R4-A.2: Rapid Similarity Prediction, Forensic Search & Retrieval in Video

I. PARTICIPANTS

<table>
<thead>
<tr>
<th>Faculty/Staff</th>
<th>Name</th>
<th>Title</th>
<th>Institution</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Venkatesh Saligrama</td>
<td>Co-PI</td>
<td>BU</td>
<td><a href="mailto:srv@bu.edu">srv@bu.edu</a></td>
</tr>
<tr>
<td></td>
<td>David Castañón</td>
<td>Co-PI</td>
<td>BU</td>
<td><a href="mailto:dac@bu.edu">dac@bu.edu</a></td>
</tr>
<tr>
<td></td>
<td>Ziming Zhang</td>
<td>Post-Doc</td>
<td>BU</td>
<td><a href="mailto:zzhang14@bu.edu">zzhang14@bu.edu</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Graduate, Undergraduate and REU Students</th>
<th>Name</th>
<th>Degree Pursued</th>
<th>Institution</th>
<th>Month/Year of Graduation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gregory Castañón</td>
<td>PhD</td>
<td>BU</td>
<td>5/2015</td>
</tr>
<tr>
<td></td>
<td>Yuting Chen</td>
<td>PhD</td>
<td>BU</td>
<td>12/2016</td>
</tr>
<tr>
<td></td>
<td>Marc Eder</td>
<td>MS</td>
<td>BU</td>
<td>5/2016</td>
</tr>
</tbody>
</table>

II. PROJECT DESCRIPTION

A. Project Overview

This project develops video analytics for maintaining airport and perimeter security. Our objectives include real-time suspicious activity detection, seamless tracking of individuals across sparse multi-camera networks, and the forensic search for individuals and activities in years of archived data. Surveillance networks are becoming increasingly effective in the public and private sectors. Generally, use of these surveillance networks falls into either a real-time or forensic capacity. For real-time use, the activities of interest are known a-priori, and the challenge is to detect those activities as they occur in the video. For forensic use, the data is archived until a user chooses an activity to search for. Forensic use calls for a method of content-based retrieval in large video corpuses, based on user-defined queries. In general, identifying relevant information for tracking and forensics across multiple cameras with non-overlapping views is challenging. This is difficult given the wide range of variations, from the traditional pose, illumination, and scale issues to spatio-temporal variations of a scene, itself.

The significance of a real-time activity monitoring effort to the Homeland Security Enterprise (HSE) is that these methods will enable the real-time detection of suspicious activities and entities throughout an airport by seamlessly tagging and tracking objects. Suspicious activities include baggage drops, unusual behavior, and abandoning objects. The forensic search capability will significantly enhance current human-driven and relatively short horizon forensic capabilities, and allow for an autonomous search that matches user-defined activity queries in years of compressed data for detecting incidents such as a baggage drop, and identifying who/what was involved in that incident over large time-scales. Boston Logan International Airport (BOS) currently has the capability to store ~1 month’s data, and much of the forensics requires significant human involvement. Our proposed research will generate new techniques for real-time activity recognition and tracking with higher probability of correct detection and reduced false alarms. Furthermore, it will enable the rapid search of historical video for the enhanced detection of complex activities in support of security
applications.
We will describe ongoing efforts related to both real-time monitoring and forensic search in more detail below. We propose to develop robust techniques for a variety of environments including unstructured, highly-cluttered, and occluded scenarios. A significant focus of the project is the development of robust features. An important consideration is that the selected features should not only be informative and easy to extract from the raw video but should also be invariant to pose, illumination, and scale variations. Traditional approaches have employed photometric properties. However, these features are sensitive either to pose, illumination, or scale variations, or are sensitive to clutter. Moreover, they do not help capture the essential patterns of activity in the field of view. Consequently, they are not sufficiently informative for generalization within a multi-camera framework.

A.1. Real-time activity monitoring
Real-time activity monitoring requires both short-term and long-term surveillance. Short-term threat detection involves the detection of baggage drops, abandoned objects, and other types of sudden unusual behaviors. On the other hand, long-term monitoring could involve identifying and tagging and tracking individuals associated with short-term threats in order to determine precursors, such as who met these targets, etc. Ongoing efforts include suspicious activity detection coupled with person re-identification (re-id) to ensure multi-camera tagging and tracking of individuals across camera networks.

A.2. Forensics
It is worth touching upon the different characteristics of the forensic and real-time problem sets. In both problems, given the ubiquity of video surveillance, it is a fair assumption that the video to be searched grows linearly with time and will stream inconsistently. This mandates an ability to detect a predetermined activity in data as quickly as it streams in, for the real-time model. In the forensic model, this massive data requirement means that: (1) whatever representation is archived is computable as quickly as the data streams in; and (2) the search process scales sub-linearly with the size of the data corpus. If there is a failure to fulfill the first requirement, the system will fall behind. If there is a failure to fulfill the second, a user will have to wait too long for his results when searching a large dataset.

B. Biennial Review Results and Related Actions to Address

B.1. Strengths
On the technical merits, the committees recognized the soundness of the technical approach for person re-id as well as forensic search. They also pointed out improvements in performance over the previous year by reducing the feasible solution space. For the activity detection part, the committees recognized that this project addresses several knowledge gaps in video analytics (searching, computer vision, autonomous detection, etc.). The technical committee also recognized the expertise of the researchers carrying out this work. The FCC recognized that this project addresses an important, emerging challenge of how to consolidate into cohesive capabilities as the number of cameras and sensing modalities grows. They stressed the importance of addressing an incomplete dataset problem, which occurs when cameras are not providing contiguous coverage, resulting in a system having to re-establish an id. They also recognized the importance of the targeted search of video to many stakeholders.

