R4-A.3: Human Detection and Re-Identification for Mass Transit Environments

I. PARTICIPANTS

<table>
<thead>
<tr>
<th>Faculty/Staff</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Title</td>
<td>Institution</td>
<td>Email</td>
</tr>
<tr>
<td>Rich Radke</td>
<td>PI</td>
<td>RPI</td>
<td><a href="mailto:rjradke@ecse.rpi.edu">rjradke@ecse.rpi.edu</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Graduate, Undergraduate and REU Students</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Degree Pursued</td>
<td>Institution</td>
</tr>
<tr>
<td>Srikrishna Karanam</td>
<td>PhD</td>
<td>RPI</td>
</tr>
</tbody>
</table>

II. PROJECT DESCRIPTION

A. Project Overview

Large networks of cameras are ubiquitous in urban life, especially in densely populated environments such as airports, train stations, and sports arenas. For cost and practicality, most cameras in such networks are widely spaced so that their fields of view are non-overlapping. Automatically matching humans who re-appear across different cameras in such networks is a critical problem in homeland-security-related surveillance applications.

This issue is highly related to the computer vision research problem of human re-identification or “re-id”. Given a cropped rectangle of pixels representing a human in one view, a re-id algorithm produces a similarity score for each candidate in a gallery of similarly cropped human rectangles from a second view. Computer vision research in re-id largely focuses on two challenging issues. The first is feature selection: determining effective ways to extract representative information from each cropped rectangle to produce descriptors. The second issue is metric learning: determining effective ways to compare descriptors from different viewpoints. These work together so that images of the same person from different points of view yield high similarity, while images of different people yield low similarity.

However, feature selection and metric learning only represent two aspects of creating an effective real-world re-id algorithm. In practice, a re-id system must be fully autonomous from the point that an end user draws a rectangle around a person of interest to the point that candidates are presented to them. This implies that the system must automatically detect and track humans in the field of view of all cameras with speed and accuracy. The candidates in the re-id gallery in practice are, thus, automatically generated and are typically much lower-quality than the hand-curated gallery of a benchmark dataset; in fact, many candidate rectangles may not even represent humans. Furthermore, in a typical branching camera network, the camera in which the target reappears is unknown, so there are actually several separate galleries to search. The timing of the reappearance is also unknown; the galleries will be constantly updated with new candidates over the course of minutes or hours instead of being presented to the algorithm all at once. Finally, real-world re-id maps naturally onto a multi-shot problem. That is, there are multiple images available to describe both the target and the matching candidates, since after a target of interest is detected in the field of view of one camera, he/she is usually tracked until leaving the current view.
This project addresses the design and deployment of real-world re-id algorithms specifically designed for mass transit environments. This involves:

- The design and analysis of new computer vision algorithms for human detection and tracking, feature selection, and metric learning problems for re-id; as well as
- The evaluation of the suitability of such algorithms for real-world homeland security applications, taking into account tracking/detection errors, latency/congestion, and human-computer interfaces to software systems.

The ideal end-state of the research is a suite of re-id algorithms that are directly applicable to the homeland security enterprise (HSE) and ready for large-scale system integration. The project will also produce an up-to-date assessment of the re-id state of the art, which will inform Department of Homeland Security (DHS) stakeholders about what is technologically feasible in this area, thus informing policies and technology solicitations.

B. Biennial Review Results and Related Actions to Address

The Biennial Review team commended the project on the rigor of its research, the extensive and challenging real-world experiments and evaluation, and the dissemination of its results in top computer vision conferences and journals. The reviewers also singled out the ability of the research team to translate theoretical research into real-world practice in an actual end-to-end surveillance system. On the other hand, the reviewers claimed that the work did not explore some recent trends in the re-id literature, relying on standard approaches and methods (see response below), and that more specific system requirements are necessary.

One identified weakness (that the camera views must be contiguous) is a misinterpretation of our actual approach.

In Year 4, we will address the perceived weaknesses in two ways. First, the Rensselaer Polytechnic Institute (RPI) research group, in collaboration with Dr. Camps’ group at Northeastern University (NEU, project R4-A.1), is about to disseminate an exhaustive community-wide evaluation standard and benchmarking dataset for re-id, investigating hundreds of combinations of feature extraction algorithms, metric learning strategies, and multi-shot ranking methods. Each combination is applied to more than fifteen benchmarking datasets, including a new, challenging dataset extracted from the research camera testbed at the Cleveland Hopkins International Airport (CLE) in Year 3. This evaluation and benchmarking effort will: (1) Suggest the overall most promising combinations of algorithms for different kinds of data; (2) Characterize achievable performance of re-id to set end-user expectation; and (3) Suggest potentially fruitful avenues for research that are likely to yield the most significant improvements in performance. It is likely that the methods and datasets referred to by the reviewers have already been incorporated into our extensive evaluation. The second approach to addressing the perceived weaknesses is to investigate re-id methods based on deep convolutional neural networks (CNNs), which have recently achieved unprecedented success for many computer vision problems. We will evaluate CNN and related cutting-edge approaches using our evaluation and benchmark, integrating them into our pipeline as they prove themselves against the state of the art.

C. State of the Art and Technical Approach

C.1. State of the art

The traditional paradigm for solving the person re-id problem is to extract appearance features of the target and each candidate and then compare the feature vectors using a distance metric. This has given rise to two different research paths: appearance modeling and metric learning. Most re-id algorithms describe the appearance using texture and color histograms [1, 2]. To learn distance metrics, most methods focus on learning Mahalanobis-like distances [3-5]. However, these methods are designed for the single-shot setting; i.e., they rely on comparing the feature vector of one probe image with the feature vector of one gallery image.
The naive way to extend such methods to the multi-shot setting is to compare every possible pair of probe and gallery images and aggregate the results.

