data streams captured from multiple sensors, techniques to detect and locate people are applied. Once people are found, a parsing of the human body is performed to segment its subparts (in streams where this is possible, such as video data). The parsing of the subparts may be further refined as long as resolution is sufficient to recognize the relevant details, in order to obtain a multiscale model of a person. Multiple sensors in different resolutions may be combined to allow parsing in increasing levels of detail.

Attributes. Finally, attributes from segments of the parsed body are extracted. They describe and form the signature of a person associated with a timestamp, and queries are made by specifying those. Considering an odor sensor, fragrance types can identify a person. In video cameras, considering the entire body as a single part, attributes such as height, body shape, actions and behavior detected during a short period could be useful cues for finding specific people. In a more detailed level, by parsing the body into head, torso and legs, the colors of the torso and legs provide information about the clothes that a person was wearing. Face analysis algorithms may be used to infer the person's gender, ethnicity and face shape, attributes that can be searched for. By increasing the resolution of the parsing, face subparts can be analyzed to extract additional information, such as facial hair type (in the mouth region), type of eyewear (in the eyes region), hair type and skin color. Similarly, hat and hair colors, and eyewear shape, could be further obtained.

3.1. Implementation Details

We have created a system that implements the framework described in the previous section. A particular case of the general framework is considered, where the sensor is a low-resolution (320x240) color video camera. In this section, we describe our implementation of a body parsing algorithm and the attribute detectors. In the current system, we have chosen to use a simple body parser for the purpose of demonstrating our search framework, but more advanced methods [16, 17, 21] could also have been applied. More sophisticated attribute classifiers [7] could have been used as well. The description in this section deals with attribute extraction from individual video frames or static images. In Section 4, we will then show how those components are applied to video and integrated with other modules of the surveillance system.

People are first detected by applying a face detector, and are further parsed to determine three major regions: face, torso and legs. We further divide the face region into three subregions: upper part, middle part, and lower part. Starting from this segmentation of the body, we extract the following attributes:

- Upper face part: hair type (bald, hair, wearing a hat);
- Middle face part: eyewear type (sunglasses, eyeglasses, absence of glasses);
- Lower face part: facial hair type (beard, mustache, absence of facial hair);
- Torso: dominant color;
- Legs: dominant color.

Figure 2 illustrates the parts and attributes from our implementation. We now proceed to the description of the techniques used to locate and analyze those parts in video frames.

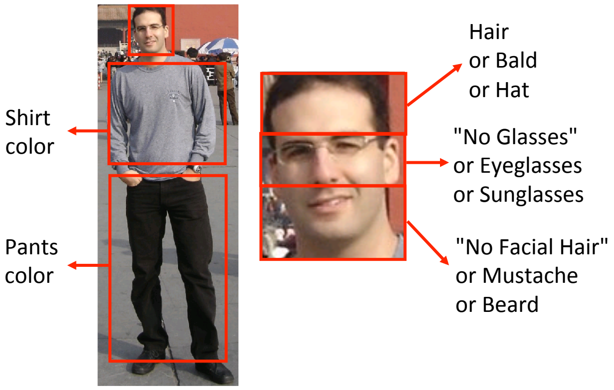

Shirt color
Pants color
Hair or Bald or Hat
"No Glasses" or Eyeglasses or Sunglasses
"No Facial Hair" or Mustache or Beard

Figure 2. Body parts and attributes considered in our implementation.

Face. For face detection, we use a cascade of Adaboost classifiers (see [24] for more details) based on Haar features, trained from image patches (of 20x20 size) containing faces. To extract facial attributes, we have trained nine additional Viola-Jones detectors, one for each attribute. A dataset of about 9,800 frontal face images (with slight pose variations) collected from the Labeled Faces in the Wild dataset [8] and from Internet sites has been created for this purpose. The images have been manually labeled by indicating bounding boxes for specific attributes (Figure 3). Table 1 shows the attribute distribution within the labeled images. Starting from the labeled images, slight variations of

Table 1. Number of the labeled images in the training set, and patch sizes used for training the detectors. For each sample, three variants are generated by random rotation and translation. Thus, the number of available training examples for each facial attribute was four times the value shown in this table.

