Multitarget Tracking in Nonoverlapping Cameras Using a Reference Set

Xiaojing Chen, *Student Member, IEEE*, Le An, and Bir Bhanu, *Fellow, IEEE*

***Abstract*— Tracking multiple targets in nonoverlapping cameras are challenging since the observations of the same targets are often separated by time and space. There might be significant appearance change of a target across camera views caused by variations in illumination conditions, poses, and camera imaging characteristics. Consequently, the same target may appear very different in two cameras. Therefore, associating tracks in different camera views directly based on their appearance similarity is difficult and prone to error. In most previous methods, the appearance similarity is computed either using color histograms or based on pretrained brightness transfer function that maps color between cameras. In this paper, a novel reference set based appearance model is proposed to improve multitarget tracking in a network of nonoverlapping cameras. Contrary to previous work, a reference set is constructed for a pair of cameras, containing subjects appearing in both camera views. For track association, instead of directly comparing the appearance of two targets in different camera views, they are compared indirectly via the reference set. Besides global color histograms, texture and shape features are extracted at different locations of a target, and AdaBoost is used to learn the discriminative power of each feature. The effectiveness of the proposed method over the state of the art on two challenging real-world multicamera video data sets is demonstrated by thorough experiments.**

***Index Terms*— Multi-target tracking, reference set, surveillance.**

I. Introduction

AS THE demand for surveillance cameras at public areas (e.g., airports, parking lots, and shopping malls) is rapidly growing, a major effort has been underway in the vision community to develop effective and automated surveillance and monitoring systems [1]–[5]. In most cases, it is not feasible to use a single camera to cover a complete area of interest, and using multiple cameras with overlapping field-of-views (FOVs) has high cost in both economical and computational aspects. Therefore, camera networks with non-overlapping FOVs are preferred and widely adopted in real world applications.

Fig. 1. Sample frames from each camera view of the MultiCam dataset. Bounding boxes with the same color indicate the same target. Note that illumination may change drastically within camera and across cameras. As a result, the appearance of the same target may vary significantly.

Multi-target tracking is an extensively exploited topic in the surveillance domain, as it is the foundation for many higher level applications, such as anomaly detection, activity detection and recognition [6], and human behavior understanding [7]. The goal of multi-target tracking is to estimate the trajectories of all moving targets and keep their identities consistent from frame to frame. In single camera tracking, successive observations of the same target often have a large proximity in appearance, space and time [8], [9]. However, it is not the case for tracking people across cameras with non-overlapping FOVs. The appearance of the same target may have a large difference even in two adjacent cameras due to a sudden change in illumination (e.g., from outdoor to indoor). Other aspects, such as variations in pose (e.g., frontal view to rear view) and camera imaging conditions (e.g., low resolution and noise) further complicate the tracking task in multiple cameras. In Fig. 1 some sample frames are shown in which the appearance of the same target in different camera views differs significantly.

A possible way to tackle the appearance difference in multiple cameras is to learn a Brightness Transfer Function (BTF) [10]–[15] that is a mapping of color models

Manuscript received August 21, 2014; revised November 1, 2014 and January 3, 2015; accepted January 6, 2015. Date of publication January 16, 2015; date of current version March 25, 2015. This work was supported in part by the National Science Foundation under Grant 0905671 and Grant 1330110, and in part by the Office of Naval Research under Grant N00014-09-C-0388 and Grant N00014-12-1-1026. The associate editor coordinating the review of this paper and approving it for publication was Dr. Brian C. Lovell. *(Corresponding author: Le An.)*

X. Chen is with the Department of Computer Science, University of California at Riverside, Riverside, CA 92521 USA (e-mail: xchen010@ucr.edu).

L. An is with the Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA (e-mail: lan004@unc.edu).

B. Bhanu is with the Center for Research in Intelligent Systems, University of California at Riverside, Riverside, CA 92521 USA (e-mail: bhanu@cris.ucr.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/JSEN.2015.2392781

between a pair of cameras. However, BTF is not suitable for a camera network that has a large *within* camera illumination change. For example, suppose camera i and camera j both have dark and bright regions in their camera views. A BTF that is able to map colors in dark region of camera i (low brightness) to colors in bright region of camera j (high brightness) will not work well for mapping colors in bright region of camera i (high brightness) to dark region of camera j (low brightness).

To address this problem, we propose a novel reference set based appearance model to estimate the similarity of multiple targets in different cameras. Given the intra-camera tracking results of all involved cameras, the goal of multi-target tracking across cameras is to associate tracks in different cameras that contain the same person. Our method is inspired by the recent advances in face verification/recognition [16]–[18] and person re-identification [19] in which an external reference set or a library is used to facilitate the matching process of the same objects imaged under different conditions. The reference set contains the appearance of individuals in different camera views under different imaging conditions. Namely, there are multiple appearance instances for each individual in the reference set. During tracking, instead of comparing the appearance of two targets directly, targets from different cameras are compared to the individuals in the reference set. The individuals in the reference set act like basis functions and for a given target, its similarity to each of the individuals in the reference set are used as its new feature representation instead of the original low level color or texture features.

In order to create a comprehensive representation for each target, besides color features, we also extract shape and texture features from different locations on the body of a target. The discriminative power for each feature is learned using the reference set, and features with high discriminative power contribute more to the similarity score.

The rest of this paper is organized as follows: an overview of the related work and contributions of this paper are provided in Section II. Section III describes the proposed reference set based appearance model for multi-target tracking across non-overlapping cameras. Experimental results are shown in Section IV. Finally, Section V concludes this paper.

II. Related Work and Contributions

A. Related Work

In general, methods for tracking multiple targets in multiple cameras can be categorized into two groups according to the structure of camera networks: methods for overlapping FOVs and methods for non-overlapping FOVs. Techniques used for tracking in these two groups differ significantly. For instance, tracking in cameras with overlapping FOVs normally require explicit camera calibration [20]–[23] while it is not a necessity for tracking with non-overlapping FOVs. As this paper focuses on inter-camera tracking with non-overlapping FOVs, related work for tracking in overlapping camera views is not discussed.

To cope with the illumination change in different camera views, BTF has been studied extensively [10]–[15]. An incremental unsupervised learning method is proposed in [10] to model color variations and posterior probability distributions of spatial-temporal links between cameras in parallel. The model becomes more accurate over time with accumulated evidence. In [11] a cumulative BTF is proposed to map color between different cameras and significant improvement over other BTF-based methods is reported. In [12] the inter-camera relationships is learned using multivariate probability density of space-time variables. It is shown that BTFs from one camera to another camera lie in a low dimensional subspace and this subspace is learned for appearance matching. In [13], BTFs are obtained from the overlapped area during tracking to compensate for the color difference between camera views. In addition, the perspective difference is compensated with tangent transfer functions (TTFs) by computing the homography between two cameras. In [14] different methods are compared to evaluate the color BTFs between non-overlapping cameras and experimental results show BTFs have limitations in people association when a new person enters in camera's FOV. In [15], to track people across non-overlapping cameras, a camera link model including BTF, transition time distribution, region mapping matrix/weight, and feature fusion weight is estimated in an unsupervised manner.

In [24] a combined maximum a posteriori (MAP) formulation is proposed to jointly model multi-camera reconstruction and global temporal data association, in which a flow graph is constructed to track objects. In [25] information from a crowd simulation is integrated into a multi-camera multi-target tracking framework to improve the tracking accuracy. In [26] a data association approach based on principal axis and a joint probabilistic model are applied for multi-object tracking in multi-cameras to overcome occlusion in camera views. In [27] a metric based on three performance indices is developed to evaluate the performance of multi-camera tracking algorithm based on Rao-Blackwellized Monte Carlo data association (RBMCDA). In [28] a track-before-detect particle filter (TBD-PF) is used to increase track consistency against noisy data for multi-camera multi-target fusion and tracking. In [29] a modified Social Force Model (SFM) with a goal-driven approach for multi-camera tracking is proposed. This work takes into account key regions as potential intersections where people can change the direction of motion. In [30] inter-camera transfer models containing spatio-temporal cues and appearance cues are proposed, which are learned by a topology recovering method and a color characteristic transfer (CCT) method for tracking across non-overlapping cameras.

Recently, the reference-based idea has been used in different fields of computer vision, for example, face verification [16], face recognition [17], and person re-identification [19]. The reference-based framework is data-driven in which different entities to be matched are first described using the samples in the reference set and then reference-based descriptors are generated. Therefore, a direct comparison of objects (e.g., faces at different poses) is avoided. In [16], pose, illumination, and expression invariant face verification is achieved by using a library of faces in various appearances to describe a given face based on the insight that it is most meaningful to compare faces with the same imaging conditions. In [17] an

"Associate-Predict" model is proposed which is built on a generic identity data set that contains multiple images with large intra-person variation. Given a face, it is first associated to alike identities in the data set and then its appearance under settings of another input face is predicted. In this way the intra-personal variation is handled. Recently, to improve person re-identification in different camera views, a reference set is used to generate reference-based descriptors for probe and gallery subjects, bypassing the need to direct compare the features from subjects with significant appearance change [19].

