Person Re-identification for Improved Multi-person Multi-camera Tracking by Continuous Entity Association

Neeti Narayan, Nishant Sankaran, Devansh Arpit, Karthik Dantu, Srirangaraj Setlur, Venu Govindaraju

University at Buffalo

University at Buffalo
The State University of New York

Overview

- **Introduction** - Person analysis (Re-ID, MPMCT), Challenges
- **Related Work** - Existing methods
- **Motivation and Goals**
- **Proposed Approach** - Continuous entity association for person tracking
- **Future Work** - Spatio-temporal based tracking approach incorporating visual appearance and location together

Introduction

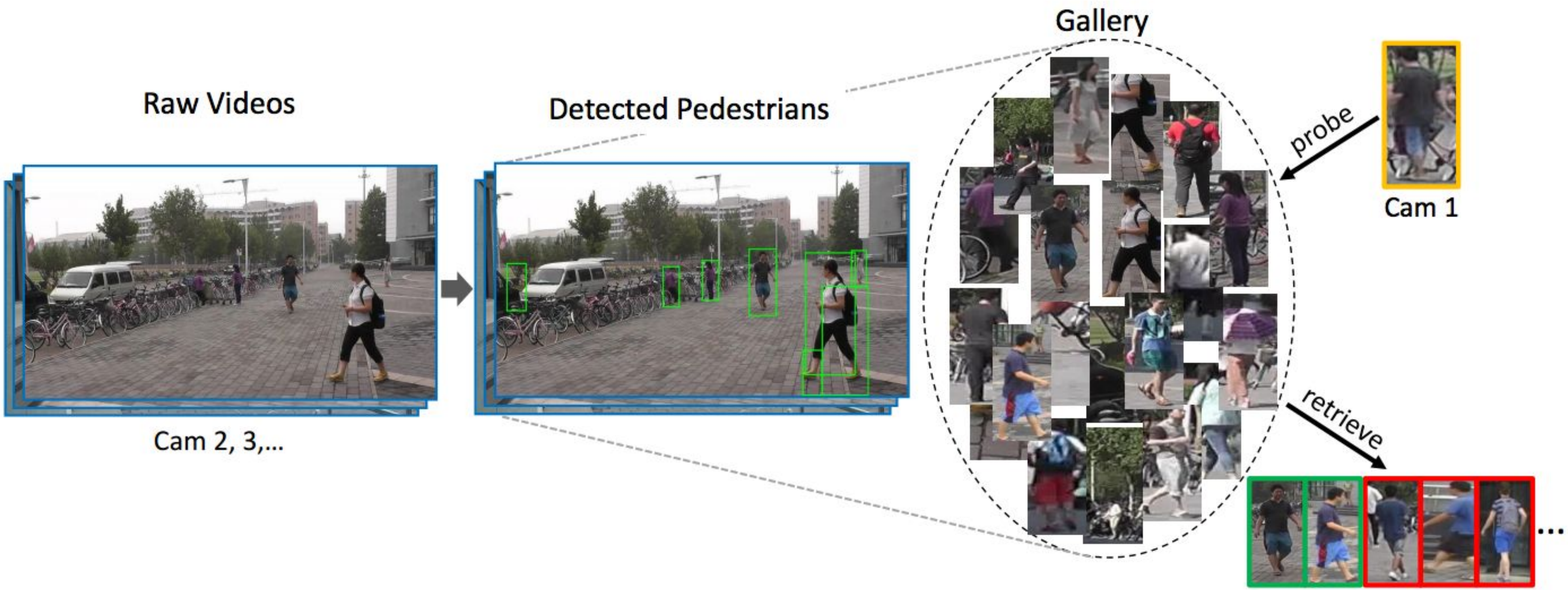
Automated Analysis of Large Video Data

- Video surveillance
- Activity and behavior characterization
- Increase in number of deployed cameras
  - Increase in the workload of video operators
  - Decrease in efficiency
- Growing demand of automated analysis and understanding video content
- Key person analysis tasks: Person recognition, verification and tracking

University at Buffalo
The State University of New York

Person Re-identification (ReID)

- Target re-identification retrieves all and only the gallery images of the same target as the query.