20x20		30x20		20x20	
Beard	1,028	Sunglasses	677	Bald	562
Mustache	1,094	Eyeglasses	618	Hair	1,106
No Facial Hair	1,717	No Glasses	2,273	Hat	682

the samples were created by randomly rotating and shifting the original images, resulting in a set with approximately 39,000 images. The upper and lower face regions overlap with the middle region so that the eye region is included in both, as the eyes constitute important features for alignment while detecting the upper and lower region attributes.

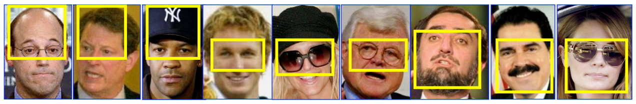


Figure 3. Examples of labelings for each facial attribute. From left to right: bald, hair, hat, "no glasses," sunglasses, eyeglasses, beard, mustache, "no facial hair."

To train each classifier, the labeled image patches were cropped, converted to grayscale and rescaled to the size listed in Table 1 in order to generate positive examples. Negative examples were generated from images of the two other classes from the same face region by cropping a slightly enlarged version of the labeled area (*e.g.*, while training the beard detector, negatives were selected from the mustache and "no facial hair" sets). In the feature selection process, Haar features based on gray-level intensities were used. Each level of the cascade was designed to reject about half of the number of negative samples while correctly accepting 99.5% of the true positive examples. About 3,000 positive examples were used for each classifier when this number was available (this holds for most attributes – see Table 1); otherwise, we used all available examples. 4,000 negative samples were obtained through bootstrap for training each level of the cascade.

When applying the Viola-Jones attribute detectors to the images, a sliding window is scanned near the corresponding regions of the detected faces. Since the application of the classifiers is constrained to regions near faces, training the classifiers with negative examples constrained in the same way (as described in the previous paragraph) led to better results when compared to training against examples that included, besides faces, general background images.

Torso and legs. To find the torso and legs and extract their colors, we first compute estimates of their location as rectangles with fixed aspect ratio, whose position and size are roughly estimated from the position and size of the detected face. Since people have different heights and body shapes, the rectangles may not precisely outline the torso and legs region. However, this is not a problem, as for the sake of this implementation we are interested only in obtaining the dominant color attribute for these parts.

Our dominant color classifier method is based on acquiring a normalized color histogram for each body part (torso or legs) in bi-conic (hue, saturation, luminance) HSL space. Euclidean distances in this color space closely map to human perception color differences.

We first quantize the HSL space into 8 colors – red, green, blue, yellow, orange, purple, black, and white. Given the estimated region for a body part, each pixel inside the region is quantized into one of these colors and a histogram with 8 bins is built. The color corresponding to the bin which receives the majority of votes is then assigned as the dominant color for the body part.

In the process above it is left to explain how we quantize the HSL space. This is done in a two-stage process. In the first stage, we partition the color space using hue information alone, by computing the hue angular cutoffs between the colors. Then, in the second stage, pixels may be relabeled as either white or black depending on whether they lie outside the lightness/saturation curve above or below the horizontal mid-plane. This is related to earlier work in color segmentation performed by Tseng and Chang [22].

When dealing with multiple cameras, we refer to the works on color matching by [28] and [10], which learn the inter-camera brightness transfer functions.

4. Video Surveillance System Architecture

In this section, we describe the architecture of our video surveillance system. The system can be divided into three main components: the *analytics engine*, which implements the computer vision algorithms described in Section 3.1; the *database backend*, which indexes the videos based on the attributes detected by the analytics engine; and the *search interface*, which connects to the database backend and allows the user to issue queries. Figure 4 illustrates the multiple components of the system architecture and their interactions. We now proceed to a more detailed description of the system.