B.2. Weaknesses
The technical committee did not fully grasp the methodology adopted, and questioned the significance of the person re-id research with respect to the existing literature.
For the project on activity analysis and search, while the committees recognized the merits of the approach, they felt that the current project did not adequately consider the context and semantics that are necessary for relevant analysis and retrieval. They also felt that the approach could benefit from learning and updating models over time. The FCC committee felt that while there were several stakeholders that had expressed interest; they felt that none were actively involved currently. They also felt that specific goals, milestones, and as how to achieve them were lacking.

B.3. Proposed plan

The problem of person re-id has received much recent attention, but without much success. While our algorithms currently rank among the best performing in the literature, the state-of-the-art performance of person re-id algorithms on benchmark datasets has about 40% accuracy for the top ranking hypothesis to be correct. In order to improve this performance, we will investigate the development of deep learning algorithms for person re-id. Unlike existing approaches that typically rely on hand-crafted features, deep learning algorithms are capable of automatically learning feature representations that are most relevant to the recognition task.

In order to increase the applicability of our results to different security applications, we will develop approaches for re-id in the open world. The current literature on person re-id is constrained to a closed-world setting, and is somewhat unrealistic; by “closed world re-id”, we mean the tracking of a fixed set of persons appearing in known ways in multiple cameras. Our proposed research is to address requirements of unconstrained airport surveillance where unknown persons appear in an unknown subset of cameras. We propose to extend our person re-id algorithms to deal with track discontinuities and the arrival and exit of individuals from the camera network by imposing global constraints on the network.

In the forensic search, we propose to focus on the recognition and retrieval of unusual, unscripted, and abnormal activities that are based on semantic information provided by user queries. This manner allows users to build their own models for what they are interested in searching. In this context, we will develop a class of zero-shot learning algorithms that attempt to recognize semantic actions, retrieve matches, and detect unusual incidents in the absence of models.

C. State of the Art and Technical Approach

C.1. Activity monitoring in real-time: person re-id

C.1.a. Related work

While re-id has received significant interest [1-3], much of this effort can be viewed as methods that seek to classify each probe image into one of a gallery of images. Broadly, re-id literature can be categorized into two themes, with one focusing on cleverly designing local features [4-19], and the other focusing on metric learning [20-32]. Typically, local feature design aims to find a re-id specific representation based on some properties among the data in re-id, e.g. symmetry and centralization of pedestrians in images [7], color correspondences in images from different cameras [17 and 18], spatial-temporal information in re-id videos/sequences [6 and 8], discriminative image representation [4, 5 and 11], and viewpoint invariance prior [19]. Unlike these approaches, that attempt to match local features, our method attempts to learn changes in appearance or features to account for visual ambiguity and spatial distortion. On the other hand, metric learning aims to learn a better similarity measure using, for instance, transfer learning [23], dictionary learning [24], distance learning/comparison [25, 27 and 28], similarity learning [29], dimension reduction [30], template matching [31], and active learning [32]. In contrast to metric learning approaches that attempt to find a metric such that features from positively associated pairs are close in distance, our algorithm learns similarity functions for imputing similarity between features that naturally undergo appearance changes.
C.1.b. Technical approach

Many surveillance systems require the autonomous long-term behavior monitoring of pedestrians within a large camera network. One of the key issues in this task is re-id, which deals with how to maintain entities of individuals as they traverse through diverse locations that are surveilled by different cameras with non-overlapping camera views. Re-id presents several challenges. From a vision perspective, camera views are non-overlapping, and so conventional tracking methods are not helpful. Variation in appearance between the two camera views is so significant, due to the arbitrary change in view angles, poses, illumination, and calibration, that features seen in one camera are often missing in the other. Low resolution of images for re-id makes biometrics-based approaches often unreliable [1]. Globally, the issue is that only a subset of individuals identified in one camera (location) may appear in the other.

We have proposed PRISM: Person Re-identification via Structured Matching. PRISM is a weighted bipartite matching method that simultaneously identifies potential matches between individuals viewed in two different cameras. Figure 1 on the next page illustrates our pipeline re-id system with two camera views. At the training stage, we extract low-level feature vectors from randomly sampled patches in training images and then cluster them into codewords to form a codebook, which is used to encode every image into a codeword image. Each pixel in a codeword image represents the centroid of a patch that has been mapped to a codeword. Further, a visual word co-occurrence model (descriptor) is calculated for every pair of gallery and probe images, and the descriptors from training data are utilized to train our classifier using structured learning. We perform re-id on the test data by first encoding images using the learned codebook, then computing descriptors, and finally structurally matching the identities. During testing we have an image from one camera view (probe) that needs to one of the images from the gallery (second camera view). Graph matching requires edge weights, which correspond to similarity between entities viewed from two different cameras. We learn to estimate edge weights from training instances of manually labeled image pairs. We formulate the problem as an instance of a structured learning [33] problem. While structured learning has been employed for matching text documents, re-id poses new challenges. Edge weights are obtained as a weighted linear combination of basis functions. For texts, these basis functions encode shared or related words or patterns (which are assumed to be known a priori) between text documents. The weights for the basis functions are learned from training data. In this way, during testing, edge weights are scored based on a weighted combination of related words. In contrast, visual words (i.e. vector representations of appearance information, similar to the words in texts) are suffering from well-known visual ambiguity and spatial distortion. This issue is further compounded in the re-id problem, where visual words exhibit significant variations in appearance due to changes in pose, illumination, etc.
To handle the visual ambiguity and spatial distortion in re-id, we propose new basis functions based on co-occurrence of different visual words. We then estimate weights for different co-occurrences from their statistics in training data. While co-occurrence-based statistics have been used in some other works [34 and 35], ours has a different purpose. We are largely motivated by the observation that the co-occurrence patterns of visual codewords behave similarly for images from different views. In other words, the transformation of target appearances can be statistically inferred through these co-occurrence patterns. We observe that some regions are distributed similarly in images from different views and robustly in the presence of large cross-view variations. These regions provide important discriminant co-occurrence patterns for matching image pairs. For instance, statistically speaking, “white” color in one camera can change to “light blue” in another camera. However, “light blue” rarely changes to “black”.

