Title

Context-Aware Hypergraph Modeling for Re-identification and Summarization

Journal

IEEE Transactions on Multimedia, 18(1)

ISSN

1520-9210

Authors

Sunderrajan, Santhoshkumar

Manjunath, BS

Publication Date

2016

DOI

10.1109/tmm.2015.2496139

Peer reviewed

Context-Aware Hypergraph Modeling for Re-identification and Summarization

Santhoshkumar Sunderrajan, *Member, IEEE*, and B. S. Manjunath, *Fellow, IEEE*

***Abstract*—Tracking and re-identification in wide-area camera networks is a challenging problem due to non-overlapping visual fields, varying imaging conditions, and appearance changes. We consider the problem of person re-identification and tracking, and propose a novel clothing context-aware color extraction method that is robust to such changes. Annotated samples are used to learn color drift patterns in a non-parametric manner using the random forest distance (RFD) function. The color drift patterns are automatically transferred to associate objects across different views using a unified graph matching framework. A hypergraph representation is used to link related objects for search and re-identification. A *diverse hypergraph ranking* technique is proposed for person-focused network summarization. The proposed algorithm is validated on a wide-area camera network consisting of ten cameras on bike paths. Also, the proposed algorithm is compared with the state of the art person re-identification algorithms on the VIPeR dataset [1].**

***Index Terms*—Camera network, person re-identification, search, summarization.**

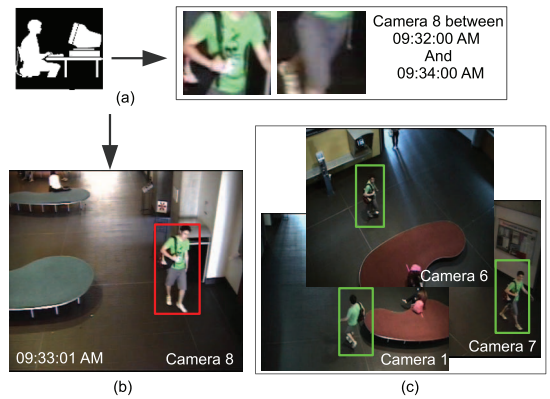

Fig. 1. (a) Initial searching query is specified using clothing colors and time interval. (b) Person of interest is marked by a red bounding box. (c) Summary of the retrieved results.

I. Introduction

DISTRIBUTED camera networks pose several challenges to data analytics, including the large amount of video that needs to be communicated. Given the bandwidth and network constraints, de-centralized approaches taking full use of limited computational power in individual cameras have been gaining prominence [2]–[4]. Person tracking across multiple cameras is one of the preliminary steps in many distributed camera applications. However, due to varying lighting, complex shape and appearance changes, the performance of person tracking is still far from the ideal. To this extent, there have been several works in the past that perform clothing based re-identification to associate a person across multiple cameras to solve the tracking problem [5], [6].

In surveillance videos, it is difficult to parse the clothing information for reliably associating a person across multiple camera views due to unknown transformation in color and overall appearance of the person (people occupy few pixels relative to the entire image frame). Also, frequent exchange of raw image data to the central server imposes network bottlenecks and is not scalable for large networks. Existing approaches tackle the problem either with appearance based features alone or in combination with spatial-temporal information. However, appearance based features are insufficient for matching due to viewpoint, pose and lighting changes. Spatial-temporal information helps to an extent, and it requires complete knowledge of the network. Furthermore, existing approaches assume that global trajectories are available and accurate [7], [8]. With the state-of-the-art person detectors and trackers, this is not a reasonable assumption.

In the context of tracking in a wide-area camera network, we envision that the human analyst would be interested in the following:

1) *searching* [see Fig. 1(a)] people based on clothing colors and time interval;
2) *re-identifying* [see Fig. 1(b)] a specific person of interest across the entire network using appearance and spatial-temporal information; and
3) *summarizing* [see Fig. 1(c)] events using person-focussed non-redundant snapshots across the network.

The above-mentioned tasks arise in several fields such as e-commerce (fashion) [6], [5], surveillance [9], [10]. Towards this extent, we propose a distributed framework that enables users to re-identify person across multiple camera views, and provides a summary of persons's activity. At individual camera nodes, moving people are detected and tracked. Each camera sends an abstracted record consisting of appearance and spatial-temporal information per person to the central server. At

Manuscript received July 21, 2015; accepted October 12, 2015. Date of publication October 29, 2015; date of current version December 14, 2015. This work was supported by the ONR under Grant N00014-12-1-0503 and by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Changsheng Xu.

The authors are with the Department Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: santhosh@ece.ucsb.edu; manj@ece.ucsb.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2496139

the central server, a time evolving graph is built using tracklets. In our analysis, tracklets form the basic unit. Tracklets are represented using LAB space color histogram. We learn multi-view color drift i.e., unknown color based appearance transformation using annotated tracklets in the training set. During testing, a superpixel based person representation is used to associate a person across multiple views. Specifically, we use a graph matching algorithm that explicitly takes color drift patterns during the association. The proposed color drift aware person re-identification algorithm and the spatial-temporal topology information are used to build relationship between tracklets for the hypergraph model. Finally, we rank the nodes according to the query provided by the human operator based on re-identification and summarization criteria. The following are the main contributions of this work.

1) A data driven approach to *learning color drift patterns in a large scale camera network* with annotated training samples. We pose the problem of learning color drift as a distance metric learning and solve efficiently using Random Forest Distance (RFD) function that is inherently non-parametric and robust to imputations.
2) Given the color drift model, we propose a clothing context-aware appearance based association using a unified *graph matching framework that explicitly takes color drift patterns into account for computing node-to-node similarity*. More importantly, edge-to-edge similarity explicitly captures clothing appearance difference/context.
3) A hypergraph based representation for encoding contextual relationship. Hypergraph provides an efficient strategy to encode group relationship between tracklets for re-identification. Final candidate records are obtained by ranking nodes in the hypergraph. For summarization, *we emphasize on diversity within the hypergraph retrieval*. Extensive experimentation on a large scale camera network and person re-identification datasets.

II. Related Works

Person re-identification/recognition is a well studied topic in the computer vision literature. Satta *et al.* [11] provide a detailed review of appearance based descriptors for person re-identification. Sankaranarayanan *et al.* [12] discuss in detail about object tracking and recognition in smart cameras.

A. Smart Camera Networks

Javed *et al.* [13] assign globally unique ID for observations collected from remote cameras and associate objects based on appearance and spatial information. Existing multi-view object re-identification methods work by matching features across different views based on a best matching criterion. [8] learns brightness transfer function on a low-dimensional subspace for matching. In [7], object's appearance and motion are jointly modeled. Arth *et al.* [14] used a PCA-SIFT based vocabulary tree in a camera network setup to re-identify objects. Gheissari *et al.* [15] proposed a triangulated model fitting to people that addressed association with articulation. Park *et al.* [16] proposed a similar system and they assumed fixed body configuration for extracting HSV based color features for matching. Most of the appearance based methods fail due to viewpoint, pose and lighting changes. Prosser *et al.* [17] propose a Cumulative Brightness Transfer Function for mapping color between cameras located at different physical locations. Mazzeo *et al.* [18] compared different methodologies to evaluate the color Brightness Transfer Function for non-overlapping tracking. Ni *et al.* [2] introduced graph based models for object search and retrieval in camera networks. Xu *et al.* [3] implicitly introduced contextual links into a graph model but they do not use appearance based information. Sunderrajan *et al.* [4] explicitly modeled appearance, spatial-temporal and scene contexts into the graph model to improve the accuracy. This paper extends these by modeling color drift in such networks and including such contextual information within a graph ranking procedure for person re-identification tasks.

B. Person Re-identification

Zhao *et al.* [19] find salient regions in image patches and compute pairwise similarity between images. They do not account for varying lighting conditions and it may fail when object's appearance changes due to illumination changes. In [20], authors proposed a saliency matching based structured RankSVM and it might fail when applied on a wide area-camera network with large lighting changes. In [21], authors proposed a symmetry driven local features for pairwise matching. They computed Maximally Stable Color Regions, Weighted color histograms and Recurrent Highly Structured Patches over the symmetry regions for pairwise matching. Gray *et al.* [22] proposed a localized feature selection methodology through an ensemble learner. Kostinger *et al.* [23] proposed a large scale metric learning for recognition using equivalence constraints. Zheng *et al.* [24] proposed a probabilistic relative distance comparison for computing the distance metric for person re-identification such that the probability of getting a right match is greater than getting a wrong match. In [25]–[27], person re-identification is modeled as a ranking problem and focus directly on improving the ranking scores. [28] provides a discriminative re-ranking strategy to improve the person re-identification accuracy. Compared to the above-mentioned algorithms, the proposed methodology provides a unified strategy for computing saliency based graph matching with distance learning using Random Forest Distance function.