Analytics Engine. This module is responsible for all computer vision operations. The video input can be obtained from sources such as a file in the hard disk, a camera or an Internet URL. Each frame of the video stream is then processed in the following way: first, a background subtraction algorithm [20] is applied to detect objects in motion. A Viola-Jones face detector is then scanned through the image, but its application is constrained to the motion regions. Detected faces are tracked through multiple frames in the following way: if, for a given frame, there is a detection from the Viola-Jones detector, the face location is assumed to be at the detected location; otherwise, the most recent

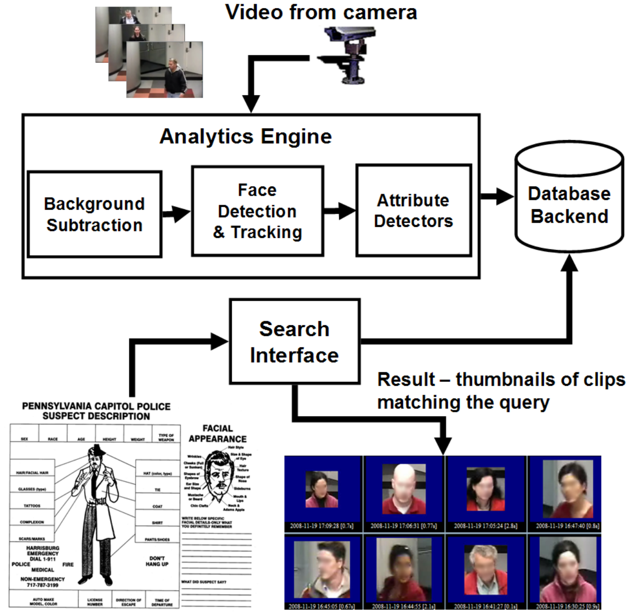


Figure 4. High-level view of our surveillance system's architecture. The faces in the result have been intentionally blurred for privacy.

face detection result is matched to its neighboring regions based on correlation. The tracker determines that a person left the scene when the face detector fails to detect a face through a given number of frames (typically 10). Each person who leaves the scene is assigned a unique identifier ("track ID"). This is important, as we are interested in combining the results of individual attribute detectors at multiple frames in order to assign a single set of attributes to a given person.

Given a face's bounding box, the torso and legs detector described in Section 3.1 is run to find the corresponding bounding boxes. The attribute detectors are then applied to the regions delimited by each box, *i.e.*, the facial feature classifiers are applied to the three face subregions, and the color histograms are computed from the regions that correspond to the torso and legs. An integrator module collects the results from all detectors and keeps track of the detection history for the multiple frames of each track ID. Once the face tracker determines that a given person left the scene, the integrator combines the attribute results from all frames from that track to produce the final classification for that person. For the facial attributes, the final result is determined by voting, by choosing, for each face subregion, the attribute with largest number of detections during the entire track. The integrator also collects the color histograms of each body part and builds a single cumulative histogram (for each part) to determine the final dominant colors. The final result is a tuple represented by (track ID, a face keyframe, a short video of the track, facial hair type, eyewear type, hair type, torso color, legs color), which is sent to the database backend using a metadata format in order to be stored along with a timestamp.

Table 2. Distribution of the images used for testing the facial attribute detectors.

Lower Region	858	Middle Region	1044	Upper Region	796
Beard	316	Sunglasses	308	Bald	268
Mustache	302	Eyeglasses	318	Hair	284
No Facial Hair	240	No Glasses	418	Hat	244

While the Viola-Jones face detector achieves real-time performance (25/30Hz) running on 320x240 images, the facial attribute detectors are composed of a larger number of features and cannot be run in real-time. To overcome this problem, a multithreading architecture could be implemented, where a main thread would execute the background subtraction, face detection and face tracking steps, and other threads running in parallel would receive the bounding boxes for each frame and would apply the attribute detectors. In this way, the result would be generated with a very short delay, but the face tracker and capture subsystem would run with real-time performance, avoiding frame drops.