B. Contributions of This Paper

The contributions and novelty of this paper are:

- A reference set based appearance model is proposed to mitigate track association ambiguities caused by cross camera illumination and pose variations.
- Each track is divided into several subtracks based on time constraint and appearance similarity to provide multiple appearance instances of a target.
- Various appearance features are extracted from different locations of a target, and their discriminative powers are learned and used to build a robust appearance model.
- Two real-world surveillance datasets are used for evaluation and extensive experiments are carried out to validate the effectiveness of the proposed method.

A preliminary version of this work appeared in [31]. In this paper, we have the following major changes and improvements as compared to [31]. (1) We extend the related work section and more recent advances are discussed. (2) We use an improved track division technique to provide multiple concise appearance representations of a target. (3) We incorporate a learning scheme to assign a weight to each extracted appearance feature so as to achieve a robust appearance model. (4) We conduct in-depth experiments on more data, provide a comparison to state-of-the-art method, and include more result analysis for the proposed method.

III. Technical Approach

A. Formulation of the Multi-Camera Tracking Problem

Suppose we have m cameras $C_1, C_2, \ldots, C_m$ with non-overlapping FOVs. Given the tracking results in each single camera, we can generate a set $T = \{T_1, \ldots, T_N\}$ that contains all within-camera tracks. A track T_i is a consecutive sequence of detections that contain the same target, in a time interval $[t_i^{begin}, t_i^{end}]$, and its corresponding camera is denoted as $C(T_i)$. The problem of tracking across cameras is to find out tracks that contain the same target, given certain spatio-temporal constraints. Let association a_{ij} define the hypothesis that track T_i and T_j contain the same target, with T_i occurring before T_j and $C(T_i) \neq C(T_j)$ (associating tracks that contain the same target in the same camera is not considered in this paper). A valid association matrix A is defined as follows:

$$A = \{a_{ij}\}, \quad a_{ij} = \begin{cases} 1 & \text{if } T_i \text{ is associated to } T_j \\ 0 & \text{otherwise} \end{cases}$$
$$\text{s.t.} \sum_i a_{ij} = 1 \text{ and } \sum_j a_{ij} = 1 \tag{1}$$

The constraints for matrix A indicate that each track should be associated to and associated by only one other track.

The cost S_{ij} for linking track T_i and T_j is based on time, appearance, and camera topology constraints, as defined below:

$$S_{ij} = Time(T_i, T_j) + Topo(T_i, T_j) + Appr(T_i, T_j) \tag{2}$$

where $Time(\cdot)$, $Topo(\cdot)$, and $Appr(\cdot)$ are the time, topology, and appearance models, respectively. The time model is defined as:

$$Time(T_i, T_j) = \begin{cases} 0 & \text{if } 0 < Gap_{ij} < GAP \\ \infty & \text{otherwise} \end{cases} \tag{3}$$

where Gap_{ij} is the time difference between T_i and T_j, and only when Gap_{ij} is smaller than the pre-defined maximum allowed gap GAP the two tracks can be linked. The topology model is similar to the time model, which gives the restriction that T_i can be associated with T_j only when there is a path allowing people to walk between camera $C(T_i)$ and $C(T_j)$ without entering the view of any other cameras.

Let Ω be the set of all possible association matrices, the task of multi-target tracking in non-overlapping camera views is formulated as the following optimization problem:

$$A^* = \arg\min_{A \in \Omega} \sum_{ij} a_{ij} S_{ij} \tag{4}$$

This assignment problem can be solved by Hungarian algorithm [32] in polynomial time. In order to reduce the computational cost, a pre-defined time sliding window is used, and the association is carried out independently in each time sliding window. Instead of using the cost matrix S directly, we use the augmented matrix S' (details for the augmented matrix can be found in [8]) as the input for the Hungarian algorithm. This augmented matrix enables us to set a threshold for association, a pair of tracks can only be associated when their cost is lower than the threshold. In the following, we present the reference set based appearance model in detail.

B. Reference Set Based Appearance Model for Across Camera Tracks

The basic idea of reference set based appearance model is illustrated in Fig. 2. A reference set $RefSet_{ij}$ is constructed for a pair of cameras C_i and C_j. It contains a set of reference subjects $R = \{R_1, R_2, ..., R_n\}$ that appear in both C_i and C_j. The tracks for all the reference subjects that appear in C_i form $RefSet_{ij}^i$, and the tracks for all the reference subjects that appear in C_j form $RefSet_{ij}^j$, as shown in Fig. 2. Given two tracks T_p and T_q with T_p captured in the view of camera C_i and T_q captured in the view of camera C_j, the appearance similarity between these two tracks is not computed by comparing T_p and T_q directly. Instead, T_p is compared with all the tracks in $RefSet_{ij}^i$ and T_q is compared with all the tracks in $RefSet_{ij}^j$, and their similarities with the reference set are used to calculate the similarity of T_p and T_q. In other words, track T_p and T_q are compared with other tracks that undergo the same illumination conditions as T_p and T_q, and if they are the tracks of the same target, they should

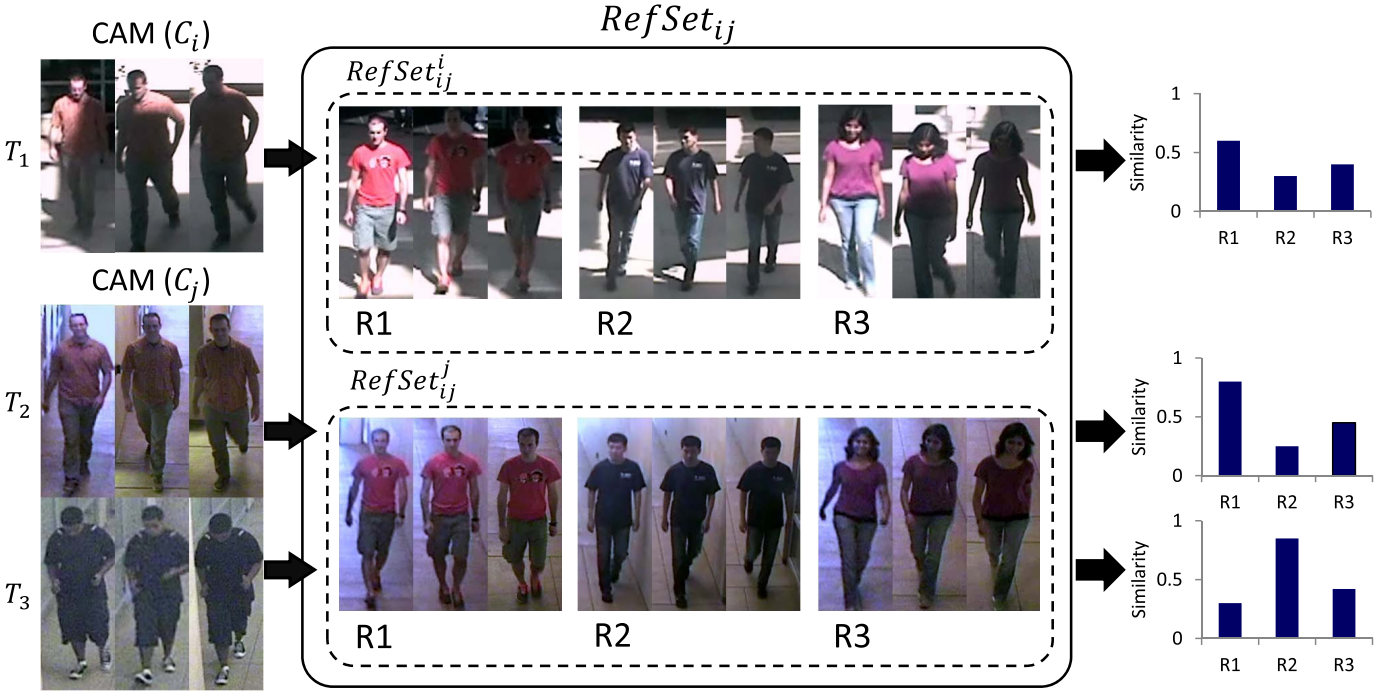


Fig. 2. Illustration of the reference set based appearance model. For a pair of cameras C_i and C_j, a reference set $RefSet_{ij}$ (the middle part) is constructed containing a number of reference subjects appearing in both C_i and C_j. When comparing track T_1 in C_i with tracks T_2 and T_3 in C_j using their color histograms directly, T_3 is more likely to be matched with T_1. Even though they contain totally different targets, the significant illumination change in C_i makes T_1 looks much darker than its actual appearance. Instead of comparing the tracks directly, each input track is described by all the reference subjects. The description is a vector of similarities ordered by the identities of reference subjects, and each similarity is generated by comparing the input track with one reference subject. The right part of this figure shows the similarity plots obtained by comparing T_1, T_2, T_3 with R_1, R_2, and R_3, respectively. Note that both the input and the reference subjects have multiple appearance instances (only three instances are shown for illustration purpose) that cover the appearance changes of corresponding targets in a particular camera. This indirect match enables us to handle within camera illumination and pose variation. After representing T_1, T_2 and T_3 using the reference set (the right part), it is clear that T_1 is more similar to T_2 than to T_3.

have high similarities with the same set of reference subjects. Otherwise, they are more likely to be the tracks that contain different targets.

