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

I. PARTICIPANTS INVOLVED FROM JULY 1, 2019 TO JUNE 30, 2020

Faculty/Staff			
Name	**Title**	**Institution**	**Email**
Rich Radke	PI	RPI	rjradke@ecse.rpi.edu
Graduate, Undergraduate and REU Students			
Name	**Degree Pursued**	**Institution**	**Month/Year of Graduation**
Meng Zheng	PhD	RPI	5/2020

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 spaced widely so that their fields of view are nonoverlapping. Automatically matching humans that reappear 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. Early re-id system designs focused on two major problems: (1) feature extraction, in which the goal is to determine effective ways to extract representative information from each cropped rectangle to produce feature representations, and (2) metric learning, in which the goal is to produce high similarity scores for feature representations from the same person, and low similarity scores for feature representations for different persons. With the recent explosion of Convolutional Neural Networks (CNNs) and deep learning, deep networks have been gradually adopted in re-id system design and have achieved great success. The most recent research in re-id has been focused on developing various deep CNN-based networks that combine feature extraction and metric learning as an integrated module to facilitate end-to-end training. In practice, a re-id system must be fully autonomous from the point that a user draws a rectangle around a person of interest to the point that candidates are presented to them; thus, the system must also automatically detect and track humans in all cameras with speed and accuracy.

This project addresses the design and deployment of real-world re-id algorithms specifically designed for large-scale surveillance systems in mass transit environments. This involves:

- The design and analysis of new state-of-the-art computer vision algorithms for human detection and tracking, feature representation learning, and deep CNN model design for cross-view matching;
- 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 design and dissemination of new experimental protocols and datasets that more closely resemble the types of re-id problems practitioners will encounter in real-world deployments; and
- The design of novel algorithms that can be directly integrated into real-world surveillance systems for practical Department of Homeland Security (DHS) use for person re-id and trajectory reconstruction.

The latter three aspects differentiate our approach from most related re-id research. That is, in academic research the re-id problem is often reduced to comparing one candidate rectangle of pixels to another, which is only a small part of a fully automated re-id system. In addition, we take into account that the candidate rectangles are likely to be generated by an automatic (and possibly inaccurate) human detection and tracking subsystem; that many cameras may be involved; that new subjects are constantly appearing in the target camera(s); that the overall system needs to operate in real time; and that the system may be in operation for very long periods of time. We provided practical solutions that can be directly integrated into real-world surveillance systems to automatically calculate motion trajectories of persons of interest, given the cropped rectangles of the person's first appearance. Our solution differs from conventional re-id algorithms, which typically require a substantial amount of manual intervention to achieve the same goal.

The end-state of the research is a suite of re-id/trajectory recovery algorithms that are directly applicable to the homeland security enterprise (HSE) and ready for large-scale system integration, as well as several re-id benchmark databases of interest to the homeland security community. The project also produced an up-to-date assessment of the re-id state of the art that can help DHS stakeholders understand what is technologically feasible in this area, informing policies and technology solicitations.

B. State of the Art and Technical Approach

To date, the majority of deep CNN-based person re-id research has been focused on image-based re-id, which considers the cross-view similarity between single person images as discussed above. However, recent progress in big data and deep learning has facilitated the storage and processing of large-scale data, which recently influenced the re-id research community. Person tracklets (or image sequences), which consist of consecutive frames of cropped person images, typically contain richer temporal information than single frames and are more helpful for re-identification tasks. Thus, we proposed a deep CNN-based network architecture, with spatiotemporal consistent attention for cross-view person tracklet matching, to deal with the video-based re-id problem.

As with image-based re-id, illumination and viewpoint changes, occlusions, and misalignment issues are critical challenges for cross-view person tracklet matching. However, the availability of multiple image frames for each identity under the same or different camera views raises additional problems for video-based re-id. Specifically, how to extract cross-view invariant spatiotemporal feature representations and effectively aggregate temporal information along the image sequence are critical questions in video scenarios. In our work in Year 6 [1], we showed the effectiveness of introducing attention as a principled part of the training process of the deep re-id network, which is able to automatically spatially localize regions of interest while ignoring irrelevant noisy backgrounds of person images. In Year 7, we extended the involvement of attention and attention-consistency supervision from the spatial dimension to the temporal dimension, to automatically discover 3D attentive regions and learn inter-view and intra-view consistent feature representations for same-person image sequences.

