CAN: Composite Appearance Network and a Novel Evaluation Metric for Person Tracking

Neeti Narayan, Nishant Sankaran, Srirangaraj Setlur, Senior Member IEEE, and Venu Govindaraju, Life Fellow, IEEE

Abstract—Tracking multiple people across multiple cameras is an open problem. It is typically divided into two tasks: (i) single-camera tracking (SCT) - identify trajectories in the same scene, and (ii) inter-camera tracking (ICT) - identify trajectories across cameras for real surveillance scenes. Many methods cater to SCT, while ICT still remains a challenge. In this paper, we present a feature aggregation architecture called Composite Appearance Network (CAN) to address the above problem. The key structure of this architecture is called EvalNet that pays attention to each feature vector and learns to weight them based on gradients it receives for the overall template for optimal re-identification performance. We demonstrate the efficiency of our approach with experiments on the challenging multi-camera tracking dataset, DukeMTMC. We also survey existing tracking measures and present an online error metric called “Inference Error” (IE) that provides a better estimate of tracking/re-identification error, by treating SCT and ICT errors uniformly.

Index Terms—Person re-identification, Person tracking, Feature aggregation, Error metric

1 INTRODUCTION

PERSON tracking is one of the fundamental problems in the field of computer vision. Determining multi-camera person trajectories enable applications such as video surveillance, sports, security, behavior analysis and anomaly detection. Maintaining the correct trajectory of a person across cameras is difficult because of occlusion, illumination change, background clutter, and blind-spots due to non-overlapping cameras often placed far apart to reduce costs. Methods like [11] focus on target tracking using a low-level handcrafted feature. Although they achieve good tracking performance, they are still inefficient in solving the mentioned obstacles and are limited to scenarios where targets are known a-priori. Some model people entry and exit points on the ground or image plane explicitly [3], [7]. Few others [2], [26] exploit data fusion methods from partially overlapping views or completely overlapping views as a pre-processing step.

Person re-identification for the purpose of estimating the similarity of person images which deals with data comprised of only image pairs, one for the probe and one for the gallery, is relatively straightforward with the use of metric learning methods [13], [25] and feature representation learning [12]. However, in unconstrained tracking (Figure 1), person tracklet contains multiple images and therefore requires a way to fuse features/attributes to a single feature vector representative of the tracklet. We design a spatio-temporal attention model, Composite Appearance Network (CAN), where the network looks at each individual feature vector of a trajectory in the gallery and predicts how important it is to be a part of the final representative feature vector by incorporating metadata such as person bounding box and camera information. We also discuss the Inference Error (IE) evaluation criterion to measure the performance of online person tracking.

The main contributions of this paper include:

- Finding the optimization function that best exploits the variances between the appearance features (extracted from a CNN or any other embedding system) using additional metadata. This would result in optimal aggregation weights for pooling.
- Study the performance measures of a tracking system and demonstrate how they are different when evaluating/characterizing online and real-time tracking performance. We formally introduce the “Inference Error” metric and emphasize on de-duplication.

2 RELATED WORK

Multi-person pose estimation in videos has been studied in [10]. They jointly perform pose estimation and tracking, but is limited to single camera videos. For multi-camera object tracking, input detections can be regarded as a graph and weight edges between nodes (detections) based on similarity [4]. When merging tracklets, trackers are dependent on the feature extractor/representation technique to avoid high false negatives. For example, some trackers adopt LOMO [13] appearance features and hankel matrix based IHTLS [6] algorithm for motion feature. A hard mining scheme and an adaptive weighted triplet loss is proposed in [27] to learn person appearance features from individual detections.

