Reference Set Based Appearance Model for Tracking Across Non-overlapping Cameras

Xiaojing Chen, Le An, Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside 92521, USA
xchen010@ucr.edu, lan004@ucr.edu, bhanu@cris.ucr.edu

***Abstract*—Multi-target tracking in non-overlapping cameras is challenging due to the vast appearance change of the targets across camera views caused by variations in illumination conditions, poses, and camera imaging characteristics. Therefore, direct track association is difficult and prone to error. In most previous methods the appearance similarity is computed either using color histograms directly or based on pre-trained Brightness Transfer Function (BTF) that maps color between cameras. In this paper, we propose a novel reference set based appearance model to improve multi-target tracking in a network of non-overlapping video cameras. Unlike previous work, a reference set is constructed for a pair of cameras, containing targets appearing in both camera views. For track association, instead of comparing the appearance of two targets in different camera views directly, they are compared to the reference set. The reference set acts as a basis to represent a target by measuring the similarity between the target and each of the individuals in the reference set. Besides color histograms, other soft-biometric features are also integrated into the feature representation of a target. The effectiveness of the proposed method over the baseline models on challenging real-world multi-camera video data is validated by the experiments.**

I. Introduction

Recently, a major effort has been underway in the vision community to develop effective and robust multi-target tracking systems. It is the foundation for many higher level applications, such as anomaly detection, activity detection and recognition [1], human behavior understanding [2], and surveillance and monitoring [3]. 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 large proximity in appearance, space and time [4]. However, it is not the case for tracking people across cameras with non-overlapping field-of-views (FOVs). The appearance of the same target may have large difference even in two adjacent cameras due to a sudden change in illumination conditions (e.g., from outdoor to indoor). Other aspects, such as variations in pose (e.g., frontal view to rear view) and camera imaging conditions further complicate the tracking task in multiple cameras. In Fig. 1 some sample frames are shown in which the appearance of same target in different camera views differs significantly.

A possible way to tackle the appearance difference in multiple cameras is to learn Brightness Transfer Function (BTF) [5] [6] [7] [8] [9] [10] that is a mapping of color models between a pair of cameras. However, BTF is not suitable for a camera network that has a large within camera illumination change. For example, camera1 and camera2 both have dark and bright regions in their camera views. A BTF that is able to map colors in dark region of camera1 (low brightness) to colors in bright region of camera2 (high brightness) will not work well for mapping colors in bright region of camera1 (high brightness) to dark region of camera2 (low brightness).

Fig. 1. Sample frames from each camera view. Bounding boxes with the same color indicate the same target. Notice that illumination may change drastically within camera and across cameras, the appearances of the same target has significant variations.

To address this problem, we propose a novel reference set based appearance model to estimate the similarity of multiple targets in different cameras. Based on the tracking results from single camera, the goal is to associate tracks in different cameras that contain the same person. Our method is inspired by the recent advances in face verification/recognition [11] [12] and person re-identification [13] in which an external reference set or library is used to facilitate the matching process of the same objects in different imaging conditions. The reference set contains the appearance of individuals in different camera views under different imaging conditions. For tracking, instead of comparing 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 representation rather than the original low level color or texture features.

In addition to color histogram we integrate other soft-

biometric features which are invariant to view and illumination changes into the feature representation of a target. Soft-biometrics are characteristics that can be used to describe a person [14], for instance height, weight, gender, hair color and clothes color. Although each one of them is not discriminative enough to uniquely identify an individual, when bundled as a whole they can provide coarse representation of a target. Because soft-biometrics can be directly acquired from surveillance videos without any target's cooperation, they are suitable for constructing appearance model for tracked targets. Soft-biometrics have been widely used for retrieval and recognition tasks on image datasets [14][15], recently they are also applied for identifying a specific target in surveillance videos [16]. However, to the best of the authors' knowledge, soft-biometrics have never been used for improving tracking performance.

The rest of this paper is organized as follows: an overview of the related work is provided in Section 2. Section 3 describes the proposed reference set based appearance model for multi-target tracking across non-overlapping cameras. Experimental results are shown in Section 4. Finally, Section 5 concludes this paper.