Recently, the success of 2D CNN designs have attracted broad attention in the computer vision community [2, 3]. These designs are critical to various computer vision tasks such as image classification, detection, and semantic segmentation. 3D CNNs, on the other hand, due to the lack of large-scale 3D pretraining data comparable to ImageNet [4], are more limited in application for computer vision tasks. However, recent progress [5-7] in 3D CNNs has shown superior performance in various video-based tasks, such as action

recognition compared to 2D CNNs, due to their advantage in involving temporal cues in the convolution procedure, enabling spatiotemporally correlated feature representations. Despite the superiority of vanilla 3D CNN architectures in dealing with video-based data, real-world difficulties in re-id systems, such as occlusions and background noise as mentioned above, will deteriorate the effectiveness of the learned features. Thus we incorporate the gradient-based attention technique [8] with a vanilla 3D CNN architecture, which generates spatiotemporal attention maps to automatically discover attentive image regions along the temporal dimension. Both spatial and temporal consistency constraints are applied to lead the network to find consistent attentive regions along same-person image sequences under different camera views.

For temporal aggregation of feature vectors, we suspected that the conventional approach of average or max pooling [9] could deteriorate re-identification performance if person images along the sequence are misaligned. Instead, we applied an additional temporal convolution layer [10] to automatically find consistent information along the feature sequence, at the same time including useful temporal variations to obtain aggregated feature representations for person image sequences. To this end, we proposed a 3D CNN network called the Spatiotemporal Consistent Attentive Network (SCAN) as a video extension to our work [1] for image data. The pipeline of the proposed SCAN method is shown in Figure 1. The proposed SCAN method is learned in an end-to-end way, with 3D attention maps explicitly regularized for spatiotemporal invariant feature learning, producing robust cross-view matching for same-person tracklets. Experimental results show that our SCAN method achieved 82.7%, 93.4%, and 93.3% rank-1 performance on the Motion Analysis and Re-identification Set (MARS) [11], DukeMTMC-VideoReID (Multi-Target, Multi-Camera) [12], and Person Re-ID (PRID) [13] benchmark datasets, which are comparable to the state-of-the-art video-based re-id algorithms [12, 14-15].

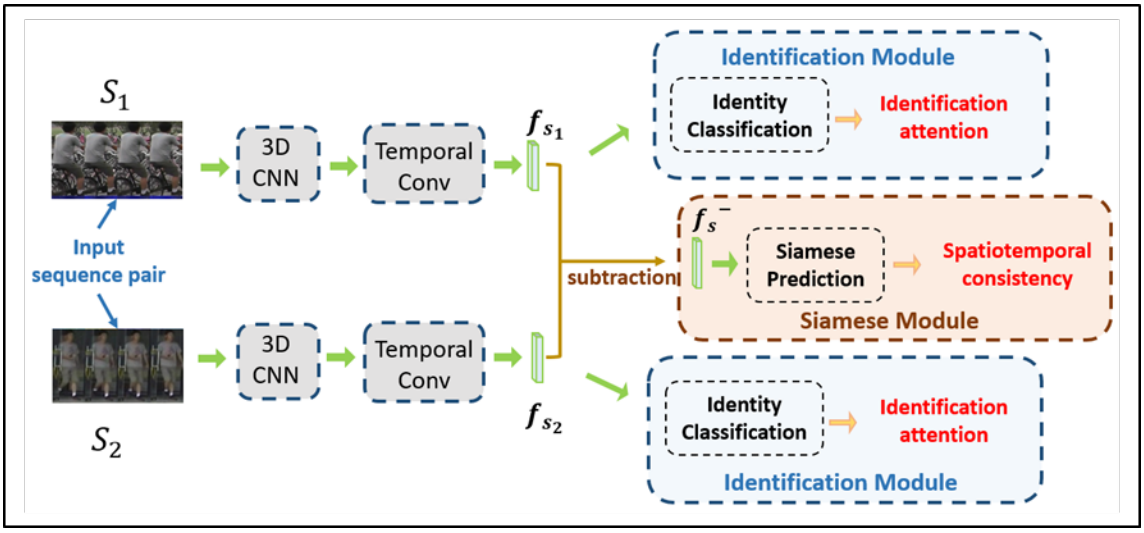

Figure 1: Pipeline of the proposed SCAN method for video-based re-id.