C. Clothing Appearance-Aware Recognition

Clothing information provides a strong cue for re-identification when properly segmented. Gallagher *et al.* [5] perform clothing segmentation using a subset of labeled examples and recognize people. Compared to [6], proposed methodology does not require explicit parsing of clothing and we leverage superpixel co-occurrence to match object appearance across views (clothing segmentation is difficult due to uncontrolled environment and partial occlusion in surveillance videos). Most of the existing approaches operate on controlled environments, hence, they cannot be directly applied to surveillance videos. Yang *et al.* [9] proposed clothing recognition in surveillance videos. Recently, Fulkerson *et al.* [29] demonstrated superiors results using superpixel neighborhood based representation for object localization.

D. Diverse Graph Ranking

In addition to the importance and relevance of the retrieved candidates, several works in the past stressed on diversity for summarization [30]–[32]. Grasshopper [30] algorithm is based on random walks with absorbing nodes. DivRank [31] uses a vertex-reinforced random walks to improve the diversity and it does not guarantee relevance. MRSP [32] proposed a manifold ranking algorithm with sink points to emphasize diversity. However, all these methods operate on pairwise relationship based graphs and group relationship is not explicitly taken into account. We propose a diverse ranking with hypergraphs that explicitly encode group relationship into the graph model and emphasize on relevance, importance and diversity in a unified manner.

The rest of the paper is structured as follows: Section III gives a brief overview about the proposed methodology. Section III-A2 explains the manner in which multi-camera color drift pattern is learned and color drift aware appearance matching. Section III-B presents spatial temporal topology modeling. Section IV gives a detailed information about hypergraph based ranking. Section V presents experimental results from a real 10-camera outdoor network and VIPeR datasets. Finally, Section VI concludes this paper.

III. Proposed Methodology

Consider a camera network with M distributed and static cameras. We assume that the entire network is time-synchronized and no calibration information is available. At each camera node, a background subtraction based tracker is used to segment foreground pixels and automatically detect moving objects (pedestrians and bikers) using connected component analysis [33]. Each camera assigns a unique local ID to every person and sends an abstracted record consisting of: camera ID, timestamp, person's bounding box on the image plane and the person's image data (obtained from the frame with the maximum person area) to the central server. *The central server builds a time-evolving hypergraph with a decaying memory using the observations obtained from remote cameras*. Let $w_{ij}^{Appearance}$ and $w_{ij}^{Spatial-Temporal}$ be appearance and spatial-temporal topology based similarity matrices where i and j are tracklet/node indices. In the following we describe appearance based similarity computations by explicitly modeling color drift patterns. Then, spatial-temporal weighting using network topology. Given this setup, we envision that the central server could be queried to infer network events and provide a smart summarization. An example query would be *"Find people wearing green shirt and grey shorts in camera 8 between time 9:32am and 9:34am"*.

We leverage superpixels defined over clothing regions to improve the robustness of appearance based matching. Most importantly, existing approaches do not explicitly model color drift patterns and they are not applicable for multiple views. We propose a unified framework for matching person's appearance between multiple views by explicitly modeling color drifts in a discriminative manner. We model spatial-temporal topology using ground truth associations and build a hypergraph representation for modeling relationships. We pose the problem of person re-identification as hypergraph ranking. Finally, we propose a diverse hypergraph ranking algorithm for summarization. Table I summarizes the notations used in this paper.

TABLE I
Summary of Notations

Symbol	Description
i, j	Person/tracklet Indices.
t	Time, Iteration Index.
M	Number of Static Cameras.
m	Camera Index.
D	Dimensionality of Color Histogram.
d	Color Histogram Feature Index.
$\mathbf{f}, \mathbf{f}'$	Absolute Color Difference Vector.
U	Number of trees in RFD.
u	Tree Index.
l	x-y Position on the Image Plane.
O	Number of Ground Truth Trajectories.
td	Time Delay.
$\mathbf{y}$	Ground Truth Spatial Association.
η, μ, Σ	Weight, Mean and Co-variance of Gaussians.
$\mathbf{z}$	LAB space Color Histogram.
G_i, G_j	Superpixels based Graphs.
S_i	Superpixel set.
n_i, n_j	Number of Superpixel Graph Nodes.
m_i, m_j	Number of Superpixel Graph Edges.
K	Maximum Number of Salient Superpixels.
$k, k', k1, k2, k1', k2', k*$	Superpixel indices.
Q	Superpixel Graph Affinity Matrix.
ϕ	Node-to-node Superpixel Similarity.
ψ	Edge-to-edge Superpixel Similarity.
G	Network Graph.
V, E	Network Graph Node and Edge Sets.
v, e	Network Graph Nodes and Edges.
H	Hypergraph Incidence Matrix.
A	Hypergraph Similarity Matrix.
D_v	Hypergraph Vertex Degree Matrix.
D_e	Hyperedge Degree Matrix.
W	Hyperedge Weight Matrix.
f	Graph Rankings Score Vector.
$\mathbf{r}$	Graph Ranking Label Vector.
γ, β	Graph Ranking Constants.
$r_1, r_2 \ldots$	Ranked Vertices.
q	Query Index.
$\mathcal{M}_q$	Query Node Set.
$\mathcal{X}_r$	Set of Nodes to be Ranked.
$\mathcal{X}_a$	Set of Absorbing Nodes.

A. Appearance Modeling

1) Dense Color Histogram: For every moving person, a multi-dimensional ($D = 288$) LAB color histogram is extracted. A patch size of 10×10 is sampled on a grid with a step size of 4 pixels. We used 32-bins in L, A and B channels respectively. Also, we down-sampled the image into 3-levels with scaling factors of 0.5, 0.75 and 1.0. Finally, we average pool the $L2$-normalized dense color histogram to represent regions. Each person bounding box is divided into three regions (1/8 for head, 3/8 for torso and 4/8 for legs) and dense color histogram for torso and legs are extracted for learning color drift pattern. In addition to dense color histogram, we compute culture color histogram for indexing top and bottom portions of the person [34]. The following subsections explain the manner in which color drift patterns are learned in a non-parametric manner using ground truth associations. Finally, color drifts patterns are automatically transferred for appearance matching.

2) Learning Multi-View Color Drift Patterns: During the training stage, we assume that ground truth associations between tracklets in two different views are available. Let $\mathbf{z}_i^{torso}$ and $\mathbf{z}_j^{torso}$ be dense color histograms computed for torsos of people i and j respectively (similarly, $\mathbf{z}_i^{legs}$ and $\mathbf{z}_j^{legs}$ are computed for legs of people). Euclidean distance provides a simple

Fig. 2. Original (left) versus salient superpixels (right) for camera views cam_a (row 1) and cam_b (row 2) from VIPeR dataset. Salient superpixels are highlighted in red. Best viewed in color.

method for computing distances and it is often insufficient for capturing the underlying data representation. Traditionally, Mahalanobis based distance metric is used for learning a globally optimal metric based on the equivalence constraints [23], [35], [36]. It is limited in its capacity to capture complex data representation and it is computationally inefficient as the dimension of data increases. To overcome this, several multi-metric techniques have been proposed [37], [38] and they are often affected by the memory storage complexity since the matrixes need to be stored per point or the subset of points. In contrast, Random Forest Distance function [39] provides a non-parametric and non-linear distance metric that achieves the efficiency of both global and multi-metric techniques.

