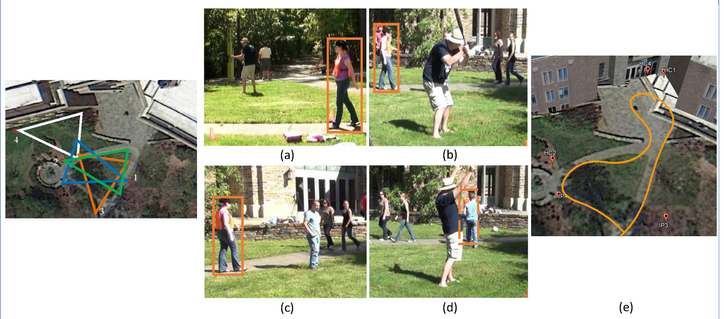
Multi-view People Tracking via Hierarchical Trajectory Composition

Yuanlu Xu[1], Xiaobai Liu[2], Yang Liu[1] and Song-Chun Zhu[1]
[1]Department of Computer Science and Statistics, University of California, Los Angeles (UCLA)
[2]Department of Computer Science, San Diego State University (SDSU)

Overview

Motivation

a) Most existing multi-view trackers exploit a particular cue, e.g., 3d localization,, appearance, motion to obtain the tracking results, while different cues may dominate different periods over object trajectories, especially for complicated scenes.
b) We are interested in automatically discovering the optimal compositional hierarchy for object trajectories from various cues, in order to handle a wider variety of tracking scenarios.

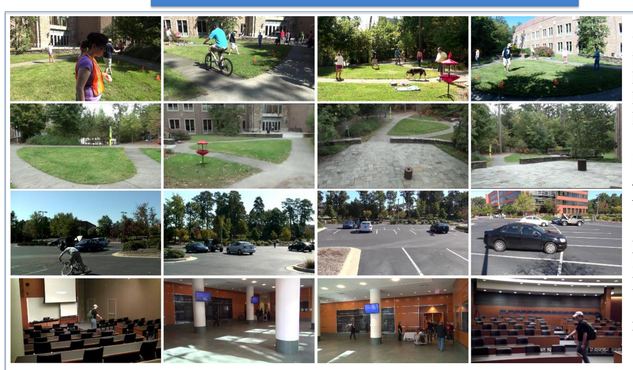


For example, in (a), since no occlusion happens, we apply an appearance based tracker to get a 2D tracklet; in (b) and (c), since the subject can be fully observed from two different views, we can group these two boxes into a 3D tracklet by their 3D locations; in (d), since the subject is fully occluded in this view, we consider sampling its position from the 3D trajectory curve constrained by background occupancy.

Overview

a) We formulate multi-view multi-object tracking as a structure optimization problem described by a hierarchical composition model.
b) We define a set of composition criteria. Each criterion focuses on utilizing one certain cue.
c) We learn parameters of each criteria separately by gradient descent and infer the compositional structure by a greedy pursuit algorithm.
d) We establish a new dataset to provide richer information and cover more challenges in multi-view people tracking.

Dataset



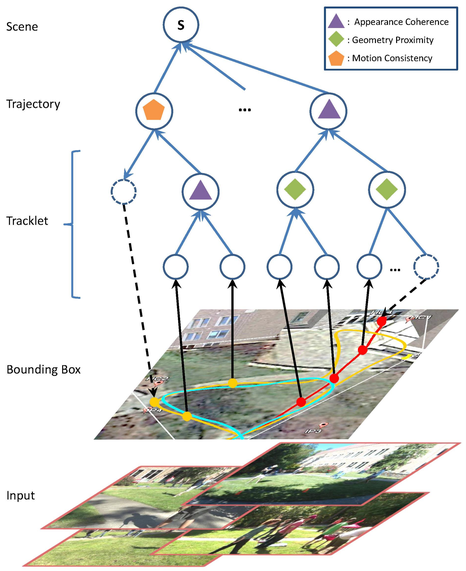
Collected from DARPA MSEE Project.

Specs: 4 sync. views, 10-20 people, 3-4 min. long, 1920x1080 res.

Features of our CAMPUS Dataset:
a) Dense foreground (around 15-25 objects, frequent conjunctions and occlusions).
b) Complex scenarios (objects conducting diverse activities, dynamic background, interactions between objects and background).
c) Various object scales (tracking targets sometimes either too tiny or huge in certain cameras).

Hierarchical Composition Model

The proposed hierarchical composition model for a scene. We start from structureless tracklets, i.e., object bounding boxes, and gradually compose them into tracklets of larger size and finally into trajectories. Each trajectory entity may be observed in single view or multiple views. The composition process is guided by a set of criteria, which describe the composition feasibility in the hierarchical structure.