In [21], an online method for tracking is proposed by using Long Short-Term Memory (LSTM) networks and multiple cues such as appearance, motion, and interaction. However, this solution is for single camera tracking and there is no proof of how well the system scales for multi-camera environment. In [17], authors address the multi-person multi-camera tracking problem by reformulating it as a pure re-identification task. The objective is to minimize misassociations at every timestep. For this reason, the authors propose a new evaluation metric called the “Inference Error” (IE). The method, however, does not encode long-term spatial and temporal dependencies which are critical for correcting data association errors or recovering from occluded state. We discuss in detail the advantages of using IE and how it adapts to both online and offline tracking systems. An online tracking method, which is an extension of [17], is presented in [18]. LSTM-based space-time tracker is developed using history of location and visual features to learn from past association errors and make
of trajectories, we propose to use orthogonal attributes such as metadata, that are learnt using different target objectives.

The framework proposed in this paper is inspired from [22]. The authors propose set based feature aggregation network (FAN) for the face verification problem. By generating representative template features using metadata like yaw, pitch and face size, their system could outperform traditional pooling approaches. Instead of building a siamese network, we focus on learning the pooled feature for gallery template (trajectory) while coupling the probe metadata beside gallery metadata. This way, we derive each trajectory’s representation by considering the impact or influence of the probe detection in the context of data association.

3 Our Approach

The input to the multi-person multi-camera tracking (MPMCT) system is a set of videos $V = \{v_1, v_2, ..., v_k\}$, where $k$ is the number of cameras. The output is a set of trajectories $T = \{t_1, t_2, ..., t_n\}$ across $k$ cameras. Here, we take the tracking solution a step further by proposing a simple feature aggregation approach that can well distinguish identical pair of features from non-identical ones to maintain a margin between within-identity and between-identity distances.

3.1 Preliminary Appearance Cues

Given the input person detections in frame-level $d = d^1, d^2, ..., d^T$ in a single camera or across multiple cameras over a time period $T$, we extract the appearance feature maps set

$$A = A^1, A^2, ..., A^T$$

using DenseNet [9]. We train DenseNet on 25% of the training set (called trainval) of DukeMTMC dataset. Features of length 1024 are extracted from the last dense layer. The DenseNet used has 4 blocks with a depth of 121, compression/reduction of 0.5 and growth rate of 32. We run for 15 epochs with a learning rate starting at 0.1 and we train using stochastic gradient descent with batch size 96 and momentum 0.9. As the number of remaining epochs halves, we drop learning rate by a factor of 10 and drop by a factor of 2 at epoch 14.

3.2 CAN Architecture

The Composite Appearance Network (CAN) is shown in Figure 2. The main component of CAN is EvalNet which evaluates the quality of every feature vector (FV) as being a part of the template features using metadata like yaw, pitch and face size, that are learnt using different target objectives. The terms “gallery” and “probe” in the tracking scenario apply to tracklets identified dynamically as time progresses and a new detection to be associated with an existing tracklet respectively. The gallery FV list is one of the inputs to the network. It is comprised of a list of appearance attributes corresponding to a trajectory in the gallery. Gallery metadata list is the other input to the network. Metadata for person detections include bounding box information ($BB_{x_i}, BB_{y_i}, BB_{w_i}, BB_{h_i}$) as well as other external details such as camera number ($cam_{id}$). The probe metadata is concatenated with gallery metadata, which in-turn constitutes the gallery metadata list. Each vector in the list is of length 10. In this way, the probe’s influence is also taken into account for learning weights for aggregation. Probe FV is the probe detection’s appearance attribute. Consider $n$ detections in a trajectory and corresponding appearance attributes of length 1024. Then, the gallery FV list is of shape $(n \times 1024)$, probe FV has shape $(1 \times 1024)$ and metadata is $(n \times 10)$. 
EvalNet is a Fully Connected Network (FCN) with 4 fully connected layers, batch normalization and ReLU as the activation function. This FCN block produces weights $w_k$ such that $\sum_{k=1}^{n} w_k = 1$, where $k$ is the $k^{th}$ feature vector and $n$ is the number of detections belonging to a tracklet. These weight predictions are then applied on the corresponding appearance features to obtain the aggregated appearance feature for every tracklet. The similarity between probe and aggregated FV is obtained using cosine similarity. The cosine similarity loss layer and softmax loss layer are optimized against the match label and class label respectively. During real-time tracking, we do not use entire CAN and instead, use only EvalNet to obtain aggregated FV given appearance attributes and metadata information for every tracklet.