An end-to-end person ReID system that includes person detection and re-identification

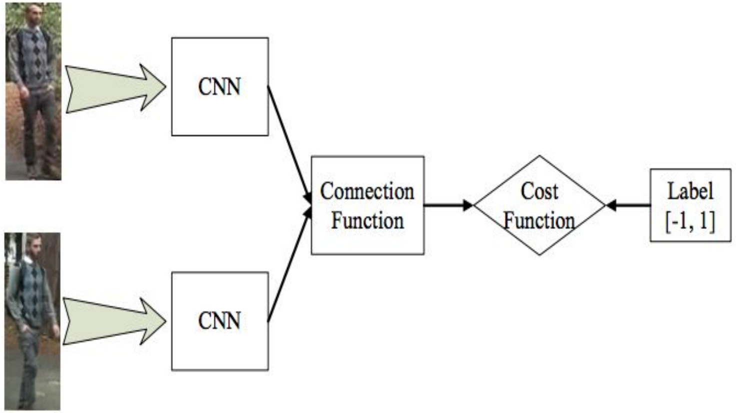
University at Buffalo
The State University of New York

Multi-Person Multi-Camera Tracking (MPMCT)

Related Work

Person Re-identification

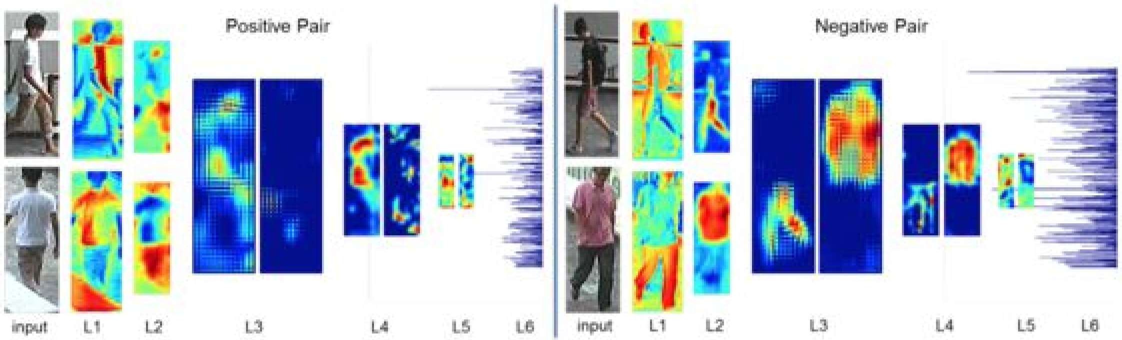
- Siamese deep neural network for person re-identification
  - + Learns similarity metric from image pixels directly.
  - + People need not be enrolled.
  - - No motion modeling.
  - - Real-world scenarios not modeled.



(Yi, Dong, Zhen Lei, and Stan Z. Li. “Deep metric learning for practical person re-identification”, 2014)

Person Re-identification

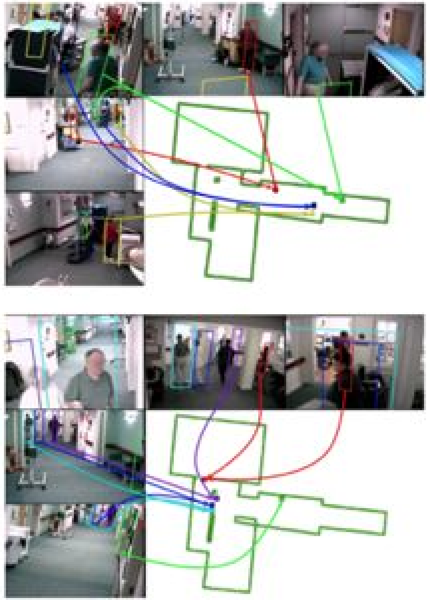
- An improved deep learning architecture for person re-identification
  - \+ Addresses re-identification problem
  - \+ Does not require persons to be enrolled
  - \- Additional modalities like face, gait etc are not utilized
  - \- No motion modeling / location based association
  - \- Does not handle real world scenarios of needing to associate multiple simultaneous observations



(Ahmed, Ejaz, Michael Jones, and Tim K. Marks. “An improved deep learning architecture for person re-identification.” Proceeding of the IEEE CVPR 2015)

Person Tracking

- Harry Potter’s Marauder’s Map: Localizing and Tracking Multiple Persons-of-Interest by Nonnegative Discretization
  - \+ Person localization and tracking.
  - \+ Complex indoor scenarios.
  - \+ Uses color, person detection & non-background info.
  - \- No spatial locality constraint enforced (person at multiple places simultaneously).
  - \- Persons to be tracked must be enrolled previously.