In order to handle within camera illumination and pose variation, each track is further divided into several short subtracks (details for track division are presented in Section III-C) such that detections in each subtrack are visually very similar. After track division, each subtrack is an appearance instance for the target under certain illumination condition. Features extracted from each detection in the subtrack are fused into a single set of features, which is used as one representation for the target contained in the subtrack. By this means, we generate multiple representations for each target that covers the appearance changes of that target in a certain camera.

To represent a track by its corresponding reference set, we need to formulate a way of comparing tracks that are obtained in the same camera. When comparing the similarity of two tracks T_a and T_b in the same camera, every subtrack in T_a is compared with every subtrack in T_b. Let t_a^k denotes the k-th subtrack in track T_a, $sim(t_x, t_y)$ be the similarity of two subtracks, and N_a and N_b be the number of subtracks in T_a and T_b, respectively. The similarity score for T_a and T_b is defined as follows:

$$Sim(T_a, T_b) = \frac{1}{N_a} \sum_{i=1}^{N_a} max(\{sim(t_a^i, t_b^j),\ j \in [1, N_b]\}) \quad (5)$$

Concretely, each t_a^i is compared with all subtracks in T_b, and the maximum score is used as the similarity between t_a^i and T_b. Similarity between T_a and T_b is the average of all these maximum scores. The appearance model used to compute $sim(t_a^i, t_b^j)$ is explained in detail in Section III-D.

In the reference set, each reference subject may have several tracks in the same camera (e.g., walking towards and away from the camera). The similarity between a track T_i and a reference subject R_l is the maximum of the similarities of T_i and all the tracks for R_l. This lays the strength of our reference set based appearance model-tracks from different cameras that contain the same target under various pose and illumination conditions have a chance to get high similarity scores with similar reference subjects. In other words, each reference subject is an indirect feature that describes some characteristics of the target's appearance, and having the tracks in two different cameras compared to the same set of reference subjects enables us to better compare the similarity of these two tracks. Besides variation in illumination conditions, difference in poses are also taken care of by the presence of various appearance instances in each reference subject.

After comparing tracks T_p and T_q with each reference subject in its corresponding reference set, we get two vectors of similarities ordered by the identities of reference subjects, as shown in Fig. 2. Let $Ref_{ij}^i(T_p)$ and $Ref_{ij}^j(T_q)$ be the representations of T_p and T_q by the reference set $RefSet_{ij}$, the similarity of T_p and T_q is computed using cosine similarity.

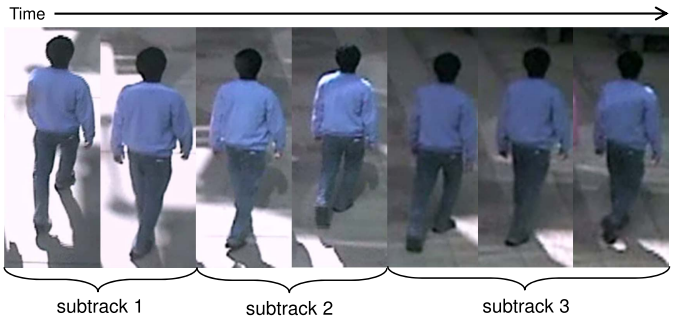


Fig. 3. An example of track division (only detections on key frames are shown). Detections in the same subtrack have higher appearance similarities as compared to the detections in other subtracks.

As it is widely used in tracking, negative logarithm is applied to similarity/linking probability to obtain the linking cost, which is then minimized [9], [33], [34]. In order to get the appearance model, we use the negative logarithm function to calculate the cost, as defined in Eq. (6):

$$Appr(T_p, T_q) = -\ln(cos(Ref_{ij}^{i}(T_p), Ref_{ij}^{j}(T_q))) \quad (6)$$

where $cos(\cdot,\cdot)$ is the cosine similarity between two vectors. We also tested other similarity/distance measures (i.e., χ^2 distance, l_2 norm). Among them, cosine similarity and l_2 norm provide comparable performance and are better than χ^2 distance. As cosine similarity is computationally more efficient, it is chosen as the similarity measure in our experiments.

C. Track Division

In a track, the appearance of a target may vary with time (see Fig. 3), but the detection responses that are obtained in consecutive frames often possess high visual similarity. For efficient computation and to create concise representation of a track, we further divide each track into several subtracks and consider every subtrack as an appearance instance of a target. An example of track division is shown in Fig. 3.

We assume that a target cannot have large pose variation in a very short period of time Δt (0.5s in our experiments). Track division starts from the beginning of a track and subtracks are generated one by one. Let len be the number of frames included in Δt (about 10 in our experiments). The first detection of a track is the appearance reference to form the current subtrack. Specifically, with respect to the detection in the first frame as the reference detection, the detections from the following frames within Δt are compared to the reference detection. As long as the detection similarities are above a pre-defined threshold (0.9 in our experiments), the corresponding frames are kept in this subtrack. Thus, the number of detections in a subtrack is smaller or equal to len. Once a detection's similarity to the reference detection is below the threshold or the number of detections in the current subtrack exceeds len, this detection becomes a new reference detection and detections in latter frames are compared to this reference detection to form a new subtrack. Here color histogram is used to measure the detection similarity.

Fig. 4. Features (HSV color histogram, LBP, HOG) are extracted from different locations of the detection response: torso, legs and body. The torso part is the upper half of the detection and the legs part is the lower half of the detection.

D. Appearance Model for Within Camera Subtracks

To build a comprehensive and strong appearance representation, different local and global features are extracted to describe a tracked target. Three kinds of widely used appearance features: HSV color histograms [35], Local Binary Pattern (LBP) [36], and Histogram of Gradient (HOG) [37], are used to capture color, texture, and shape information of a target. Given a detection response, each feature is extracted at different scales and locations to increase the descriptive ability. Specifically, each detection is divided into an upper and a lower part with equal height to provide coarse representations of the torso and the legs of the contained target, as shown in Fig. 4. Therefore, nine feature descriptors (BodyHSV, BodyLBP, BodyHOG, TorsoHSV, TorsoLBP, TorsoHOG, LegsHSV, LegsLBP, and LegsHOG) are extracted from each detection response. As there are several detection responses in one subtrack, features of the same type are averaged to construct a concise representation for each subtrack.

Given two subtracks t_a and t_b, we can obtain a similarity score by comparing one of the nine appearance feature descriptors. Let x_i denotes a pair of subtracks (t_a, t_b), a feature vector $f(x_i)$ is generated by concatenating the nine similarity scores. We consider each element in this vector as input to a weak classifier. For color histograms and HOG features, we use Bhattacharyya coefficient [38] to measure the similarity. For LBP features, χ^2 distance is used as measurement.