II. Related Work

To cope with the illumination change in different camera views, BTF has been studied extensively [5] [6] [7] [8] [9] [10]. An incremental unsupervised learning method is proposed in [5] 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 [6] a cumulative BTF is proposed to map color between different cameras and significant improvement over BTF-based methods is reported. Javed *et al.* [7] learn the inter-camera relationships 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 [8], BTFs are built from the overlapping area during tracking to compensate for the color difference between camera views. In addition, the perspective difference is compensated for with tangent transfer functions (TTFs) by computing the homography between two cameras. Different methods are compared to evaluate the color BTFs between non-overlapping cameras and experimental results show BTFs limitations in people association when a new person enters in one camera's FOV [9]. In [10], to track people across non-overlapping cameras, a camera link model including BTF, transition time distribution, region mapping matrix/weight, and feature fusion weight are estimated in an unsupervised manner.

Recently, the reference-based idea has been used in the field of computer vision, for example, face verification [11], face recognition [12], and person re-identification [13]. The reference-based framework is data-driven and different entities to be matched or compared are first described using the elements in the reference set and reference-based descriptors are generated. Therefore, direct comparison of objects with different modalities (e.g., faces at different poses) is avoided. In [11], pose, illumination, and expression invariant face verification is achieved 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. Yin *et al.* [12] proposed an "Associate-Predict" model 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 like 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, An *et al.* [13] used a reference set to generate reference-based descriptors for probe and gallery subjects, bypassing the need to direct compare the features from subjects with significant appearance change.

III. Technical Approach

A. Formulation of the Multi-Camera Tracking Problem

Suppose we have m cameras $C_1, C_2, ..., C_m$ with non-overlapping FOVs. Given the tracking results in each single camera, we can generate a set $T = \{T_1, .., T_N\}$ that contains all the within-camera tracks. A track T_i is a consecutive sequence of detections that contain the same target, its time interval is denoted as $[t^i_{begin}, t^i_{end}]$, and its corresponding camera is denoted as C^i. The problem of tracking across cameras is essentially to find out tracks that contain the same target, given certain spatial-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 $C^i \neq C^j$ (associate 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:

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

The constraints for matrix A indicate that each track cannot be associated to more than one track.

The cost S_{ij} for linking track T_i and T_j is based on time, appearance, and camera topology constraints, as defined in Equ. (2).

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

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 within 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^i and C^j without entering the view of any other cameras.