### 3.3 CAN Training

Let $g_t$ and $M_t$ be the $t^{th}$ gallery FV and corresponding metadata vector in a trajectory. If $F$ denotes the final composite or aggregated appearance FV, it can be defined as:

$$ F = \sum_{i=1}^{n} H_{\theta}([g_t, M_t])g_t $$

where $H_\theta$ is the EvalNet function on gallery FV and metadata, parameterized by $\theta$, predicting a weight that evaluates the relative importance of the feature vector using accompanying metadata. $n$ is the number of feature vectors/detections in the trajectory. We use orthogonal information such as probe metadata and gallery metadata to yield additional context that can help discriminate appearance features of an identity.

If $P$ is the probe FV, then the mean-squared loss function is defined as:

$$ L_{mse}^{(i,j)} = \left( \frac{P_i, F_j}{\|P_i\|^2 \|F_j\|^2} - Y(P_i, F_j) \right)^2 $$

and the categorical cross-entropy loss function is defined as:

$$ L_{cce}^{(j)} = -\sum_{k=1}^{K} y_k^{(j)} \log \hat{y}_k^{(j)} $$

where $Y(F_q, P) \in \{0, 1\}$ is the ground-truth match label, $K$ is the number of classes (unique identities), $y_k$ is the ground-truth class label of the one-hot encoded vector $y$ and $\hat{y}$ is the normalized probability assigned to the correct class. Given gallery-probe vectors, the objective of CAN is to find the optimal parameters $\theta$ that minimize the following cost function:

$$ J(\theta) = \frac{1}{m_p m_g} \frac{1}{m_p m_g} \sum_{i=1}^{m_p} \sum_{j=1}^{m_g} L_{mse}^{(i,j)} + L_{cce}^{(j)} $$

where $m_p$ is the number of examples in probe set and $m_g$ is the number of examples in gallery set.

**Batch processing:** Since each trajectory may be comprised of a variable number of person detections, we would require CAN to be trained on a single gallery trajectory in every iteration as making batches of trajectories would be difficult. However, we work around this problem by coupling an additional input - trajectory index $t_i$, that corresponds to the unique identity in the batch that each person (features and metadata) would be mapped to. This allows us to gather multiple sets of trajectories as a batch, enabling CAN to converge faster. The aggregated trajectory representation is computed using only the corresponding features as indexed by $t_i$.

### 3.4 Tracking

We also include the motion property of each individual in our framework for within-camera tracking. This would help address one key challenge - handling noisy detections - because velocities of an individual can be non-linear due to occlusion or detections being noisy. Here, we propose to use a simple method to aggregate detections into single camera trajectories and further merge these trajectories using multi-camera CAN framework.

#### 3.4.1 Single-camera Motion and Appearance Correlation

Single-camera trajectories are computed online by merging detections of adjacent frames in a sliding temporal window with an intersection of 50%. Let $i$ denote a previous detection and $j$ denote a current detection at timestep $t$. If two bounding boxes, $B_i$ and $B_j$, detected in neighboring windows are likely to belong to one target, the factors that effect this likelihood include: (i) appearance similarity and (ii) location closeness and is defined as $score_{ij}$ computed as:

$$ score_{ij} = T - Dis(A_i, A_j) - f(B_i, B_j) $$

where $Dis(A_i, A_j)$ is considered as the appearance similarity between two detections - given by cosine distance formulation $Dis(A_i, A_j) = 1 - \frac{\|A_i\|^2 \|A_j\|^2}{\|A_i\|^2 \|A_j\|^2}$ and $f(B_i, B_j)$ is defined as:

$$ f(B_i, B_j) = \begin{cases} 0, & \text{if } IoU(B_i, B_j) \geq 0.5. \\ \infty, & \text{otherwise.} \end{cases} $$

The output of this step will include a set of single-camera trajectories $T_s = \{s_{k1}, s_{k2}, ..., s_{kn}\}$ from all $k$ cameras. The number of trajectories $n$ is not a constant and varies from one camera to another.