Given the set of training pairs with equivalence constraints i.e. $\{(\mathbf{z}_i, \mathbf{z}_j)\}$, we learn a distance metric using the Random Forest Distance function: A set of similar (label 1) and dis-similar (label 0) samples are generated from the training set. Dis-similar samples are generated by randomly pairing color descriptors from different persons. A random forest classifier is trained using the training samples from similar and dis-similar sets and the color drift probability is given by

$$p^{\text{color}-\text{drift}}(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{U}\sum^{U} f_u(\mathbf{z}_i, \mathbf{z}_j) \tag{1}$$

where U is the number of trees (in our experiments we set $U = 800$) and f_u is the classification output of the tree (similar vs dis-similar). The above-mentioned strategy provides a simple yet powerful technique for modeling color drift patterns in surveillance scenarios wherein pixel-wise multi-view correspondence is not possible. Also, the RFD method scales well for increasing dimensionality compared to Kernel Density Estimation and they are inherently non-linear.

3) Salient Superpixel Graph: During the testing stage, we over-segment torso and legs regions of tracklets into several non-overlapping segments using SLIC superpixel algorithm [40]. Let $S_i = \{s_{ik}\}$ be the set of superpixels for the ith tracklet where k is the superpixel index. We represent the superpixels using dense color histograms and choose top K superpixels based on the saliency. We compute pairwise feature distance between superpixels and order the median-th distance in the increasing order and choose the top K superpixels. We define a complete graph $G_i = (V_i, E_i)$ over the salient superpixels where V_i is the set of n_i superpixel nodes such that $n_i \leq K$ and E_i is the set of edges defined between superpixels. Similarly, we define a complete graph $G_j = (V_j, E_j)$ for the jth tracklet in a different view with n_j superpixel nodes ($S_j = \{s_{jk'}\}$ where k' is the superpixel index). Fig. 2 highlights the salient superpixels obtained with sample images on the VIPeR dataset.

4) Clothing Context-Aware Appearance Matching: We propose a novel clothing context-aware appearance based association using a graph matching framework. Let $Q \in \mathcal{R}^{n_i n_j \times n_i n_j}$ be the global affinity matrix that defines node and edge affinities such that

$$Q_{k_1 k_1', k_2 k_2'} = \begin{cases} \phi(k_1, k_1'), & \text{k}_1 = k_2 \text{ and } k_1' = k_2' \\ \psi(e_{k_1,k_2}, e'_{k_1',k_2'}), & k_1 \neq k_2 \text{ and } k_1' \neq k_2' \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\phi(k_1, k_1')$ encodes node-to-node similarity. For computing node-to-node similarity, color drift pattern is taken into account and it is given by

$$\phi(k_1, k_1') = p^{\text{color}-\text{drift}}(\mathbf{z}_{k_1}, \mathbf{z}_{k_1'}) \overbrace{\sum_{d}^{D=288} min(\mathbf{z}_{k_1}(d), \mathbf{z}_{k_1'}(d))}^{\text{Histogram Intersection}} \tag{3}$$

where $\mathbf{z}_{k_1}$ and $\mathbf{z}_{k_1'}$ are LAB space color histograms extracted over superpixel regions. $\psi(e_{k_1k_2}, e'_{k_1'k_2'})$ encodes edge-to-edge similarity and takes clothing context into account. Edge $e_{k_1k_2}$ is represented by a feature difference vector computed using culture color histograms of nodes k_1 and k_2 i.e., $\mathbf{f} = |\mathbf{z}_{k_1} - \mathbf{z}_{k_2}|$. Similarly, edge $e'_{k_1'k_2'}$ is represented by feature difference vector $\mathbf{f}'$. By this, we capture clothing context aware edge representation i.e., edge between torso and leg superpixels captures clothing difference. $\psi(e_{k_1k_2}, e'_{k_1'k_2'})$ is given by

$$\frac{1}{D}\sum_{d}^{D=288} \begin{cases} 1, \text{if } \mathbf{f}(d), \mathbf{f}'(\text{d}) = 0 \\ max\left[I(\mathbf{f}(d) = \mathbf{f}'(d)), \frac{min(\mathbf{f}(d), \mathbf{f}'(d))}{max(\mathbf{f}(d), \mathbf{f}'(d))}\right] \end{cases} \tag{4}$$

where I is the indicator function. Intuitively, ψ equals one when the difference histogram matches exactly. Let x be a

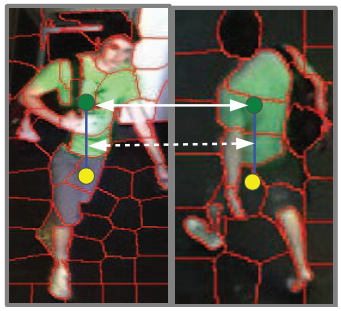

Fig. 3. Clothing context-aware appearance matching. Superpixel nodes in top and bottom clothing are marked by green and yellow dots, respectively. Node-to-node matching between two views is marked by a solid white arrow and it takes color drift into account. Edge-to-edge matching between two views is marked by a dashed white arrow and it takes clothing context into account.

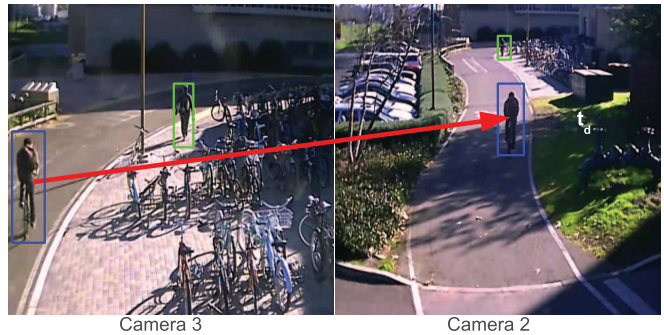

Fig. 4. Spatial-temporal topology modeling. Ground truth association between every pair of views is used to learn mixture of Gaussians distribution for spatial-temporal topology. The person marked by the blue bounding box is associated between camera views 2 and 3. Centroids of the bounding boxes and time delay are used to define spatial-temporal topology model.

binary vector such that $x \in \{0,1\}^{kk'} : x_{kk'} = 1$ iff superpixel nodes k and k' match. For one-to-one matching, we constrain $\sum_k x_{kk'} = 1$ and $\sum_{k'} x_{kk'} = 1$. The appearance based weighting is obtained by solving the graph matching cost

$$\max_x \left(\frac{x^T Q x}{x^T x} \right), \qquad s.t.\ Px = b \tag{5}$$

where the affine constraint $Px = b$ enforces one-to-one matching constraints. Equation (6) could be formulated as a maximization of Rayleigh quotient under affine constraint and computed using spectral methods. Interested readers refer to [41] for more details. Given the pairwise superpixel correspondence based on graph matching, we compute the appearance based weighting using the node-to-node similarity

$$w_{ij}^{\text{Appearance}} = \text{median}(\{\phi(k, k^*)\}_1^{n_i}) \tag{6}$$

where k^* is the matching superpixel index in graph G_j obtained by graph matching. Fig. 3 illustrates clothing context-aware appearance matching.

5) Computational Complexity: The total complexity of graph matching is proportional to the number of non-zero elements in Q. Let m_i and m_j be the number of edges in graphs G_i and G_j respectively. The complexity is given by $O(m_i m_j)$ for full matching. For surveillance videos, people appear relatively small and hence number of superpixels is also small.

B. Spatial-Temporal Topology Modeling

For a fixed camera network, spatial-temporal topological information provides a strong cue for association. However, pedestrians exhibit irregular motion patterns in different parts of the scene. We leverage trajectories of historical motion patterns to model the relationship between different regions in the image plane on a given view with respect to other camera views. At the training stage, we assume that ground truth associations between person trajectories in multiple views are available. For every pair of camera views, we use ground truth trajectories to estimate the spatial-temporal density for modeling motion of the i-th person from a given location, l_{it}^m, on the image plane of m-th camera view to the location, $l_{jt}^{m'}$, on the image plane of m'-th camera view within a time delay td_{it}^m (person i and j are associated using ground truth and t is the time index). At m-th camera view, we build a five dimensional model from a set of training samples $\{\{\mathbf{y}_{it}^{mm'}\}_t\}_i^O$ where $\mathbf{y}_{it}^{mm'} = [l_{it}^m, l_{jt}^{m'}, td_{it}^m]$, and O is the number of ground truth trajectories in the m-th view. As a pre-processing step, we perform whitening transformation on the training data and project it on to a lower dimensional sub-space using PCA (in our experiments, we used a four dimensional sub-space).