(Yu, Shoou-I., Yi Yang, and Alexander Hauptmann. "Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.)

Motivation and Goals

Motivation

- Target ReID and MPMCT are clearly different. But, they share several common aspects as well.
    - Assume semantic notion of “identity”.
    - Some components of the solution to one problem can be used to solve the other.
    - ReID involves associating object hypotheses, hence, possible to draw some parallels to tracking as well.
- Tracking failures can be effectively recovered by learning from historical visual semantics and tracking associations.

Goals

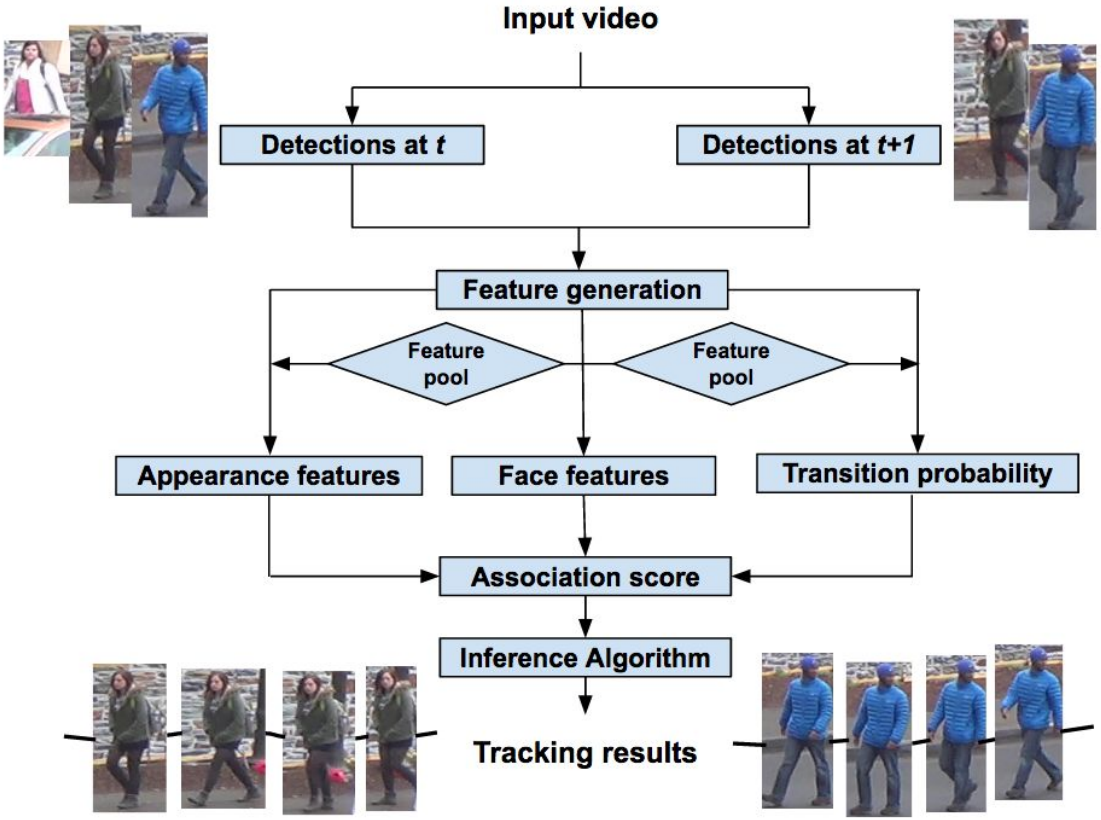
- Automated analysis of video data
  - Do not rely on constant human interaction.
- Within the context of data association, introduce a learning perspective to person tracking
  - Address influence of human appearance, face biometric and location transition on person re-identification.
- Minimal assumptions
  - Do not assume enrollment of people.
- High tracking accuracy
  - Do not compromise on tracking accuracy.
  - Track all people in the scene very efficiently with minimum identity switches.
- Design metrics to quantify the tracking system.

