Department of Computer Science

Coupling Detection and Data Association for Multiple Object Tracking

Zheng Wu, Ashwin Thangali, Stan Sclaroff, Margrit Betke

Abstract

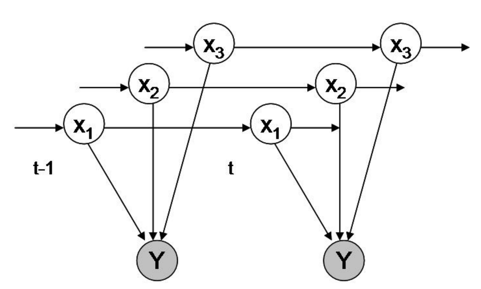
We present a novel framework for multiple object tracking in which the problems of object detection and data association are expressed by a **single** objective function. The framework follows the Lagrange dual decomposition strategy, taking advantage of the often complementary nature of the two subproblems. The advantages of our coupling framework are:

- No problem of error propagation from which traditional "detection-tracking approaches" to multiple object tracking suffer.
- No need to apply "non-maximum suppression" during detection stage.

Traditional Tracking System (Detection-Tracking): Object Detection → Data Association

Our System (Coupling): Object Detection ⇄ Data Association

Bayesian Formulation

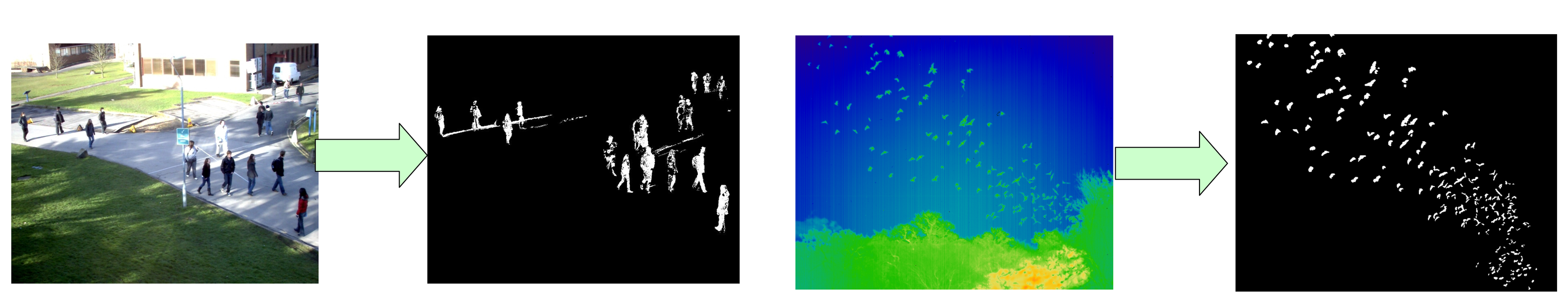


MAP Estimation

$$\max_X p(X|Y) = \max_X p(Y|X)p(X) = \max_X \prod_t p(Y_t|X_t)[p(X_1)\prod_t p(X_t|X_{t-1})] \approx \max_X \prod_t p(Y_t|X_t)\prod_i p(x_{i,1})\prod_t p(x_{i,t}|x_{i,t-1})$$

X is the joint state vector of **all** objects in the scene

Y is the observation vector for the **entire** image, which depends on the states of all objects. Example of Y: binary image obtained after background subtraction



General Form of the Objective Function

$$\max_X p(X|Y) \quad \Rightarrow \quad \min_{X_1,X_2} g(X_1,Y)+h(X_2) \quad s.t\ X_1 = q(X_2)$$

- Function g(.) can be seen as the objective function in detection problem (image likelihood term); h(.) is the objective function in data association problem (temporal smoothness term); q(.) is the coupling constraint to enforce **agreement** of the solutions between two sub-problems.
- The choices of g, h, q and their combination are flexible. Typically, we want g, h are relatively easy to optimize.
- For traditional detection-tracking scheme with independence likelihood assumption (h(.) does not model the joint image likelihood), there is no need to introduce coupling constraint; the objective function is equivalent to classic tracker such as Multiple-Hypothesis-Tracking or Network-Flow-Tracker.
- The overall optimization can be solved through **Dual Decomposition**