A mixture of Gaussians is used to model the spatial-temporal distribution between every pair of camera views. We set the number of Gaussians, C, using a non-parametric distortion method [42]. Expectation maximization algorithm is used for learning the mixture model. Spatial-temporal topology based graph weighting, $w_{ij}^{\text{Spatial-Temporal}}$, is given by

$$w_{ij}^{\text{Spatial-Temporal}} = p^{ST}(\mathbf{y}; \{\eta_c\}_c^C, \{\mu_c\}_c^C, \{\Sigma_c\}_c^C) \tag{7}$$

where c, η, μ and Σ are component index, weight, mean and co-variance respectively. The proposed spatial-temporal topology modeling relies on the training data to predict motion patterns in different parts of the scene. The proposed approach differentiates motion in bike paths when compared to non-bike paths. Also, it differentiates biker's motion with respect to walking pedestrians by learning a separate Gaussian component for the mixture model. During unseen motion pattern, appearance based matching plays an important role. Over the time, spatial-temporal topology is updated with the new information. Fig. 4 illustrates the spatial-temporal topological model for a pair of camera views.

IV. Hypergraph Representation

A pairwise simple graph is insufficient to represent relationship between vertices. In multi-graph based representation, multiple edges are constructed between two vertices. In hypergraph based person re-identification, multiple edges are constructed between three or more vetices. Also, a hypergraph accounts for relationship between three or more vertices containing local grouping information and also models higher order relationship between vertices. Compared to [4], we implicitly introduce group relationship between tracklets using hypergraph representation. A hypergraph contains a set of vertices defined as a weighted hyperedge; the magnitude of hyperedge weight denotes the probability of a vertex belonging to the same cluster.

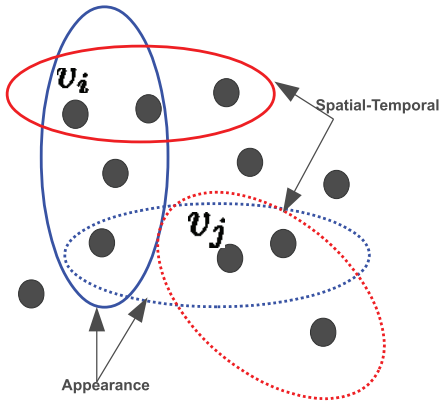

Fig. 5. Hypergraph representation of the network. A set of hyperedges are formed with each node as the centroid. Hyperedges for the nodes v_i and v_j are represented by solid and dashed ellipses, respectively. The blue ellipse represents the appearance-based hyperedge and the red ellipse represents the spatial-temporal hyperedge.

Also, in the proposed approach, a pair of relationship is available between every two nodes i.e., appearance and spatial-temporal. Conventionally, weights are averaged to form a simple graph or multiple edges are formed between two nodes to create a multi-graph. However, in this paper, we propose a novel network wide hypergraph representation $G = (V, E, W)$ such that for every vertex $v \in V$, we create a pair of hyperedges, $E^{App}, E^{ST} \in E$, using k-nearest neighbors based on appearance and spatial-temporal weighting matrices (see Fig. 5) and co-occurring people within the same camera view (in our experiments, we set threshold for co-occurring people to 5 seconds) where W is the diagonal hyperedge weight matrix. A probabilistic hypergraph is represented using a $|V| \times |E|$ incidence matrix H such that

$$H(v_i, e_j) = \begin{cases} A(i,j), & \text{if } v_i \in e_j \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $A(i,j)$ denotes the probability that node v_i belong to edge e_j (computed using the similarity between the centroid node $v_j \in e_j$ and v_i). Let $w(e_j) = \sum_{v_i \in e_j} A(i,j)$ be the hyperedge weight and $d(v) = \sum_{e \in E} w(e)H(v,e)$ be the degree of the node $v \in V$. For a hyperedge $e \in E$, let $\delta(e) = \sum_{v \in e} H(v,e)$ be the edge degree. Given the diagonal matrix of vertex degrees (D_v), hyperedge degrees (D_e) and hyperedge weights matrix (W), in the following subsection we explain the manner in which the user query is used to rank items based on hypergraph ranking framework proposed in [43] that effectively minimizes the cost function given by (9), shown at the bottom of the page, where $f \in \mathcal{R}^{|V^n| \times 1}$ is the ranking score. Intuitively, by minimizing the cost function, vertices sharing many hyperedges obtain similar ranking scores.

A. Hypergraph-Based User Query Ranking

Consider the query "*Find people wearing green top and grey bottom in camera 8 between time t1 and t2*" [see Fig. 1(a)]. In the database, each tracklet is indexed with a metadata consisting of top and bottom clothing colors (computed using dominant component in culture color histogram) and corresponding timestamp. The proposed system will retrieve those tracklets that are closely related to the supplied query. Given the set of matching nodes, $\{\mathcal{M}_q\}$, and $|\mathcal{M}_q|$ be the number of matching nodes, the system assigns uniform matching scores (positive labels) for those nodes in the query set, i.e.

$$\mathbf{r}_i = \begin{cases} 1/|\mathcal{M}_q|, & \text{if } i \in \{\mathcal{M}_q\} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $\mathbf{r} \in \mathcal{R}^{|V| \times 1}$ is the preference/label vector. For similar appearance and spatial-temporal searching [see Fig. 1(b)], user supplies a query similar to find *"Find all people related to the selected person at region C from camera C at time t1"*, which corresponds to ith node in the hypergraph. The preference vector is set as $\mathbf{r}_j = 1$ with all other entries set to 0. Given the preference/label vector, $\mathbf{r}$, Algorithm 1 that minimizes cost function in (9) is used to rank items based on relevance. Finding the ranking score is equivalent of solving the linear equation $((1+\beta)I - \Theta)f = \beta\mathbf{r}$ and $\gamma = \frac{1}{1+\beta}$. β is the regularization parameter that forces the ranking score to approach initial labeling $\mathbf{r}$.

Algorithm: 1 Ranking camera observations using Hypergraph Ranking

Input: Hyperedge weight matrix (W), vertex degrees (D_v), hyperedge degrees (D_e) and preference vector ($\mathbf{r}$).
Output: Top-ranked vertices $\{r_1, r_2, r_3, \ldots\}$.
1: Compute $\Theta = (D_v)^{-\frac{1}{2}} HW(D_e)^{-1}(H)^T(D_v)^{-\frac{1}{2}}$
2: Given the initial label vector, $\mathbf{r}$, compute ranking scores $f = (1-\gamma)(I - \gamma\Theta I_f)^{-1}\mathbf{r}$
3: Rank items according to their ranking scores, f, in descending order.

$$\begin{aligned}
\Omega(f) &= \frac{1}{2}\sum_{e\in E}\sum_{u,v\in V}\frac{w(e)H(u,e)H(v,e)}{\delta(e)}\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2 \\
&= \frac{1}{2}\sum_{e\in E^{App}}\sum_{u,v\in V}\frac{w(e)H(u,e)H(v,e)}{\delta(e)}\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2 \\
&\quad + \frac{1}{2}\sum_{e\in E^{ST}}\sum_{u,v\in V}\frac{w(e)H(u,e)H(v,e)}{\delta(e)}\left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2
\end{aligned} \tag{9}$$

B. Hypergraph Ranking With Diversity for Network Summarization

In addition to finding relevant candidates, a good ranking algorithm needs to avoid redundant candidates. There are several algorithms proposed to improve the diversity of the retrieved candidates within pairwise graph relationships. Diverse ranking algorithm within group relationship (hypergraphs) is still unexplored. We introduce hypergraph ranking with absorbing nodes in this paper. We propose an iterative algorithm that turns ranked (relevant) points into absorbing nodes at every iteration and repeat (Algorithm 2). Algorithm 2 computes the ranking score in an iterative manner, i.e.

$$f^{(t+1)} = \gamma \Theta I_f f^{(t)} + (1-\gamma)\mathbf{r} \tag{11}$$

where I_f is an identity matrix and t is the iteration index. Equation (11) converges to the ranking scores obtained by the analytical solution i.e. $f = (1-\gamma)(I-\gamma\Theta)^{-1}\mathbf{r}$ (see the proof of convergence in Appendix A). Absorbing nodes are points that take a minimum ranking score (zero in our case) during the ranking process. Therefore, these absorbing nodes prevent ranking scores from spreading to the neighbors. This prevents the redundant candidates from attaining higher ranking scores.

