R4-A.2: Rapid Forensic Search & Retrieval in Video Archives

Abstract— Person re-identification (re-id) is an important problem for monitoring & surveillance at airports. This is emerging as a critical problem with the pervasive use of camera networks in surveillance systems. Re-id deals with maintaining identities of individuals traversing different cameras. The goal of person re-id is to maintain the identity of an individual in diverse locations through different, non-overlapping camera views. During this first year, we have focused on the two-camera re-id problems. We have developed a novel machine-learning algorithm based on structured prediction. Our methodology has led to dramatic improvements in re-id rates. Our performance on benchmark datasets significantly outperforms the current state-of-art by 8.76% and 28.24%, on VIPER and CUHK benchmark re-id datasets respectively.

I. PARTICIPANTS

<table>
<thead>
<tr>
<th>Faculty/Staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Venkatesh Saligrama</td>
</tr>
<tr>
<td>Mohamed Rohban</td>
</tr>
<tr>
<td>Ziming Zhang</td>
</tr>
<tr>
<td>Mohamed El-Gharib</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Gregory Castanon</td>
</tr>
<tr>
<td>Yuting Chen</td>
</tr>
<tr>
<td>Phil Tran</td>
</tr>
</tbody>
</table>

II. PROJECT OVERVIEW AND SIGNIFICANCE

The proposed project will develop systematic techniques using algorithms for tagging, tracking and handoff in a multi-camera scenario. The significance of this project to DHS is that we propose methods that will seamlessly track individuals as they move through the airport, across sparse multi-camera networks. In general, identifying relevant information for tracking across multiple cameras with non-overlapping views is challenging. This is difficult given the wide range of variations, ranging from the traditional pose, illumination and scale issues to spatio-temporal variations of a scene itself. We propose to develop robust techniques for a variety of environments including unstructured, highly-cluttered and occluded scenarios. A significant focus of the project will be 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 and 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.
III. RESEARCH ACTIVITY

A. State-of-the-art and technical approach

A.1 Related work

Existing work on person re-id and appearance modeling can be roughly categorized into distance learning and local feature matching. In distance learning, person re-id is formulated into a metric learning problem, to learn the optimal similarity measure between a pair of person images. In local feature matching, person re-id is based on the matching score of carefully designed local features.

The theme of local features for matching is related to our kernel-based similarity measures. To ensure locality, [5] models the appearances of individuals using features from horizontal strips [11] clusters pixels into similar groups and the scores are matched based on correspondences of the clustered groups. Histogram features that encode both local and global appearance are proposed in [4]. Saliency matching [38, 39], one of the-state-of-the-art methods for re-id, uses patch-level matching to serve as masks in images to localize discriminative patches. More generally, low-level features such as color, texture, interest points, co-variance matrices and their combinations have also been proposed [1, 3, 8, 11, 14, 24, 26, and 31]. In addition, high-level structured features that utilize concatenation of low-level features [26] or deformable part models (DPMs) [28] have been proposed. Metric learning methods have been proposed for re-id (e.g. [6, 23, 27, 40]).
In [17, 30], distance metrics are derived through brightness transfer functions that associate color-levels in the two cameras [41] proposes distance metrics that lend importance to features in matched images over the wrongly matched pairs without assuming presence of universally distinctive features. Low-dimensional embeddings using PCA and local FDA have also been proposed [29]. Supervised methods that select relevant features for re-id have been proposed by [14] using Boosting and by [31] using RankSVMs.

A.2 Technical approach

As described earlier, our theme is fundamentally different from the existing state-of-art. We view re-id as an instance of bipartite graph matching. We simultaneously match all or a sub-collection of individuals in one camera to those in the other. This is natural for many surveillance contexts, such as in airports, where multiple entities are viewed in a camera at any time.

We use Structured Prediction to learn matches based on manually labeled matchings of training data. Employing structured prediction for learning graph matchings is a well-studied topic [33]. We follow this approach and represent the matching objective as a weighted combination of basis functions and learn the relative importance of the different basis functions. In many scenarios, these basis functions encode shared or related words or patterns. Re-id demands a different approach on account of the significant variation of appearance due to changes in pose and illumination. Many visual words are missing and not common even among the ground-truth matched images.

We encode pairwise co-occurrences of visual words in our basis functions. The use of co-occurrence patterns is not new but our purpose is different. 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. The structured learning problem is to determine important co-occurrences while being robust to noisy co-occurrences. Indeed, as seen in Figure 1 on the next page, 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.
Pairwise co-occurrences of visual words can be modeled in many ways. However, it has to be semantically meaningful, namely, it has to capture changes in similar things; i.e., shirt-with-shirt, skirt-with-skirt, etc. We encode images with a sufficiently large codebook to account for different visual patterns. We then map pixels into codewords (i.e. visual words) and embed the resulting spatial distribution of pixels belonging to a codeword into a kernel space through kernel mean embedding [32] with latent-variable conditional densities [18] as kernels. In this way, we obtain locality sensitive co-occurrence measures that model semantically meaningful appearance changes. Alternatively, we can also interpret our approach (see Fig. 1) as a means to transfer the information (e.g. pose, illumination and appearance) in image pairs to a common latent space for meaningful comparison.