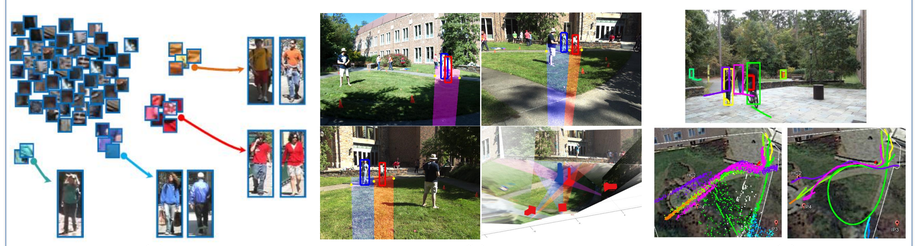
Our model is defined as:

$$\mathbf{G} = (V_N, V_T, S, X, \Theta)$$

V_N: the set of non-terminal nodes
V_T: the set of terminal nodes
S: the root node representing the scene
X: the set of states of all nodes
Θ: parameters of composition criteria

Composition Criteria

a) **Appearance Coherence**: similar to bag-of-words encoding, we use a DCNN to code a person image with common people appearance templates. We fine-tune a CaffeNet (trained on ImageNet) using people image samples. The identity label is assigned by clustering. The output is regarded as an appearance descriptor.
b) **Geometry Proximity**: the geometric distance is measured by people's feet positions on the 3D ground plane. We learn a kernel to handle the problem of tracklet misalignment and inaccurate feet estimation.
c) **Motion Consistency**: we model the motion information of a tracklet as a continuous function of its 3D ground positions w.r.t. time. We employ the b-spline function to represent the motion pattern of the trajectory, constrained by people's relative positions.



Inferring Hierarchy

The optimization should accomplish two goals: i) composing hierarchical structures, and ii) estimating states for terminal and non-terminal nodes.
We approximate the construction of the hierarchical structure by an iterative greedy pursuit algorithm:

a) In each iteration, finding two tracklets O_i and O_j with maximum probability to be composed into a new tracklet O_n, using criterion δ_n:

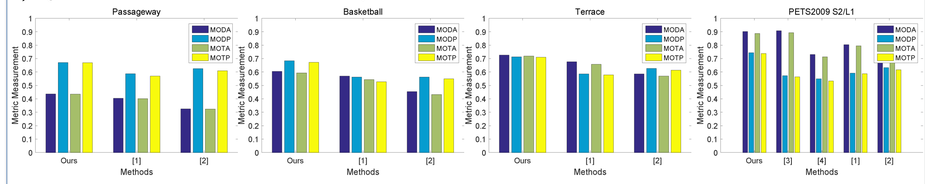
$$\max_{O_i, O_j, \delta_n} p(x(O_n)) \cdot \prod_k p_k(O_i, O_j)^{\delta_n == k}$$

b) Merging O_i and O_j into their parent node O_n, updating its state $x(O_n)$.
c) Repeat until connect all the top non-terminal nodes to the root node S.

Thank you for your time !

Experimental Results

a) Quantitative results on CAMPUS, EPFL and PETS 2009 datasets



Sequence	Method	MODA(%)	MODP(%)	MOTA(%)	MOTP(%)	MT(%)	PT(%)	ML(%)	IDSW	FRAG
Garden1	Our-full	**49.30**	72.02	**49.03**	71.87	31.25	62.50	6.25	299	200
	Our-3	44.63	72.35	44.36	72.20	18.75	68.75	12.50	296	202
	Our-2	42.10	71.08	41.69	70.97	12.50	75.00	12.50	448	296
	Our-1	41.21	71.06	37.21	70.94	12.50	75.00	12.50	4352	4390
	[1]	30.47	62.13	28.10	62.01	6.25	68.75	25.00	2577	2553
	[2]	24.52	64.28	22.43	64.17	0.00	56.25	43.75	2269	2233
Garden2	Our-full	**27.81**	71.74	**25.79**	71.59	21.43	78.57	0.00	94	73
	Our-3	23.39	71.13	22.50	71.08	14.29	85.71	0.00	92	72
	Our-2	18.76	70.20	17.27	70.12	14.29	78.57	7.14	142	97
	Our-1	17.68	70.12	10.24	70.11	14.29	78.57	7.14	700	733
	[1]	24.35	61.79	21.87	61.64	14.29	85.71	0.00	268	249
	[2]	16.51	63.92	13.95	63.81	14.29	78.57	7.14	241	216
Auditorium	Our-full	**20.84**	69.26	**20.62**	69.21	33.33	55.56	11.11	31	28
	Our-3	18.83	68.99	18.62	68.95	22.22	61.11	16.67	30	28
	Our-2	18.02	68.32	17.29	68.25	16.67	66.67	16.67	104	94
	Our-1	17.78	68.33	14.11	68.28	16.67	66.67	16.67	523	536
	[1]	19.46	59.45	17.63	59.29	22.22	61.11	16.67	264	257
	[2]	17.90	61.19	16.15	61.02	16.67	66.67	16.67	249	235
ParkingLot	Our-full	**24.46**	66.41	**24.08**	66.21	6.67	66.67	26.67	459	203
	Our-3	19.23	66.50	18.84	66.38	0.00	53.33	46.67	477	191
	Our-2	12.85	65.70	12.23	65.61	0.00	46.67	53.33	754	285
	Our-1	10.86	65.77	8.74	65.72	0.00	46.67	53.33	2567	2600
	[1]	14.73	58.51	13.99	58.36	0.00	53.33	46.67	893	880
	[2]	11.68	60.10	11.00	59.98	0.00	46.67	53.33	828	812

Baseline:
Our-1: initial tracklets only.
Our-2: only appearance cue.
Our-3: geometry and appearance cues.
Our-full: all three cues.

Reference:
[1] J. Berclaz et al, TPAMI11
[2] F. Fleuret et al, TPAMI 08
[3] A. Andriyenko et al, CVPR 12
[4] L. Leal-Taixe et al, CVPR 12

b) Qualitative results on CAMPUS, EPFL and PETS 2009 datasets