Sparsity-constrained Detection

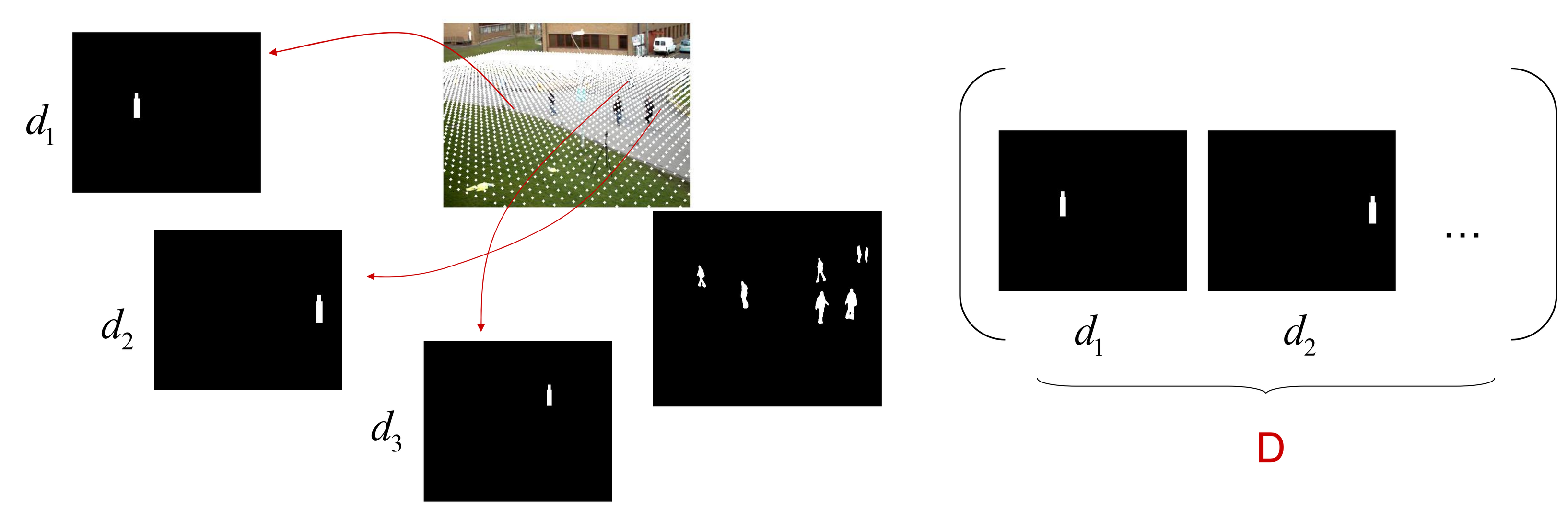
Given a dictionary D that encodes the shape and spatial information of objects in image, instantiate binary templates at selected positions through selector X such that the generated image (DX) looks similar to the observation Y.