Database Backend. The database backend module receives metadata results from the analytics engine. Those are indexed into a relational database to enable efficient search.

Search Interface. A HTML search interface where the user has access to controls that allow him/her to describe a person's attributes, in a way similar to a suspect description form, was developed. The user can also constrain the search results to match a given time period. The interface issues requests to a web server, where Java servlets receive the information and issue queries to the database backend. The results are then presented to the user. Thumbnails of the detected faces are displayed, and the user can click on them to view a video clip of the selected person. Figure 1 shows results of queries for "bald people" and "red shirts."

5. Experimental Results

We performed experiments using static images collected from the Internet, and a 2.5-month long video surveillance stream. Also, initial experiments on the use of infrared imagery to detect attributes have been conducted.

5.1. Static Images

We collected 2,698 quasi-frontal face images from Internet sites for evaluating our facial attribute detectors. The images in this set are distinct from the images used for training. The set was split into three parts, one for testing the lower region detectors, one for the middle region detectors, and one for the upper region. Table 2 shows the distribution of attributes within the test sets.

Our facial attribute detectors from Section 3.1 were applied to every image. Figure 5 displays a plot with the

ROC curves for all detectors. This demonstrates the point that machine learning techniques in conjunction with large amounts of training data can be used to design reliable detectors for facial attributes. We used well-known Viola-Jones detectors trained from frontal faces, based on Haar features and grayscale images. However, more sophisticated and robust classifiers [7] could have been used, and viewpoint variations could be handled with multiple detectors for each of the views, along the same lines of multiview face detection [25].

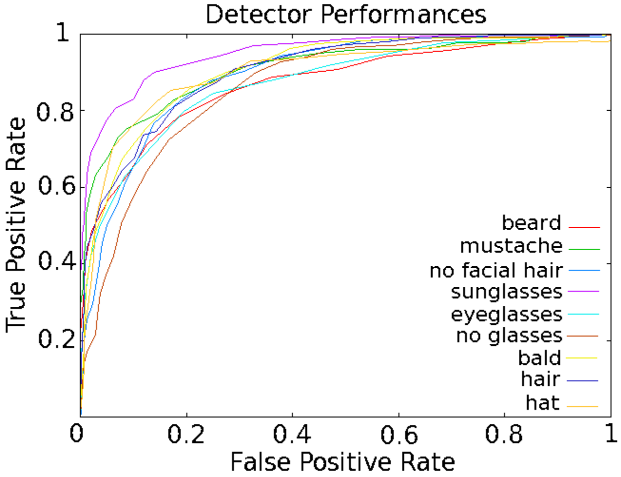


Figure 5. Receiver Operating Characteristic curves for the facial attribute detectors, evaluated on a set of face images collected from the Internet (best viewed in color).

5.2. Surveillance Video

We have developed a complete surveillance system that implements our parts-and-attributes search framework, and used it to process a 320x240 video stream during a 2.5 month period. The video was captured by a camera placed at the entrance of the IBM Research Center in Hawthorne, New York. The camera is located at the ceiling and points at a door from where people enter the building. The environment has strong direct illumination (which casts shadows), and people who walk through the scene show high variability in face pose – many are facing left or right, or looking down. This can prevent the camera from seeing all facial features in some cases. In other words, the statistical nature of the face images captured at this scene (such as the images in Figures 1 and 6) greatly differs from that of the images used for training our detectors (Figure 3). Our experience with the application of face recognition systems in surveillance (including a few of the best systems from the Face Recognition Vendor Test [4]) shows that the algorithms fail miserably in this scenario, due to variations in pose and lighting. Some face recognition systems cannot even be run because the minimum resolution requirements are not met.

Table 3. Confusion matrices for the facial attribute classifiers in surveillance video, along with their precision and recall statistics.