Our goal is to design a discriminative appearance model that gives high similarity for a pair of subtracks that contain similar target while assigning low similarity for two subtracks that contain dissimilar targets. Multiple feature learning algorithms are evaluated (see Section IV-C). Due to its superior performance, AdaBoost is selected to learn the appearance model for within camera subtracks, namely, $sim(\cdot)$ in Eq. 5. AdaBoost consists of a number of weak classifiers and adaptively learns a strong classifier that is a linear combination of all weak classifiers and minimizes the overall error. In our appearance model, the similarity computed from each feature is used in a weak classifier, and AdaBoost assigns a weighting parameter for each feature during the learning process. We formulate the

Algorithm 1 Learning Feature Discriminality

Input:

$\mathcal{S}^+ = \{(x_i, +1)\}$: positive samples
$\mathcal{S}^- = \{(x_i, -1)\}$: negative samples
$\mathcal{F} = \{f(x_i)\}$: feature pool
T: number of iterations
K: number of weak classifiers

Set $w_i = \frac{1}{2|\mathcal{S}^+|}$, if $x_i \in \mathcal{S}^+$; $w_i = \frac{1}{2|\mathcal{S}^-|}$, if $x_i \in \mathcal{S}^-$
Set $t = 1$, $k = 1$
for $t \leq T$ **do**
  for $k \leq K$ **do**
    $r = \sum_i w_i y_i h_k(x_i)$
    $\alpha_k = \frac{1}{2} ln(\frac{1+r}{1-r})$
  end for
  Choose $k^* = argmin_k \sum_i w_i exp[-\alpha_k y_i h_k(x_i)]$
  Set $\alpha_t = \alpha_{k^*}$ and $h_t = h_{k^*}$
  Update $w_i \leftarrow w_i exp[-\alpha_t y_i h_t(x_i)]$
  Normalize w_i
end for

Output:

$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$

learned appearance model as follows:

$$sim(t_a, t_b) = H(t_a, t_b) = \sum_{t=1}^{T} \alpha_t h_t(t_a, t_b) \tag{7}$$

where t indicates the iteration index, α_t is the weighting parameter and $h_t(t_a, t_b)$ is a weak classifier based on one of the features extracted from subtracks t_a and t_b.

For each camera, we train such a discriminative model using data in the reference set collected from the corresponding camera. A pair of subtracks $x_i = (t_x, t_y)$ is a positive sample if $t_x, t_y \in T_i$, and $t_x \neq t_y$, and it is a negative sample when $t_x \in T_i$, $t_y \in T_j$ and $T_i \neq T_j$. The feature of a training sample is an 9-dimensional vector as explained above. We summarize the learning procedure in Algorithm 1.

IV. Experiments

Compared to tracking in single camera, there are fewer publicly available datasets designed for real-world multi-camera tracking. In this work, we use two datasets, MultiCam dataset and VideoWeb dataset [39], to evaluate the performance of our proposed model.

A. Implementaion Details

Targets in each frame are detected via the discriminatively trained deformable part models [40]. We use the multi-target tracking method in [41] to produce reliable intra-camera tracks. It is a hierarchical association approach. First, tracklets are generated by connecting detections in consecutive frames that have high similarity in position, appearance and size using a two-threshold strategy. Then, these tracklets are further associated based on more complex affinity measures to recover the full trajectory of a target.

TABLE I
RMSE Comparison of Different Feature Learning Algorithms on MultiCam and VideoWeb Datasets

	AdaBoost	GentleBoost	LogitBoost	RUSBoost	MKL
MultiCam	**0.228**	0.303	0.235	0.249	0.240
VideoWeb	**0.387**	0.456	0.422	0.468	0.394

B. Baseline Models and Metrics

In this evaluation, our main focus is to associate tracks that contain the same target in different camera views given certain spatio-temporal constraints. We apply our reference set based appearance model with weighted features (RefSet2) on the test set. We introduce three baseline models for comparison: (1) using Bhattacharyya distance of holistic color histograms directly to measure the appearance similarity (Color); (2) generating appearance model based on the BTF model in [12] (BTF); (3) our proposed reference set based appearance model with only holistic color histograms as appearance feature (RefSet1).

For each model, various thresholds (ranging from 0.2 to 0.6) are tested for the augmented cost matrix, and the best result is chosen. Two metrics are used for evaluation:

$$ErrorRate = \frac{Error}{N_{result}}, \quad MatchRate = \frac{Match}{N_{GT}} \tag{8}$$

where $Error$ and $Match$ are the number of incorrectly and correctly associated track pairs in the result. N_{result} and N_{GT} are the number of track associations in the result and the ground-truth, respectively.

C. Evaluation of Feature Learning Algorithm

In order to find a suitable learning algorithm to build discriminative appearance models for within camera subtracks, we compare the performance of multiple alternatives, including: AdaBoost [42], GentleBoost [42], LogitBoost [42], RUSBoost [43], and Multiple Kernel Learning (MKL) [44]. The reference set is used as the dataset to test each algorithm. Root Mean Squared Error ($RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}$) is used for performance evaluation, it measures the differences between the ground truth y_t and the prediction results $\hat{y}_t$ generated by the learned appearance model. The final result is the average of five-fold cross-validation. Comparison of different algorithms on MultiCam and VideoWeb datasets are shown in Table I. As can be seen, AdaBoost gives the smallest error on both datasets. With respect to model training time, on MultiCam dataset, the boosting algorithms take less than 2 seconds, and the average training time for MKL is 181 seconds. On the VideoWbe dataset, the training time for MKL is also about two orders of magnitude more than that of the boosting algorithms. Taking both performance and computational time into account, we select AdaBoost as the feature learning algorithm for the appearance model.

D. Results on MultiCam Dataset

We use five cameras (four indoor and one outdoor) to build a real-world non-overlapping multi-camera network,

Fig. 5. Detection examples of participants that appear in both the reference set and the test set for MultiCam dataset.

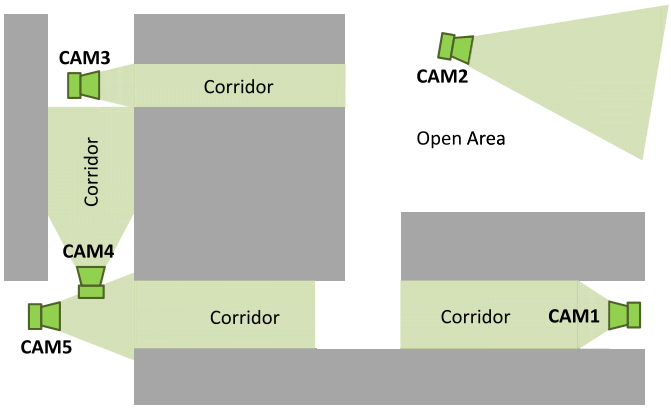


Fig. 6. Topology for cameras used in MultiCam dataset.

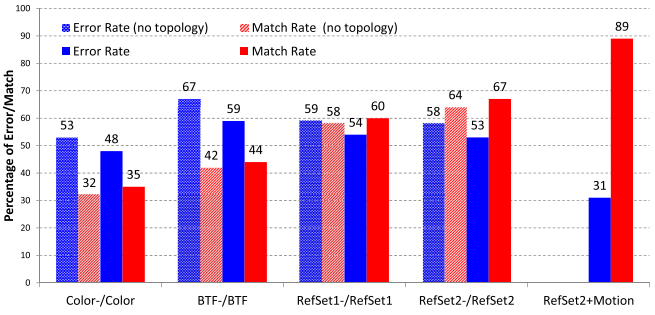


Fig. 7. Comparison of the proposed method and other baseline models on MultiCam dataset. The minus sign (−) indicates no camera topology knowledge is used for linking cost computation.

the topology of this camera network is presented in Fig. 6 and sample frames from each camera are shown in Fig. 1. All the videos (five in total) are taken during the same time period and the length of each video is about 20 minutes. The resolution of each frame is 704 × 480 and the frame rate is 20fps. The number of participants involved in each video ranges from 7 to 10. We refer this dataset as the *MultiCam dataset* in this paper.

The setting of this dataset is very challenging for multi-camera tracking due to the following reasons:

1) The outdoor camera view contains drastic illumination changes, and there exists lighting variations for indoor camera views as well. This makes it unreliable to use a single transformation to map colors in a pair of cameras, such as BTFs [12].
2) The number of cameras involved in this dataset is greater than most of the previous work that normally use 2-3 cameras [11], [12].

In order to construct the reference set, another set of data is used. It is collected using the same camera network and under similar illumination condition but with participants either not included in the test set or included in the test set but with very different clothes. There are two participants that appear in both the reference set and the test set, as shown in Fig. 5. As the appearances of the same participant have a great difference even in the same camera, each of the trained appearance model classifies them as negative (i.e., two different people) with more than 90% confidence. The data collected for each reference subject contains the appearance change of the target under different illumination conditions and various poses. The number of participants involved in each reference set ranges from 9 to 11. We manually labeled the ground-truth which consists of 220 track associations (there are 368 single camera tracks in total).