Algorithm: 2 Hypergraph Ranking with Absorbing Nodes

Input: Hyperedge weight matrix (W), vertex degrees (D_v), hyperedge degrees (D_e) and preference vector ($\mathbf{r}$).
Output: Top-ranked vertices $\{r_1, r_2, r_3, \ldots\}$.
1: Initialize the absorbing node set $\mathcal{X}_a = \emptyset$
2: Initialize the set of points to be ranked $\mathcal{X}_r$ with the set of nodes in the graph.
3: Compute $\Theta = (D_v)^{-\frac{1}{2}} HW(D_e)^{-1}(H)^T(D_v)^{-\frac{1}{2}}$
4: **while** $|\mathcal{X}_a| < \#NumberCandidatesReturned$ **do**
5: Iterate $f^{(t+1)} = \gamma\Theta I_f f^{(t)} + (1-\gamma)\mathbf{r}$ until convergence where $0 \leq \gamma < 1$ and I_f is diagonal indicator matrix with (i,i)-entry equal to zero if the item belongs to the absorbing set $\mathcal{X}_a$.
6: Rank points $x_r \in \mathcal{X}_r$ according to their ranking scores f^*.
7: Pick top ranked points $\{x_t\} \in \mathcal{X}_r$ and turn them into new absorbing nodes by moving them from $\mathcal{X}_r$ to $\mathcal{X}_a$.
8: **end while**
9: Return nodes in the order in which they were selected into $\mathcal{X}_a$ from $\mathcal{X}_r$.

C. Computation Complexity of Hypergraph Ranking

The direct minimization of the cost function defined in (9) scales as $O(n^3)$. Whereas, the iterative process [(11)] would reduce the cost to $O(n^2)$.

V. Experiments

The proposed methodology is evaluated on a wide area camera network consisting of ten cameras along a university bike path and publicly available VIPeR dataset [1]. VIPeR dataset is captured by two cameras in an outdoor environment containing two images of the same person in two different viewpoints. It contains 632 pedestrian pairs and images are normalized to 128×48 for the experiments. People undergo significant viewpoint, pose and illumination changes. This is one of the difficult person re-identification datasets and it reflects most of the real world challenges.

For the university dataset, Linksys wireless IP cameras (WVC2300) are used to record videos (640x480, approximately 20 frames per second with a variable frame rate) for approximately 30 minutes (09:30:00 AM to 10:00:00AM) during a busy school day. This is a challenging dataset wherein most of the off-the-shelf person detection and tracking algorithms fail due to the following reasons:

1. *illumination variations:* video feeds are captured at different locations wherein lighting changes abruptly within a few seconds;
2. *large object scale variations:* object scale varies vastly within a given period of time. Object size varies from 10's of pixels to a few 100 pixels; and
3. *wireless packet losses:* Captured data is corrupted by wireless packet losses and the quality of video does not remain constant.

A. Comparison Metrics

We use Cummulative Matching Characteristics curve (CMC) [44] to compare person re-identification algorithms in VIPeR dataset. For camera network dataset, we use scope v/s recall to compare graph based ranking algorithms for search and retrieval. Finally, we compare diverse graph ranking algorithms using the mean average precision and recall based on the summarization criterion defined with respect to the given ground truth data.

B. Multi-Camera Person Search and Re-identification

In first set of experiments, we compare the proposed hypergraph based implicit context representation with explicit context representation proposed in [4]. Fig. 6 shows the retrieval results for two different searching queries. Fig. 7 shows the scope v/s recall for the implicit, explicit and combined modeling for the university bike-path dataset. The combined weight matrix is obtained by averaging appearance and spatial-temporal weight matrices. We used hypergraph ranking for the combined modeling. In all our experiments, we set the number of nearest neighbors in hypergraph construction to 5 and the hypergraph parameter $\gamma = 0.1$. As seen in the figure, the proposed methodology outperforms the explicit and combined contextual modeling for the following reasons: a) Hypergraph representation provides a higher order contextual information whereas [4] provides only explicit contextual information through pair-wise relationship. b) Also, the appearance context proposed in [4] suffers from color drift and the proposed methodology explicitly models the color drift using Random Forest Distance function. c) With combined modeling, the importance of appearance and spatial-temporal information is lost during averaging.

C. Improvements due to Clothing Context

We use the VIPeR dataset to test the accuracy of the proposed clothing context-aware person re-identification algorithm. We randomly split the dataset into training and testing sets as described in [21]. Fig. 8 shows the comparison between the

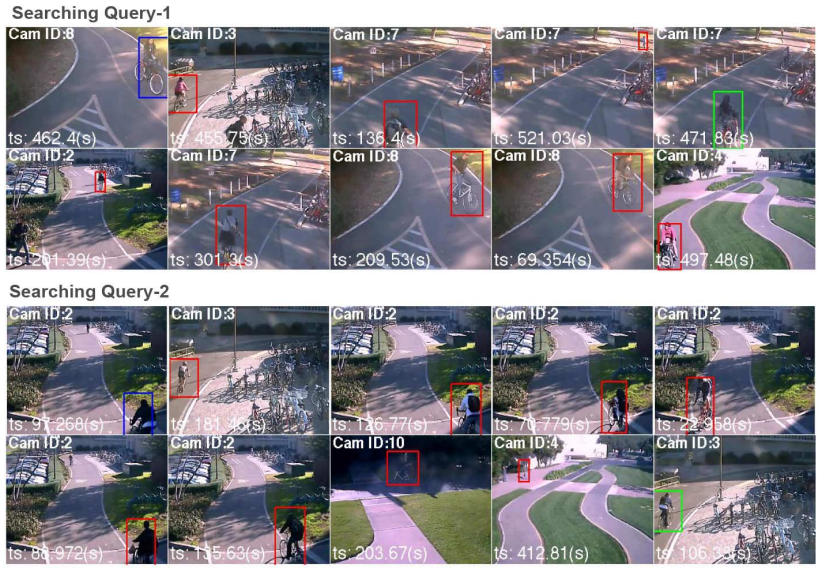

Fig. 6. Searching query results. The query person is by blue bounding box. Search results are ordered column-wise with corresponding camera ID and timestamps (ts). The positive results are highlighted with green bounding boxes and the negative results are highlighted in red bounding boxes.

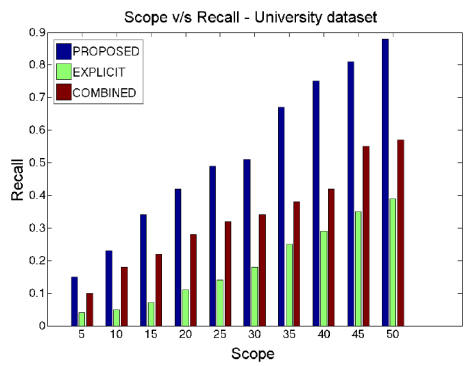

Fig. 7. Implicit versus explicit versus combined modeling. Average scope versus recall for 25 different searching queries with the proposed implicit (hypergraph) versus the explicit [4] versus the combined modeling.

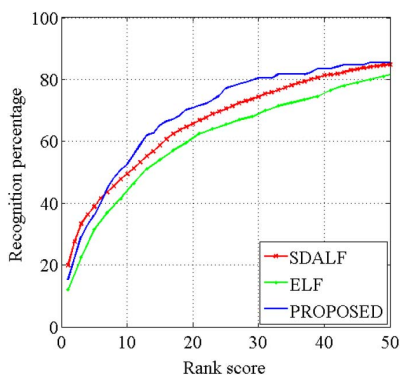

Fig. 8. Improvements due to clothing context. CMC scores in VIPeR dataset.

proposed person re-identification algorithm with respect to SDALF [21] and ELF [22]. Compared to SDALF and ELF, we use only color based features to model the appearance and outperform on most part of the curve. As shown in Fig. 8, the proposed approach outperforms SDALF and ELF in terms of CMC ranking scores due to the following reasons: a) We explicitly model the person saliency using the saliency graph and perform graph matching to find the matching candidate superixels. SDALF computes symmetry regions that might not accurately align during large pose variations. b) Feature variations (clothing context) between superpixel regions is explicitly modeled. Whereas, SDALF does not model feature variations between symmetry regions during the pairwise matching. c) ELF performs feature accumulation on the entire image for pairwise matching, neither feature variations nor saliency is taken into account during the pairwise matching.