We leverage and build on our work [4] on a novel visual word co-occurrence model to capture such important patterns between images. There, we first encode images with a sufficiently large codebook to account for different visual patterns. Pixels are then matched into codewords or visual words. The resulting spatial distribution for each codeword is embedded into a kernel space through kernel means embedding [36], with latent-variable conditional densities [37] as kernels. The fact that we incorporate the spatial distribution of codewords into appearance models provides us with locality sensitive co-occurrence measures. Our approach can also be interpreted as a means to transfer the information (e.g. pose, illumination, and
appearance) in the image pairs to a common latent space for meaningful comparison.

In this perspective appearance, change corresponds to the transformation of a visual word viewed in one camera into another visual word in another camera. Particularly, our method does not assume any smooth appearance transformation across different cameras. Instead, our method learns the visual word co-occurrence patterns statistically in different camera views to predict the identities of persons. The structured learning problem in our method is to determine important co-occurrences while being robust to noisy co-occurrences.

To illustrate the basic mathematics involved in our approach, we are given N probe entities (Camera 1) that are to be matched to M gallery entities (Camera 2). Figure 2 depicts a scenario where entities may be associated with a single image (single-shot), multiple images (multi-shot), and is unmatched to any other entity in the probe/gallery. Existing methods could fail here for the reason that entities are matched independently based on pairwise similarities between the probes and galleries, leading to the possibility of matching multiple probes to the same entity in the gallery. Our approach, based on structured matching, is a framework that can address some of these issues.

To build intuition into our method, consider $y_{ij}$ as a binary variable denoting whether or not there is a match between the $i$th probe entity and the $j$th gallery entity. Denote $s_{ij}$ as their similarity score. Our goal is to predict the structure $y$ by seeking a maximum bipartite matching:

$$
\max_{Y \in \mathcal{Y}} \sum_{i,j} \bar{y}_{ij} s_{ij}, \quad \text{s.t.} \quad \bar{y} = [\bar{y}_{ij}]_{i,j} \in \mathcal{Y}
$$

where $\mathcal{Y}$ could be the sub-collection of bipartite graphs accounting for different types of constraints. For instance, it would account for the relaxed constraint to identify at most $r_i$ potential matches from the gallery set for probe $i$, and at most $g_j$ potential matches from the probe set for gallery $j$. Hopefully the correct matches are among them.

Equation 1 needs a similarity score $s_{ij}$ for every pair of probe $i$ and gallery $j$, which is a priori unknown and could be arbitrary. Therefore, we seek similarity models that can be learned from training data based on
minimizing some loss function. Structured learning [33] formalizes loss functions for learning similarity models that are consistent with testing goals, as in Equation 1. To map it into this setting, we denote probes and gallery images as documents. These documents are a collection of visual words that are obtained using K-means (see [4] and the Phase 2, Year 1 ALERT annual report). We propose similarity models based on cross-view visual word co-occurrence patterns to learn similarity weights.

Our key insight is that aspects of appearance that are transformed in predictable ways, due to the static camera view angles, can be statistically inferred through pairwise co-occurrence of visual words. In this way, we allow the same visual concepts to be mapped into different visual words and account for visual ambiguity. We present a probabilistic approach to motivate our similarity model in Figure 2b. We let the similarity \( s_{ij} \) be equal to the probability that two entities are identical,

\[
\begin{align*}
\text{\( s_{ij} \) & \triangleq p(j_{ij} = 1|I_{1}^{(1)}, I_{j}^{(2)}) \)
= \sum_{u \in U, v \in V, h \in H} p(j_{ij} = 1|u, v)p(u, v|h, I_{1}^{(1)}, I_{j}^{(2)})p(h) \\
& = \sum_{u \in U, v \in V} p(j_{ij} = 1|u, v) \left[ \sum_{h \in H} p(u, v|h, I_{1}^{(1)}, I_{j}^{(2)})p(h) \right]
\end{align*}
\]

where \( I_{1}^{(1)}, I_{j}^{(2)} \) denotes two images from camera view 1 (left) and 2 (right), respectively, \( u, v \) denotes the visual words for view 1 and view 2, and \( h \) denotes the shared spatial locations.

Following along the lines of the text-document setting, we can analogously let \( w_{uv} = p(y_{ij} = 1|u,v) \) denote the likelihood (or importance) of co-occurrence of the two visual words among matched documents. This term is data-independent and learned from training instances. The term \( p(u, v | h, I_{1}^{(1)}, I_{j}^{(2)}) \) must be empirically estimated, and is a measure of the frequency with which two visual words co-occur after accounting for spatial proximity. To handle spatial distortion of visual words, we allow the visual words to be deformable, similar to a deformable part model [12].

In summary, our similarity model handles both visual ambiguity (through co-occurring visual words) and spatial distortion simultaneously. We learn parameters, \( w_{uv} \), of our similarity model along with analogous structured loss functions that penalize deviations of predicted graph structures from ground-truth annotated graph structures.