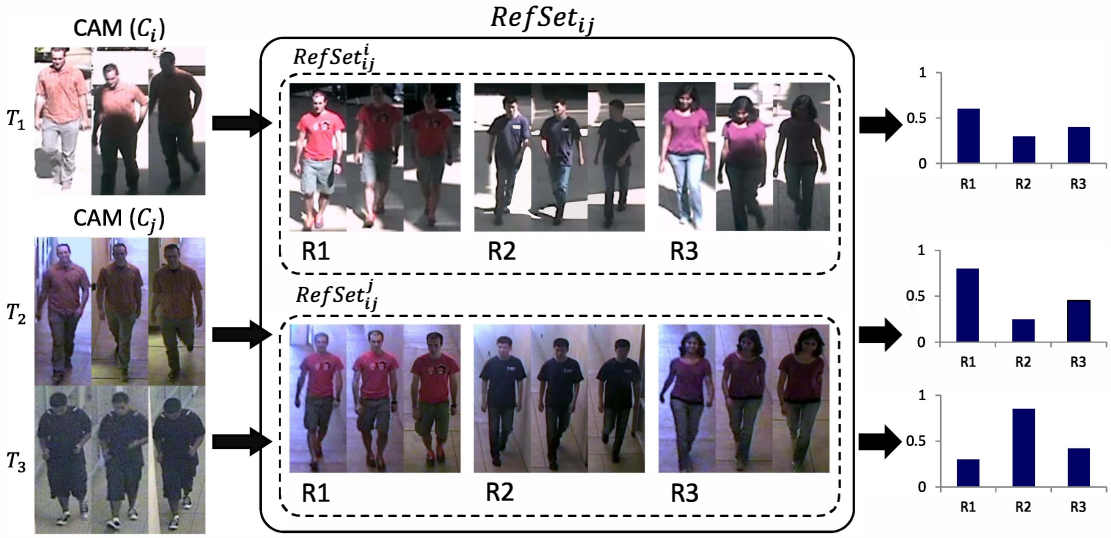

Fig. 2. Illustration of the reference set based appearance model. For a pair of cameras C_i and C_j, a reference set $RefSet_{ij}$ (the middle part) is constructed that contains a number of reference targets appearing in both cameras. When comparing track T_1 in C_i with tracks T_2 and T_3 in C_j, by using their color histograms directly, T_3 are 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 more darker than its actual appearance. Instead of comparing directly, each input track is described by all the reference targets. The description is a vector of ordered similarities, and each similarity is generated by comparing the input track with one reference target. 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. Notice that, both the input tracks and the reference targets have multiple appearance instances that cover all the appearance changes of corresponding targets in a particular camera, this enables us to handle within camera illumination variation. After representing T_1, T_2 and T_3 by the reference set, it is clear that T_1 and T_2 are more similar than T_1 and T_3.

Let Σ be the set of all possible association matrixes, the task of multi-target tracking across cameras is formulated as the following optimization problem:

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

This assignment problem can be solved by Hungarian algorithm [17] 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. Normally, there is a 50% overlap for the neighboring two time sliding windows. Instead of using the cost matrix S directly, we use the argmented matrix S' (details for the augmented matrix can be found in [4]) as the input for the Hungarian algorithm. This 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 section, we present the reference set based appearance model in detail.

B. Reference Set Based Appearance Model

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 targets $R = \{R_1, R_2, ..., R_m\}$ that appear in both C_i and C_j. The tracks for all the reference targets appear in C_i form $RefSet^i_{ij}$, and the tracks for all the reference targets appear in C_j form $RefSet^j_{ij}$, 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 are not computed by comparing T_p and T_q directly. Instead, T_p is compared with all the tracks in $RefSet^i_{ij}$ and T_q is compared with all the tracks in $RefSet^j_{ij}$, and their similarity 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 have high similarities with the same set of reference targets. Otherwise, they are more likely to be tracks that contain different targets.

In order to handle within camera illumination variation, each track is further segmented into small subtracks according to a pre-defined subtrack length (e.g., 5 frames) so that detections in each subtrack are visually very similar. After track segmentation, 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 all the appearance changes of that target in a certain 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^k_a denotes the k-th subtrack in track T_a, $simi(t_x, t_y)$ be the similarity of two subtracks (described in the following part), 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:

$$Simi(T_a, T_b) = \frac{1}{N_a} \sum_{i=1}^{N_a} max(\{simi(t^i_a, t^j_b), j \in [1, N_b]\}) \tag{5}$$

Namely, 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.

In the reference set, each reference target may have several tracks in the same camera (e.g., walking towards and away from the camera). The similarity between a track T_l and a reference target R_n is the maximum of the similarities of T_l and all the tracks for R_n. This lays the strength of our reference set based appearance model - the 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 targets. In other words, each reference target 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 targets enables us to compare the similarity of these two tracks. In addition, the reference set based appearance model does not require any extra training process. Besides variation in illumination conditions, difference in poses are also taken care of by the various appearance instances in each reference target.

After comparing tracks T_p and T_q with each reference target in its corresponding reference set, we get two vectors of ordered similarities, 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 Ref_{ij}, the similarity of T_p and T_q is computed by the Kendall tau Correlation Coefficient [18], and is further normalized to the range of $[0, 1]$. In order to get the appearance model, we use the negative logarithm function to calculate the cost, as defined in Equ. (6):

$$Appr(T_p, T_q) = -log(\tau'(Ref_{ij}^i(T_p), Ref_{ij}^j(T_q))) \quad (6)$$

where $\tau'(\cdot)$ is the normalized Kendall tau Correlation Coefficient.

C. Soft-biometrics Fusion and Subtrack Similarity

The soft-biometric features extracted from a detection response are shown in Table I, where the potential values for each feature are also listed. These soft-biometric features can be categorized into three types: symbolic, scalar-valued, and vector-valued. A confidence level which scales from 0 to 1 is associated with each feature to indicate the prediction confidence. In order to generate concise representation for a given subtrack, we design a fusion method that can combine common soft-biometric features extracted from several detections into a single one. In the remainder of this paper, fn represents the feature name, $fval$ is the feature value, and fc is the confidence level.