Single Template $g: \min_X \| Y - DX \|_0, X \in \{0,1\}^N$

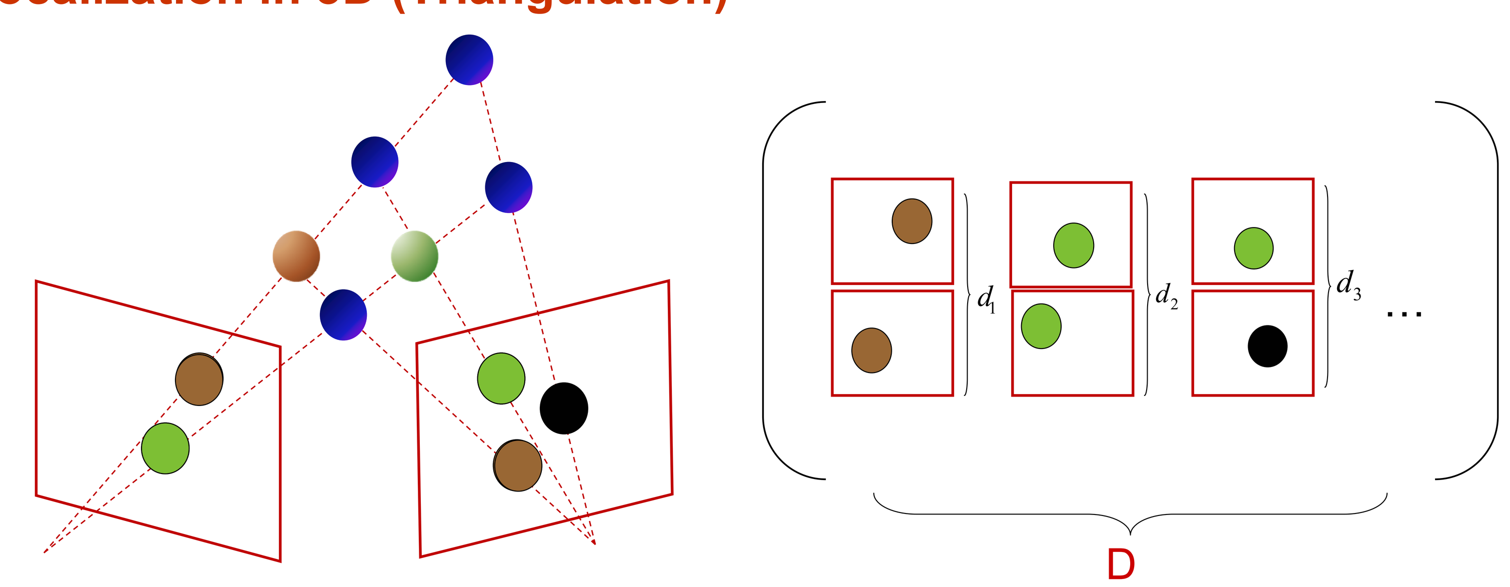
Multiple Templates $g: \min_X \| Y - \sum_i D_i X_i \|_0,$

$s.t. \quad \sum_i X_i \le 1, \quad X_i \in \{0,1\}^N$

Localization in 2D (Ground Plane)

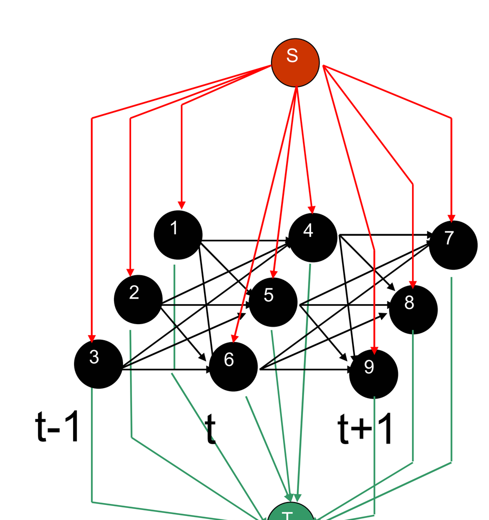


Localization in 3D (Triangulation)



Min-cost Flow Data Association

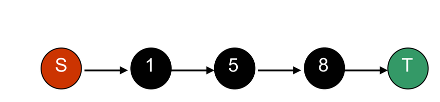
Each edge on the network is associated with a flow variable f and a cost c to measure how likely an object moves from one location to another. Push certain amount of flows into the network with minimum cost such that each path of the flow connects a set of detections across time to form a unique track.



$$h: \min_f \sum_t \sum_i \sum_j c_{i,j}^{(t)} f_{i,j}^{(t)}$$

$$s.t. \quad \sum_i f_{i,n}^{(t)} = \sum_j f_{n,j}^{(t)}, \forall t \forall n$$

source sink Object location



Coupling Detection and Data Association

$$\min_{X,f} \sum_t \| Y_t - D_t X_t \|_0 + \sum_t \sum_i \sum_j c_{i,j}^{(t)} f_{i,j}^{(t)}$$

$$s.t. \quad \sum_i f_{i,n}^{(t)} = \sum_j f_{n,j}^{(t)}, \quad \forall t \forall n$$

Flow conservation: If there is a flow coming into a node, there is a flow coming out.

$$X_{t,n} = \sum_j f_{n,j}^{(t)}, \quad \forall t \forall n$$

Variable agreement: If there is a detection at a node, there is a flow going through.