### C.1.c. Results

We start by interpreting our learned model parameters. A typical learned co-occurrence matrix is shown in Figure 3 on the next page, with 30 visual words per camera view. Recall that \( w_{uv} = p(y_{ij} = 1|u,v) \) denotes how likely two images come from the same person according to the visual word pairs and our spatial kernel returns non-negatives, indicating the spatial distances between visual word pairs in two images from two camera views. As we see in Figure 3, by comparing the associated learned weights, “white” color in camera A is likely to be transferred into “light-blue” color (with higher positive weight), but very unlikely to be transferred into “black” color (with lower negative weight), in camera B. Therefore, when comparing two images from cameras A and B, if within the same local regions, the “white” and “light-blue” visual word pair from the two images occurs, and it will contribute to identifying the same person. On the other hand, if “white” and “black” co-occur within the same local regions in the images, it will contribute to identifying different persons.
C.1.d. Experiments and comparison with the state of the art

C.1.d.i. Single-shot learning

For single-shot learning, each entity is associated with only one single image, and re-id is performed based on every single image pair. In the literature, most of the methods are proposed under this scenario. Table 1 on the next page lists our comparison results on the three datasets, where the numbers are the matching rates over different ranks on the Cumulative Match Curve (CMC). Overall, fusion methods achieve better performance than those (including ours) using single type of features, which is very reasonable, but our method is always comparable. At rank-1, our performance in terms of matching rate is 9.2% on VIPeR and 1.4% on CUHK01, lower than [38]. Using single types of features on VIPeR, “Mid-level filters+LADF” from [30] is the current best method, which utilizes more discriminative mid-level filters as features with a powerful classifier; ”SCNCD_{final}(ImgF)” from [31] is the second, which utilized only foreground features. Our results are comparable to both of them. However, PRISM always outperforms their original methods significantly when either the powerful classifier or the foreground information is not used. On CUHK01 and iLIDS-VID, PRISM performs the best. At rank-1, it outperforms our previous work [4] and [30] by 8.0% and 11.8%, respectively. Compared with our previous work (see [4] and the Phase 2, Year 1 ALERT annual report), our improvement here mainly comes from the structured matching in testing by precluding the matches that are probably wrong (i.e. reducing the feasible solution space).

C.1.d.ii Multi-shot learning

For multi-shot learning, each entity is associated with at least one image, and re-id is performed based on multiple image pairs. How to utilize the redundant information in multiple images is the key difference from single-shot learning. We extend the visual word co-occurrence model for single-shot scenarios to the multi-shot scenario. To do this, we compute a feature vector corresponding to each super-pixel location across all the shots. This feature vector is typically characterized as a histogram of the different codewords found at that location across the multiple shots of the person in that camera view.
For multi-shot learning, since VIPeR does not have multiple images per person, we compare our method with others on CUHK01 and iLIDS-VID only, and list the comparison results in Table 1. Clearly, PRISM beats the state of the art significantly by 36.7% on CUHK01, and 27.5% on iLIDS-VID, respectively, at rank-1. Our multi-shot CMC curves on CUHK01 are also shown in Fig. 4 for comparison.

<table>
<thead>
<tr>
<th>Rank $r =$</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CUHK01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAFT [60]</td>
<td>31.4</td>
<td>58.0</td>
<td>68.3</td>
<td>74.0</td>
<td>79.0</td>
<td>83.0</td>
</tr>
<tr>
<td>LDM [63]</td>
<td>12.1</td>
<td>31.7</td>
<td>41.7</td>
<td>48.3</td>
<td>54.0</td>
<td>58.0</td>
</tr>
<tr>
<td>Ours: PRISM-I</td>
<td>66.5</td>
<td>82.9</td>
<td>87.9</td>
<td>90.7</td>
<td>92.0</td>
<td>93.4</td>
</tr>
<tr>
<td>Ours: PRISM-II</td>
<td>68.1</td>
<td>80.7</td>
<td>85.8</td>
<td>88.7</td>
<td>90.3</td>
<td>92.2</td>
</tr>
<tr>
<td>Ours: PRISM-III</td>
<td>64.6</td>
<td>79.6</td>
<td>85.8</td>
<td>89.5</td>
<td>90.5</td>
<td>92.0</td>
</tr>
</tbody>
</table>

|             | iLIDS-VID |     |     |     |     |     |
| MS-SDALF [14] | 6.3 | 18.8 | 27.1 |    | 37.3 |    |
| MS-Colour&LBP+RSVM | 23.2 | 44.2 | 54.1 | 68.8 |  |    |
| DVR [59]     | 23.3 | 42.4 | 55.3 | 68.4 |  |  |
| MS-SDALF+DVR | 26.7 | 49.3 | 61.0 | 71.6 |  |  |
| MS-Colour&LBP+DVR | 34.5 | 56.7 | 67.5 | 77.5 |  |    |
| Salience [23]+DVR | 30.9 | 54.4 | 65.1 | 77.1 |  |  |
| Ours: PRISM-I | 60.7 | 86.7 | 89.3 | 94.7 | 96.0 | 96.7 |
| Ours: PRISM-II | 62.0 | 86.0 | 90.0 | 94.7 | 96.0 | 96.7 |
| Ours: PRISM-III | 62.0 | 86.0 | 90.0 | 94.7 | 96.0 | 96.7 |

Table 1: Matching rate comparison (%) for multi-shot learning, where “-” denotes no result reported for the method.