Several methods specifically tackle the multi-shot re-id problem. For example, Cong et al. [6] used image sequences to build aggregated appearance descriptors. Wang et al. [7] proposed an algorithm that selects discriminative fragments to learn a video ranking function. Li et al. [8] learned discriminative random forests and aggregated classification scores for all the available images for each person to make a decision. Image sequences have also been used to perform direct sequence matching. Simonnet et al. [9] used dynamic time warping to perform temporal sequence matching. The multi-shot re-id problem has also been formulated as a gait recognition problem [10], where person discrimination is based on the walking style.

C.2. System description

The key computer vision aspects of our deployed system include human detection and tracking, feature selection, and descriptor comparison for re-id. Figure 1 illustrates the main steps of the process.

C.3. Detection and tracking

The first step is using mixtures of Gaussians (MoG) [11] to detect foreground pixels and group them into blobs; the bounding boxes of these blobs define regions of interest (ROIs). ROIs with small sizes or impossible locations are discarded. Each viable ROI is input to the aggregated channel features pedestrian detector of Dollar et al. [12], as illustrated in Figure 2 on the next page. This detector uses a boosted decision tree classifier to rapidly generate pedestrian candidates. We found it was important to train a specific classifier for each camera in the network to obtain good results, which was accomplished using 500 pedestrian images from each camera and randomly sampled background images (to create negative samples). The pedestrian detection runs at several scales within each ROI, resulting in a set of candidate detections of different sizes within each foreground blob. Because our system must run in real time, it was critical to restrict the candidate search to only viable ROIs, resulting in a human detector that runs at about 100 frames per second.
Our approach to tracking the detected human candidates is twofold. First, we perform tracking-by-detection in each frame as described above. Second, another set of candidate bounding boxes is generated in each frame by predicting the bounding box locations of tracked pedestrians from the previous frame. This prediction is made by detecting low-level FAST corner features [13] in each previous bounding box, removing features estimated to belong to the background [14], estimating the motion vector for each feature with the Kanade-Lucas-Tomasi feature tracker [15], and averaging the resulting motion vectors to update the location of the bounding box in the current frame.

The tracking-by-detection and motion-prediction bounding boxes are merged at the current frame to produce a final set of human detections as follows. We compute the intersection of each tracking-by-detection bounding box with each motion-prediction bounding box and find the maximum ratio between the area of intersection and the area of the smaller bounding box. The new tracking-by-detection box is associated with the corresponding motion-predicted box if this ratio is above a predefined threshold (in our experiments, we used 0.8); otherwise, it is used to initialize a new track. Motion-predicted bounding boxes not matching any tracking-by-detection box in the previous frame are retained if both their aspect ratio and location in the frame are plausible.

C.4. Re-identification

The re-id process has three key steps. First, a feature descriptor needs to be extracted from each candidate detection. Second, given a pair of descriptors $X_{\text{target}}$ and $X_j$ (one from the tagged target and the other from the $j^{th}$ candidate detection), we must compute an appropriate similarity score $s_j = f(X_{\text{target}}, X_j)$. Finally, by ranking the similarity scores $s_j, j = 1, 2, \ldots, n$ in each frame, an ordered list of “preferred” candidates to be shown to the user is generated.

For feature extraction, we use texture and color histograms, which are popular descriptors for person re-id [16]. Following the approach of Gray and Tao [17], we divide the image into six horizontal strips. In each strip, we first compute filter responses of 13 Schmid and 6 Gabor filters. The filter responses are then used to compute a histogram with 16 bins. To describe the color information in each strip, we compute the 16-bin histograms in the whitened RGB space, the HSV space, and the YCbCr space. This results in a 432-dimensional feature vector for each strip. The feature vectors for all of the 6 strips are concatenated to form a 2592-dimensional feature vector.

Given the feature vectors $g_j$ for the gallery images and $p_j$ for the probe images, where $j$ denotes the $j^{th}$ image of the $i^{th}$ unique person, computed as described above, we then learn feature transformation using local Fisher discriminant analysis (LFDA) [18]. For the sake of notational convenience, let us define the matrix $F$ of all the feature vectors $g_j$ and $p_j$ as $F = \{g_j\} \{p_j\}$. The traditional Fisher discriminant analysis (FDA), which minimizes the within-class and maximizes the between-class scatter, fails to give satisfactory results if the
input data is multi-modal. Indeed, in the multi-shot re-id problem, the data is multi-modal since each person in the gallery view and the probe view has multiple images. To this end, we employ Local Fisher discriminant analysis (LFDA), wherein locality preserving projections [19] are used to ensure that the feature vectors of each person are close in the embedding space, thereby preserving the local structure of the data. Specifically, we first define an affinity matrix $A$ that captures the closeness of the feature vectors $F_a$ and $F_b$, where $F_a$ is the $a^{th}$ column of $F$. The value is $A_{ab} = 1$ if $F_a$ and $F_b$ are close to each other; otherwise it is set to 0. Here, we use the k-nearest neighbors rule with $k=7$ to determine this closeness.