A quantitative comparison between the proposed model and baseline models is presented in Fig. 7. It can be observed that when using the reference set based appearance model with weighted features, we achieve the highest match rate and the lowest error rate compared to all the baseline models. Compared with BTF, the RefSet2 model increases the match rate by 23% and reduces the error rate by 6%. Even with color histograms only, the reference set based appearance model (RefSet1) provides better performance than BTF in terms of both the error rate and the match rate. The comparison between RefSet1 and RefSet2 demonstrates that by using features of various types and extracted at different locations we can get more information than using global color hitograms only,

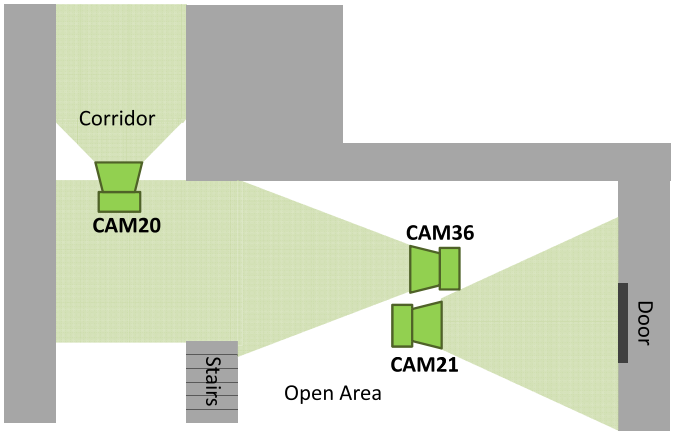


Fig. 8. Topology for cameras used in VideoWeb dataset.

as they capture the appearance information that is overlooked by color hitograms. It is worth noting that although the error rate is high even for RefSet1 and RefSet2 (more than 50%), these results are obtained by using appearance information as the only visual cue.

In addition, to evaluate the contribution of camera topology knowledge to the overall tracking system, we further conducted experiments in which the topology information is not used for computing the linking cost in Eq. 2. With such relaxation, more track pairs are included in each time sliding window as potential association candidates. Therefore, the error rates increase by at least 5% for all the methods, and the match rates decrease by 2.5% on average, as shown in Fig. 7. These results demonstrate the importance of camera topology information as prior knowlege for a tracking system, as it is helpful for mitigating unnecessary track association ambiguities ahead of time.

As another kind of clue, motion information plays an important role in multi-target tracking. For example, in a time sliding window, a track in CAM4 can be associated with tracks in both CAM3 and CAM5 based on the camera topology (Fig. 6). Given the knowledge that the target is walking away from CAM4, we can easily eliminate tracks in CAM5 from possible associations. When a motion model that measures the walking direction of a target is integrated into the tracking system (RefSet2+Motion), the error rate is greatly reduced to 31%. Also, with motion information our proposed method can correctly associate almost 90% track pairs. Comparison between BTF and RefSet2 on some challenging cases are shown in Fig. 11, which validates the robustness of our method.

E. Results on VideoWeb Dataset

In order to further validate our method, we carried out experiments on a public dataset, the VideoWeb dataset [39]. Three cameras CAM20, CAM21, and CAM36 with disjoint views are selected to form the multi-camera network, the topology of which is shown in Fig. 8. Three sets of videos are selected as the test set, each videos is about 6 minutes, the resolution of a frame is 640×480, and the frame rate is 30fps. Under the same setting, videos from Day3 are used to generate the reference set and videos from Day2 are used to build the test set. Participants involved in Day2 are either not included in Day3 or they are included but with different clothes. There are four participants appearing in both videos from Day3 (reference set) and videos from Day2 (test set), as shown in Fig. 9. However, due to significant appearance differences, tracks in the reference set and tracks in the test set, even from the same target and captured by the same camera, are considered to contain two different people (with more than 85% confidence) according to the prediction by the trained appearance model. There are 10 participants involved in the test set, and the number of reference subjects in each reference set ranges from 9 to 12. We manually labeled the ground-truth which consists of 66 track associations (there are 222 single camera tracks in total).

A quantitative comparison between the proposed model and baseline models is presented in Fig. 10. Using the reference set based appearance model with weighted features (RefSet2) we obtain a match rate of 64% and an error rate of 44%, which is better than the performance of all the other baseline models. Comparison between RefSet1 and BTF (both of them use global color histograms only as appearance feature), further demonstrate the superiority of the reference set based appearance model as it provides a better method to handle track association ambiguities caused by illumination and pose variations across cameras.

We also tested the tracking system without using camera topology information on this dataset. Similar to the observations from Fig. 7, with no camera topology information as prior knowledge to reduce unnecessary track associations, higher error rates and lower match rates are observed, as illustrated in Fig. 10. However, the impact on the error rate is not as significant as it is on the MultiCam dataset, the error rates increase by at most 3% for all the methods on this dataset. This is probably due to the following two reasons: 1) this dataset has less number of cameras compared to the MultiCam dataset; 2) this dataset contains fewer scenarios in which participants exist in the FOVs of all the three cameras simultaneously. Therefore, the numbers of potential track associations generated by our tracking system with and without camera topology information would be close on this dataset.

Results from Fig. 7 and Fig. 10 suggest that the performance of our method is consistent on both datasets, which validate the robustness of the reference set based appearance model with weighted features. Note that the ViedoWeb dataset is originally designed for complex real-world activity recognition, participants in this dataset have more non-linear motion and heavy interactions than that in the MutliCam dataset. Therefore, the overall tracking performance on this dataset is not as good as that on the MultiCam dataset. Also, non-linear motion and interactions among individuals make it difficult to predict accurate motion direction of a target. Thus, after integrating motion model with RefSet2, the improvement on both error rate and match rate is small. Comparison between BTF and RefSet2 on some challenging cases of VideoWeb dataset are illustrated in Fig. 12.

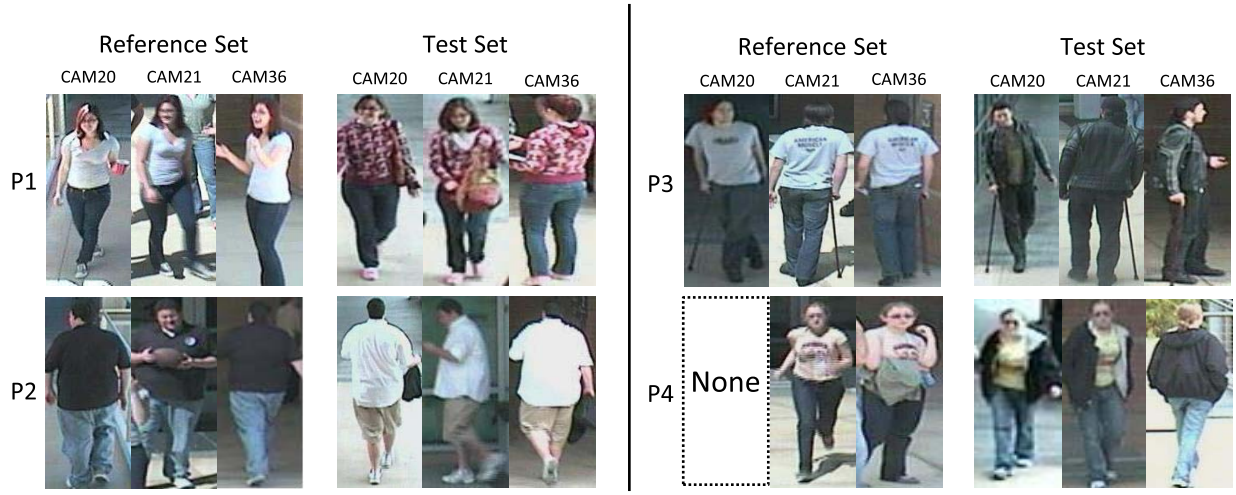


Fig. 9. Detection examples of participants that appear in both the reference set and the test set for VideoWeb dataset. The "None" box indicates the corresponding participant (P4) generates no track in camera 20 for the reference set.

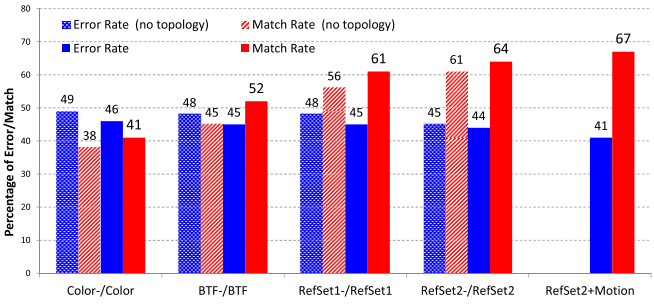


Fig. 10. Comparison of the proposed methods and other baseline models on VideoWeb dataset. The minus sign (−) indicates no topology knowledge is used for linking cost computation.