In our Year 6 ALERT-funded paper [1], we introduced the Consistent Attentive Siamese Network (CASN), which is able to generate attention from classification-based re-id models to spatially localize informative regions and provide visual explanations given person identity label information, showing superior performance compared to state-of-the-art re-id algorithms. In particular, depending on the dataset, we reported 5%–7% improvements in rank-1 accuracy compared to general competing algorithms, and substantial 10%–30% improvements in rank-1 accuracy compared specifically to attention-based methods. In Year 7, we took a step further to develop a similarity attention generation technique that produces attention from similarity distances between feature points in the deep embedding space. The generated

attention is able to explain why the input data points (pairs/triplets of person images) are similar/dissimilar by localizing meaningful spatial regions in the input person image set. Compared to CASN, the proposed similarity attention generation model has a much cleaner and more simplified architecture that is able to achieve comparable/superior performance in person re-identification tasks. More importantly, it is the first method proposed to interpret the reasoning behind CNN-based metric learning models by means of gradient-based attention.

Existing deep similarity predictors for person re-id [16-18] are trained in a distance-learning fashion where the big-picture goal is to embed features of same-person data points close to each other in the learned embedding, while also pushing features of data from other person classes further away. Consequently, most techniques distill this problem into optimizing a ranking objective that respects the relative ordinality of pairs [19], triplets [20], or even quadruplets [18] of training examples. However, a key limitation of these approaches is their lack of decision reasoning (i.e., explanations for why the model predicts the input set of person images is similar or dissimilar). An illustrative example of our proposed Similarity Attention Network is shown in Figure 2. A more detailed presentation of the approach is described in our paper [21], which is currently under review for the European Conference on Computer Vision 2020 (ECCV). As we demonstrate in our paper, our method offers not only model explainability but also decision reasoning that can further be infused into the model training process, in turn helping bootstrap and improve the generality of the trained re-id model.

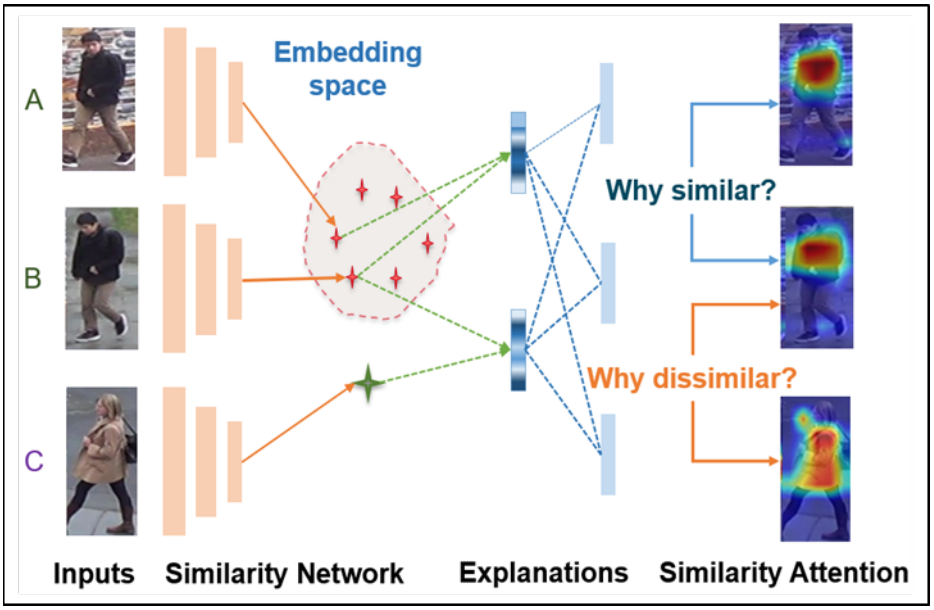

Figure 2: We present the first gradient-based similarity attention and similarity mining framework, giving a new attention-driven learning paradigm for similarity predictors that results in similarity model explainability and improved downstream performance.

The output of existing re-id algorithms is typically a list of candidate images ranked by the corresponding similarity scores computed by the algorithm. However, for real-world applications of re-id, this is not enough. For example, a suspect typically will need to be consistently tracked throughout a wide-area camera network, not just a single camera, so that the user of the system (e.g., a security officer) can monitor and recover the spatial and temporal information about his/her travel path. To achieve this with conventional re-id algorithms, system users would have to manually, incrementally select correct matches from the camera-to-camera ranked lists produced by the algorithm and then reconstruct the person of interest's path based on the selected appearances. With the list of candidate images likely to be large in dense mass-transit

environments (e.g., airports) where such systems are typically installed, the burden on the system operator to manually go through even a part of the list and recover the travel path can be immense. To address this key practical necessity, in Year 7 we introduced spatiotemporal trajectory recovery, the problem of automatically reconstructing the timestamped spatial trajectory of a person of interest in a camera network. We illustrate the problem and its difference to standard re-id in Figure 3. Given a person of interest, a traditional re-id algorithm will produce ranked candidate lists for each camera in the network. On the other hand, our proposed trajectory recovery involves a complete spatial and temporal path reconstruction of the person of interest as the individual moves through the camera network. Specifically, as shown in Figure 3, the output is a timestamped sequential list, including camera identity and image sequence, that shows where the person of interest was in the camera network and at what time. End users of a surveillance system can then use this data to easily retrieve the desired forensic information for the person of interest. To the best of our knowledge, this is the first work in the re-id community proposed to solve the trajectory recovery problem.