#### 3.4.2 Multi-camera

Once we have associated detections to tracklets or trajectories, we use CAN as the next step for tracking. Tracking in multi-camera environment is challenging owing to variation in lighting conditions, change in human posture, presence of occlusion or blind-spots. To handle all the above challenges in a principled way, we use CAN to weaken the noisy features corresponding to a trajectory and narrow down variances for an identity, thus making the trajectory representation more discriminative.

The input during testing is the set of trajectories $T_s$ and the output of this step is an association matrix, whose scores are determined using CAN. For every pair of trajectories, appearance feature maps and metadata sequences of one are considered as gallery FV and gallery metadata respectively; while the other forms the probe FV. Probe metadata is appended to each occurrence of gallery metadata. Using EvalNet, we obtain the aggregated gallery FV that corresponds to gallery trajectory’s representation. The association score between aggregated FV and probe FV is determined using cosine similarity. In this manner, we perform pairwise trajectory feature similarity comparisons. Final matching/ associations are based on a simple greedy algorithm [17] which is equivalent to selecting the largest association score sequentially. This controls proliferation of fragmented identities i.e. a subject is associated with one identity only.

### 4 Experiments and Results

#### 4.1 Performance Measures

In this section, we survey the many existing tracking measures and discuss their use cases. There are two criteria crucial to a MPMCT
Fig. 2: Composite Appearance Network architecture

system: (i) single-camera (within-camera) tracking module and (ii) inter-camera (handover) tracking module. Few metrics evaluate the two separately, some handle both simultaneously. Also, based on the application, measures can be classified into (i) event-based: count how often a tracker makes mistakes and determine where and why mistakes occur, and (ii) identity-based: evaluate how well computed identities conform to true identities.

4.1.1 Multi Object Tracking Accuracy (MOTA)
MOTA (Multiple Object Tracking Accuracy) is typically used to evaluate single camera, multiple person tracking performance. It is defined as:

\[
\text{MOTA} = 1 - \frac{\text{FN} + \text{FP} + \text{Fragmentation}}{\text{T}} 
\]

where \(\text{FN}\) is the number of false negatives i.e. true targets missed by the tracker, \(\text{FP}\) is the number of false positives i.e. an association when there is none in the ground truth, Fragmentation is the ID switches, and \(\text{T}\) is the number of detections. However, MOTA penalizes detection errors, under-reports across-camera errors \([4]\) and can be negative due to false positives.

4.1.2 Identity-based Metrics
We evaluate our framework on DukeMTMC using measures proposed in \([19]\): Identification F-measure \((\text{IDF}_1)\), Identification Precision \((\text{IDP})\) and Identification Recall \((\text{IDR})\). Ristani et al. \([19]\) propose to measure performance by how long a tracker correctly identifies targets, as some users may be more interested in how well they can determine who is where at all times. Few critical aspects such as where or why mismatches occur are disregarded. \(\text{IDP}\) is the fraction of computed detections that are correctly identified. \(\text{IDR}\) is the fraction of ground truth detections that are correctly identified. \(\text{IDF}_1\) is the ratio of correctly identified detections over the average number of ground-truth and computed detections.