For multi-shot learning, since VIPeR does not have multiple images per person, we compare our method with others on CUHK01 and iLIDS-VID only, and list the comparison results in Table 1. Clearly, PRISM beats the state of the art significantly by 36.7% on CUHK01, and 27.5% on iLIDS-VID, respectively, at rank-1. Our multi-shot CMC curves on CUHK01 are also shown in Fig. 4 for comparison.

Figure 4: CMC Curve Comparison on the CUHK01 dataset. The curves show significant improvement in performance with multi-shot images (labeled MS) over single-shot (labeled SS).

The improvement of our method for multi-shot learning mainly comes from the multi-instance setting of our latent spatial kernel (see Eq. 10 in [39]). By averaging all the gallery images for one entity in multi-shot learning, the visual word co-occurrence model constructed is more robust and discriminative than that for single-shot learning, leading to better similarity functions that are beneficial for structured matching in test time and, thus, significant performance improvement. It has been clearly demonstrated that there is an improvement, as we compare our performances using single-shot learning and multi-shot learning on both CUHK01 and iLIDS-VID, with improvements of 16.1% and 40.0%, respectively. Similar to single-shot learning, PRISM-I works the best among all the variants of PRISM.
C.2. Forensics

A video search system operates in two modalities: archival and search. During the archival step, we process each video in a video corpus and extract features from a pre-defined feature vocabulary. These features are all local, as in they are associated with a certain area or location in space/time in a specific video. We store these locations in an inverted index by feature and feature value. So, if the system needs to find all places in a video corpus where we found the color red, it would go to the color index and look in the “red” bin, which would contain all location/video pairs where that color was found. Our main contributions are

1. Inverted indices for efficient archive downsampling: We introduce an inverted hashing scheme for simple features. We use these and other indexing techniques to dramatically downsample a video corpus to the set of features that are potentially relevant to a given query. This allows us to efficiently reason over large video corpora without prior knowledge and without that each corpus being subdivided into small videos.

2. Sub-graph matching in video search: We introduce temporal relationships on simple features to find a wide variety of user-driven queries using a novel dynamic programming approach. We expand this approach to include spatial relationships and search for arbitrary graphs in large videos. In particular, we use a subgraph matching approach to render our method agnostic to background noise. Other approaches use bipartite matching or require a tree-based query [40], and are thus unable to represent activities with a similar degree of structural complexity.

3. Tree-matching for space-downsampling in video search: We introduce a novel method for successive search-space reduction based on selecting the Maximally Discriminative Spanning Tree. We extend this method to iteratively reduce the search space based on the statistics of the dataset. This approach significantly outperforms contemporary algorithms for search space reduction in subgraph matching, like random trees.

C.2.a. Related work

All approaches to video-based exploration aim to serve the same purpose: to reduce a video to the sections which are relevant to the user’s interest. The simplest form of this is video summarization, which focuses on broad expectations of what interests a user. Videos, which have scene transitions and sparse motion, are good candidates for these broad expectations; scene transitions are interesting, and absence of motion tends to be uninteresting. Recent approaches [41 and 42] divide the video into “shots” and summarize based on the absence of motion. More complex models of human interest rely on input from the user to denote activities that matter to them. In real-time problems, approaches based on previously-specified exemplar videos are extremely popular. Most approaches try to construct a common feature representation for the exemplar videos corresponding to each topic [43, 44 and 45]. Others try to learn hidden variables, such as rules (e.g. traffic lights, left-turn-lanes and building entries), which govern behaviors in the training videos. These rules and behaviors are called “topics”, and common topic modeling techniques include Hidden Markov Models (HMMs) [46 and 47], Bayesian networks [48], context free grammars [49], and other graphical models [50, 51 and 52]. Most of these approaches are primarily employed in a real-time context; models are defined before the archive data begins streaming, and are detected as the data streams in. This is necessary because training complex models from exemplar video is time-consuming. Likewise, the features that are used [43 and 45] are memory-intensive and often over complete. Many of these techniques [46 and 50] also rely on tracking, which can be difficult to perform on large datasets given obscuration, poor resolution, and changes in lighting conditions. Once a model has been created for activities or topics, the classification state can be used to retrieve these patterns [47].

Forensics poses fundamental technical challenges, including:

- Data lifetime: Since video is constantly streamed, there is a perpetual renewal of video data. This calls for
a model that can be updated incrementally as video data is made available. The model must be capable of substantial compression for efficient storage. Our goal is to leverage the relatively stationary background and exploit dynamically changing traffic patterns to realize 1000X compression.

- Unpredictable queries: The nature of queries depends on the field of view of the camera, the scene, the type of events being observed, and the user’s preferences. The system should support queries that can retrieve both recurrent events, such as people entering a store, as well as infrequent events, such as abandoned objects or aimless lingering.

- Unpredictable event duration: Within semantically equivalent events, there is significant variation. Events start anytime, vary in length, and overlap with other events. The system is nonetheless expected to return complete events regardless of their duration and whether or not other events occur simultaneously.

- Clutter: Events in real surveillance videos rarely happen in isolation. Videos have a vast array of activities, so the majority of a video tends to be comprised of activities unrelated to any given search. This “needle in a haystack” quality differentiates exploratory search from many standard image and video classification problems.

- Occlusions: Parts of events are frequently occluded or do not occur. Trees, buildings, and other people often get in the way and make parts of events unobservable.

The challenges of search can be summarized as big data, unknown query when the data arrives, numerous false alarms, and poor data quality. To tackle these challenges, we utilize a three-step process that generates a graphical representation of an activity, downsamples the video to the potentially relevant data and then reasons intelligently over that data. This process is shown in Figure 5.