TABLE II
TRACKING RESULTS WITH DIFFERENT REFERENCE SET SIZES. "N" IS THE NUMBER OF REFERENCE SUBJECTS IN THE ORIGINAL REFERENCE SET. "MATCH" AND "ERROR" STAND FOR MATCH RATE AND ERROR RATE

$RefSet$ size →	n = N		n = 2/3*N		n = 1/2*N	
Results →	Match	Error	Match	Error	Match	Error
MultiCam	0.67	0.53	0.45	0.49	0.27	0.39
VideoWeb	0.64	0.44	0.43	0.41	0.21	0.32

TABLE III
TRACKING RESULTS WITH AND WITHOUT PARTICIPANTS APPEARING IN BOTH REFERENCE AND TEST SETS. THE ASTERISK SIGN (*) INDICATES DATASET WITHOUT IDENTITY OVERLAP IN REFERENCE AND TEST SETS

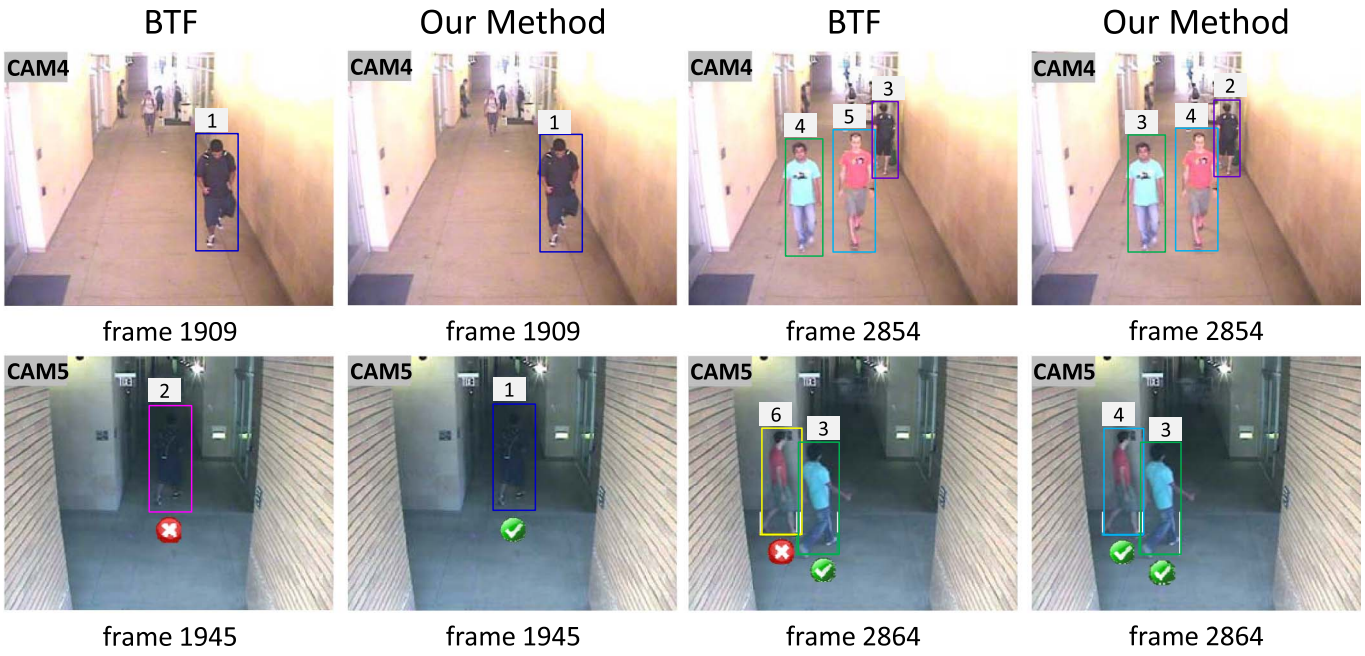
	GT	ErrorRate	MatchRate
MultiCam	220	0.31	0.89
MultiCam*	174	0.32	0.89
VideoWeb	66	0.41	0.67
VideoWeb*	43	0.43	0.65

F. Reference Set Analysis

In Table II we evaluate the performance of our reference set based appearance model with reduced reference subjects. Each time, a subset of the original reference set is randomly selected as a new reference set. The reported results are the average of 10 runs. It is observed that as the size of the reference set reduces the match rate degrades. As less number of track associations are produced in the result, the error rate also decreases. Therefore, the results suggest that for small test sets (about 10 subjects) in order to get good performance from the reference set based appearance model, it is better to make the number of reference subjects comparable to the number of targets in the test set. However, as the size of reference set increases, more redundancy together with more diversity are introduced. Different methods can be used to select a subset from the entire reference candidate pool in order to maintain discriminality while reducing redundancy for better efficiency. For example, in [45] for face recognition, reference set selection is proposed from a low-rank decomposition point of view. In [46] for biometric pattern retrieval, rule-based methods are suggested for reference set selection, including max-variation, max-mean, and min-correlation.

Moreover, the appearances of subjects in a reference set should be as diverse as possible, so that each reference subject can be used to capture some unique characteristic of a target. If there are highly similar subjects in the reference set, there will be redundant information in the reference set based appearance descriptor. When such redundancy increases, the performance of the model will be adversely affected.

To evaluate the effect of having participants existing in both reference and test set, we removed the overlap and carried out experiments on both MultiCam and VideoWeb datasets, the results are shown in Table III. As can be seen, the numbers of track associations in the ground-truth decreased as we removed some participants from the test sets, but there was no significant difference in both the error rate and the match rate for datasets with and without overlapping identities. The results justify the rationality of our experiments



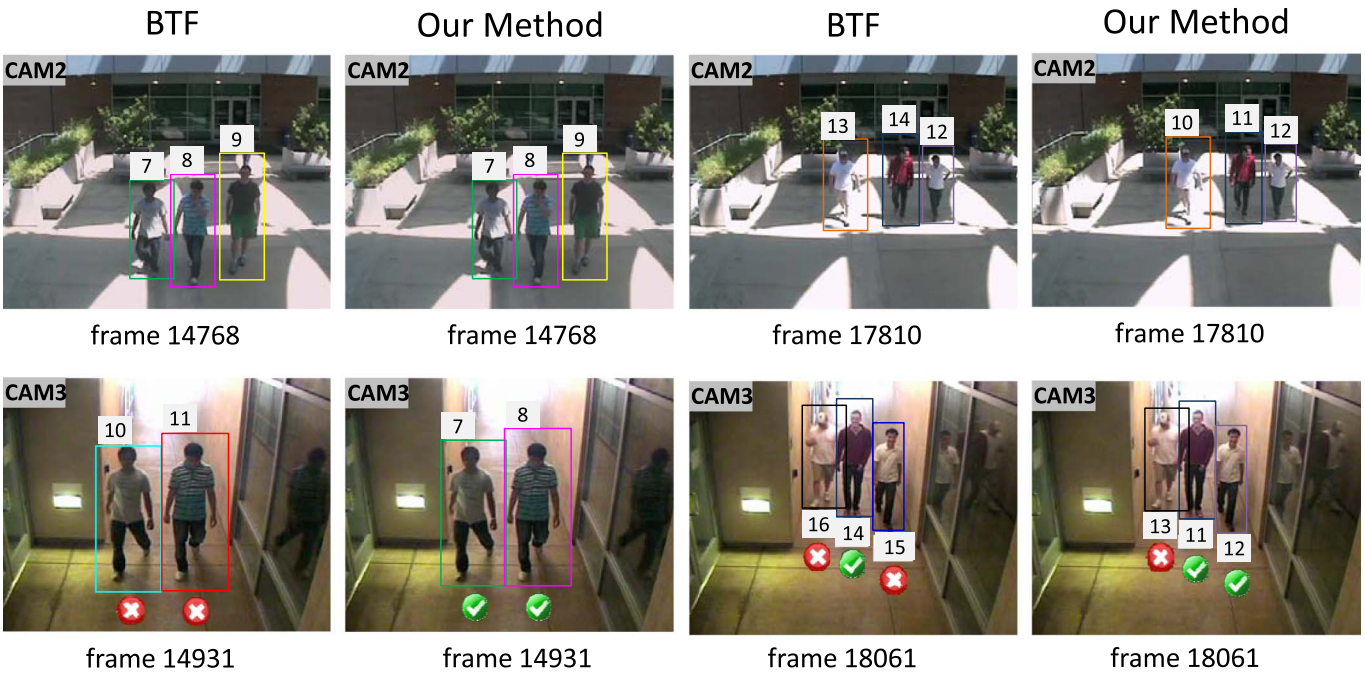


Fig. 11. Example tracking results on MultiCam dataset. The first and the third columns are the results obtained by using BTF in [12], the second and the fourth columns are the results by the proposed method using reference set based appearance model with weighted features (RefSet2). With the reference set, our method is able to match most of the targets even with the presence of drastic within camera and across camera illumination variations. The method in [12] fails to associate tracks that contain the same target in these challenging situations. Best viewed in color. (a) Tracking between CAM4 and CAM5. (b) Tracking between CAM2 and CAM3.

in Section IV-D and IV-E, that is to say, having participants appearing in both reference and test sets but with very different clothes did not impact the performance of the proposed method greatly. This is because only the appearance of reference subjects matters, not the real identities of those particular subject.