We then define the local within-class and between-class scatter matrices $S_w$ and $S_b$ as:

$$S_w = \left( \frac{1}{2} \right) \sum_{a,b=1}^{N} A_{ab} (F_a - F_b)(F_a - F_b)^T$$

$$S_b = \left( \frac{1}{2} \right) \sum_{a,b=1}^{N} A_{ab} (F_a - F_b)(F_a - F_b)^T$$

where $N$ is the total number of available images and $A_{ab}^w$ and $A_{ab}^b$ are defined as:

$$A_{ab}^w = \begin{cases} 
\frac{N}{n_c}, & \text{class}(F_a) = \text{class}(F_b) = c \\
0, & \text{class}(F_a) \neq \text{class}(F_b) 
\end{cases}$$

$$A_{ab}^b = \begin{cases} 
\frac{1}{N}, & \text{class}(F_a) = \text{class}(F_b) = c \\
\frac{1}{N}, & \text{class}(F_a) \neq \text{class}(F_b) 
\end{cases}$$

where $n_c$ is the number of images available for person with index $c$. The feature transformation $T$ is then learned as:

$$T = \arg\max_T \text{trace}\left((T^T S_w T)^{-1} T^T S_b T\right)$$

Now, we project the feature vectors of the gallery and probe images using this transformation matrix $T$, and compute the mean feature vector to form the average descriptors for each person in each of the gallery and probe cameras. The next step is to find a metric to accurately quantify similarity. Many metric learning techniques have been proposed for re-id [20, 3, and 12]. In our implementation, we applied the RankSVM (support vector machine) method [16] to maximize the norm of a weight vector $W$ subject to the constraints that if $d_{same} = \left| x^a_i - x^b_i \right|$ is the absolute difference of two descriptors of the same person $i$, and $d_{diff} = \left| x^a_i - x^b_j \right|$ is the absolute difference of descriptors of two different people $i$ and $j$, then $W^T d_{same} < W^T d_{diff}$ for all possible pairs from same and different people. The idea is to minimize the norm of $W$ that satisfies the following ranking relationship

$$W^T \left( \left| x^a_i - x^b_i \right| - \left| x^a_i - x^b_j \right| \right) > 0, \quad i, j = 1, 2, ..., P \text{ and } i \neq j$$

where $x^a_i$ is the mean feature vectors of person $i$ in camera $a$, projected by the learned transformation matrix $T$ as described above, and $P$ is the total number of training subjects. The RankSVM method finds $W$ by solving the problem

$$\arg \min_{W, \xi} \left( \frac{1}{2} \left| W \right|^2 + C \sum_{i=1}^{P} \xi_i \right)$$

$$s.t. \ W^T \left( \left| x^a_i - x^b_i \right| - \left| x^a_i - x^b_j \right| \right) \geq 1 - \xi_i, \quad \xi_i \geq 0$$
where $\xi_i$ is a slack variable. The re-id distance function between the two descriptors $X_{\text{target}}$ and $X_j$ is then computed as $f(X_{\text{target}} X_j) = W^T |X_{\text{target}} - X_j|$. 

In the approach described above, the idea of computing the mean feature vector, given the descriptors corresponding to multiple images of the same person, is naïve and does not fully exploit the available discriminative information for each person. In Year 3, we developed specialized algorithms to address this multi-shot re-id problem, which are described next.

### C.5. Discriminative dictionary learning for person re-id

Dictionaries learned from data have recently achieved impressive results in several classification and recognition problems [21]. This can be attributed to their strong representational power. Here, we present a technique to learn a dictionary that is capable of discriminatively encoding the feature vectors of different people. Given representative feature vectors computed from the available images for each person, we learn a single dictionary capable of adapting to the variations across camera views. Additionally, we incorporate explicit constraints on the feature representations, with respect to the dictionary, into our problem formulation, providing the dictionary with a strong discriminative ability.

#### C.5.a. Problem specification

Let $p_i$ be the average feature vector computed from all the available feature vectors for the person with index $i$ in the probe view and $g_j$ be the corresponding feature vector in the gallery camera view. Consider three such feature vectors $p_i, g_j,$ and $g_j$. We seek to learn a dictionary $D$ that is capable of discriminating the representations corresponding to the feature vectors $p_i, g_j,$ and $g_j$. Specifically, let $a_i, b_j,$ and $b_{g_j}$ be the representations of these feature vectors with respect to the dictionary $D$, respectively. To clarify the notation, the representation of a probe feature vector $p_i$, with respect to the dictionary, is denoted $a_i$. The representation of a gallery feature vector $g_j$, with respect to the dictionary, is denoted $b_j$. We compute $a_i$ (and equivalently $b_j$) by solving the following problem:

$$a = \text{argmin}_a \| p - Da \|^2_2 + \lambda \| a \|^2_2$$

which has a closed-form solution, given by $a = (D^TD + \lambda I)^{-1}D^Tp$.

Since $p_i$ and $g_j$ are the feature vectors, albeit in different camera views, of the same person, our hypothesis is that $b_j$ will have a smaller Euclidean distance to the gallery representation $a_i$ than $b_j$. The intuition here is that the images of the same person in different camera views should have similar representations with respect to the learned dictionary.