D. Improvements due to Learning Color Drift

The proposed color drift computation is related to distance metric learning in the machine learning literature. Towards this extent, Fig. 9 compares the proposed approach with distance metric based person re-identification algorithms such as KISSME [23], ITML [36], Mahalanobis, Identity (Euclidean), LDML [45] and LMNN [35]. We randomly split the VIPeR dataset into training (474 image pairs) and testing (158 image pairs) sets. As shown in the figure, Random Forest Distance function based distance metric-learning outperforms other distance metric learning methodologies due to the following reasons: a) RFD is non-parametric and it does not make assumptions about the underlying data model and hence it provides inherently a non-linear model. b) RFD scales very well for high-dimensional data ($576 - dimension$ in our case) compared to other Mahalanobis distance based methods. c) More importantly, RFD learns both locally and globally optimal distance metric by taking the position of the data point into account. Figs. 10 and 11 show the comparison between the proposed algorithm with and without learning color drift using RFD for the university bike-path and VIPeR datasets respectively. As shown in the figure, learning color drift performs better for the majority of the ranking scores. For the VIPeR dataset, learning color drift marginally improves the overall accuracy since there is no considerable viewpoint

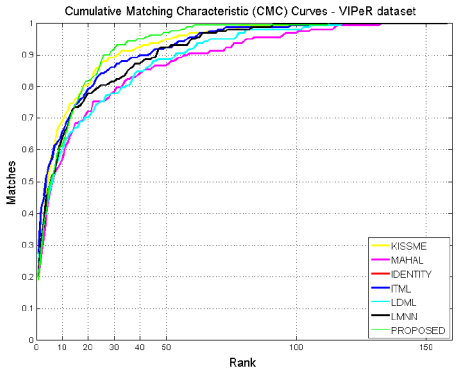

Fig. 9. Comparison with distance metric learning. CMC scores for 158 different subjects in VIPeR dataset with different distance metric learning algorithms.

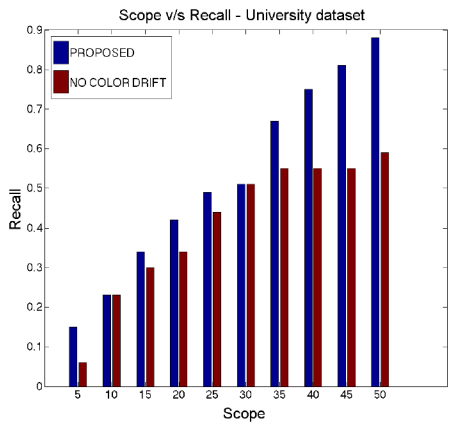

Fig. 10. Improvements due to learning color drift. Average scope versus recall for 25 different searching queries with (proposed) and without learning color drift model.

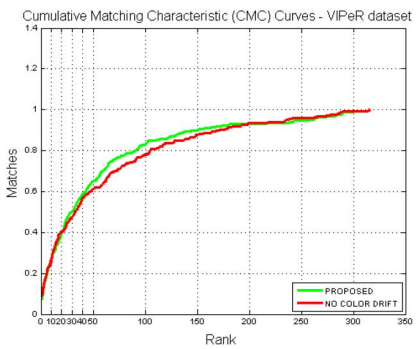

Fig. 11. Improvements due to learning color drift. CMC scores for 316 different subjects in VIPeR dataset with (proposed) and without learning color drift model.

changes compared to the university dataset wherein the color drift patterns drastically improves the accuracy.

E. Comparison With Graph Ranking Algorithms

In this set of experiments, we compare with some of the state of the art graph ranking algorithms such as Manifold Ranking (MR) [46], Similarity Ranking (SR) and PageRanking (PR) [47] with an emphasis on relevance. Fig. 12 shows comparison between various graph ranking algorithms. For pair-wise graph ranking algorithms, the overall weight matrix is obtained by averaging appearance and spatial weight matrices. As seen, the

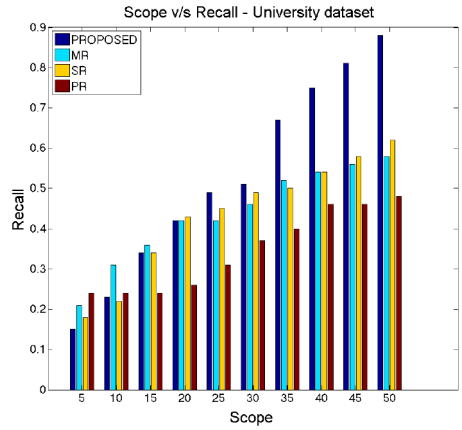

Fig. 12. Comparison of graph ranking algorithms. Average scope versus recall for 25 different searching queries with different graph ranking algorithms.

TABLE II
MEAN AVERAGE PRECISION/RECALL/F-MEASURE FOR 20 DIFFERENT QUERIES

	Mean Average Precision	Mean Recall	F-measure
HRAN	**0.34**	**0.42**	**0.37**
MRSP	0.057	0.074	0.0648
GH	0.045	0.059	0.512

proposed hypergraph based modeling outperforms other graph ranking algorithms in the university dataset due to the local grouping information in hypergraph that is missing in pair-wise matching based graph ranking algorithms.

F. Emphasis on Diversity for Network Summarization

In this set of experiments, we compare the proposed Hypergraph ranking with absorbing (HRAN) with different graph-based ranking algorithms such as Grasshopper (GH) [30], Manifold ranking with sink points (MRSP) [32] that emphasize on diversity. We formulated our evaluation of the summarization results based on the following insights that mainly focus on re-identification, topology and common activity between people:

1. How many camera views that the target passed?
2. How many co-occurring people that traveled with the target?
3. How many people traveled through the marked location within a temporal window?
4. What is the estimated time-delay between two camera views?

Given the ground truth of the target and its co-occurring people across multiple views, we use mean average Precision and Recall within a scope of 10 to evaluate different queries. By limiting the scope to 10, we are able to verify the "diversity" and "importance" of retrieved snapshots with respect to summarization task. Table II shows the mean average precision and recall with corresponding F-measures. As seen in Fig. 13, with the retrieved items for HRAN, a network analyst would be able to answer questions posed with respect to the summarization criterion compared to items retrieved by MRSP and GH. Also, HRAN outperforms MRSP and GH due to the following reason: a) Co-occurring people are implicitly

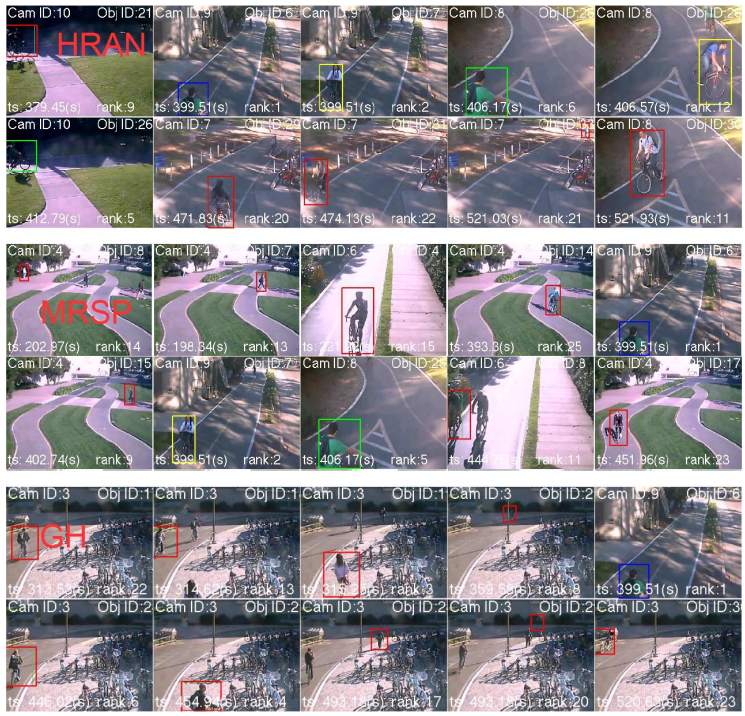

Fig. 13. Person-focused summarization: comparison of different diversity emphasized graph ranking algorithms. The query person is marked by a blue bounding box. The re-occurrence of the query person in the other camera views is marked by green bounding boxes and its corresponding co-occurring people are marked by yellow bounding boxes. Retrieved results are time-ordered.