G. Feature Discriminality Analysis

In addition, we further carried out experiments to analyze the discriminative power of all the nine features (HSV, LBP, and HOG extracted on body, torso and legs, respectively) used in the appearance model for within camera subtracks.

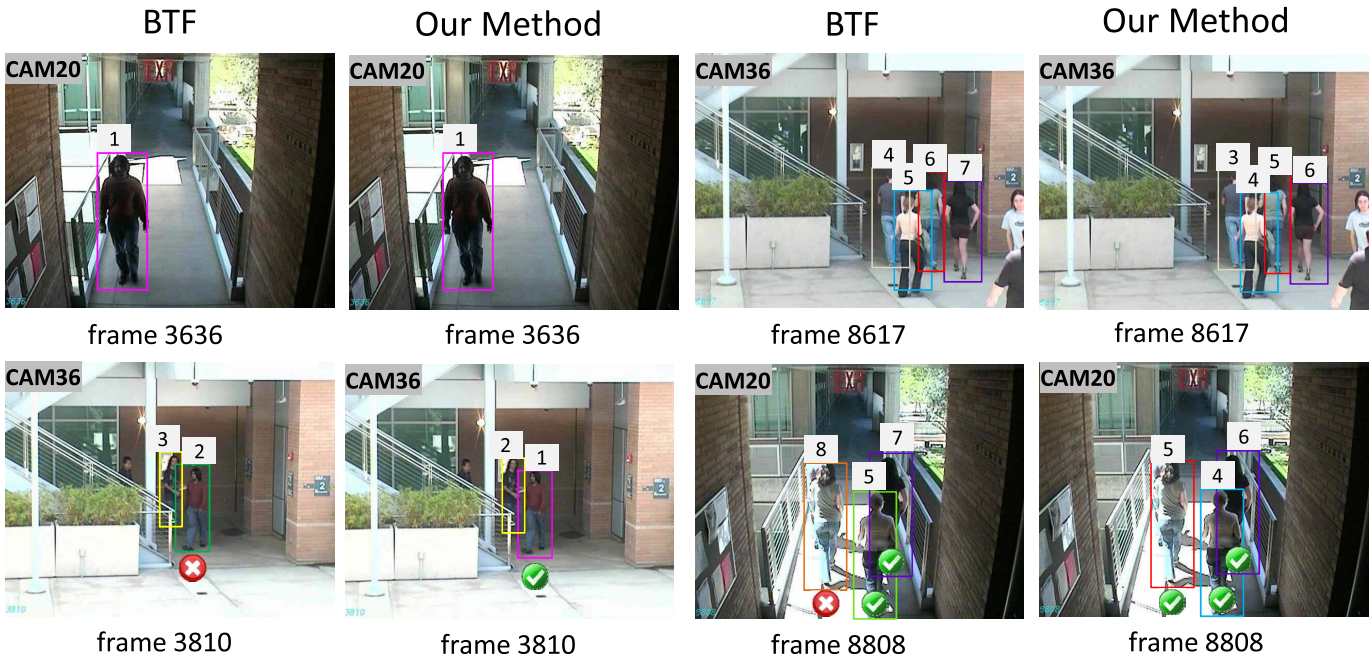


Fig. 12. Example tracking results on VideoWeb dataset. The first and the third columns are the results obtained by using BTF in [12], the second and the fourth columns are the results by the proposed method using reference set based appearance model with weighted features (RefSet2). Best viewed in color.

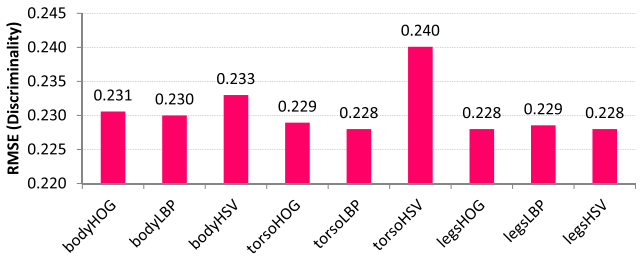


Fig. 13. Feature discriminality analysis for MultiCam dataset. Each column represents the RMSE (representing discriminality) when the corresponding feature is *removed* from feature pool of the appearance model.

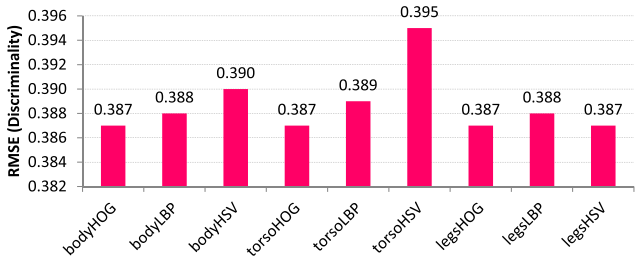


Fig. 14. Feature discriminality analysis for VideoWeb dataset. Each column represents the RMSE (representing discriminality) when the corresponding feature is *removed* from the appearance model.

For each feature, the RMSE obtained when that feature is "removed" from the appearance model is considered as its discriminality measurement. A more discriminative feature would produce a higher RMSE when it is discarded from the feature pool, therefore, it is more important for the learned appearance model. The experimental results for MultiCam and VideoWeb datasets are shown in Fig. 13 and Fig. 14, respectively. For both datasets, it is clear that HSV are more discriminative than HOG and LBP, and torsoHSV is the most discriminative one. Also, legs carry less information for the appearance model compared to body and torso, probably because most of the participants are wearing jeans with similar color. Since HOG features are not pose invariant, discarding HOG features does not increase the RMSE in the VideoWeb dataset where participants interacted heavily, indicating that in this case HOG features do not have high discriminality. On the other hand, in the MultiCam dataset, we observe slightly higher RMSE for bodyHOG and torsoHOG, as participants in this datasets are less active and their poses remained relatively stable during the data capturing process. Moreover, the RMSE obtained by using all the nine features is 0.228 and 0.387 for MultiCam and VideoWeb datasets, respectively. These RMSE values are equal or smaller than the RMSE values obtained with one of the nine features removed from the feature pool. It is observed that the removal of some features, such as legsHSV and legsHOG, have no impact on the RMSE results. This is plausible since in practice often the upper body dress of the subject being tracked is more distinctive (e.g., shirts with various color and patterns) compared to the lower body dress (e.g., jeans) which is more uniform, as shown, for example, in Fig. 1. Although not effective on the datasets used in our experiments, these less discriminative features may contribute to the tracking accuracy when the appearance captured by these features is more discriminative.

V. Conclusion

In this paper, we propose a novel reference set based appearance model with weighted features for multi-target tracking in a camera network with non-overlapping FOVs. In order to deal with track association ambiguities caused by

illumination and pose variations across cameras, we generate multiple appearance instances for each track and make indirect comparison of two tracks obtained in different cameras by utilizing a reference set. The experimental results demonstrate the superiority of the combination of reference set based appearance model and weighted features over the baseline models on two challenging real-world video datasets. A future work would be testing the proposed reference set based tracking method on larger datasets with more analysis on reference subjects selection.

REFERENCES

[1] I. Saleemi, K. Shafique, and M. Shah, "Probabilistic modeling of scene dynamics for applications in visual surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1472–1485, Aug. 2009.

[2] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 663–671, Apr. 2006.

[3] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, 2013.

[4] N. T. Siebel and S. J. Maybank, "The advisor visual surveillance system," in *Proc. Eur. Conf. Comput. Vis. Workshop (ECCVW)*, pp. 103–111, 2004.

[5] Z. Qin, C. R. Shelton, and L. Chai, "Social grouping for target handover in multi-view video," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.

[6] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 91–101, Feb. 2013.

[7] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behavior-recognition algorithms," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 206–224, Mar. 2010.

[8] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1972–1978.

[9] X. Chen, Z. Qin, L. An, and B. Bhanu, "An online learned elementary grouping model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1242–1249.

[10] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *Proc. 9th Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 125–136.

[11] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2008, pp. 64.1–64.10.

[12] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Comput. Vis. Image Understand.*, vol. 109, no. 2, pp. 146–162, 2008.

[13] C.-T. Chu, J.-N. Hwang, K.-M. Lan, and S.-Z. Wang, "Tracking across multiple cameras with overlapping views based on brightness and tangent transfer functions," in *Proc. 5th ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Aug. 2011, pp. 1–6.