Proposed Approach

Analysis of People in Public Spaces

- A model for multi-camera tracking
- Continuous entity association
    - Between current and previous timestamp detections
- Steps in learning detection associations and tracking people:
    - Person detection
    - Feature extraction based on human appearance, biometric and location constraints
    - Association probability matrix
    - Most probable associations - Linear programming problem

Flowchart

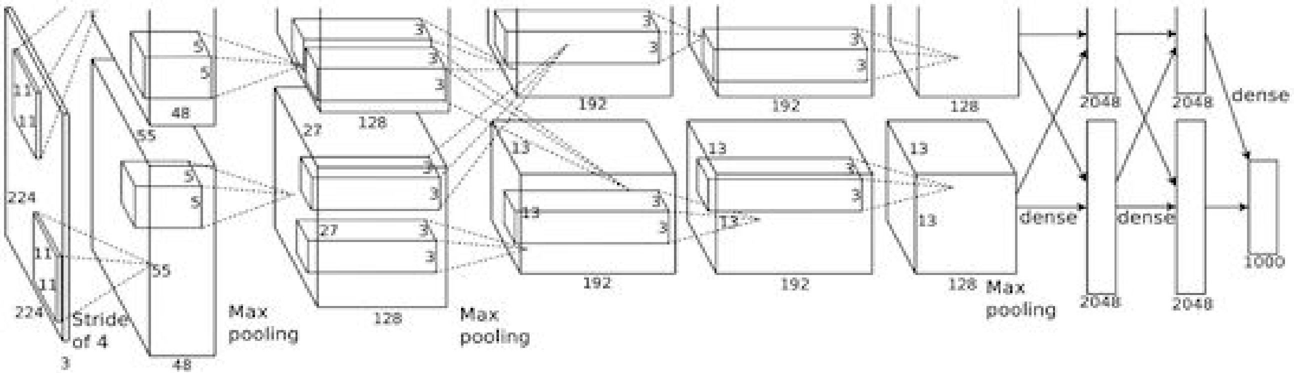

Input video
Detections at t
Detections at t+1
Feature generation
Feature pool
Feature pool
Appearance features
Face features
Transition probability
Association score
Inference Algorithm
Tracking results

Feature Extraction

- Desirable Properties
  - Robust to inherent variations.
  - Good discriminative ability.
- Features Explored
  - Appearance features
    - Feature length = 4096; AlexNet feature
  - Face features
    - Feature length = 4096; VGG-16 feature
  - Location transition
    - Feature length = 9 x 9 x num of cameras.

Appearance Features

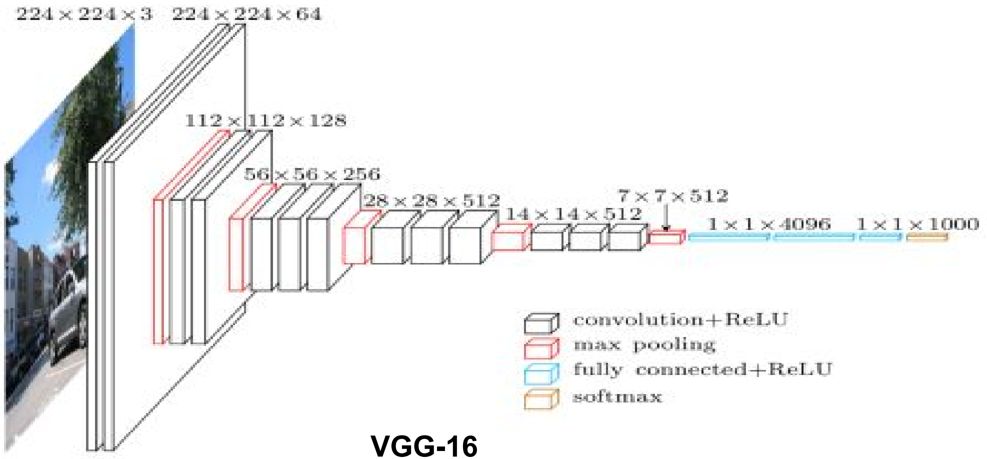
- Input: Person BB
- Output: Appearance-based features from last FC layer