	Classification result				
Ground truth	Beard	Mustache	No facial hair	Precision	Recall
Beard	139	6	59	56%	68%
Mustache	1	28	23	50%	54%
No facial hair	108	22	1863	96%	93%
Ground truth	Sunglasses	Eyeglasses	No glasses	Precision	Recall
Sunglasses	23	1	41	85%	35%
Eyeglasses	0	45	463	90%	8.9%
No glasses	4	4	1668	77%	99.5%
Ground truth	Bald	Hair	Hat	Precision	Recall
Bald	31	68	6	100%	42%
Hair	0	1906	183	97%	91%
Hat	0	0	55	29%	100%

Facial Attributes. We computed the confusion matrix for the classification results for 2095 people captured by the face detector. It is shown in Table 3, along with precision and recall statistics for each face attribute. This evaluation should be interpreted as an approximation, as in some cases it is impossible to tell (even for a human observer) the difference between some attributes, such as eyeglasses and "no glasses," due to the surveillance conditions.

We can see in the table that some detectors perform well, while others are not at the same level. Below, we analyze the main causes for failure for these detectors, and propose solutions to overcome their issues.

The main problem faced by the eyeglasses detector is the lack of resolution. Thus, faces in this category had often been mistaken for "no glasses." For sunglasses, beards and mustaches, shadows cast by the direct light source are the main source of errors. Notice that the illumination can cause shadows along the neck, generating a pattern which is very similar to a beard. Also, many faces of people looking slightly down are found due to the camera's placement, and clothing can be similar in appearance to a beard. The light sources can also cast a small shadow along the nose, generating features that resemble mustaches. In the upper face part, we notice that many fringe hairs are misclassified as hats. The pose variability in this environment was also very different than in our training set. Figure 6 shows the main causes for failure cases.

These issues may be addressed by training multi-view attribute detectors, and by combining features in the infrared domain with features in the visible domain to aid in the disambiguation of failure cases. Figure 7(a) shows two visible/thermal infrared image pairs of people with different facial attributes. In the infrared image, the transparency of the eyeglasses is no longer an issue, as they become similar to sunglasses, while the visible domain image can still be used to distinguish between sunglasses and eyeglasses. In addition, the infrared image accentuates the contrast between skin-color and hair regions and is less sensitive to

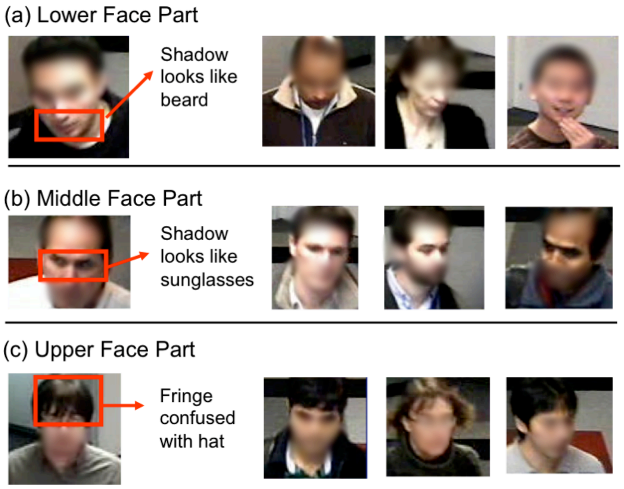
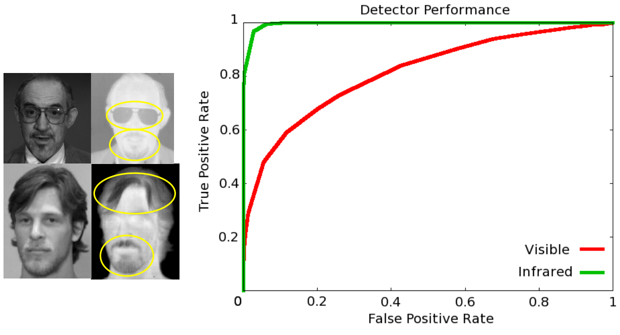


Figure 6. Examples of failure cases in our system. Notice the differences in pose and lighting as compared to the images from Figure 3, representative of our training set. The images have been intentionally blurred for privacy.

problems caused by shadows. While facial attributes such as eyeglasses and facial hair have traditionally been seen as a nuisance in face recognition methods, and IR imagery has been used to deal with those (*e.g.*, [11] uses infrared images to explicitly remove eyeglasses as a preprocessing step for recognition), in our method infrared imagery enables us to improve the accuracy of our attribute detectors, which can then be used as a positive source of information in attribute-based search applications.