Figure 5: In the archival step, we take incoming data, extract attributes and relationships and store them in hash tables. In the query creation step, a user utilizes our GUI to create a query graph that is used to extract the coarse graph $C$ from archive data. In the Maximally Discriminative Subgraph Matching (MDSM) step, we calculate the maximally discriminative spanning tree (MDST) from the query graph, retrieve matches to it, and assemble them into ranked search results for the user.

Due to data magnitude, the first step of any approach to a large-scale video search has to be an efficient storage mechanism for the raw video data. To this end, we define a broad feature vocabulary that is extracted in real time as data streams in. For raw video, we extract activity, object size, color, persistence, and motion. Given a tracker, we also identify object types, such as people and vehicles. To facilitate $O(1)$ recovery of these features, we store discrete features in hash tables and continuous-valued features in fuzzy hash tables using
Locality Sensitive Hashing (LSH) [53]. This step addresses a number of the aforementioned challenges. Data reduction is achieved because feature locations are stored, rather than feature or raw pixel values. The imprecision of the features, as well as the quantization via fuzzy hashing, serves to mitigate the noisiness of the data. Finally, because of the hash table structure, features can be extracted at a fixed cost, which allows us to construct a set of potentially relevant features if we can identify which bins in the hash table correspond to a given query.

Next, we acquire a query from a user. Most video search approaches rely on exemplar videos to detect a given activity. In the context of large-scale video search for complex actions, this becomes difficult to do; complex activities require a great number of clean examples to learn models from. These models are frequently hard to come by.

Instead, we leverage the fact that our features are simple and semantically meaningful, and provide the user with a Graphical User Interface (GUI) to build their own query in the form of a graph. This graph takes the form of a series of features (nodes) and relationships (edges) between those features that he expects to find in the video. The relationships come from a separate vocabulary; common examples include spatial and temporal [54] relationships. However, not all relationships need to be as structured or simplistic; given a matching engine which compares feature profiles of identified people, “likely the same” could be a relationship as well.

These features and relationships comprise the query graph, G_q=(V_q,E_q), a representation of the query of interest (see Fig. 5). Given this graph, our goal in the second step of our approach is to find the features and relationship in the archive data. The archive data can also represented as a graph, G_c=(V_c,E_c), albeit a large one. Our task is to find a subgraph with maximum similarity to the query graph. We define a distance metric from the ideal query graph to a given set of features in the archive, which encompasses missing elements (deletions) as well as displaced elements (distortions). Computing an approximate subgraph isomorphism is NP-complete and, thus, computationally infeasible. Our approach is to solve this problem using a novel random sample-tree auction algorithm, which solves a series of dynamic programming problems to rank candidate matches in descending order of similarity.

![Graphical representation of an object deposit event.](image)

We solve for matching function M: V_q \rightarrow V_c, where M is a one-to-one function. As exact computation of an optimal subgraph matching is known to be NP-hard, we instead select a spanning tree Q_t of the graph to search for and solve a tree-matching problem via dynamic programming (DP).

C.2.b. Tree selection and search

Because this is a search problem, the spanning tree selected has significant run-time implications. Because
the search is iterative, starting at the root and moving down, placing discriminative nodes and edges near the top of the trees pays continuous dividends throughout the search. To this end, we select $Q_t$ to minimize the total number of look-ups. Given a tree $Q$, and a breadth-first ordering of the nodes $v_0, v_1, ..., v_i$ with $V_0$ being the root we define scores $S(v)$, and $S(v1,v2)$ for vertices and edges, respectively, these scores denote the percentage of the archive which matches vertex $v$ or edge $(v1,v2)$. This problem can be formulated as an All-Source Acyclic Longest Path problem, which is NP-hard to solve precisely. However, we have found that, in practice, random sampling is highly likely to yield a near-optimal tree.

C.2.c. Experiments and comparisons

We explored the VIRAT 2.0 street surveillance dataset from building-mounted cameras. This is a popular surveillance dataset containing 35 gigabytes (GB) of video and represented in a graph of 200,000 nodes and 1 million edges. These are relatively standard 2-megapixel surveillance cameras acquiring image frames at 30 frames per second. Because of the smaller field of view, there are far more pixels on target, enabling basic object recognition to be performed. As such, we define A to include object type (e.g. person, vehicle and bag) as well as the attributes. The VIRAT ground dataset contains 315 different videos covering 11 scenes rather than a single large video covering one scene.

We demonstrate run-time of our approach in Table 2 on the next page. This demonstrates the futility of solving a large-scale graph search problem without performing intelligent reduction first. It should not be surprising that an algorithm which must explore all $|Vq|$-sized subsets of the data will take a long time to run on a large dataset. More relevant is how long it takes us to compute the Maximally Discriminative Spanning Trees approach (MDST) and downsample the data to the relevant subset. We observe that this is less than a second with pre-hashed relationships (all examples except meetings) and 80 seconds when we do not have pre-hashed relationships. When we do not have hashed relationships, our algorithm must compute pair-wise relationships, which is expensive to do even when the data is significantly reduced.