A.3 Visual word co-occurrence models

We generally face two issues in visual recognition problems: (1) visual ambiguity [30] (i.e. the appearance of instances which belong to the same thing semantically can vary dramatically in different scenarios), and (2) spatial displacement [31] of visual patterns. While visual ambiguity can be somewhat handled through codebook construction and quantization of images into visual words, our goal of matching humans in re-id imposes additional challenges. Human body parts exhibit distinctive local visual patterns and these patterns systematically change appearance locally. Our goal is to account for this inherent variability in appearance models through co-occurrence matrices that quantify spatial and visual changes in appearance.

A.3.1 Locally sensitive co-occurrence designs

We need co-occurrence models that not only account for the locality of appearance changes but also the random spatial and visual ambiguity inherent in vision problems. Therefore, we construct a codebook $Z$ with $M$ codewords. Our codebook construction is global and thus only carries information about distinctive visual patterns. Nevertheless, for a sufficiently large codebook, distinctive visual patterns are mapped to different elements of the codebook, which has the effect of preserving local visual patterns. Specifically, we map each pixel at each 2D location $\pi \in \Pi_f$ image $I$ into (at least one) codewords to cluster pixels. To emphasize local appearance changes, we look at the spatial distribution of each codeword. Concretely, we let $C(I; z)$ denote the set of pixel locations associated with codeword $z$ in image $I$ and associate a spatial probability $p(\pi | z, I)$ over this observed collection. In this way, visual words are embedded into a family of spatial distributions. Intuitively, it should now be clear that we can use the similarity (or distance) of two corresponding spatial distributions to quantify the pairwise relationship between two visual words. This makes sense because our visual words are spatially locally distributed and small distance between spatial distributions implies spatial similarity.
locality. Together, this leads to a model that accounts for local appearance changes.

While we can quantify the similarity between two distributions in a number of ways, the kernel mean embedding method is particularly convenient for our task. The basic idea to map the distribution, $p$, into a reproducing kernel Hilbert space (RKHS) is

$$ p \rightarrow \mu_p(\cdot) = \sum K(\cdot, \pi)p(\pi) \stackrel{\text{def}}{=} E_p(K(\cdot, \pi)). $$

For universal kernels, such as RBF kernels, this mapping is injective; i.e., the mapping preserves the information about the distribution [7]. In addition, we can exploit the reproducing property to express inner products in terms of expected values.

The co-occurrence matrix (and hence the appearance model) is the inner product of visual words in the RKHS space, namely,

$$ \phi(x_{ij})_{mn} = \left\langle \mu_{p(\cdot|z_m, I^{(1)}_i)}, \mu_{p(\cdot|z_m, I^{(2)}_j)} \right\rangle = \sum_{\pi_u} \sum_{\pi_v} K(\pi_u, \pi_v)p(\pi_u|z_m, I^{(1)}_i)p(\pi_v|z_n, I^{(2)}_j), $$

where we have used the reproducing property in the last equality.

We develop novel latent spatial kernels for this purpose. Figure 2 illustrates computation of visual basis functions with latent appearance model given the codeword images. Here, each codeword is represented as a collection of codeword slices. For each codeword slice, the max operation is performed at every pixel location to search for the spatially closest codeword in the slice. This procedure forms a distance transform image, which is further mapped to a spatial kernel image. It allows each peak at the presence of a codeword to be propagated smoothly and uniformly. To calculate the matching score for a codeword co-occurrence, the spatial kernel from a probe image and another from a gallery image are multiplied element-wise and then summed over all latent locations. This step guarantees that our descriptor is insensitive to the noise data in the codeword images. This value is a single entry at the bin, indexing the codeword co-occurrence in our descriptor for matching the probe and gallery images. As a result, we have generated a high dimensional sparse appearance descriptor.

Figure 2: Illustration of our visual word co-occurrence model generation process. Here, the white regions in the codeword slices indicate the pixel locations with the same codeword. A and B denote two arbitrary pixel locations in the image domain. $\Sigma$ denotes a sum operation which sums up all the values in the point-wise product matrix into a single value $\Phi(x_{ij})_{mn}$ in our model.
A.4 Implementation

We illustrate the schematics of our method in Figure 3. 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, performing re-id on the test data.

Specifically, for each image a 672-dim ColorSIFT [2] feature vector is extracted for a 10*10 pixel patch centered at every possible pixel. Further, we decorrelate each feature using the statistics learned from training data.

For codebook construction, we randomly sample 1000 patch features per image in the training set, and cluster these features into a codebook using K-Means. Then we encode each patch feature in images from the probe and gallery sets into a codeword whose Euclidean distance to the patch feature is the minimum among all the codewords. As a result, each image is mapped into a codeword image whose pixels are represented by the indices of the corresponding encoded codewords. We also normalize our appearance descriptors using min-max normalization. The min value is for our descriptors is always 0, and the max value is the maximum among all the codeword co-occurrence bins over every training descriptor. This max value is saved during training and utilized for normalization during testing. For learning classifiers, we employ LIBLINEAR, an efficient linear SVMs solver, with the `2 norm regularizer. The trade-off parameter c in LIB-LINEAR is set using cross-validation. Our pipeline approach is described in Figure 3.