With standard re-id evaluation measures not directly relevant for the problem, we introduced three new evaluation metrics that consider various spatiotemporal aspects in the context of practical, real-world use of our algorithm. Specifically, we consider (1) the percentage of recovered reappearance duration comparing to ground-truth reappearance duration; (2) similarity between recovered trajectory and ground-truth trajectory; and (3) the percentage of recovered tracking lifetime (time difference between the first and last identified appearance) with respect to ground-truth tracking lifetime to evaluate our algorithm. The experimental results on our RPIfield dataset [22] show that our trajectory recovery algorithm is able to automatically reconstruct trajectories of the persons of interest with 80.1%, 82.5% and 97.0% accuracy evaluated with the above three metrics respectively. The detailed description of this algorithm and experimental results are under review at the journal *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

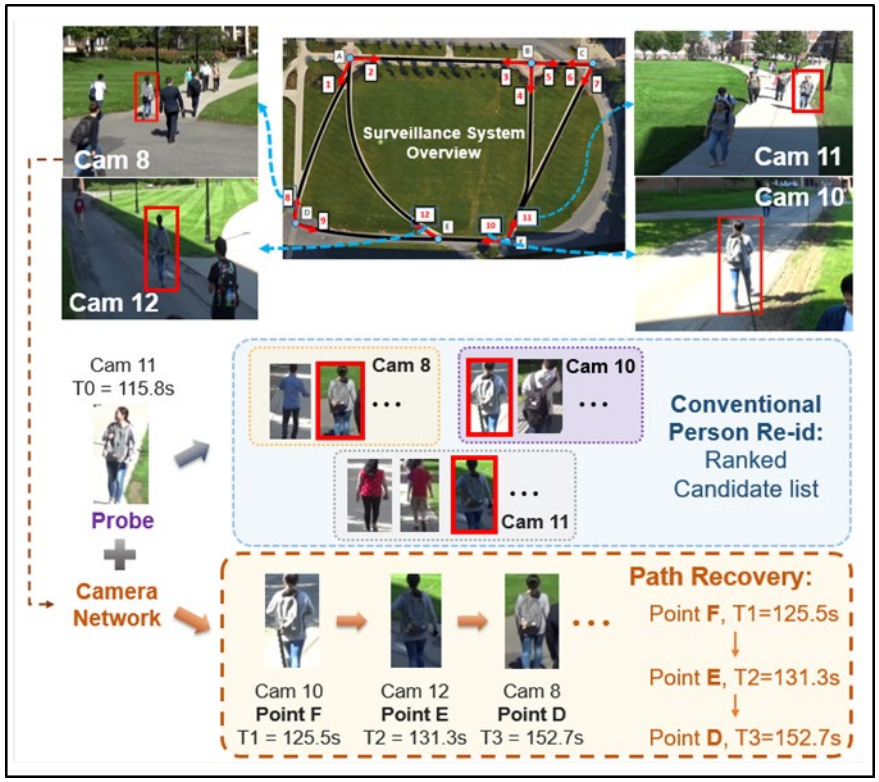

Figure 3: An illustration of the trajectory recovery problem in a wide-area multi-camera system, compared to conventional person re-identification.

C. Major Contributions

As described in more detail in the previous sections, the major contributions from Year 7 include:

- (Year 7) We proposed a novel network based on 3D convolution called the Spatiotemporal Consistent Attentive Network (SCAN) for video-based person re-id. The proposed SCAN is able to localize complete spatial regions along the temporal dimension for person tracklets by means of gradient-based 3D CNN attention, while learning intra- and inter-view consistent attentive sequence features by utilizing 3D attention as the guidance for network training. We performed extensive experiments on three benchmark video re-id datasets along with a functionality study of the individual attention modules in our new method, demonstrating the superiority of our proposed 3D attention and attention-consistency mechanism in performing video-based re-id tasks.
- (Year 7) We presented new techniques to explain and visualize, with gradient-based attention, predictions of similarity models. We showed our resulting similarity attention is generic and applicable to many commonly used similarity architectures. We presented a new paradigm for learning similarity functions with our similarity-mining learning objective, resulting in improved downstream model performance. We also demonstrated the versatility of our framework in learning similarity-attention-driven models for related surveillance applications (e.g., anomaly detection) in addition to person re-id tasks. This work is currently under review for the ECCV 2020. An application of the proposed algorithm in anomaly detection and disentangled latent representation, with which we collaborated with Northeastern University, was selected for oral presentation at the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
- (Year 7) We proposed the first method to automatically reconstruct the trajectory of persons of interest by re-identifying them through a network of cameras. To efficiently compute the motion path of the persons of interest, we developed an incremental pruning strategy with a novel confidence score calculation strategy based on appearance similarity and transition-time modeling. Experimental results on the RPIfield dataset [22] show that the proposed trajectory recovery algorithm is able to consistently retrieve the crucial appearances of the persons of interest starting from their first appearance until their last reappearance, presenting the user with easy-to-read spatial and temporal reconstruction information. This work is under second review for a Minor Revision at *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

Milestones achieved in the project since its inception include:

- (Year 6) We presented a novel deep architecture, called the Consistent Attentive Siamese Network (CASN), for image-based person re-id, which achieves state-of-the-art performance on popular benchmark datasets. The proposed technique makes spatial localization of the person of interest a principle part of the learning process, providing supervision only by means of person identity labels. This makes spatial localization end-to-end trainable and automatically discovers complete attentive regions. The proposed CASN defines a new framework that enforces attention consistency as part of the learning process, providing supervision that facilitates end-to-end learning of consistent attentive regions of images of the same person. This is the first proposed Siamese attention mechanism that jointly models consistent attention across similar images through a Siamese learning framework, resulting in a powerful method that can help explain the reasoning behind the network's prediction. This work was presented at the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
- (Year 6) We developed a preliminary “chained re-id” algorithm that can recover the spatial path of a tagged person by following them through a network of cameras. Previously, all of our re-id research and development had focused on re-identifying a target person in a second camera (and then stopping).

In contrast, we investigated the operationally important problem of reconstructing the timestamped reappearances of people across many cameras (e.g., Person X appeared in Camera 1 at 12:30, Camera 4 at 12:33, Camera 2 at 12:35, etc.). We developed a human-in-the-loop Graphical User Interface program to completely recover the trajectory of the persons of interest. This approach was refined and automated in Year 7, as discussed above, to remove the human from the loop.

- (Year 5) We created a new large-scale real-world re-id dataset, RPIfield [22], using 12 disjoint surveillance cameras placed around an outdoor field at RPI. RPIfield contains the largest number of camera views in the context of benchmarking datasets in the re-id research community. It includes 112 known "actors" walking along predefined paths among ~4000 distractors, which simulates the mass-transit environments of interest to DHS. With timestamp information preserved for every person image, the dataset allows extensive temporal analysis of re-id algorithms that was not previously addressable by existing benchmarking datasets.
- (Year 5) Temporal attribute analysis of re-id algorithms is mostly ignored in academic re-id research; however, these considerations are essential to real-world deployment. Specifically, for a real-world re-id system that automatically detects and tracks people for a long time (e.g., several days), the gallery sets will be continuously populated with incoming candidate images. Consequently, existing measures adopted to evaluate re-id algorithms, such as CMC curves, fall short because they ignore such time-varying behavior of the gallery. In Year 5, we extended the idea of Rank Persistence Curve (RPC) proposed in Year 4, extensively evaluating and analyzing the temporal performance of benchmarking re-id algorithms on our new RPIfield [22] dataset. This kind of analysis is critical for bridging the gap between real-world system application and academic research in this area.
- (Year 4) In Year 4, we initially proposed the Rank Persistence Curve (RPC) methodology. We designed both qualitative and quantitative evaluation metrics that are generic and can be used to evaluate the "temporal robustness" of any re-id algorithm. We assessed preliminary results using a custom re-id dataset we built specifically to model such temporal aspects of real-world re-id, constructed from the camera views collected by the ALERT team at the Greater Cleveland Rapid Transit Authority (GCRTA) in Years 2 and 3. However, this dataset was quite limited and necessitated the collection of a larger dataset in Year 5 as described above.
- (Years 4–5) The ALERT Airport Re-Identification Dataset was curated and released on ALERT's website, generating substantial interest from the research community. The dataset has been requested and downloaded 280 times since it was made available, and 13 industrial organizations have made use of the data, including Bosch, Intel and Mitsubishi. See more information below.
- (Year 4) We began to investigate the transition of the best-performing re-id algorithms identified by our extensive benchmarking analysis (see below) to a large environment equipped with multiple pan-tilt-zoom (PTZ) cameras. The goal is to actively orient the cameras in conjunction with real-time re-id, for example, to keep promising candidates in sight by panning and acquire higher resolution images by zooming. The problem is complicated by the issue that there may be more candidates than cameras, requiring time-sharing schemes to be developed to entertain multiple hypotheses.
- (Years 3–5) 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 Cleveland Hopkins International Airport (CLE) as well as existing publicly available datasets. This study is the largest and most comprehensive re-id benchmark to date, and has been accepted and published online at *IEEE Transactions on Pattern Analysis and Machine Intelligence* in February 2018.