#### C.5.b. Problem formulation

Given the feature vectors of the $N$ people in both cameras $p_i$ and $g_j$, $i = 1, \ldots, N$, our goal is to learn a dictionary $D$ such that the representations of $p_i$ is closer to $g_j$ when compared to $g_j, j \neq i$. From these feature vectors, we gather the pairs $(p_i, g_j), i = 1, \ldots, N$, corresponding to the same people (the “positive” pairs) and the pairs $(p_i, g_j), i = 1, \ldots, N, j = 1, \ldots, N, i \neq j$, corresponding to different people (the “negative” pairs). We now define two matrices, $P = [p_1, p_2, \ldots, p_N, p_1, p_2, \ldots, p_N]$ and $G = [g_1, g_2, \ldots, g_N, g_1, g_2, \ldots, g_N]$, where $K$ is the number of negative pairs considered. Note that the first $N$ columns of $P$ come from the probe points of the positive pairs, and the next $K$ columns come from the probe points of the negative pairs $N$. A similar notation holds for $G$. Letting $P$ and $G_i$ respectively denote the $i^{th}$ column of $P$ and $G$, we see that $P$ and $G$ have a one-to-one column-to-column correspondence. For example, $P_i$ and $G_i$ represent a positive pair whereas $P_{N+i}$ and $G_{N+i}$ represent a negative pair. With this background and $T=N+K$, we formulate the following unconstrained nonlinear optimization problem:
Here, $P_i - Da_i$ is a reconstruction term that ensures each of the probe and gallery points are well represented by the learned dictionary $D$. $\lambda(\|a_i\|_2^2 + \|b_i\|_2^2)$ is a regularization term on the representations of each of the probe and gallery feature vectors with respect to the dictionary, and $c_i(a_i - b_i\|_2^2)$ is the term that ensures discriminability of the dictionary. Specifically, if $a_i$ and $b_i$ are representations corresponding to the same person, we set $c_i$ to a large value. In our experiments, we use $c_i = 10$ for this case. On the other hand, if $a_i$ and $b_i$ are representations corresponding to different people, we set $c_i$ to a relatively small value. In our experiments, we use $c_i = 0.01$ for this case. Intuitively, $c_i$ is a weight factor determining the extent to which we would like $\|a_i - b_i\|_2^2$ to be minimized. For a large value of $c_i$, in the case of representations corresponding to the same person, we place more emphasis on minimizing this term, in which case we seek to learn representations that are close. Similarly, for a small value of $c_i$, in the case of representations corresponding to different people, we place relatively less emphasis on minimizing this term, thereby seeking to learn representations that are relatively far.

Defining the two matrices $A$ and $B$, with $a_i$ and $b_i$ as their $i^{th}$ columns, the above optimization problem can be restated as:

$$\min_{D,A,B} \| P - Da \|_2^2 + \| G - Db \|_2^2 + \lambda(\|a\|_2^2 + \|b\|_2^2) + \sum_{i=1}^{p} c_i \|a_i - b_i\|_2^2$$

which can be solved using the alternating directions minimization approach by fixing two of the three variables $D, A, B$ and optimizing for the third.

**C.5.c. Re-identification**

Given the gallery feature vectors $g_i$, $i = 1, \ldots, p$, we propose the following steps to re-identify a person represented by a probe feature vector $p_u$:

- For each gallery feature vector $g_i$, compute the corresponding representations with respect to the learned dictionary $D$ as $b_i = (D^T D + \lambda I)^{-1} D^T g_i$.
- Similarly, compute the representation $a_u$ of the unknown probe feature vector $p_u$ with respect to $D$.
- Assign the index $i$ of the representation $b_i$ with the least Euclidean distance to $a_u$.

This procedure is summarized in Figure 3 on the next page.
C.6. Learning discriminative affine hull representations

The multi-shot re-id problem raises the following two critical questions:

• How do we describe the available “multiple instance” data for each person?
• How do we exploit this data description to learn feature representations that are sufficiently discriminative to perform accurate re-id?

Representing multiple instance data in the context of recognition has been a longstanding problem in machine learning, and is typically studied as a multiple instance learning (MIL) problem [22]. While traditional MIL represents data as bags of feature points and recognizes a bag as positive if it contains at least one positive instance, we need a different interpretation in the context of re-id. In re-id, we have multiple feature points corresponding to a single person, all of which are positive instances. Subsequently, the representation of this data as an “image set” is more appropriate. Developing recognition algorithms based on image sets has been an active research area, with several approaches based on constructing affine or convex hulls of the data and considering the distance between the closest points on these hulls [23–25]. In the following, we briefly describe the method of describing data using affine hulls and the workflow of typical recognition algorithms that are based on constructing affine hulls of data.

C.6.a. Describing data using affine hulls

Given the set of feature vectors \( P = \{p_1, p_2, \ldots, p_n\} \), corresponding to \( n \) images of a certain person, the affine hull of this data is the smallest affine subspace containing the data. Formally, if \( \mu \) is the mean vector of the data and \( U \) is the matrix representing the set of the orthonormal bases describing the data, the affine hull of \( P \) can be written as the set \( H(P) = \{x = Uv + \mu | v \in \mathbb{R}^d\} \). This is illustrated in Figure 4 on the next page.
Given a probe image set $P$ and a gallery image set $G$, image set based recognition algorithms typically first construct the affine hulls of these two sets. To compute the extent of similarity/dissimilarity between $P$ and $G$, the general workflow is to determine the two points, one on each of the two affine hulls, that are closest to each other. Subsequently, the distance between these two points is used to represent the distance between the two sets $P$ and $G$. An illustration of this concept is provided in Figure 5.

Formalizing this notion, if $s$ and $t$ represent the two nearest points on the affine hulls $H(P) = \{x_p = U_p v_p + \mu_p | v_p \in \mathbb{R}^t\}$ and $H(G) = \{x_g = U_g v_g + \mu_g | v_g \in \mathbb{R}^t\}$, we solve the following optimization problem to find them:

$$\min_{v_p, v_g} \| x_p - x_g \|_2^2$$

The closest points are then given by $s = U_p v_p + \mu_p$ and $t = U_g v_g + \mu_g$. The distance between $P$ and $G$ is then simply $\| s - t \|$. Most hull distance algorithms differ in how they formulate the optimization problem shown above. While the AHISD (affine-hull-based image set distance) algorithm [23] uses the same formulation, algorithms like SANP (sparse approximated nearest points) [24] and RNP (regularized nearest points) [25] incorporate some kind of regularization into the problem formulation, typically based on the $l_1$ or $l_2$ norm to determine the closest points. Once the closest points are determined, computing the distance between the image sets reduces to the same Euclidean distance computation as above.