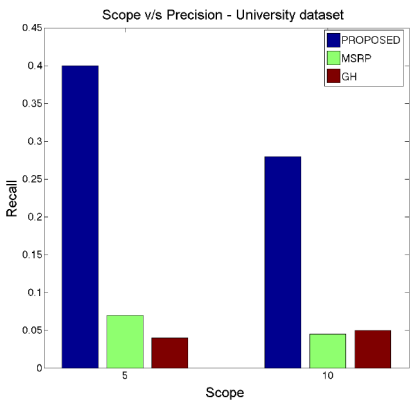

Fig. 14. Precision versus scope. Comparison of different graph-based summarization techniques.

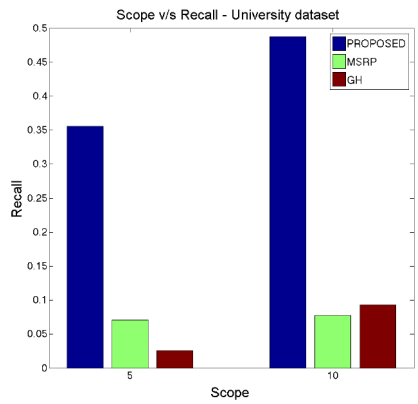

Fig. 15. Recall versus scope. Comparison of different graph-based summarization techniques.

modeled within hyperedges and by turning highly scored items into absorbing nodes, co-occurring people in other views are emphasized more. b) Whereas in MRSP and GH, co-occurring people in the query view are introduced through explicit links within the simple graph and hence co-occurring people in other views are not properly emphasized during the diversification procedure.Figs. 14 and 15 shows average precision and recall for 20 different queries within a scope of 10.

G. *Effect of Parameters*

In this set of experiments, we test the efficacy of various parameters used in the proposed algorithm on university bike-path dataset. Fig. 16 compares different components in the proposed algorithm for 25 different queries in University bike-path dataset. As seen in the figure, without clothing context and color-drift based appearance matching, average recall drops the most since the clothing-context with color-drift model helps in varying lighting conditions. Without spatial-temporal context,

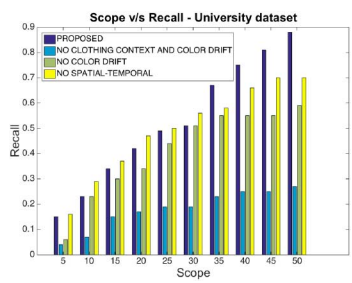

Fig. 16. Effect of various components in the proposed algorithm for the university bikepath dataset.

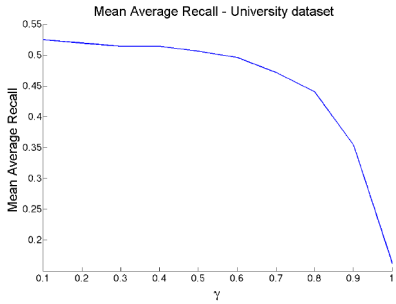

Fig. 17. Effect of hypergraph ranking constant (γ). Mean average recall for different values of γ.

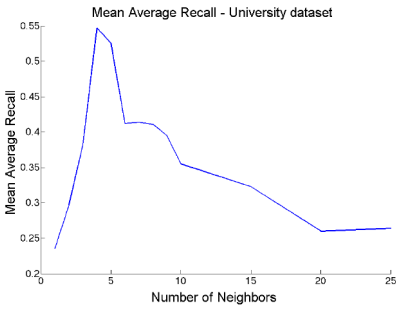

Fig. 18. Effect of number of neighbors in hyperedge. Mean average recall for different number of neighbors to construct hyperedges.

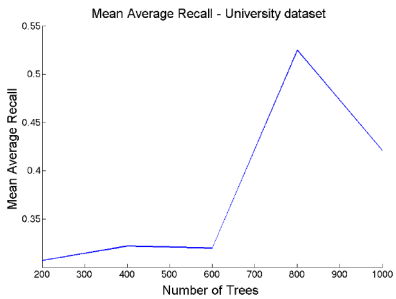

Fig. 19. Effect of number of trees in RFD function (U). Mean average recall for different number of trees in random forest distance for learning color drift.

the average recall drops marginally since the spatial-temporal context is more meaningful for the cameras with overlapping field of views and approximately similar direction of sensing. Fig. 17 shows the overall mean average recall (mean recall for 25 different queries averaged over 10 different scopes i.e. [5], [10], [15], [20], [25], [30], [35], [40], [45]) for different values of γ. Similar to [43], $\gamma = 0.1$ performs the best in the university dataset. Fig. 18 shows the overall mean average recall for different number of neighbors used to create the hyperedge. As the neighbors increase more than 5, the mean average recall drops. This could be due to the following: a) Due to the irregular and low lighting conditions, many people appear similar in the given metric space and hence accuracy drops down when disparate people combine to form an hyperedge. b) At any point of time, there are not more than 3 - 5 co-occurring people in a given scene. Fig. 19 shows the overall mean average recall for different number of trees in RFD function for learning color drift. As seen, the overall accuracy attains the maximum when $U = 800$. With the training pairs of dimension 576

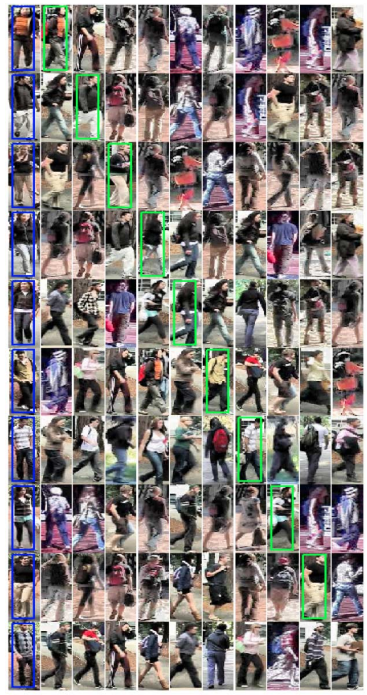

Fig. 20. Person re-identification. Visual results for person re-identification with 11 different queries. Query people are highlighted with blue bounding boxes and the positive results are highlighted with green bounding boxes. Results are column-wise rank ordered.

(288-dimensional LAB color histogram), $U < 576$ under-fit the color drift distribution and $U > 800$ over-fit the color drift distribution. Fig. 20 shows re-identification results obtained in ViPeR dataset.

H. Graph Size and Efficiency

At the central server, we build a time-evolving graph using the tracklets received from different camera views and it grows over the time. In our network analysis, the graph consisted of 288 nodes with 576 hyperedges. For memory optimization, we delete nodes that were added before a certain period of time (say 2 hours) and re-organize the graph. Also, we can neglect the connectivity between the nodes from distant camera views. The complete system (Appearance Modeling + Spatial-Temporal Modeling + Hypergraph Representation) took 0.70 seconds per query on average (based on 288 queries in the University Bike Path Dataset) to run on MATLAB with 2.6 GHz and 16 GB RAM. For this evaluation, we assume that hypergraph representation is already computed.

VI. Conclusion

We proposed a novel color drift and clothing context-aware person search and re-identification method for a wide-area camera network with an emphasis on summarization. The proposed methodology is validated with extensive experiments on some real-world large-scale camera network dataset with 10 cameras along the university bike path. We showed significant improvements over state-of-the-art distance metric based appearance matching and graph ranking techniques. Furthermore, the proposed clothing context-aware person re-identification algorithm is compared with the state-of-the-art person re-identification algorithms in VIPeR dataset.