AlexNet

Face Features

- Input: Face BB
- Output: Face features from last FC layer



VGG-16

Transition Probability

- Predict most probable paths within and across cameras

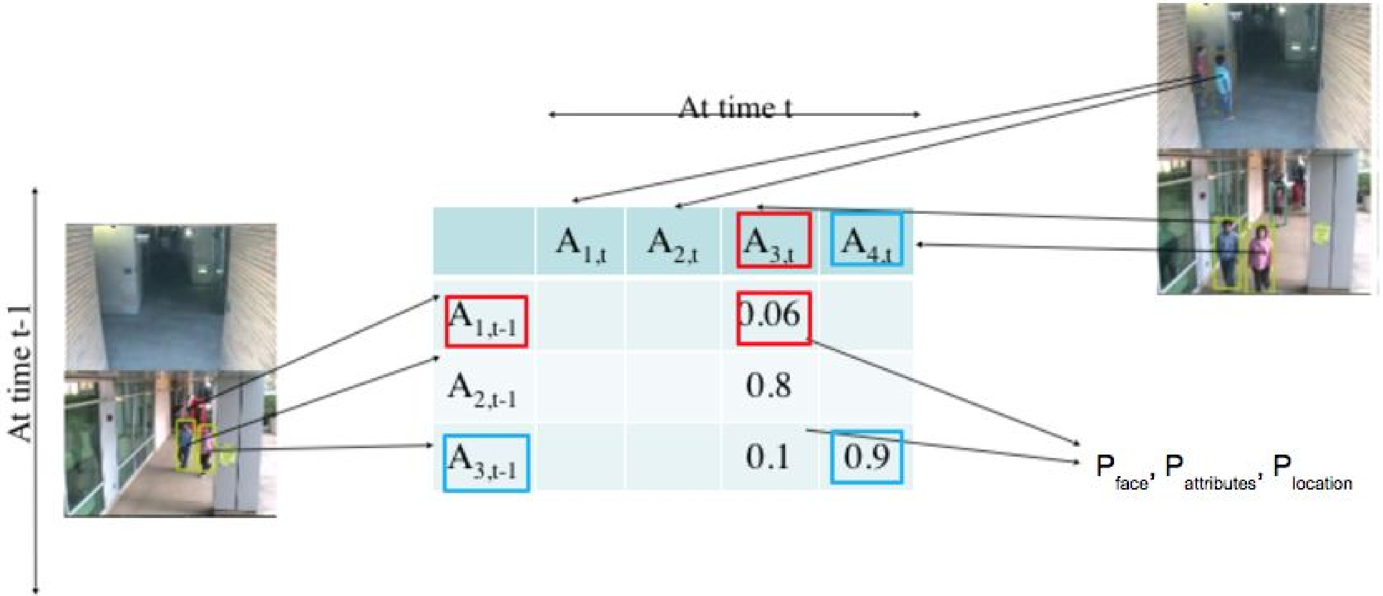
$$N = n \times k.$$

$$\forall S_i, S_j, P_{S_i,S_j} \in [0,1]$$

$$\forall S_i, \sum_{j=1}^{N} P_{S_i,S_j} = 1$$

$$P(S_i, S_j) = \Pr(X_t = S_j | X_{t-1} = S_i)$$
$$= \frac{|X_t = S_j \wedge X_{t-1} = S_i|}{\sum_{k=1}^{N} |X_t = S_k \wedge X_{t-1} = S_i|}$$