We hypothesize that such a distance computation between affine hulls is not discriminative enough for re-id. Consequently, a natural question to address is: what are better representations of these affine hulls that make subsequent reasoning more accurate? In the previous section, we presented an algorithm to learn feature representations with respect to discriminatively trained dictionaries. However, even this method was based on the traditionally-used approach of averaging available multiple feature points. Can we learn such discrimi-
native dictionary-based feature representations in the context of affine hulls? Traditional dictionary learning is typically based on solitary feature points, and extending it to the case involving affine hulls of data is a non-trivial problem. In this regard, how do we directly learn discriminative dictionaries using affine hulls? It is natural to expect computational difficulties when dealing with affine hull representations of large amounts of data. So, is there a computationally efficient way of learning such discriminative dictionaries?

In Year 3, we developed an approach that addresses all questions raised above in a principled and intuitive manner. Specifically, we focus on the following aspects:

- **Data description**
  
  We tackle the problem of describing the multiple instance data inherent in real-world re-id by constructing affine hulls. Such a mathematical representation provides for an intuitive description of the available data.

- **Learning discriminative affine hull representations**
  
  We hypothesize that the traditionally-used approach of computing the distance between affine hulls in algorithms based on image sets is not sufficiently discriminative in the context of re-id. This hypothesis is motivated by the fact that these affine hull distance algorithms are unsupervised, giving sub-optimal performance. We propose to address this issue by learning discriminative dictionary-based feature representations directly from these affine hulls. We also study the efficacy of such representations in comparison with distance metrics learned using the traditional approach of taking the average feature point as the data exemplar.

- **Affine hull dictionary learning**
  
  We develop a technique to train dictionaries discriminatively and, more crucially, directly using affine hulls of data. We propose to do this efficiently by formulating the associated optimization problem in a manner that results in closed-form updates in each iteration of the algorithm. This enables us to design a training scheme that provably converges to the globally optimal solution.

- **Learning affine hull representations**
  
  Before describing the main algorithm, we lay out the notation used in the subsequent sections. We use to denote the set of feature vectors corresponding to the images of the person with index \( i \) in the probe camera of the training set. Similarly, \( G_i \) denotes the set of feature vectors corresponding to the images of the same person in the gallery camera of the training set. Let \((s_i, t_j)\) be the pair of closest points on the affine hulls of \( P_i \) and \( G_j \).

  Our key insight is that directly computing the distance between the closest points on the affine hulls of \( P_i \) and \( G_j \) will not lead to accurate re-id results because this would be an unsupervised, suboptimal approach. To this end, we propose to learn discriminative dictionary-based representations of these affine hulls. Essentially, the formulation is in the same spirit as traditional metric learning algorithms that formulate pairwise constraints on the available feature points. However, the key idea is that we now formulate these constraints on representations computed from a dictionary learned using affine hulls of the sets of image data available for each person. To make this more clear, let \( P_i, G_j \), and \( G \) be three sets of feature vectors. Let \((s_i, t_j)\) and \((s_i, t_j)\) be the pairs of closest points on the affine hulls of \( \{ P_i, G_j \} \) and \( \{ P_i, G_j \} \) respectively, computed using some hull distance algorithm. Our goal is to learn a dictionary \( D \) such that the representation of \( s_i \) with respect to \( D \) is closer to that of \( t_j \) when compared to that of \( t_j \). To determine the representation of a point \( a \) with respect to \( D \), we use the following optimization problem:

\[
a = \arg\min_{a} \| s - Da \|_2^2 + \lambda \| a \|_2^2
\]

which has a closed-form solution, given by \( a = (D^TD + \lambda I)^{-1}D^Ts \).
We note that there is a key difference formulation presented in the previous section. In particular, while the previous method learns a dictionary using the average feature point of the image set, we learn a dictionary using the entire image set by constructing affine hulls of data. This approach is more favorable since we are now preserving the available discriminability in the image set.

C.6.b. **Affine hull dictionary learning**

Given the feature sets of the \( N \) people in both cameras \( P_i \) and \( G_i \), \( i = 1, \ldots, N \), we first construct affine hulls for every pair \( \{ P_i, G_j \} \) and determine the closest points on these pairs of affine hulls using a hull distance algorithm, such as AHISD, RNP or SANP discussed in the previous section. Let \( (s_i, t_j) \) be the pair of closest points. From these pairs of points, we gather the pairs \( \{(s_i, t_i)\} \), \( i = 1, \ldots, N \), corresponding to the same people (the “positive” pairs) and the pairs \( \{(s_i, t_j)\}, i = 1, \ldots, N, j = 1, \ldots, N, i \neq j \), corresponding to different people (the “negative” pairs). Using these pairs of points, we construct matrices \( P \) and \( G \) in a manner similar to that described previously. Subsequently, we use the same mathematical framework and the alternating directions minimization algorithm to learn the dictionary \( D \).

The detailed results of applying the proposed approach to benchmark datasets and comparing it with the state of the art are reported in a conference paper under review.