4.1.3 Multi-camera Object Tracking Accuracy (MCTA)
Another existing multi-camera evaluation metric is multi-camera object tracking accuracy (MCTA) \([4]\). Here, both SCT and ICT are considered equally important in the final performance measurement. So, all aspects of system performance are condensed into one measure and defined as:

\[
\text{MCTA} = F_1 \cdot \left( 1 - \frac{M^s}{T^s} \right) \left( 1 - \frac{M^i}{T^i} \right) 
\]

where the first term \(F_1\)-score measures the detection power (harmonic mean of precision and recall). The second term penalizes within-camera mismatches \((M^s = \text{ID-switches})\) and normalized using true positive detections \((T^s)\). The final term penalizes inter-camera mismatches \((M^i)\) normalized by true inter-camera detections \((T^i)\). MCTA factors single-camera mismatches, inter-camera mismatches, false-positives and false negatives through a product instead of adding them. In addition, error in each term is multiplied by the product of other two terms which might drastically change the performance measure.

4.1.4 Inference Error
Traditional biometric measures \([15]\) such as FMR (False Match Rate), FNMR (False Non-match Rate) assume that the occurrence of an error is a static event which cannot impact future associations. However, in a tracking by re-identification system, the reference gallery is dynamically evolving, as new tracks are created (following “no association” outcomes) or existing tracks are updated (following “association” outcomes).

For this reason, we propose the “Inference Error” measure for tracking by continuous re-identification evaluation and is defined as:

\[
E = \frac{1}{T} \sum_{t=1}^{T} \frac{\text{fail}_t}{\text{Det}_\text{curr}} 
\]

where \(\text{fail}_t\) is the number of misassociations at time \(t\) and \(\text{Det}_\text{curr}\) is the number of current detections. The inference error is a real-time evaluation paradigm which measures how often a target is incorrectly associated. We normalize it by the number of detections. It handles multiple simultaneous observations in any given time and also handles identities reappearing in the same camera or in different cameras. In \([5]\), similar metrics such as false dynamic match (FDM) and false dynamic non-match (FDNM) are proposed for biometric re-identification systems. We do not claim
that one measure is better than the other, but only suggest that different error metrics are suited for different applications.

The main intuition behind counting the number of misassociations at every timestep is to reduce creating duplicate trajectories for an identity. Consider one true identity ID1 as illustrated in Figure 3. Let time be represented in the horizontal direction. In case (i), a tracker mistakenly identifies two trajectories (Traj1 and Traj2) for the same person. While in case (ii), three trajectories (Traj1, Traj2 and Traj3) are mistakenly identified for the same person. If Traj1 covers 80% of ID1’s path, identification measure would charge 20% of the length of ID1 to each of the two cases. However, inference error would assign a higher penalty to case (ii) owing to multiple fragmentations. Given a real-time large-scale tracking scenario with multiple simultaneous observations across multiple cameras, system designers and security professionals will want a tracker that maintains the identity of a person and effectively updates the reference gallery without growing exponentially. So, identifying where and when trajectories are broken is essential for de-duplication. De-duplication is analogous to tracking framework, with the advantage of describing duplication errors (i.e., instances of incorrect “non-match” matching outcomes resulting in duplicate data, that would otherwise maintain the same identity) better.

4.2 Experiments on DukeMTMC

4.2.1 DukeMTMC Dataset

It consists of more than 2,000 identities, with over 2 million frames of 1080p, 60fps video, approximately 85 minutes of video for each of the 8 cameras [19]. We report results across all 8 cameras to demonstrate the efficiency of our approach for MPMCT. We use the ground-truth available for the training set (called trainval) of DukeMTMC for evaluation. Only 25% of this set (we call this ‘net-train’) is used for CNN and CAN training and the remaining 75% (‘net-test’) is used for testing.

4.2.2 Comparison with Baseline

The reference (baseline) approach [19] for DukeMTMC is based on correlation clustering using a binary integer program (BIPCC). The task of tracking is formulated as a graph partitioning problem. Correlations are computed from both appearance descriptors and simple temporal reasoning.