Figure 3: The pipeline of our method, where “codebook” and “classifier” are learned using training data, and each color in the codeword image denotes a codeword.

A.5 Results & comparison with state-of-the-art

The VIPeR dataset is a benchmark dataset for person re-id. VIPeR is comprised of 632 different pedestrians captured in two different camera views, denoted by CAM-A and CAM-B respectively. Many cross-camera image pairs in the dataset have significant variations in illumination, pose and viewpoint, and each image is normalized to 128_48 pixels. In order to compare with other person re-id methods, we followed the experimental set up described in [38]. The dataset is split in half randomly, one partition for training and the other for testing. In addition, samples from CAM-A form the probe set, and samples from CAM-B form the gallery set. Figure 4 on the next page shows our matching rate comparison with other methods on this dataset. When the codebook size is 100, which is pretty small, our performance is close to that of SalMatch [38]. With increase of the codebook size, our performance improves significantly, and outperforms SalMatch by large margins. For instance, at rank-15, our best matching rate is 10.44% higher. Using larger sizes of codebooks,
the codeword representation of each image has finer resolution because the quantization in the feature space is reduced. However, it seems that when the codebook size is beyond 500, our performance is saturated. Therefore, in the following experiments, we only test our method using 100/200/500 codewords. Figure 4 illustrates some codeword co-occurrence examples with relatively high positive/negative weights in the learned weighting matrix.

Figure 4: Matching rate comparison between different methods on (a) VIPeR and (b) CUHK Campus datasets. Numbers following ours in the legends denote the size of the codebook used in each experiment. Except for our results, the other CMC curves are cited from [38].

The CUHK Campus dataset is a relatively new person re-id dataset explored by two state-of-the-art approaches outlined in [38] and [22]. This dataset consists of 1816 people captured from five different camera pairs, labeled P1 to P5. Each image contains 160_60 pixels. Following the experimental settings from [38] and [22], we use only images captured from P1 as our dataset. This subset contains 971 people in two camera views, with two images per view, per person. One camera view, which we call CAM-1, captures people either facing towards or away from the camera. The other view, CAM-2, captures the side view of each person. For our experiments, we adopt the settings described in [2] for comparison. We randomly select 485 individuals from the dataset and use their 4 images for training, and the rest are used for testing. The gallery and probe sets are formed by CAM-1 and CAM-2 respectively. To re-identify a person, we compare the probe image with every gallery image, leading to 486 times 2=972 decision scores. Then, per person in the gallery set, we average the 2 decision scores belonging to this person as the final score for ranking later. Figure 4 summarizes our matching rate comparison with some other methods. Clearly, using only 100 codewords, our method has already outperformed others dramatically, and it works better when using larger sizes of codebooks, similar to the behavior in Figure 4. At rank-15, our best performance is 22.27% better than that of SalMatch.

B. Major contributions

1. Developed a new conceptual framework for multi-tag and track based on discriminative learning.
2. Our approach is robust to appearance changes such as pose, calibration and illumination. Our approach does not require estimation of pose or modeling of illumination artifacts or correcting for camera parameters.
3. Empirical performance of our method on benchmark re-id datasets (VIPeR [12] and CUHK Campus [38]) achieves accuracy rates of 38.92% and 56.69%, at rank-1 on the so-called Cumulative Match Characteristic curves and beats state-of-the-art results by 8.76% and 28.24% respectively. Cumulative Match Characteristic (CMC) curve is a standard metric for re-id performance. The CMC curve displays an algorithm's recognition rate as a function of rank. For instance, a recognition rate at rank-r on the
CMC curve denotes what proportion of queries were correctly matched to a corresponding gallery individual at rank-r or better.

C. **Future plans**

During the following year we plan to work on the following extensions:

1. Extend our two-camera setup to handle multiple cameras. We will need to model multi-camera topology to account for spatio-temporal dependencies.
2. Extend our current framework to videos. Our current method matches frames across two cameras. However, in practice, cameras record videos and it would be important to exploit temporal features provided by videos to improve performance.
3. Computational Efficiency- While our current method beats the state-of-art, it requires complex features in the form of co-occurrence representations. Our goal is to develop new representations that are fast to compute but do not sacrifice statistical performance.

IV. **RELEVANCE AND TRANSITION**

A. **Relevance of your research to the DHS enterprise**

The proposed project will develop systematic techniques algorithms for tagging, tracking and handoff in a multi-camera scenario. The significance of this project to DHS is that we propose methods that will seamlessly track individuals as they move through the airport, across sparse multi-camera networks.

B. **Anticipated end-user technology transfer**

We have been and will continue to work with the Cleveland Airport to transition technology.

V. **PROJECT DOCUMENTATION AND DELIVERABLES**

A. **Student theses or dissertations produced from this project**

   1. Phillip Tran, Spatially Closest Codeword Co-occurrence Descriptors for Person Re-ID, May 2014, MS Thesis.

VI. **REFERENCES**


This page intentionally left blank.