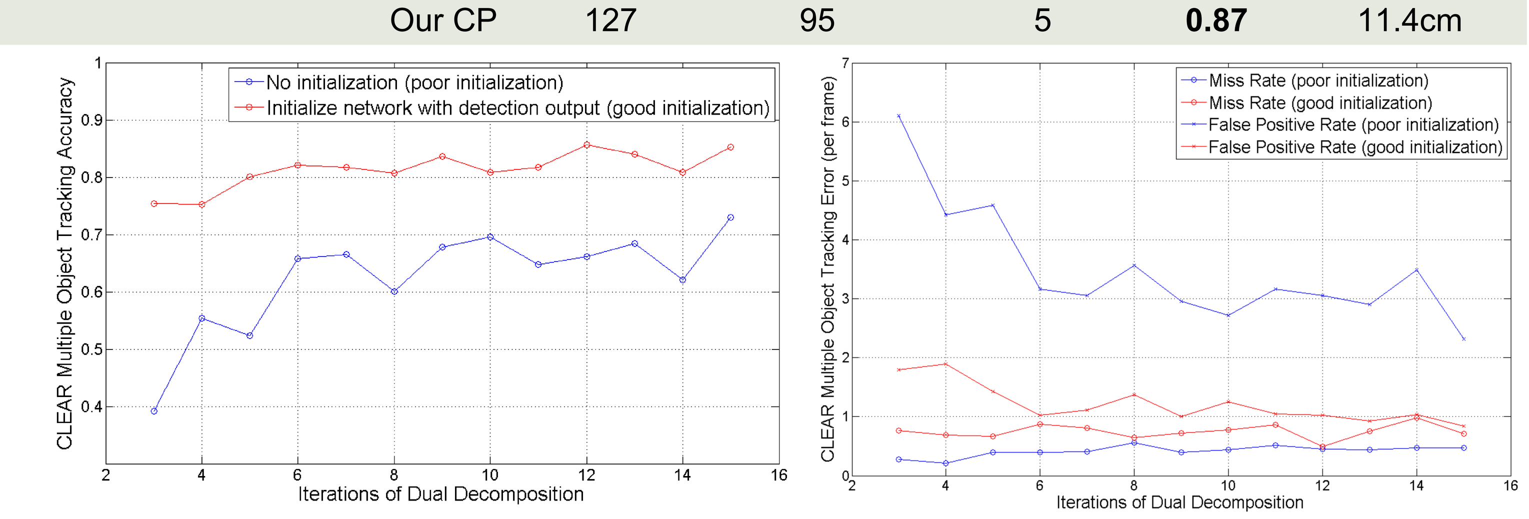
Dual Decomposition

$$g(\lambda) = \min_X \sum_t (\| Y_t - D_t X_t \|_0 + \lambda^T X_t) \qquad h(\lambda) = \min_f \sum_t \sum_i \sum_j (c_{i,j}^{(t)} - \lambda_{t,i}) f_{i,j}^{(t)}$$

Experiment

Datasets: PETS2009 (single view), Infrared Video (three views)

Sequence	Density	Method	#Objects	Mostly Track	Mostly Lost	MOTA	MOTP
PETS S2L1	Low	OM[2]	23	20	1	0.88	0.76
		ILP[3]	23	20	8	0.26	0.67
		Our CP	23	22	0	**0.94**	0.70
PETS S1L1-2	Median	OM[2]	36	20	7	0.64	0.67
		Our CP	36	24	2	**0.89**	0.61
Infrared S1	Low	RT[1]	19	19	0	0.80	9.0cm
		Our CP	19	19	0	**0.90**	9.5cm
Infrared S2	Median	RT[1]	75	68	0	0.51	9.9cm
		Our CP	75	71	1	**0.92**	9.7cm
Infrared S3	High	RT[1]	127	60	8	-0.34	11.6cm
		Our CP	127	95	5	**0.87**	11.4cm



Reference

[1] Z. Wu, N. I. Hristov, T. H. Kunz, and M. Betke. Tracking-reconstruction or reconstruction-tracking? Comparison of two multiple hypothesis tracking approaches to interpret 3D object motion from several camera views. In *IEEE Wkshp Motion and Video Computing (WMVC)*, 2009

[2] A. Andriyenko, S. Roth, and K. Schindler. An analytical formulation of global occlusion reasoning for multi-target tracking. In *11th IEEE Intl. Workshop on Visual Surveillance*, 2011.

[3] A. Andriyenko and K. Schindler. Globally optimal multi-target tracking on a hexagonal lattice. In *11th European Conf. on Computer Vision*, 2010.