C.7. **Deployment at CLE**

Over the course of Years 1-3, and as part of an associated multi-university Task Order, the ALERT team designed two re-id testbeds at CLE. ALERT-designed algorithms successfully ran in real time, generating “line-ups” of re-id candidates that could be reviewed by the user. Unfortunately, CLE chose to remove the ALERT cameras and computer systems during their checkpoint lane renovations in 2015, and the on-site testbed is no longer active. Research will continue on the substantial amount of archived data collected during the multi-year experiment, and this data forms the basis of a new, highly-realistic benchmark evaluation for re-id to be distributed by ALERT in Year 4.

D. **Major Contributions**

The key achievements of the project to date (most recent first) include:

- **(Year 3)** The public release of several datasets and code for vision algorithms has facilitated rapid progress in re-id research over the past decade. However, directly comparing re-id algorithms reported in the literature has become difficult since a wide variety of features, experimental protocols, and evaluation metrics are employed. In order to address this need, we undertook an extensive review and performance evaluation of single- and multi-shot re-id algorithms. The experimental protocol incorporates the most recent advances in both feature extraction and metric learning. All approaches were evaluated using a new large-scale dataset created using videos from CLE as well as existing publicly-available datasets. This study is the largest and most comprehensive re-id benchmark to date, and is currently under review at the European Conference on Computer Vision to be held in October 2016. Reports from this evaluation will be summarized in the Year 4 annual report.

- **(Year 3)** We refined and improved the end-to-end system solution for the re-id problem installed in CLE in Year 2. We constructed a new large-scale dataset that accurately mimics the real-world re-id problem using videos from CLE and conducted several new experiments in the concourse testbed. The overall system architecture and the challenges of bringing academic re-id research to a real-world deployment were described in an overarching journal paper that should be quite valuable to both the academic and industrial research communities. This work appeared online in IEEE Transactions on Circuits and Systems for Video Technology in April 2016, and will be published later in 2016.

- **(Year 3)** We introduced an algorithm to describe image sequence data using affine hulls and to learn
feature representations directly from these affine hulls using discriminatively trained dictionaries. While existing metric learning methods typically employ the average feature vector as a data exemplar, this discards the rich information present in the sequence of images available for a person. We show that using affine hull representations computed with respect to the learned dictionary results in superior re-id performance when compared to using the average feature vector as done in existing methods. This work is under review at the European Conference on Computer Vision to be held in October 2016.

- (Year 3) We proposed a new approach to address the person re-id problem in cameras with non-overlapping fields of view. Unlike previous approaches, that learn Mahalanobis-like distance metrics in some embedding space, we propose to learn a dictionary that is capable of discriminatively and sparsely encoding features representing different people. To tackle viewpoint and associated appearance changes, we learn a single dictionary in a projected embedding space to represent both gallery and probe images in the training phase. We then discriminatively train the dictionary by enforcing explicit constraints on the associated sparse representations of the feature vectors. In the testing phase, we re-identify a probe image by simply determining the gallery image that has the closest sparse representation to that of the probe image in the Euclidean sense. Extensive performance evaluations on two publicly-available multi-shot re-id datasets demonstrate the advantages of our algorithm over several state-of-the-art dictionary learning, temporal sequence matching, spatial appearance, and metric-learning based techniques. This work was presented at the IEEE International Conference on Computer Vision (ICCV) in December 2015.

- (Years 2-3) We introduced an algorithm to hierarchically cluster image sequences and use the representative data samples to learn a feature subspace maximizing the Fisher criterion. The clustering and subspace learning processes are applied iteratively to obtain diversity-preserving discriminative features. A metric learning step is then applied to bridge the appearance difference between two cameras. The proposed method is evaluated on three multi-shot re-id datasets, and the results outperform state-of-the-art methods. This work was presented at the British Machine Vision Conference in September 2015.

- (Year 2) We proposed a novel approach to solve the problem of person re-id in non-overlapping camera views. We hypothesized that the feature vector of a probe image approximately lies in the linear span of the corresponding gallery feature vectors in a learned embedding space. We then formulated the re-id problem as a block sparse recovery problem, and solved the associated optimization problem using the alternating directions framework. We evaluated our approach on the publicly-available person re-id (PRID) 2011 and iLIDS-VID multi-shot re-id datasets, and demonstrated superior performance in comparison with the current state of the art. This work was presented at the IEEE/ISPRS 2nd Joint Workshop on Multi-Sensor Fusion for Dynamic Scene Understanding in June 2015.

- (Year 2) We proposed a novel metric learning approach to the human re-id problem with an emphasis on the multi-shot scenario. First, we perform dimensionality reduction on image feature vectors through random projection. Next, a random forest is trained based on pairwise constraints in the projected subspace. This procedure repeats with a number of random projection bases so that a series of random forests are trained in various feature subspaces. Finally, we select personalized random forests for each subject using their multi-shot appearances. We evaluated the performance of our algorithm on three benchmark datasets. This work was presented at the IEEE Winter Conference on Applications of Computer Vision (WACV) in January 2015.

- (Year 2) An end-to-end system solution of the re-id problem was installed in an airport environment, with a focus on the challenges brought by the real-world scenario. We addressed the high-level system design of the video surveillance application and enumerated the issues we encountered during our development and testing. We described the algorithm framework for our human re-id software and discussed considerations of speed and matching performance. Finally, we reported the results of an experiment conducted to illustrate the output of the developed software, as well as its feasibility for the airport surveillance task. This work was presented at the Eighth ACM/IEEE International Conference on Distributed
Smart Cameras (ICDSC) in November 2014.