<table>
<thead>
<tr>
<th>Query</th>
<th>MDST Reduction</th>
<th>Baseline[3]</th>
<th>Brute (MDST)</th>
<th>Brute (MDST)</th>
<th>DP (MDST)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long Meeting</td>
<td>85</td>
<td>6.5</td>
<td>3e25</td>
<td>4610</td>
<td>2.3</td>
</tr>
<tr>
<td>Short Meeting</td>
<td>86</td>
<td>6</td>
<td>3e25</td>
<td>10671</td>
<td>2.5</td>
</tr>
<tr>
<td>Deposit</td>
<td>&lt;1</td>
<td>12.5</td>
<td>4.5e16</td>
<td>&lt;1</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Removal</td>
<td>&lt;1</td>
<td>14.4</td>
<td>4.7e16</td>
<td>4.9</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Mount</td>
<td>&lt;1</td>
<td>6.3</td>
<td>3.7e8</td>
<td>&lt;1</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Dismount</td>
<td>&lt;1</td>
<td>7.7</td>
<td>3.7e8</td>
<td>&lt;1</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Turn</td>
<td>&lt;1</td>
<td>2.0269</td>
<td>6.6e8</td>
<td>&lt;1</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Dismount</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Mount</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Suspicious Stop</td>
<td>&lt;1</td>
<td>&lt;1</td>
<td>32.2820</td>
<td>&lt;1</td>
<td>&lt;1</td>
</tr>
</tbody>
</table>

Table 2: The run times for baseline [3], brute force and DP (dynamic programming) algorithms on the VIRAT (top) and YUMA (bottom) datasets. When a brute force algorithm is infeasible, we estimate runtime based on a subset of solutions.

D. Major Contributions

We have made significant progress in re-id. In particular, we have proposed a new structured matching approach. The first key aspect of this approach is that, in contrast to existing methods that match each individual independently in other cameras, our machine learning algorithms incorporate the insight that two
people cannot be at two different places at the same time. This insight is enforced both in training as well the testing phases. A second contribution of our approach is that we model for appearance changes. Specifically, we incorporate the fact that aspects of appearance can be transformed in predictable ways, due to the static camera view angles. These appearance changes can be statistically inferred through pairwise co-occurrence of visual words. These two aspects are key factors in significantly improving the accuracy of our results. To summarize our contributions:

• We have proposed a new structured matching method to simultaneously identify matches across multiple cameras.
• Our framework can seamlessly deal with both single-shot and multi-shot scenarios in a unified framework.
• We account for significant changes in appearance through the design of new basis functions, which are based on visual word co-occurrences.
• We outperform the state of the art significantly on several benchmark datasets, with good computational efficiency in testing.

We have begun to explore the forensic theme by leveraging ongoing parallel efforts funded by the Department of Defense (DOD)/the National Geospatial-Intelligence Agency (NGA).

The key aspect of our approach is based on efficient retrieval for activity detection in large surveillance video datasets based on semantic graph queries. Unlike conventional approaches, our method does not require knowledge of the activity classes contained in the video. Instead, we propose a user-centric approach that models queries through the creation of sparse semantic graphs based on attributes and discriminative relationships. We then pose search as a ranked subgraph matching problem and leverage the fact that the attributes and relationships in the query have different levels of discriminability to filter out bad matches. In summary our contributions include:

• A user-centric approach to model acquisition through the creation of sparse semantic graphs based on attributes and discriminative relationships. Rather than perfectly model every aspect of an activity, we provide a user with a series of simple semantic concepts and arrange them in a graph.
• We use this query in a sub-graph matching approach to identifying activity; this allows our algorithm to effectively ignore confusing events and clutter that happen before, after, and during the activity of interest. This graphical representation also makes the approach relatively agnostic to the duration of the event, allowing it to detect events that take place over several minutes.

D.1. Quantitative outcomes in Year 3

1. We developed a graphical user interface that will take a query graph as an input. The query graph will guide the user to depict the activity that is required to be retrieved.
2. We tested and validated our approach on both indoor, near-field and airborne datasets. We plan to demonstrate archival storage efficiencies ranging from 10X to 1000X on a number of different indoor and outdoor datasets.
3. We developed software for retrieval algorithms for indoor, outdoor, and aerial video data.
4. We developed successive search space reduction techniques that will build upon MDST techniques presented this year. Our goal is to reduce search space based on successive search to improve retrieval time by as much as 10-50% over conventional tree-matching algorithm.
5. Our goal is to realize over 15% AUC precision/recall improvement over conventional feature-accumulation algorithms.
E. **Milestones**

Some of our accomplishments over the last year are:

- Multi-shot re-id was accomplished. Our goal is to extend our current multi-shot re-identification to open-world re-identification scenarios.
- Six papers were presented at top computer vision conferences (ICCV, ACM MM, CVPR, ECCV). Two journal papers have been accepted. We also developed a software library.
- One Ph.D. student, Greg Castañón, completed his degree and is currently employed at Science and Technology Research (STR Inc.) a defense contractor. One Ph.D. student will defend in Year 4.

Year 4 milestones that need to be achieved include:

- Deep re-id algorithm development and performance evaluation.
- Deep hashing algorithm development for fast retrieval with low false alarms and missed detections.
- Robust re-id work with structured output prediction for unconstrained airport scenarios.
- An efficient subgraph matching algorithm for the fast retrieval of unusual and anomalous activity for forensic search.
- Efficient retrieval algorithm development that combines with re-id to recover lost tracks, to improve retrieval performance.
- Validation of re-id algorithms using data collected at the Cleveland Hopkins International Airport (CLE).
- Presentation of forensic search algorithms to the ALERT Transition Team and Industrial Advisory Board to assist in the identification of transition paths.