Inference Algorithm



$\max_{\mathbf{W}} \mathbf{P} \cdot \mathbf{W}$

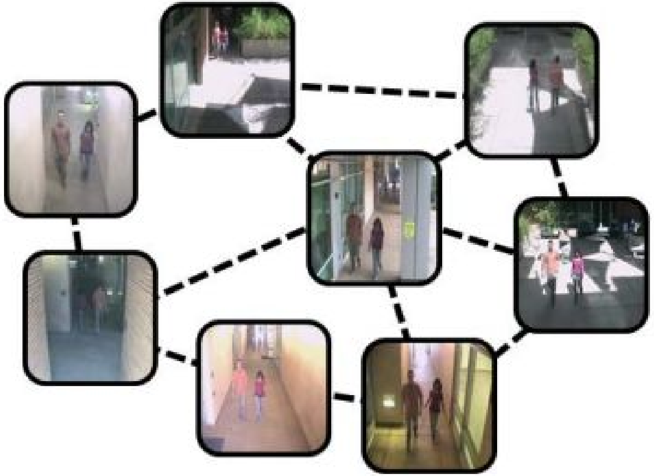
s.t. $\mathbf{W} \in [0,1]$, $\mathbf{W1} = 1$, $\mathbf{1}^T\mathbf{W} = 1$

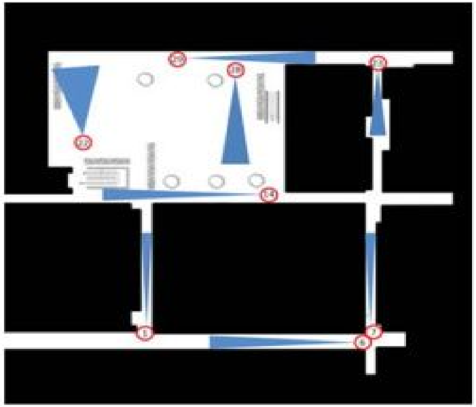
$\Updownarrow$

Greedy approach of selecting largest probability sequentially

Database and Protocols

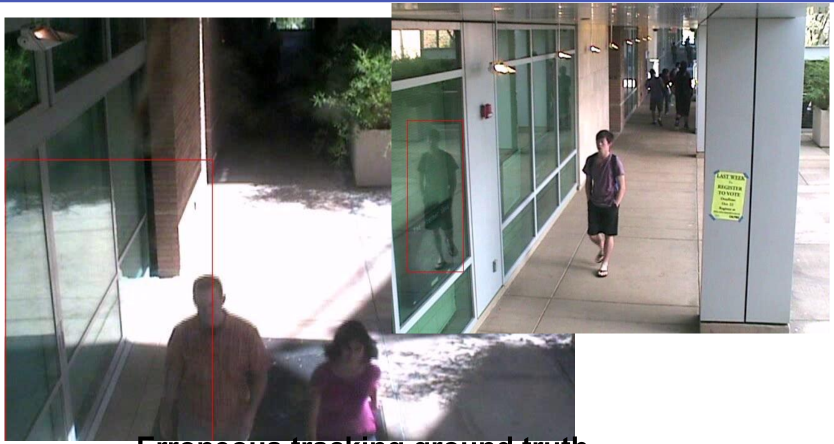
- CamNeT
  - 8 cameras covering indoor and outdoor scenes at a university, more than 16,000 images of 50 people.
  - 640x480 images, @20-30fps
- Protocol
  - Use Scenario 1.
  - IDs not unique - manual tagging performed.
  - Upto 6-7 simultaneous observations.