4.2.3 Single-camera Multi-person Tracking (SCT) Results

Our system aggregates detection responses into short tracklets using motion and appearance correlation as described in Section 4.1. These tracklets are further aggregated into single camera trajectories using CAN. In Table 1 and Table 2, SCT performances are reported on each one of the cameras. We compare quantitative performance of our method with the baseline. For a fair comparison, the input to our method are the person detections obtained from MOTChallenge DukeMTMC. The average performance (IDF1) over 8 cameras is 96.58 using our approach and 89.18 using BIPCC. An improvement of over 8% is achieved in IDF1, approximately 6% in IDP and 10% in IDR metrics. Thus, the results demonstrate that the proposed CAN approach improves within-camera tracking performance.

4.2.4 Inter-camera Multi-person Tracking (ICT) Results

We use CAN for ICT. This demonstrates the ability of CAN to scale to SCT as well as ICT. We compare our results with BIPCC using Inference Error metric and Identification scores. Tracking performance across all 8 cameras is presented in Table 3. An improvement of over 20% is achieved in IDF1, 25% in IDP and 17% in IDR metrics. Thus, the results demonstrate that the proposed CAN approach improves multi-camera tracking performance.

5 Conclusion

For representation, instead of using one random feature vector, we can use a set of feature vectors available for a trajectory. Using CAN, we can measure the relative quality of every feature map (spatial) in a trajectory for aggregation based on the location and camera information (temporal). Moreover, it can be incorporated into any existing re-identification feature representation framework to obtain an optimal template feature for SCT, ICT or both.

In addition, we study the error measures of a tracking system. We define a new measure that emphasize on tracker mismatches at every timestep. A traditional re-identification system aims at matching person images often on a fixed gallery and cite performance using CMC [16] and ROC [8] curves. However, with a dynamically growing reference set (gallery), it is important to investigate the mismatches. We hope this survey will benefit researchers working on video surveillance and tracking problems.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant IIP #1266183.

References

[1] A. Babenko and V. Lempitsky. Aggregating local deep features for image retrieval. In ICCV, 2015.
[2] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. IEEE transactions on pattern analysis and machine intelligence, 2011.
[3] Y. Cai and G. Medioni. Exploring context information for inter-camera model-based approach for multicamera object tracking. IEEE Transactions on Circuits and Systems for Video Technology, 2017.
[4] B. DeCann and A. Ross. Modelling errors in a biometric re-identification system. IET Biometrics, 2015.
[5] C. Dicle, O. I. Camps, and M. Sznaier. The way they move: Tracking and machine intelligence using k-shortest paths optimization. IEEE transactions on pattern analysis and machine intelligence, 2011.
[6] W. Chen, L. Cao, X. Chen, and K. Huang. An equalized global graph model-based approach for multicamera object tracking. IEEE Transactions on Circuits and Systems for Video Technology, 2017.
[7] A. Gilbert and R. Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In ECCV, 2006.

Fig. 3: Duplication effect
TABLE 1: SCT ID score comparison

| Method | Cam 1 | Cam 2 | Cam 3 | Cam 4 |
|--------|-------|-------|-------|-------|
| IDP    | 92.8  | 95.8  | 94.2  | 93.2  |
| IDT    | 94.8  | 88.8  | 94.1  | 91.0  |
| IDR    | 97.6  | 98.1  | 88.2  | 93.1  |
| Ours   | 94.5  | 98.3  | 96.4  | 92.6  |

TABLE 2: SCT ID score comparison

| Method | Cam 5 | Cam 6 | Cam 7 | Cam 8 |
|--------|-------|-------|-------|-------|
| IDP    | 96.0  | 81.5  | 88.1  | 80.1  |
| IDT    | 72.0  | 75.9  | 94.8  | 91.6  |
| IDR    | 93.2  | 93.2  | 93.2  | 93.2  |
| Ours   | 95.6  | 92.2  | 93.8  | 95.6  |

TABLE 3: ICT (across all 8 cameras) performance

| Method | Identification Score | Our Measure |
|--------|----------------------|-------------|
| IDP    | 76.7                 | 60.0        |
| IDT    | 90.1                 | 60.5        |
| IDR    | 93.2                 | 1.3         |
| Ours   | 96.6                 | 82.9        |

[8] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2008.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In CVPR, 2017.