- (Years 3–4) 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 was published in March 2017.
- (Years 3–4) 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 was accepted by *IEEE Transactions on Circuits and Systems for Video Technology* and appeared online in July 2017.
- (Year 3) We proposed a new approach to address the person re-id problem in cameras with nonoverlapping 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 was evaluated on three multi-shot re-id datasets, and the results outperformed 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 nonoverlapping 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 PRID (person re-id) 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.
- (Years 1–2) In collaboration with the R4-A.1 project, the design and deployment of an on-site re-id algorithm for the new branching testbed at CLE occurred, leveraging a software architecture using Data Distribution Service (DDS), including an experimental graphical user interface for tagging subjects of interest and viewing top-ranked matching candidates.
- (Years 1–2) ALERT-guided design and deployment of a new six-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 ten 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 evaluated 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.

D. Milestones

As described in our Year 6 Report, the major milestones planned for Year 7, which were achieved, included:

- The development of an effective, end-to-end training deep network for the video-based re-id problem, to learn spatiotemporal attentive and cross-view invariant feature representations for person tracklets. To demonstrate our effectiveness and improvement over the state of the art, we used the standard measure of rank-k performance with an emphasis on rank-1 performance (i.e., the percentage in which the algorithm's top match is the correct candidate). We achieved this milestone with 82.7%–93.4% rank-1 performance on several benchmark video-based re-id datasets, which is described in more detail in our section "State of the Art and Technical Approach." The outcome is a pre-trained end-to-end package that could easily be adopted by DHS practitioners.
- To continue the study of a robust "chained-re-id" algorithm to automatically recover the trajectory of a person of interest in a network of cameras, which has immediate applications to real-world multicamera surveillance environments. This milestone was achieved with the new trajectory recovery algorithm and additional novel evaluation metrics described in the previous section. We performed extensive experiments and individual algorithm module studies on the RPIfield dataset [22] with the proposed evaluation metrics. The outcome is an algorithm module able to process timestamped person image sequences from a large camera network, along with side information about the topology of the network, to produce maps and corresponding camera-by-camera thumbnails of the recovered person trajectory.

We also achieved an additional milestone in Year 7 not on the original list:

- We presented a new paradigm for learning similarity functions for person re-id using our similarity-mining learning objective, resulting in improved downstream model performance. We achieved state-of-the-art rank-1 performance compared to the most recent re-id algorithms, which is better than as published in our Year 6 ALERT-funded paper [1]. We also demonstrated the versatility of our framework in learning attention-driven models for anomaly detection, which is also a crucial problem in modern surveillance system development. We collaborated with Northeastern on developing this work, which will be presented orally at the *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

E. Final Results at Project Completion (Year 7)

The achieved final results in Year 7 include:

- The development of an effective, end-to-end training deep network for the video-based re-id problem, achieving 82.7%–93.4% rank-1 performance on various benchmark video-based re-id datasets, which matches the most recent state of the art. This end-to-end deeply learned re-id model can be directly integrated into large-scale real-world surveillance systems for cross-view person tracklet matching.
- The development of a state-of-the-art similarity-attention-driven deep re-id model, which is able to provide explicit reasoning behind the similarity predictions (i.e., why the person image pair was found to be similar or dissimilar), by means of gradient-based spatial localization, while additionally

producing state-of-the-art re-id performance in terms of rank-1 matching accuracy. This algorithm can be combined with any existing temporal aggregation method and easily plugged into existing real-world systems for direct use of person re-id, while presenting explicit visualizations of why the algorithm produces right/wrong predictions in the form of heatmaps.