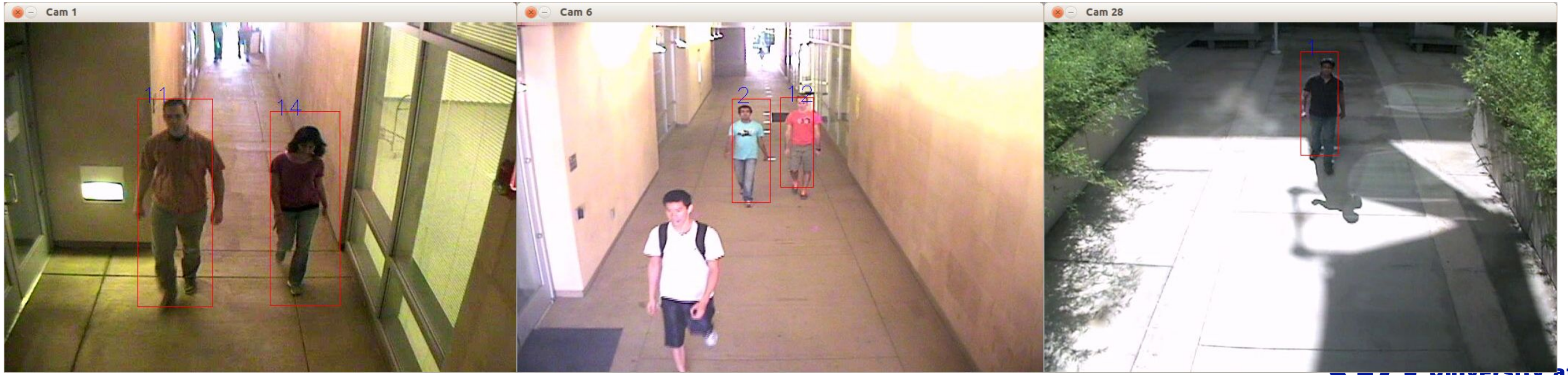
Zhang, Shu, et al. "A camera network tracking (CamNeT) dataset and performance baseline." 2015 IEEE Winter Conference on Applications of Computer Vision. IEEE, 2015.

Database and Protocols



Erroneous tracking ground truth

Incorrect ID tagging



Database and Protocols

- DukeMTMC
  - 8 cameras, more than 2M frames of 2,700 people.
  - 1920x1080 images, @60fps
- Protocol
  - Use only training data for experiments.
  - Use camera1 and camera3 video data for attribute feature evaluation.

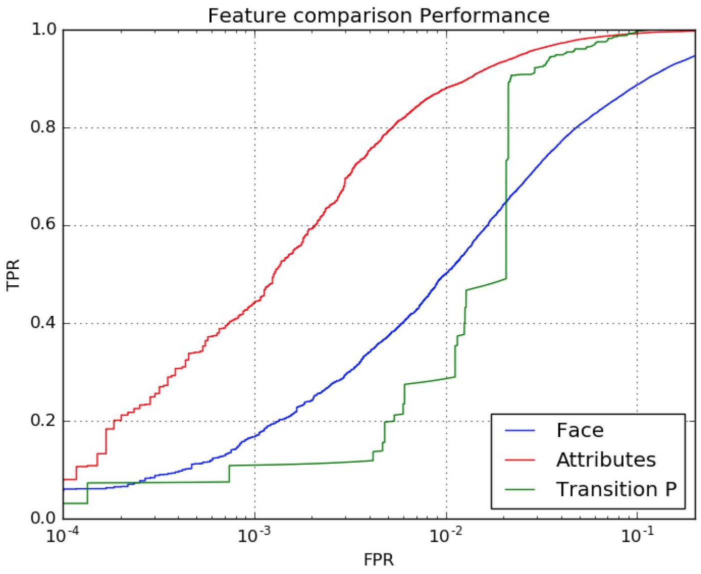
Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. E. Ristani, F. Solera, R. S. Zou, R. Cucchiara and C. Tomasi. ECCV 2016 Workshop on Benchmarking Multi-Target Tracking.

Evaluation Metric

- Use ROC curves and the area under the curve (AUC) for evaluating data association results.
- Use a continuous re-identification evaluation metric for person tracking:

$$E = \frac{1}{T}\sum_{t=1}^{T}\frac{\#misclassified\ detections\ at\ time\ t}{\#total\ detections\ at\ time\ t}$$