F. **Future Plans**

F.1. **Person Re-identification**

Our outlier detection will be coupled with longer-term semantic threat discovery. In this context, we plan to leverage our multi-camera tag and track algorithms. One issue with our algorithm is that it is computationally expensive, requiring encoding features in a joint multi-camera space. An additional issue is that it currently applies only to single-shot scenarios. Nevertheless, our framework generalizes to multi-shot and video scenarios as well, which we propose to develop in the future. We believe that this will result in a significant improvement in accuracy. Improvement in accuracy is a fundamental requirement for long-term threat detection. This is because, to overcome errors introduced in the tagging process, one usually creates multiple hypotheses. The number of hypotheses explodes combinatorially with time, and checking each hypothesis becomes intractable. Consequently, our goal would be to improve accuracy through fusion of all available information about each individual. Specifically, in this context we propose to:

- Extend single-shot re-id algorithms to multi-shot, video, and large-scale camera networks.
- Focus on algorithm speed, robustness, and transition-readiness.
- Extend re-id to new problem domains, such as open world re-id in mass transit systems.

The basic risk here is that performance of a real-world tagging and tracking system depends on many factors. While re-id is an important sub-component, the ability to tag and track in highly crowded scenarios even with a single camera is challenging. In addition, a re-id performance is also impacted by significant illumination and pose changes, particularly when we have scenarios where people constantly enter and exit the camera system. The other challenge is the ability to rapidly process high-frame-rate, high-resolution cameras as we scale the number of cameras and number of people/camera. Finally, the ability of the software/hardware infrastructure to deal with high-frame-rate systems is also a factor. One way to mitigate these risks is to
simulate their effects during algorithm development. To this end, we have begun to replicate the effects of an open-world system, with people entering and exiting the system, by creating datasets that have missing people in the gallery. In addition, we have also begun to test our algorithm against loss of tracks and developing techniques that can recover tracks based on re-id. Still, the performance offered by re-id appears to top-out at about 40% at rank-1 performance. This calls for techniques that can fuse multi-modal information, such as integrating video data with cell-phone signals to disambiguate difficult cases.

F.2. Forensic search

Our current forensic search algorithms have primarily been applied in outdoor settings.

- An immediate goal would be to develop a forensic search capability for indoor settings, with particular emphasis on airport datasets. While the outdoor surveillance setting does have clutter, there is significantly more clutter in an airport setting, especially during peak hours.

- A second thrust we propose is to identify an interesting collection of queries. For instance, one goal would be to represent counterflow as a graph and retrieve all activities in a corpus of stored video data. Another goal is to determine how our storage space scales with time. This is a critical factor for increasing the “forensic horizon” from the current setting (about a month’s worth of data at BOS) to over a years’ worth.

- On the technical side, we propose to develop new search algorithms based on MDST. The goal of MDST is to leverage statistics of archive data stored in a compressed database to improve the search algorithm. The idea is to exploit the sparsity of novel elements and calculate an optimal combination of elements to maximally reduce the archive data.

- We propose to focus on the recognition and retrieval of unusual, unscripted, and abnormal activities that are based on semantic information provided by user queries. This manner allows users to build their own models for what they are interested in searching. In this context, we will develop a class of zero-shot learning algorithms that attempt to recognize semantic actions, retrieve matches, and detect unusual incidents in the absence of models.

III. RELEVANCE AND TRANSITION

A. Relevance of Research to the DHS Enterprise

There are thousands of unmanned cameras at Department of Homeland Security (DHS) locations (airports, transit stations, border crossings, etc.). These can be exploited to enable re-id and enhanced security.

The relevant metrics for person re-id are: (a) negligible mis-identification probability, and (b) speed of re-identification with # people and # cameras.

For retrieval the metrics are: (a) storage ratio for archive data, (b) precision/recall of a desired suspicious activity, and (c) retrieval speed.

B. Potential for Transition

- We developed real-time tag and track algorithms (re-id) for deployment at Cleveland Hopkins International Airport (CLE). These algorithms can also be used to enhance safety in other mass-transit scenarios. We are working closely with PIs at Rensselaer Polytechnic Institute (RPI) and Northeastern University (NEU) to develop reliable systems.

- Forensic technology is multi-use (i.e. it can be used by DHS, NGA, or other DOD agencies).

- Interest has been expressed by BOS Massport for transition. Other airports and mass transit locations are possible.
C. Data and/or IP Acquisition Strategy

A patent for forensic search capability through BU patenting office has been submitted.

D. Transition Pathway

- At CLE, the tag and track system was deployed.
- For forensics, there has been no activity for transition by DHS but we have been approached by NGA for a potential proof of concept proposal via their NGA University Research Initiative (NURI) program.

E. Customer Connections

- Re-id: Real-time tag and track system, Transportation Security Administration (TSA) at CLE and BOS.
- Forensic search: Currently talking to companies, including TSA at BOS and Progeny Systems.

IV. PROJECT ACCOMPLISHMENTS AND DOCUMENTATION

A. Peer Reviewed Journal Articles

Pending –


B. Peer Reviewed Conference Proceedings

1. Z. Zhang, Y. Chen, V. Saligrama. “Group Membership Prediction.” International Conference on Computer Vision (ICCV15), Santiago, Chile, 11-18 December 2015.


C. Other Presentations

1. Seminars
2. Poster Sessions
   a. Z. Zhang, Y. Chen, V. Saligrama, “Group Membership Prediction.” International Conference on Computer Vision (ICCV15), Santiago, Chile, 11-18 December 2015.

D. Student Theses or Dissertations

E. Technology Transfer/Patents
   1. Patent Applications Filed (Including Provisional Patents)

F. Software Developed
   1. Algorithms
      a. Person Re-ID via Structured Matching (Matlab)
      b. MDST based Forensic Search Algorithm (MATLAB/Python)

V. REFERENCES

[10] C. Liu, S. Gong, C. C. Loy, and X. Lin, “Person re-identification: What features are important?” in
on Frontiers of Computer Vision (FCV), 2014.