APPENDIX
CONVERGENCE OF ITERATIVE HYPERGRAPH RANKING

Using the iterative ranking score given by

$$f^{(t+1)} = \gamma \Theta I_f f^{(t)} + (1-\gamma)\mathbf{r}. \tag{12}$$

We have

$$f^{(t)} = (\gamma \Theta I_f)^{(t-1)} f^{(0)} + (1-\gamma) \sum_{i=0}^{t-1} (\gamma \Theta I_f)^i \mathbf{r}. \tag{13}$$

Let $\hat{P} = (D_v)^{-1} H W (D_e)^{-1} (H)^T I_f$ be the similarity transformation of ΘI_f such that

$$\begin{aligned} \Theta I_f &= (D_v)^{-\frac{1}{2}} H W (D_e)^{-1} (H)^T (D_v)^{-\frac{1}{2}} I_f \\ &= (D_v)^{\frac{1}{2}} (D_v)^{-1} H W (D_e)^{-1} (H)^T I_f (D_v)^{-\frac{1}{2}} \\ &= (D_v)^{\frac{1}{2}} \hat{P} (D_v)^{-\frac{1}{2}}. \end{aligned} \tag{14}$$

Therefore, $\hat{P}$ and ΘI_f have the same eigenvalues. By Gershgorin circle theorem, we then have

$$|\rho| \leq \sum_{j \neq i} |\hat{P}_{ii}| \leq 1 \tag{15}$$

where ρ is the largest eigenvalue of $\hat{P}$. Hence, eigenvalues of ΘI_f is not more than one. Since $0 \leq \gamma \leq 1$ and $|\rho| \leq 1$, we have

$$\lim_{t \to \infty} (\gamma \Theta I_f)^t = 0 \tag{16}$$

and

$$\lim_{t \to \infty} \sum_{i=0}^{t-1} (\gamma \Theta I_f)^i = (I - \gamma \Theta I_f)^{-1}. \tag{17}$$

By substituting equations (16) and (17) in equation (13), we have

$$f^* = (1-\gamma)(I - \gamma \Theta I_f)^{-1} \mathbf{r}. \tag{18}$$

ACKNOWLEDGMENT

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation.

REFERENCES

[1] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Performance Eval. Tracking Surveillance*, Sep. 2007.

[2] Z. Ni, J. Xu, and B. Manjunath, "Object browsing and searching in a camera network using graph models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2012, pp. 7–14.

[3] J. Xu, V. Jagadeesh, Z. Ni, S. Sunderrajan, and B. Manjunath, "Graph-based topic-focused retrieval in a distributed camera network," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2046–2057, Dec. 2013.

[4] S. Sunderrajan, J. Xu, and B. Manjunath, "Context-aware graph modeling for object search and retrieval in a wide area camera network," in *Proc 7th Int. Conf. IEEE Distrib. Smart Cameras*, Oct.–Nov. 2013, pp. 1–7.

[5] A. C. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *Proc. IEEE Conf. IEEE Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[6] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2012*, 2012, pp. 3570–3577.

[7] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking across multiple cameras with disjoint views," in *Int. Conf. Computer Vision*, 2003, pp. 952–960.

[8] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 26–33.

[9] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *18th IEEE Int. Conf. Image Processing (ICIP), 2011*, 2011, pp. 2937–2940.

[10] N. S. Thakoor, L. An, B. Bhanu, S. Sunderrajan, and B. Manjunath, "People tracking in camera networks: Three open questions," *Computer*, vol. 48, no. 3, pp. 78–86, Mar. 2015.

[11] R. Satta, "Appearance descriptors for person re-identification: A comprehensive review," *CoRR*, 2013 [Online]. Available: http://arxiv.org/abs/1307.5748

[12] A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa, "Object detection, tracking and recognition for multiple smart cameras," *Proc. IEEE*, vol. 96, no. 10, pp. 1606–1624, Oct. 2008.

[13] O. Javed, Z. Rasheed, O. Alatas, and M. Shah, "KNIGHT: A real time surveillance system for multiple overlapping and non-overlapping cameras," in *Proc. IEEE Conf. Multimedia Expo*, Jul. 2003, vol. 1, pp. I-649–I-652.

[14] C. Arth, C. Leistner, and H. Bischof, "Object reacquisition and tracking in large-scale smart camera networks," in *Proc. 1st ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Sep. 2007, pp. 156–163.

[15] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 1528–1535.

[16] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita, "ViSE: Visual search engine using multiple networked cameras," in *Proc. 18th Int. Conf. Pattern Recog.*, Aug. 2006, vol. 3, pp. 1204–1207.

[17] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *Proc. BMVC*, 2008, vol. 8, pp. 164–1.

[18] P. L. Mazzeo, P. Spagnolo, and T. D'Orazio, "Object tracking by non-overlapping distributed camera network," in *Advanced Concepts for Intelligent Vision Systems*. Berlin, Germany: Springer-Verlag, 2009, pp. 516–527.

[19] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3586–3593.

[20] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2528–2535.

[21] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 2360 –2367.

[22] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008, pp. 262–275.

[23] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2288–2295.

[24] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 649–656.

[25] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *Proc. BMVC*, 2010, vol. 2, no. 5, p. 6.

[26] C. Liu, C. C. Loy, S. Gong, and G. Wang, "Pop: Person re-identification post-rank optimisation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 441–448.

[27] C.-H. Kuo, S. Khamis, and V. Shet, "Person re-identification using semantic color names and rankboost," in *Proc. WACV*, 2013, pp. 281–287.

[28] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*, ser. Lecture Notes in Comput. Sci.. Berlin, Germany: Springer-Verlag, 2011, pp. 91–102.

[29] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep.–Oct. 2009, pp. 670–677.
[30] X. Zhu, A. B. Goldberg, J. Van Gael, and D. Andrzejewski, "Improving diversity in ranking using absorbing random walks," in *Proc. HLT-NAACL*, 2007, pp. 97–104.
[31] Q. Mei, J. Guo, and D. Radev, "Divrank: The interplay of prestige and diversity in information networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 1009–1018.
[32] P. Du, J. Guo, J. Zhang, and X. Cheng, "Manifold ranking with sink points for update summarization," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 1757–1760.
[33] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 2–10.
[34] G. Wu *et al.*, "Identifying color in motion in video sensors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 1, pp. 561–569.
[35] J. Blitzer, K. Q. Weinberger, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1473–1480.
[36] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 209–216.
[37] A. Frome, Y. Singer, and J. Malik, "Image retrieval and classification using local distance functions," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2007, vol. 19, p. 417.
[38] L. Wu, R. Jin, S. C. Hoi, J. Zhu, and N. Yu, "Learning Bregman distance functions and its application for semi-supervised clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2089–2097.
[39] C. Xiong, D. Johnson, R. Xu, and J. J. Corso, "Random forests for metric learning with implicit pairwise position dependence," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 958–966.
[40] R. Achanta *et al.*, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[41] T. Cour, P. Srinivasan, and J. Shi, "Balanced graph matching," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, vol. 19, pp. 313–320.
[42] C. A. Sugar and G. M. James, "Finding the number of clusters in a dataset," *J. Amer. Statist. Assoc.*, vol. 98, no. 463, pp. 750–763, 2003.
[43] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3376–3383.
[44] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[45] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. Int. Conf. Comput. Vis.*, Sep. 2009, pp. 498–505.
[46] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang, "Manifold-ranking based image retrieval," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 9–16.
[47] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford Univ. InfoLab, Stanford, CA, USA, Tech. Rep. SIDL-WP-1999-0120, 1999.

Santhoshkumar Sunderrajan (S'13–M'14) received the B.E. degree from Anna University, Chennai, India, in 2007, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2009 and 2014, respectively.

He is currently a Senior Member of Technical Staff with Pinger, Inc., San Jose, CA, USA. His research interests include computer vision, large scale machine learning, multiple camera tracking, object re-identification, and activity analysis.

Mr. Sunderrajan was the recipient of the Excellent Paper Award at the ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC) in 2013.

B. S. Manjunath (S'88–M'91–SM'01–F'05) received the B.E. degree (with distinction) in electronics from Bangalore University, Bangalore, India, in 1985, the M.E. degree (with distinction) in systems science and automation from the Indian Institute of Science, Bangalore, India, in 1987, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1991.

He is currently a Professor of Electrical and Computer Engineering and Director of the Center for Bio-Image Informatics with the University of California at Santa Barbara, Santa Barbara, CA, USA. He has authored or coauthored over 250 peer-reviewed articles and is a coeditor of the book *Introduction to MPEG-7* (Wiley, 2002). His current research interests include image processing, data hiding, multimedia databases, and bio-image informatics.

Prof. Manjunath was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON, INFORMATION FORENSICS, and the IEEE SIGNAL PROCESSING LETTERS, and is currently an Associate Editor for the *BMC Bioinformatics Journal*. He is a coauthor of the paper that was the recipient of the 2013 IEEE TRANSACTIONS ON MULTIMEDIA Best Paper Award.