- (Year 1-2) In collaboration with NEU, the design and deployment of an on-site re-id algorithm for the new branching testbed at CLE occurred, leveraging a software architecture using DDS, including an experimental graphical user interface for tagging subjects of interest and viewing top-ranked matching candidates.

- (Year 1-2) ALERT-guided design and deployment of a new 6-camera branching testbed leading from the exit of the central security checkpoint in CLE to each of the three concourses.

- (Year 1) Development of a novel re-id algorithm that mitigates perspective changes in surveillance cameras. We built a model for human appearance as a function of pose, using training data gathered from a calibrated camera. We then applied this “pose prior” in online re-id to make matching and identification more robust to viewpoint. We further integrated person-specific features learned over the course of tracking to improve the algorithm’s performance. We evaluated the performance of the proposed algorithm and compared it to several state-of-the-art algorithms, demonstrating superior performance on standard benchmarking datasets as well as a challenging new airport surveillance scenario. This work was published in *IEEE Transactions on Pattern Analysis and Machine Intelligence* in May 2015.

- (Year 1) Developed an algorithm for keeping a pan-tilt-zoom (PTZ) camera calibrated. We proposed a complete model for a PTZ camera that explicitly reflects how focal length and lens distortion vary as a function of zoom scale. We show how the parameters of this model can be quickly and accurately estimated using a series of simple initialization steps and followed by a nonlinear optimization. Our method requires only 10 images to achieve accurate calibration results. Next, we show how the calibration parameters can be maintained using a one-shot dynamic correction process; this ensures that the camera returns the same field of view every time the user requests a given (pan, tilt, zoom), even after hundreds of hours of operation. The dynamic calibration algorithm is based on matching the current image against a stored feature library created at the time the PTZ camera is mounted. We evaluate the calibration and dynamic correction algorithms on both experimental and real-world datasets, demonstrating the effectiveness of the techniques. This work was published in *IEEE Transactions on Pattern Analysis and Machine Intelligence* in August 2013.

- (Year 1) Establishment of an initial tag and track testbed in CLE that included a selection of cameras leading from the parking garage to the terminal.

E. Milestones

As described in more detail in Section II.D., the major milestones accomplished in Year 3 include:

- An extensive review and performance evaluation of single- and multi-shot re-id algorithms, resulting in an experimental protocol that incorporates the most recent advances in both feature extraction and metric learning, and includes a new large-scale dataset created using videos from CLE as well as existing publicly available datasets. This study is the largest and most comprehensive re-id benchmark to date.

- A complete description of the end-to-end system solution for the re-id problem installed in CLE, including the overall system architecture and the challenges of bringing academic re-id research to a real-world deployment. The resulting paper should be quite valuable to both the academic and industrial research communities, and also to DHS entities seeking to understand the promise and limitations of re-id.

- A new multi-shot re-id algorithm that describes image sequence data using affine hulls and learn feature representations directly from these affine hulls using discriminatively-trained dictionaries.

- A new multi-shot re-id algorithm that learns a dictionary that is capable of discriminatively and sparsely encoding features representing different people.

As described in more detail in Section II.F., further fundamental research is still required on multi-shot
algorithms that maximally leverage the discriminative information available in long-person tracks.

The major milestones to be achieved for Year 4 include:

- A demonstration of the utility of the convolutional neural networks as applied to the re-id problem.
- A tracking, feature extraction, matching, and ranking re-id pipeline that can achieve 90% performance at rank 10 on our challenging multi-shot airport dataset (our current best algorithm has about 70% performance at rank 10).
- Widespread dissemination of the ALERT re-id evaluation protocol and benchmark.
- An ALERT-designed re-id contest to challenge the academic community at a workshop at a top computer vision conference in 2017.

F. Future Plans

Year 4’s research agenda includes extensions of the work motivated by the success of deep CNNs, which have not been widely applied to re-id. We will consider two situations where CNNs should be able to significantly boost re-id performance: (1) automatically selecting “important” key-frames from each tracked person to provide the best input to multi-shot re-id algorithms, and (2) learning important features for each person from large amounts of training data instead of constructing them by hand, as is done currently.

A through-line of our research in this area that differentiates it from most related academic work is our strong emphasis on the end-to-end re-id problem, as it would need to be solved in a real-world surveillance environment (e.g., assuming that probe and gallery candidates come from automatic trackers that may produce incorrect results). Our group has become known in the research community for “real world re-id”, as exemplified by the custom-designed, on-site re-id testbed at CLE deployed in Years 2 and 3. Our CLE system is fully described in a journal paper that will appear in Year 4.

However, changing infrastructure and personnel in CLE resulted in the loss of our testbed midway through Year 3; as the security areas of the airport were remodeled, all research computers and cameras were removed and shipped back to ALERT. This is a difficult obstacle to overcome, as the CLE testbed was a unique opportunity to work with real-world data, constraints, and end-users. On the positive side, the research team collected tens of hours of video from the research camera network that we continue to use to design and validate our algorithms, and new video would not look substantially different than the video already collected. This data forms a new and challenging re-id benchmark that we will disseminate to the re-id community in Year 4.

As an alternate environment, we are still hopeful that Year 4 will bring a continuation of our preliminary re-id research in the light rail stations of the Greater Cleveland Regional Transit Authority (GCRTA). These environments challenge re-id algorithms in ways not typically considered by the re-id community. For example, one goal of GCRTA police is to re-identify thieves that frequent the same rail stations on different days; thus, the re-id gallery is constantly changing over the course of tens of hours. We believe our tracking and re-id algorithms will be immediately applicable to the GCRTA environment if funds are made available for the transition task.