For binary symbolic features, the sum of confidence levels of all potential values is equal to 1. Thus, given the confidence level of one potential value, the confidence level for the other potential value can be inferred. When fusing symbolic features, the averaged confidence level for each potential value is computed and the one with the highest score is selected as the fused confidence level, and the corresponding value is the fused feature value. For scalar-valued and vector-valued features, the fused value is the weighted sum of all $fval$s, where the weights are the corresponding fc, and normalization is carried out to make the result lie in the range of $[0, 1]$. The fused confidence level is the average of all fcs.

Name	Value	Type
HairColor	Light, Dark	Symbolic
SkinColor	Caucasian, Non_Caucasian	Symbolic
Height	Centimeters	Scalar
Weight	Kilograms	Scalar
BodyColor	1-D probability distribution	Vector
TorsoColor	1-D probability distribution	Vector
LegsColor	1-D probability distribution	Vector

TABLE I. SOFT-BIOMETRICS EXTRACTED FROM DETECTION.

After soft-biometrics fusion, each subtrack is represented by a single set of soft-biometric features. The similarity of two subtrack is computed based on the similarity between common features for each feature type (symbolic, scalar, vector). For the symbolic features (HairColor, SkinColor), if the symbolic value of the two features are the same then the similarity is the average of the two confidence levels. If the symbolic values are dissimilar, then the similarity is the maximum confidence level, as defined in Equ. (7).

$$sim_1(fval_1, fval_2) = max(fc_1 \times (1 - fc_2), (1 - fc_1) \times fc_2) \quad (7)$$

For the scalar-valued features (Height and Weight), we assume that the feature values are from a normal distribution with parameters μ and σ^2. As the height accuracy is $\pm 12.7cm$ and the weight accuracy is $\pm 9kg$ (learned by analyzing soft-biometrics extracted from previous data), we define the standard deviation so that for the height $P([fval - 12.7, fval + 12.7]) = 80\%$ and for the weight $P([fval - 9, fval + 9]) = 80\%$. For the accumulated probability to be equal to 80% the range should be $(\mu - 1.28\sigma, \mu + 1.28\sigma)$. Thus, the standard deviations are equal to $1.28\sigma = 12.7$ for height and $1.28\sigma = 9$ for weight. The similarity score sim_2 is defined as:

$$sim_2(fval_1, fval_2) = 1 - \sqrt{1 - e^{\frac{-(fval_1 - fval_2)^2}{8\sigma^2}}} \quad (8)$$

For the vector-valued features (BodyColor, TorsoColor, LegsColor), the Bhattacharyya Coefficient [19] is used to measure the similarity sim_3, which approximates the amount of overlap between two probability distributions, as given in Equ. (9):

$$sim_3(fval_1, fval_2) = \sum_{i=1}^{n} \sqrt{fval_{1i} \times fval_{2i}} \quad (9)$$

IV. EXPERIMENTAL RESULTS

In order to evaluate the proposed model, five cameras (four indoor and one outdoor) are used to establish the desired non-overlapping setting, the topology is presented in Fig. 3 and sample frames from each camera is shown in Fig. 1. All the videos are taken during the same time period and each video is about 20 minutes in duration. The resolution is 704×480,

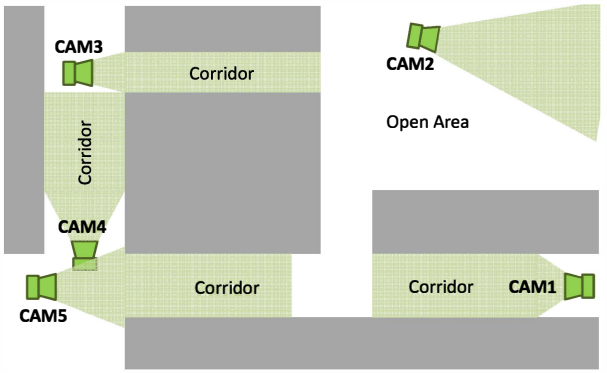

Fig. 3. Topology for cameras used in the experiments.