Results

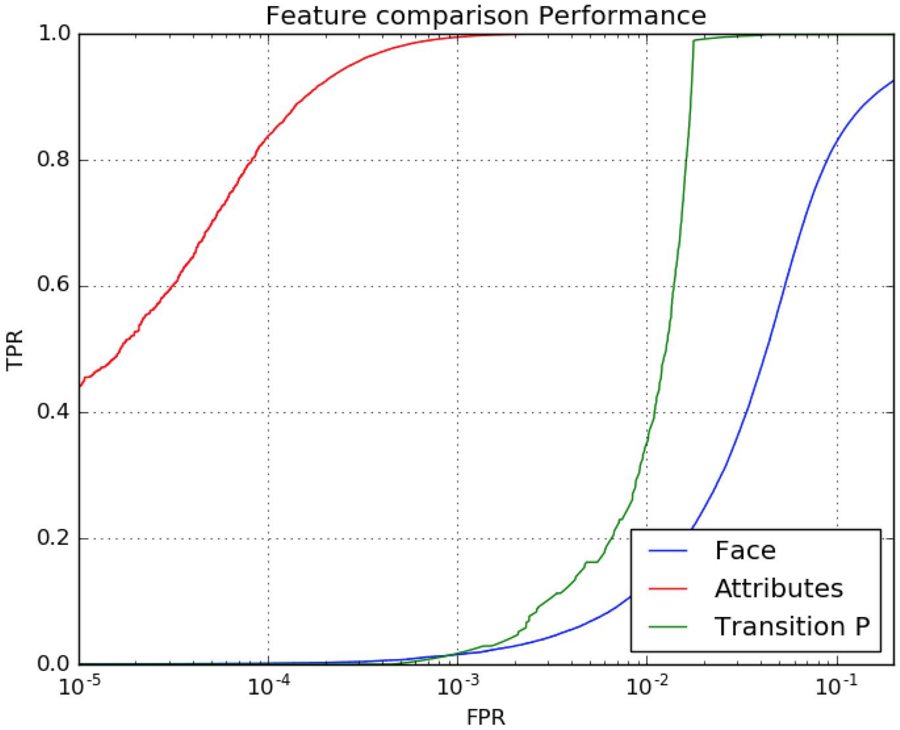


CamNeT

- **Evaluated performances of individual features for tracking achieving AUC scores of:**
  - Face features – 96.56%
  - Attribute features – 99.37%
  - Location transition – 98.28%

- Appearance features performed well even at low FAR.
- The performance of face features deteriorated because of low resolution.

Results



DukeMTMC

- **Evaluated performances of individual features for tracking achieving AUC scores of:**
  - Face features – 92.07
  - Attribute features – 99.99%
  - Location transition – 98.73%

University at Buffalo
The State University of New York

Inference Results

- **Inference error rates using proposed entity association algorithm**
  - **CamNeT:**
    - Face features – 4.67%
    - Attribute features – 2.9%
    - Location transition– 4.49%
  - **DukeMTMC:**
    - Face features – 12.07%
    - Attribute features – 0.01%
    - Location transition– 0.5%

Comparison

- CamNet dataset:
  - Crossing fragments (XFrag): The number of true associations missed by the tracking system.

Method	XFrag
Baseline results [1]	27
Method in [2]	24
Ours	**5**

[1] Shu Zhang, Elliot Staudt, Tim Faltemier, and Amit K Roy-Chowdhury. A camera network tracking (camnet) dataset and performance baseline. In WACV, 2015
[2] Bi Song and Amit K Roy-Chowdhury. Robust tracking in a camera network: A multi-objective optimization framework. IEEE Journal of Selected Topics in Signal Processing, 2008

Comparison

- DukeMTMC dataset:
  - Fragmentation: The number of identity switches in the tracking result, when the corresponding ground-truth identity does not change.

Method	Cam1	Cam2	Cam3	Cam4	Cam5	Cam6	Cam7	Cam8
Baseline results [1]	366	1929	336	403	292	3370	675	365
Ours	**34**	**47**	**102**	**42**	**69**	**84**	**139**	**12**

[1] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In European Conference on Computer Vision. Springer, 2016

Contribution

- Algorithm
    - Within the context of data association, we introduce a learning perspective to the tracking problem.
    - Does not require temporally contiguous sequence of video data.
    - Minimal assumptions
- Applications
    - The framework can be extended to a variety of data types:
        - Multimodal biometrics
        - Person wardrobe model - Clothing
- Impact
    - Pave the way towards future research in this direction.
    - Encourage incorporating other constraints like speed, travel time etc.

Future Work

- Develop a learning model to recover from association errors.
- Minimize association errors in Entry-Exit case.

Thank You!