- The development of an algorithm module to process timestamped person image sequences from a large camera network, along with side information about the topology of the network, producing maps and corresponding camera-by-camera thumbnails of the recovered person trajectory. This is the first work proposed in the re-id community that solves the problem of trajectory recovery of persons of interest moving in a wide-area camera network. Our proposed solution automatically outputs possible moving trajectories of the person of interest, which avoids intractable manual labeling effort from real-world end users required by conventional re-id. The developed algorithm module is easy to integrate into real-world systems for direct use, which does not require any extra transition effort.

  The contributions with the highest impact over the life of the project (described more fully in the previous section) include:

  - Our benchmarking, categorization, and code dissemination efforts that culminated in our highly cited survey in *IEEE Transactions on Pattern Analysis and Machine Intelligence* in 2018. This extensive review and performance evaluation of single- and multi-shot re-id algorithms defined an experimental protocol that incorporated the most recent advances in both feature extraction and metric learning.
  - Our fundamental theoretical work on viewpoint-invariant human re-identification that laid the groundwork for many further advancements in the field. In particular, two of our papers [24, 25] published in 2015 have 123 and 162 citations as of July 2020.
  - Our real-world implementation of a real-time, multicamera system at Cleveland Hopkins International Airport that demonstrated the feasibility and challenges of bringing a previously academic problem into DHS practice. This early effort formed the basis of our group's emphasis on the real-world operational aspects of re-id algorithms.

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, in particular for agencies such as the Transportation Security Administration (TSA). As ALERT evolved, video surveillance, in particular for mass transit environments, took on an increasingly prominent role, as a result of associated task orders and follow-on funding (e.g., Counterflow and Tag and Track efforts at CLE and the Correlating Luggage and Specific Passengers program at Kostas Research Institute at Northeastern University).
- For evaluating typical re-id systems, the metric we seek to maximize is the rank-k performance; that is, the percentage of tagged subjects who appear in the short list of k best matches automatically predicted by the algorithm. In Year 7, on the one hand, we developed novel, effective re-id algorithms that achieved state-of-the-art/competitive rank-1 performance on the most popular benchmark re-id datasets. These end-to-end learning algorithms can be plugged in as a submodule to real-world

surveillance systems, which will produce more accurate candidate lists in a graphical user interface than existing competitive methods. On the other hand, we developed a novel trajectory recovery algorithm which provides the first proposed solution for reconstructing a person of interest's trajectory in a network of cameras. The output of our algorithm is one or multiple possible timestamped reappearances along with spatial locations, which directly addresses the concern of real-world users who usually want to retrieve this information from conventional re-id algorithms but currently would need substantial manual selection effort to achieve this.

B. Status of Transition at Project End

In Years 1–4, the video analytics group built a strong relationship with Cleveland TSA, CLE, and the Greater Cleveland Regional Transit Authority (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. We then 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, as detailed in the joint publication [26]. Our research in Years 5–7 of developing effective deep person re-id networks could naturally be integrated and tested on simulated or real-world surveillance systems, due to its end-to-end learning architecture. Specifically in Year 7, we worked on realizing higher-level functions of re-id (i.e., systems that can automatically recover trajectories of persons of interest walking through a network of cameras). These two aspects could be easily transited to real-world applications, in the form of algorithm module development. Possible avenues for transition/industrial collaboration were discussed with the ALERT team in December 2018 (e.g., nonairport scenarios like ports/shipping, customs, campus security / public safety). It would be particularly interesting to partner with the re-id effort apparently underway at Lincoln Labs.

Another aspect of transition in which our project is strong is the creation, packaging, and dissemination of algorithms and datasets that can immediately be adopted by DHS stakeholders. These include several benchmark datasets/codebases published in the past three years (e.g., the ALERT Airport Re-Id Dataset, the RPIField Dataset, the DukeMTMC4ReID dataset, the easy-to-use GitHub codebase that accompanies our survey paper [23]).

C. Transition Pathway and Future Opportunities

With the research of four graduate students concluded over the lifetime of ALERT, the project leaves behind a large codebase with state-of-the-art re-id performance directly applicable to DHS surveillance systems. Our early experience at CLE demonstrated that interfacing our algorithms to a legacy surveillance system requires a substantial effort, as well as months of negotiations that generally have to occur above the individual project level. As a center, ALERT continues to push strongly into this space (e.g., the current effort to integrate the CLASP algorithms at KRI into a domestic airport). If these CLASP agreements are successful, the target airport would be a natural venue to transition the re-id algorithms developed in this project.

Three of the four ALERT PhD students still work closely together, first at Siemens Corporate Research in Princeton, New Jersey and now at United Imaging Intelligence in Cambridge, Massachusetts. They would be willing to advise future students or staff on concrete transitions to field sites once agreements are in place.