[14] T. D'Orazio, P. L. Mazzeo, and P. Spagnolo, "Color brightness transfer function evaluation for non overlapping multi camera tracking," in *Proc. 3rd ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Aug./Sep. 2009, pp. 1–6.

[15] C.-T. Chu, J.-N. Hwang, J.-Y. Yu, and K.-Z. Lee, "Tracking across nonoverlapping cameras based on the unsupervised learning of camera link models," in *Proc. 6th IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Oct./Nov. 2012, pp. 1–6.

[16] F. Schroff, T. Treibitz, D. Kriegman, and S. Belongie, "Pose, illumination and expression invariant pairwise face-similarity measure via Doppelgänger list comparison," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2494–2501.

[17] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 497–504.

[18] L. An, M. Kafai, and B. Bhanu, "Dynamic Bayesian network for unconstrained face recognition in surveillance camera networks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 3, no. 2, pp. 155–164, Jun. 2013.

[19] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re-identification," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2013, pp. 244–249.

[20] R. Jain and K. Wakimoto, "Multiple perspective interactive video," in *Proc. Int. Conf. Multimedia Comput. Syst.*, May 1995, pp. 202–211.

[21] T.-H. Chang and S. Gong, "Tracking multiple people with a multi-camera system," in *Proc. IEEE Workshop Multi-Object Tracking*, Jul. 2001, pp. 19–26.

[22] S. L. Dockstader and A. M. Tekalp, "Multiple camera fusion for multi-object tracking," in *Proc. IEEE Workshop Multi-Object Tracking*, Jul. 2001, pp. 95–102.

[23] J. Black, T. Ellis, and P. Rosin, "Multi view image surveillance and tracking," in *Proc. Workshop Motion Video Comput.*, Dec. 2002, pp. 169–174.

[24] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3650–3657.

[25] Z. Jin and B. Bhanu, "Integrating crowd simulation for pedestrian tracking in a multi-camera system," in *Proc. 6th ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Oct./Nov. 2012, pp. 1–6.

[26] F. Madrigal and J.-B. Hayet, "Multiple view, multiple target tracking with principal axis-based data association," in *Proc. 8th IEEE Int. Conf. Adv. Video Signal-Based Surveill. (AVSS)*, Aug./Sep. 2011, pp. 185–190.

[27] L. Marcenaro, P. Morerio, and C. S. Regazzoni, "Performance evaluation of multi-camera visual tracking," in *Proc. 9th IEEE Int. Conf. Adv. Video Signal-Based Surveill. (AVSS)*, Sep. 2012, pp. 464–469.

[28] M. Taj and A. Cavallaro, "Multi-camera track-before-detect," in *Proc. ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Aug./Sep. 2009, pp. 1–6.

[29] R. Mazzon and A. Cavallaro, "Multi-camera tracking using a multi-goal social force model," *Neurocomputing*, vol. 100, pp. 41–50, Jan. 2013.

[30] X. Chen, K. Huang, and T. Tan, "Object tracking across non-overlapping views by learning inter-camera transfer models," *Pattern Recognit.*, vol. 47, no. 3, pp. 1126–1137, 2014.

[31] X. Chen, L. An, and B. Bhanu, "Reference set based appearance model for tracking across non-overlapping cameras," in *Proc. 7th ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Oct./Nov. 2013, pp. 1–6.

[32] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.

[33] B. Yang and R. Nevatia, "Multi-target tracking by online learning a CRF model of appearance and motion patterns," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 203–217, 2014.

[34] J. F. Henriques, R. Caseiro, and J. Batista, "Globally optimal solution to multi-object tracking with merged measurements," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2470–2477.

[35] X.-Y. Wang, J.-F. Wu, and H.-Y. Yang, "Robust image retrieval based on color histogram of local feature regions," *Multimedia Tools Appl.*, vol. 49, no. 2, pp. 323–345, 2010.

[36] X. Liu and D. Wang, "Texture classification using spectral histograms," *IEEE Trans. Image Process.*, vol. 12, no. 6, pp. 661–670, Jun. 2003.

[37] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[38] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

[39] G. Denina *et al.*, "VideoWeb dataset for multi-camera activities and non-verbal communication," in *Distributed Video Sensor Networks*. London, U.K.: Springer-Verlag, 2011, pp. 335–347.

[40] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[41] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 788–801.

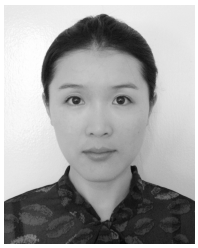
[42] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, Apr. 2000.

[43] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed," in *Proc. 19th Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2008, pp. 1–4.

[44] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 1065–1072.

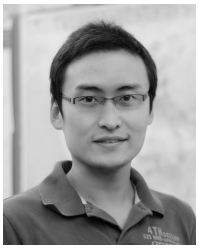
[45] M. Kafai, L. An, and B. Bhanu, "Reference face graph for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2132–2143, Dec. 2014.

[46] A. Gyaourova and A. Ross, "Index codes for multibiometric pattern retrieval," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 518–529, Apr. 2012.



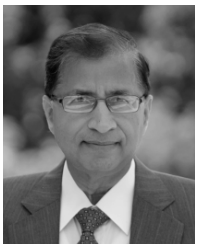
Xiaojing Chen received the B.S. degree in information management and information systems from Beijing Language and Culture University, Beijing, China, in 2007, and the M.S. (Hons.) degree in computer science from Leiden University, Leiden, The Netherlands, in 2009. She is currently pursuing the Ph.D. degree in computer science at the University of California at Riverside, Riverside, CA, USA, in USA.

Her research interests are in computer vision, pattern recognition, and machine learning. Her recent research has been concerned with multitarget tracking in surveillance cameras.



Le An received the B.Eng. degree in telecommunications engineering from Zhejiang University, Hangzhou, China, in 2006, the M.Sc. degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2008, and the Ph.D. degree in electrical engineering from the University of California at Riverside, Riverside, CA, USA, in 2014. He is currently a Post-Doctoral Research Associate with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

His research interests include image processing, computer vision, pattern recognition, and machine learning. He was a recipient of the best paper award from the 2013 IEEE International Conference on Advanced Video and Signal-Based Surveillance.



Bir Bhanu (F'95) received the B.S. (Hons.) degree from the Indian Institute of Technology (BHU, Varanasi), Varanasi, India, the M.E. (Hons.) degree from the Birla Institute of Technology and Science, Pilani, India, the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, and the M.B.A. degree from the University of California at Irvine, Irvine, CA, USA. He is currently the Distinguished Professor of Electrical Engineering and serves as the Founding Director of the Interdisciplinary Center for Research in Intelligent Systems with the University of California at Riverside (UCR), Riverside, CA, USA. He was the Founding Professor of Electrical Engineering with UCR, and served as its first Chair from 1991 to 1994. He has been the Cooperative Professor of Computer Science and Engineering since 1991, Bioengineering since 2006, Mechanical Engineering since 2008, and the Director of the Visualization and Intelligent Systems Laboratory since 1991. In addition, he serves as the Director of the NSF Interdisciplinary Graduate Education, Research and Training Program on Video Bioinformatics with UCR. He was a Senior Honeywell Fellow with Honeywell Inc., Minneapolis, MN, USA. He has been with the Faculty of Computer Science, University of Utah, Salt Lake City, UT, USA, Ford Aerospace and Communications Corporation, Newport Beach, CA, USA, the INRIA France, and the IBM San Jose Research Laboratory, San Jose, CA, USA. He has been the Principal Investigator of various programs for the NSF, DARPA, NASA, the Air Force Office of Scientific Research, the Office of Naval Research, the Army Research Office, and other agencies and industries in the areas of video networks, video understanding, video bioinformatics, learning and vision, image understanding, pattern recognition, target recognition, biometrics, autonomous navigation, image databases, and machine-vision applications.

Dr. Bhanu has co-authored seven books and edited three books. He has over 475 reviewed technical publications, including over 130 journal papers and 44 book chapters. He has been on the Editorial Board of many journals, and has edited as the lead Editor of the special issues of over 15 journals, including some of the top IEEE publications. He was the General Chair of the IEEE CVPR, AVSS, ICDSC, WACV, and DARPA IUW. He served on the IEEE Fellow Committee (2010–2012). He was a recipient of many Best Conference Papers and Outstanding Journal Paper Awards and the Industrial and University Awards for Research Excellence, Outstanding Contributions, and Team Efforts, and the Doctoral/Dissertation Advisor/Mentor Award. He is a fellow of the American Association for the Advancement of Science, the International Association of Pattern Recognition, the American Institute for Medical and Biological Engineering, and the International Society for Optical Engineering.