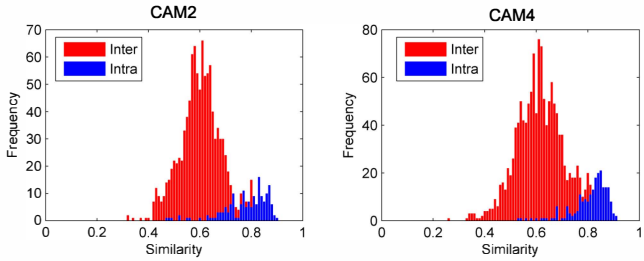

Fig. 4. Histograms of inter-class and intra-class similarities from two testing videos. Best viewed in color. "Intra" stands for tracks that contain the same target and "Inter" stands for tracks that contain different targets.

the frame rate is 20fps. The number of participants involved in each video ranges from 7 to 10. This setting is very challenging for multi-camera tracking due to following reasons. (1) The outdoor camera view contains intense 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. (2) The number of camera involved is greater than most of the previous work that normally use 2-3 cameras [6] [7]. In order to construct the reference set another set of data is used. It is collected under the same setting but with participants either not included in the testing data or they are included in the testing data but with very different clothes. There are about 10 reference targets in each reference set.

A. *Soft-biometrics verification*

As we use soft-biometrics to represent each target, the quality of soft-biometrics and the soft-biometrics similarity measurement are crucial for tracking. In the verification, for each video (captured in a single camera) we compute the similarity between any pair of tracks based on the proposed method, and these similarities are categorized into intra-class (tracks from the same target) and inter-class (tracks from different targets). The histograms for each category are plotted. Two sample plots are shown in Fig. 4. The plots suggest that most intra-class similarities are larger than most inter-class similarities and with a single threshold these two classes can be coarsely separated. Therefore, the soft-biometrics extracted from the same target have high degree of consistency.

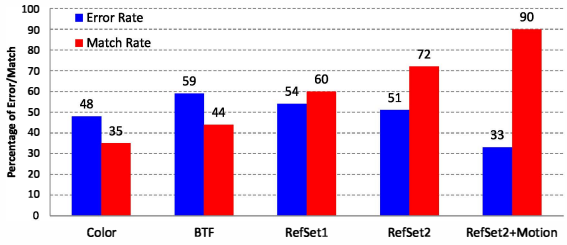

Fig. 5. Comparison of the proposed methods and other baseline models.

B. *Tracking Results*

In this evaluation, our main focus is to associate tracks that contain the same target in different camera views given certain spatial-temporal constraints. We applied our reference set based appearance model with soft-biometric features (RefSet2) on the testing data. Three baseline models are presented for comparison: (1) Use the Bhattacharyya distance of holistic color histograms directly to measure the appearance similarity (Color). (2) Generate the appearance model based on the BTF model in [7] (BTF). (3) Our proposed reference set based appearance model with only holistic color histograms as features (RefSet1).

In all our experiments, the length of subtrack is set to 10 frames. For each model, various thresholds (ranges from 0.2 to 0.6) are tested for the augmented cost matrix, and the best result is chosen. We hand labeled the ground-truth which consists of 220 track associations (there are 368 single camera tracks in total). Two metrics are used for evaluation, as defined in Equ. (10). The comparison is presented in Fig. 5.

$$ErrorRate = \frac{Error}{N_{result}}, \; MatchRate = \frac{Match}{N_{GT}} \tag{10}$$

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.

It can be observed that when using the reference set based appearance model with the soft-biometric features, we achieve the highest match rate and the lowest error rate compared with all the baseline models. Compared with BTF, the RefSet2 model increases the match rate by almost 30% and reduces the error rate by about 10%. 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 the other soft-biometric features are complementary to color histograms and reduce ambiguities, as they capture the appearnce information that is overlooked by color histograms. 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 the appearance information only.

As another kind of clue, motion information plays an important role in multi-target tracking. When a motion model that measures the walking direction of the target is integrated into the tracking system (RefSet2+Motion), the error rate is greatly reduced to about 30%. Also, with motion information