The anticipated end date for the project is early 2018, which will allow a senior Ph.D. student to complete the research and evaluation, and a junior graduate student to distill the research into actionable recommendations for the HSE.
III. RELEVANCE AND TRANSITION

A. Relevance of Research to the DHS Enterprise

Video surveillance is an integral aspect of homeland security monitoring, forensic analysis, and intelligence collection. The research projects in this area were directly motivated (and in fact, requested) by DHS officials as critical needs for their surveillance infrastructure. The presence of ALERT hardware and software on-site in CLE is expected to produce a wealth of new data and research problems of direct DHS/Transportation Security Administration (TSA) interests for several years.

The specific metric we seek to maximize for the re-id systems is rank-5 performance; that is, the percentage of tagged subjects who appear in the short list of 5 best matches automatically predicted by the algorithm. We chose the length of the shortlist under the assumption that it is difficult for a user to easily browse more than 5 candidates on a graphical interface. Depending on the dataset/scenario, we are currently able to achieve 70% - 90% rank-5 performance, which we believe meets or exceeds competitive solutions to the end-to-end problem. A continuing focus in Year 4 and the future will be a discussion with end users about the performance specifications and interface that would be required to transition the developed research into regular use, using the successful CLE deployment as a starting point for discussion.

B. Potential for Transition

Over the past three years, the video analytics group built a strong relationship with Cleveland TSA, CLE, and the GCRTA. In our first project, we transferred a set of counterflow algorithms to detect people entering the airport exit lanes, and worked with the TSA and airport officials to display the counterflow events in their coordination center for further analysis and action. In the current project, we worked with the same group to develop re-id and tracking algorithms to satisfy their needs and match their CONOPS, so that the presented results fit their operation. The developed re-id algorithms were implemented on a custom-built PC at CLE, with a working user interface.

ALERT, in collaboration with another DHS Center of Excellence (COE), Visual Analytics for Command, Control, and Interoperability Environments (VACCINE), also collaborated with the GCRTA police to address a problem related to re-id in the context of rail platforms, bus stops, and concourses. We followed a similar pathway forward with respect to problem specification and CONOPS definition. The specifications and CONOPS for the GCRTA are somewhat different (e.g., only performing re-id over a single camera, but doing so over the course of many days), making the problem easier in some ways but harder in others.

C. Data and/or IP Acquisition Strategy

ALERT has retained the services of an IP consultant to assess the feasibility of technology transfer for video analytics research and development in the Center. As new intellectual property is created, the video analytics groups write descriptions of the new property and disclose it to their respective universities. As the work matures, the disclosure will become patent disclosures and perhaps patents.

D. Transition Pathway

The described re-id research was already transitioned to CLE end-users throughout Years 1-3 as part of the associated Task Order 5, resulting in a working, on-site system, which was tested extensively and fully described in a journal publication to appear in Year 4. Unfortunately, the ALERT camera/computer network at CLE was dismantled in Year 3 and there are no immediate plans to reconstitute it. ALERT continues to pursue potential customers for re-id and related video analytics technology.

In general, the video analytics group is surrounded by practitioners/users of video. They are willing to supply real data, application ideas, use cases, and CONOPS. Included within the video analytics cohort are
representatives from Siemens Corporate Research, several divisions of which are regular contractors to the video surveillance community. We use all of this input to help us understand how to apply our research to real problems and to forge a transition pathway forward.

E. User or Customer Connections

This project historically involved regular contact with DHS, CLE, GCRTA, and law enforcement collaborators, including:

- Michael Young, former Federal Security Director, TSA at CLE
- Jim Spriggs, former Federal Security Director, TSA at CLE
- John Joyce, Chief of Police / Director of Security, GCRTA
- Don Kemer, Transportation Security Manager, Coordination Center, TSA at CLE
- Fred Szabo, Commissioner, CLE
- Michael Gettings, Lieutenant, Cleveland Transit Police

Currently contact with these individuals is sporadic, but it is hoped that the GCRTA effort will be revived in Year 4.

IV. PROJECT ACCOMPLISHMENTS AND DOCUMENTATION

A. Education and Workforce Development Activity

1. Student Internship, Job, and/or Research Opportunities
   a. Srikrishna Karanam had an internship with Siemens Corporate Research, Princeton, NJ from May to August 2016, working under the supervision of former ALERT student Ziyang Wu (RPI, Ph.D., 2013).

B. Peer Reviewed Journal Articles


Pending-


C. Peer Reviewed Conference Proceedings


3. Li, Y., Karanam, S., and Radke, R.J. “Multi-Shot Human Re-Identification Using Adaptive Fisher Dis-
D. **Software Developed**

1. **Datasets**
   - The ALERT Re-Identification Benchmark dataset, created in collaboration with Northeastern University Prof. Octavia Camps (project R4-A.1), is to be made publicly accessible in July 2016 via the ALERT website.

2. **Algorithms**
   - The full technical description of the respective algorithms listed below is contained in the corresponding papers on the PI’s webpage. Algorithm source code may be available on the PI’s webpage or by request to the PI.
   - a. Multi-shot human re-id algorithm based on discriminatively trained dictionaries.
   - b. Multi-shot human re-id algorithm based on affine hull comparison.

E. **Requests for Assistance/Advice**

1. **From DHS**
   - DHS has expressed interest in continuing the “Electronic Be On the Lookout” (eBOLO) program prototyped at the GCRTA in Years 2 and 3. This effort is being jointly pursued by both ALERT and VACCINE.

V. **REFERENCES**


ference on (pp. 373-380). IEEE.