To illustrate, we have conducted experiments with eyeglasses detection using a subset of the Equinox dataset [3], composed of face images from 20 subjects taken with varying illumination conditions and facial expressions. A total of 4500 image pairs was used in our experiments, where each pair consisted of a visible domain image and a midwave infrared image, captured from slightly different angles. From the 4500 pairs, 1796 had eyeglasses and 2704 did not have them. The images were subsampled to 60 x 80 resolution in order to simulate the lack of resolution present in our surveillance scenario. We then performed two experiments: eyeglasses detection in the visible domain, and eyeglasses detection in the infrared domain. In the first experiment, we applied the eyeglasses detector to all visible domain images while varying the detection threshold, in order to generate a ROC curve. In the second experiment, we applied the sunglasses detector to the infrared domain images (as the eyeglasses look like sunglasses in this domain), and also generated a ROC curve. Figure 7(b) shows the ROC curves for each detector. The results confirm that this problem is challenging in the visible domain, but we obtain excellent performance in the infrared domain.



(a) (b)

Figure 7. Facial attributes as a positive source of information in the infrared domain. (a) Facial attributes can be easier to identify in the IR domain; (b) ROC curves for eyeglasses detection in the visible and infrared domains (red = visible, green = infrared).

Table 4. Confusion matrix for shirt color classification, evaluated over a 2.5-month-long surveillance video stream. The leg color classification accuracy is similar.

Ground truth	Classification result							
	Red	Green	Blue	Yellow	Orange	Purple	Black	White
Red	49/51	0	0	0	0	0	2	0
Green	0	0/2	0	0	0	0	2	0
Blue	0	0	83/104	0	0	0	18	3
Yellow	0	0	0	5/7	0	0	1	1
Orange	0	0	0	0	1/1	0	0	0
Purple	0	0	0	0	0	1/3	2	0
Black	0	0	0	0	0	0	1263/1263	0
White	0	0	3	0	0	0	0	87/90

Color. In order to evaluate the performance of the color classifier, we analyzed the results obtained for torso (shirt) color. The classifier thresholds were especially tuned for the conditions of our environment. Table 4 displays the confusion matrix for colored shirts for people whose torsos were visible. The results are excellent, and the performance of the leg color classifier is similar.

6. Conclusion and Future Directions

We have proposed a novel approach for people search in surveillance data, characterized by three main elements: sensors, body parts, and their attributes. In addition, we have developed the first video-based surveillance system that implements this framework. Its usefulness is demonstrated by extensive experimentation with static images and real surveillance video. Part of this success is due to the power of learning-based object detectors trained from large amounts of data.

The search framework is closely related to several active areas of research in computer vision and machine learning, and it could greatly benefit from advances in these fields. Work on human body parsing is especially interesting. While there are existing works in the direction of general image parsing [23] and people parsing [16], not much attention has been given to extracting semantic attributes for individual parts. Only a few isolated works exist (such as [29], on hair detection), and we hope that our framework and system will inspire the community to conduct additional studies in this field.

In the future, we plan to improve the performance of the attribute detectors by retraining them using data collected from our surveillance environment. Practical issues arise, as this dataset is huge, due to the presence of multiple frames from each person. Thus, using semi-supervised learning techniques may be useful. We also would like to integrate additional attributes into the system, such as gender, age, skin color, and clothing texture, in order to expand the range of search possibilities. Activity recognition algorithms may also be incorporated, enabling search for specific actions.