[10] U. Iqbal, A. Milan, and J. Gall. Pose-track: Joint multi-person pose estimation and tracking. In CVPR, 2016.

[11] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. IEEE transactions on pattern analysis and machine intelligence, 2012.

[12] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013.

[13] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local normal occurrence representation and metric learning. In CVPR, 2015.

[14] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In CVPR, 2017.

[15] A. J. Mansfield and J. L. Wayman. Best practices in testing and reporting performance of biometric devices. CIES, Nat. Phys. Lab., Teddington, U.K., NPL Tech. Rep. CMSC 14/02, 2002. Available: http://www.npl.co.uk/upload/pdf/biometrics_bespapr_v2_1.pdf.

[16] N. Martinel and C. Micheloni. Re-identify people in wide area camera network. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, 2012.

[17] N. Narayan, N. Sankaran, D. Arpit, K. Dantu, S. Setlur, and V. Govindaraju. Person re-identification for improved multi-person multi-camera tracking by continuous entity association. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, 2017.

[18] N. Narayan, N. Sankaran, S. Setlur, and V. Govindaraju. Re-identification for online person tracking by modeling space-time continuum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018.

[19] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In ECCV, 2016.

[20] E. Ristani and C. Tomasi. Tracking multiple people online and in real time. In ACCV, 2014.

[21] A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In ICCV, 2017.

[22] N. Sankaran, S. Setlur, S. Tulyakov, and V. Govindaraju. Metadata-based feature aggregation network for face recognition. In ICB, 2017.

[23] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In ICCV, 2017.

[24] J. Yang and P. Ren. Neural aggregation network for video face recognition. In CVPR 2017.

[25] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In CVPR, 2016.

[26] S. Zhang, Y. Zha, and A. Roy-Chowdhury. Tracking multiple interacting targets in a camera network. Computer Vision and Image Understanding, 2015.

[27] E. Ristani, and C. Tomasi. Features for Multi-Target Multi-Camera Tracking and Re-Identification. In CVPR, 2018.

[28] F. Pernici, F. Bartoli, M. Bruni, and A. D. Bimbo. Memory Based Online Learning of Deep Representations From Video Streams. In CVPR, 2018.

[29] J. Zhu1, H. Yang1, N. Liu, M. Kim, W. Zhang, and M. Yang. Online multi-object tracking with dual matching attention networks. In ECCV, 2018.

Nishant Sankaran received the B.E. degree from Calicut University, Kerala, India in 2010. He has worked as a software engineer at Huawei Technologies for 3 years and as a computer vision engineer at Nervve technologies for a year. He is currently pursuing the Ph.D. degree at the University at Buffalo, The State University of New York. His research interests include deep learning, fusion of single and multi modal systems with applications in face recognition and biometrics.

Srirangaraj Setlur (M01, SM07) is a Principal Research Scientist at the Center for Unified Biometrics and Sensors, at the University at Buffalo, The State University of New York. He has over 50 publications and has co-authored the first edited book on OCR of Indic Scripts. He has held leadership positions in several academic conferences including BTAS and ISBA. He has served as PI or Co-I on research grants totaling over 35 million dollars from Federal, State and Industry sources. He is a Senior Member of the IEEE and a member of the IEEE Biometrics Council.

Venu Govindaraju (F06) is currently a SUNY Distinguished Professor of Computer Science and Engineering and the Founding Director of the Center for Unified Biometrics and Sensors at the University at Buffalo, The State University of New York. He holds four patents, has received several major professional society awards and co-authored about 425 scientific papers. He has served as General Chair of many conferences including BTAS and ISBA. He has been a PI or Co-I on research funding of over 65 million dollars. He is a Fellow of the Association of Computing Machinery, the American Association for the Advancement of Science, the International Association of Pattern Recognition, and the International Society of Optics and Photonics. He is a Life Fellow of the IEEE and has served as President of the IEEE Biometrics Council.