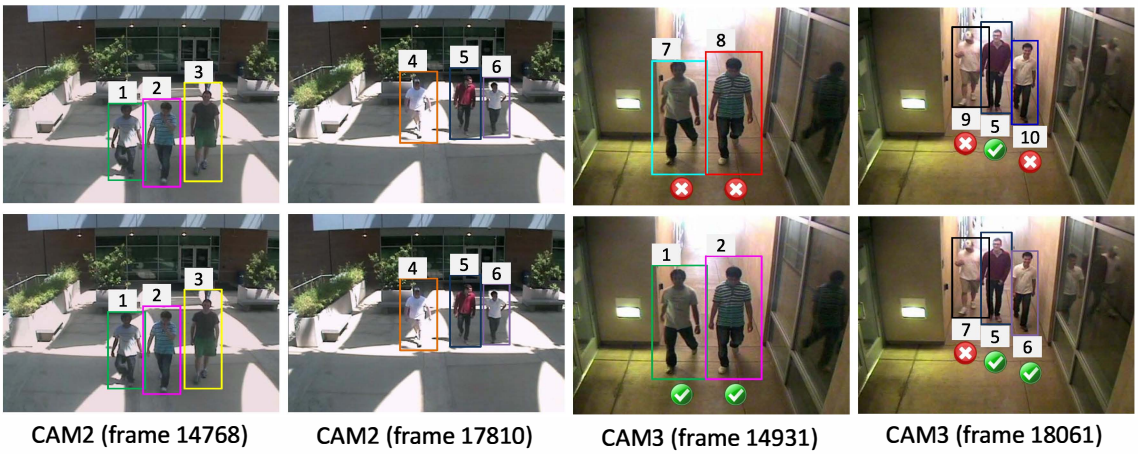

Fig. 6. Example tracking results. Best viewed in color. The first row is the results obtained by using BTF in [7], the second row is the results by the proposed reference set based appearance model (RefSet2). With the reference set, our method is able to match most of the targets where there exist drastic within camera and across camera illumination variations. The method in [7] fails to associate tracks that contain the same target under challenging conditions.

our proposed method can correctly associate 90% track pairs in the ground-truth, which further demonstrates the effectiveness of our method. Comparison between BTF and RefSet2 on some challenging cases are presented in Fig. 6, which validates the robustness of our method.

V. Conclusions

In this paper, we propose a novel reference set based appearance model for multi-target tracking in a camera network with non-overlapping FOVs. The proposed appearance model is easy to implement with zero parameters and requires no additional training process, yet provides promising results. The experimental results demonstrate the superiority of the combination of reference set based appearance model and soft-biometric features over other baseline models on a challenging real-world video data. This data set will be made publicly available in the future.

Acknowledgment

This work was supported in part by NSF grants 0641076 and 0905671. The contents and information do not reflect the position or policy of the U.S. Government.

References

[1] Y. Zhu, N. Nayak, and A. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 91–101, 2013.

[2] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behavior-recognition algorithms," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 11, no. 1, pp. 206–224, 2010.

[3] T. D'Orazio and G. Cicirelli, "People re-identification and tracking from multiple cameras: A review," in *19th IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 1601–1604.

[4] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[5] A. Gilbert and R. Bowden, "Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity," in *European Conference on Computer Vision*, vol. 3952, 2006, pp. 125–136.

[6] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *Proceedings of the British Machine Vision Conference*, 2008.

[7] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, vol. 109, pp. 146 – 162, 2008.

[8] 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 *Fifth ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2011, pp. 1–6.

[9] T. D'Orazio, P. Mazzeo, and P. Spagnolo, "Color brightness transfer function evaluation for non overlapping multi camera tracking," in *Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, 2009, pp. 1–6.

[10] 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 *Sixth International Conference on Distributed Smart Cameras (ICDSC)*, 2012, pp. 1–6.

[11] F. Schroff, T. Treibitz, D. Kriegman, and S. Belongie, "Pose, illumination and expression invariant pairwise face-similarity measure via doppelgänger list comparison," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2494–2501.

[12] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 497–504.

[13] L. An, M. Kafai, S. Yang, and B. Bhanu, "Reference-based person re-identification," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2013.

[14] A. Dantcheva, C. Velardo, A. D'angelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification : New trends and challenges," *Mutimedia Tools and Applications*, 2010.

[15] A. K. Jain and U. Park, "Facial marks: Soft biometric for face recognition," in *ICIP*, 2009.

[16] D. A. Reid and M. Nixon, "Using comparative human descriptions for soft biometrics," in *International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–6.

[17] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, 1957.

[18] H. Abdi, *Kendall Rank Correlation*. SAGE Publications, Inc., 2007, pp. 509–511.

[19] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, USA: Wiley-Interscience, 1991.