Acknowledgments

This work was supported in part by the National Science Foundation under award number IIS-0535293. We thank Lisa Capetola and Blanca Cambi for collecting and labeling the face datasets, Ioannis Pavlidis for sharing the Equinox dataset images used in the infrared experiments, and the anonymous reviewers for the constructive feedback.

References

[1] D. Anguelov, K.-C. Lee, S. B. Göktürk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, 2007. 2

[2] Cognitec. http://www.cognitec-systems.de/. 1

[3] Equinox Human Identification at a Distance Database. http://www.equinoxsensors.com/products/hid.html. 7

[4] Face Recognition Vendor Test. http://www.frvt.org/. 2, 6

[5] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, 2008. 2

[6] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst., Man, Cybern., Pt. C*, 34(3):334–352, 2004. 1

[7] C. Huang, H. Ai, Y. Li, and S. Lao. Vector boosting for rotation invariant multi-view face detection. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, Beijing, China, 2005. 2, 3, 6

[8] G. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *ECCV Workshop on Faces in Real-Life Images*, Marseille, France, 2008. 3

[9] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Intl. Conf. on Biometric Authentication*, pages 731–738, 2004. 2

[10] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, 2005. 4

[11] S. G. Kong et al. Multiscale fusion of visible and thermal IR images for illumination-invariant face recognition. *Intl. Journal of Computer Vision*, 71(2):215–233, 2007. 7

[12] N. Kumar, P. N. Belhumeur, and S. K. Nayar. FaceTracer: A Search Engine for Large Collections of Images with Faces. In *European Conf. on Computer Vision (ECCV)*, Marseille, France, 2008. 2

[13] M. H. Nguyen, J.-F. Lalonde, A. A. Efros, and F. De la Torre. Image-based shaving. *Computer Graphics Forum Journal (Eurographics 2008)*, 27(2):627–635, 2008. 2

[14] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. In *SPIE Storage and Retrieval for Image and Video Databases II*, 1994. 2

[15] D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007. 2

[16] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, 2005. 2, 3, 7

[17] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *European Conf. on Computer Vision (ECCV)*, Copenhagen, Denmark, 2002. 3

[18] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: video shot retrieval for face sets. In *International Conference on Image and Video Retrieval*, 2005. 1

[19] K. Sridharan, S. Nayak, S. Chikkerur, and V. Govindaraju. A probabilistic approach to semantic face retrieval system. *Lecture Notes in Computer Science*, 3546:977–986, 2005. 2

[20] Y. Tian, M. Lu, and A. Hampapur. Robust and efficient foreground analysis for real-time video surveillance. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, 2005. 4

[21] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2007. 3

[22] D.-C. Tseng and C.-H. Chang. Color segmentation using UCS perceptual attributes. In *Proc. Natl. Sci. Council: Part A*, volume 18, pages 305–314, 1994. 4

[23] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *Intl. Journal of Computer Vision*, 63:113–140, 2005. 7

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR)*, Kauai, Hawaii, 2001. 2, 3

[25] P. Wang and Q. Ji. Learning discriminant features for multi-view face and eye detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, California, 2005. 6

[26] Y.-F. Wang, E. Y. Chang, and K. P. Cheng. A video analysis framework for soft biometry security surveillance. In *ACM Intl. Works. on Video Surveill. & Sensor Networks*, 2005. 2

[27] B. Wu, H. Ai, and R. Liu. Glasses detection by boosting simple wavelet features. In *International Conf. on Pattern Recognition (ICPR)*, Cambridge, UK, 2004. 2

[28] G. Wu et al. Identifying color in motion in video sensors. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, New York, New York, 2006. 2, 4

[29] Y. Yacoob and L. S. Davis. Detection and analysis of hair. *IEEE Trans. on Patt. Analysis and Mach. Intelligence*, 28(7):1164–1169, 2006. 2, 7