D. 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

Of these, Michael Young was in weekly contact with the PI and students over the course of the transition project at CLE and had substantial input over the direction of the project.

Little contact was made with these individuals in Years 4–7 due to retirements and reorganizations. The proposed research in this thrust would be quite relevant to airports and rail stations, if the contacts were to be reestablished and sufficient resources made available. There is currently considerable interest and ongoing negotiations from airport officials with respect to the related CLASP project, and some of these contacts may also be interested in the re-id problem. George Naccara, retired CIO, MassPort is playing a key role in the current CLASP transition effort.

IV. PROJECT ACCOMPLISHMENTS AND DOCUMENTATION

A. Education and Workforce Development Activities

1. Student Internship, Job, and/or Research Opportunities
    a. PhD student Meng Zheng defended her thesis in March 2020 and graduated from RPI in May 2020. She joined United Imaging Intelligence in Cambridge, Massachusetts in June 2020.

B. Peer Reviewed Journal Articles

Pending –

1. Zheng, M., Karanam, S., & Radke, R.J. "Towards Automated Forensic Reconstruction in Wide-Area Multi-Camera Networks." *IEEE Transactions on Biometrics, Behavior, and Identity Science*, submitted May 2020.

C. Peer-Reviewed Conference Proceedings

1. Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R.J., & Camps, O. "Towards Automated Forensic Reconstruction in Wide-Area Multi-Camera Networks." *IEEE Conference on Computer Vision and Pattern Recognition,* 2020.

Pending –

1. Zheng, M., Karanam, S., Wu, Z., Chen, T., & Radke, R.J. "Learning Similarity Attention." *European Conference on Computer Vision*, in review, 2020.

D. Student Theses or Dissertations Produced from This Project

1. Zheng, M. "Design of Real-World Person Re-Identification Systems." PhD Thesis, Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, May 2020.

V. REFERENCES

[1] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-Identification with Consistent Attentive Siamese Networks," in *CVPR*, 2019.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012.

[4] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *ICCV*, 2015.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *CVPR*, 2015.

[7] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?" in *CVPR*, 2018.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *ICCV*, 2017.

[9] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future," in *arXiv:1610.02984*, 2016.

[10] N. McLaughlin, J. M. D. Rincon, and P. Miller, "Recurrent Convolutional Network for Video-Based Person Re-identification," in *CVPR*, 2016.

[11] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A Video Benchmark for Large-Scale Person Re-Identification," in *ECCV*, 2016.

[12] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning," in *CVPR*, 2018.

[13] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person Re-identification by Descriptive and Discriminative Classification," in *SICA*, 2011.

[14] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity Regularized Spatiotemporal Attention for Video-based Person Re-identification," in *CVPR*, 2018.

[15] Y. Liu, Y. Junjie, and W. Ouyang, "Quality Aware Network for Set to Set Recognition," in *CVPR*, 2017.

[16] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "BIER - Boosting Independent Embeddings Robustly," in *ICCV*, 2017.

[17] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon, "Attention-Based Ensemble for Deep Metric Learning," in *ECCV*, 2018.

[18] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification," in *CVPR*, 2016.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," in *CVPR*, 2014.

[20] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person Re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function," in *CVPR*, 2016.

[21] M. Zheng, S. Karanam, T. Chen, R. J. Radke, and Z. Wu, "Learning Similarity Attention," in *arXiv: 1911.07381*, 2019.

[22] M. Zheng, S. Karanam, and R. J. Radke, "RPIfield: A New Dataset for Temporally Evaluating Person Re-Identification," in *CVPR Workshops*, 2018.

[23] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R.J. Radke, "A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[24] Z. Wu, Y. Li, and R. J. Radke, Viewpoint Invariant Human Re-Identification in Camera Networks Using Pose Priors and Subject-Discriminative Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 5, pp. 1095-1108, May 2015.

[25] S. Karanam, Y. Li, and R. J. Radke, Person Re-Identification with Discriminatively Trained Viewpoint Invariant Dictionaries, International Conference on Computer Vision (ICCV), December 2015.

[26] O. Camps, M. Gou, T. Hebble, S. Karanam, O. Lehmann, Y. Li, R.J. Radke, Z. Wu, and F. Xiong, "From the Lab to the Real World: Re-Identification in an Airport Camera Network," *IEEE Transactions on Circuits and Systems for Video Technology,* special issue "Group and Crowd Behavior Analysis for Intelligent Multi-camera Video Surveillance," Vol. 27, No. 3, pp. 540-553, March 2017.