Monocular Multiview Object Tracking with 3D Aspect Parts

Yu Xiang[1,2*], Changkyu Song[2*], Roozbeh Mottaghi[1] and Silvio Saverese[1]

Stanford University[1], University of Michigan at Ann Arbor[2]

Monocular Multiview Object Tracking

❑ Inputs: video sequences from a single camera

❑ Goals:

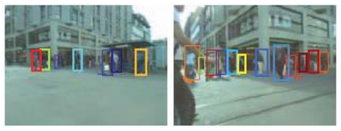
(1) Track the 2D location of the target

(2) Estimate the 3D pose of the target in time

(3) Localize the 3D parts of the target in time

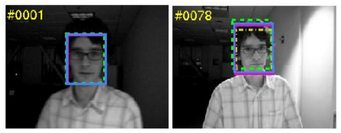
❑ Applications: Autonomous driving, robotics, augmented reality, etc.

Related Problems



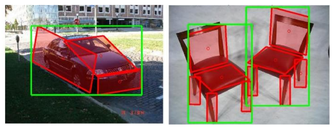
Choi et al. TPAMI' 13

Tracking by Detection



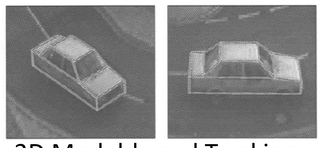
Babenko et al. TPAMI' 11

Online Object Tracking



Xiang & Savarese. CVPR'12

Multiview Object Recognition

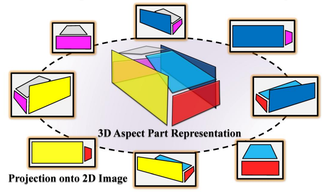


Koller et al. IJCV' 93

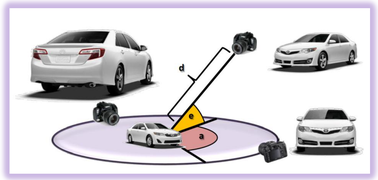
3D Model-based Tracking

Our Multiview Tracking Framework

❑ Object and Viewpoint Representation



$X_t = \{X_{it}\}_{i=1}^{n}$



$V_t = (a_t, e_t, d_t)$

❑ Posterior distribution $P(X_t, V_t \mid Z_{1:t})$

$$\propto P(Z_t \mid X_t, V_t) \int P(X_t, V_t \mid X_{t-1}, V_{t-1}) P(X_{t-1}, V_{t-1} \mid Z_{1:t-1}) dX_{t-1} dV_{t-1}$$

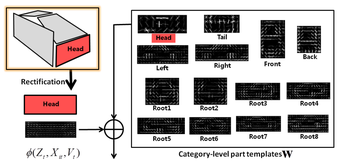
Likelihood **Motion Prior** **Posterior at t-1**

❑ Likelihood $P(Z_t \mid X_t, V_t) = \prod_{i=1}^{n} P(Z_t \mid X_{it}, V_t)$

$$P(Z_t \mid X_{it}, V_t) \propto \exp\left(\Lambda_{\text{category}}(Z_t, X_{it}, V_t) + \Lambda_{\text{online}}(Z_t, X_{it}, V_t)\right)$$

$$\Lambda_{\text{category}}(Z_t, X_{it}, V_t) = \begin{cases} \mathbf{w}_i^T \phi(Z_t, X_{it}, V_t), & \text{if visible} \\ \alpha_i, & \text{if self-occluded} \end{cases}$$

$$\Lambda_{\text{online}}(Z_t, X_{it}, V_t) = \begin{cases} \mathbf{H}_i(\psi(Z_t, X_{it}, V_t)), & \text{if visible} \\ \lambda_0, & \text{if self-occluded} \end{cases}$$



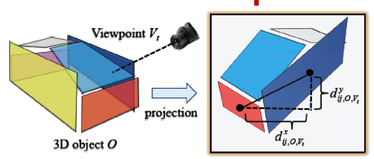
❑ Motion Prior

$$P(X_t, V_t \mid X_{t-1}, V_{t-1}) = P(X_t \mid X_{t-1}, V_t)\ P(V_t \mid V_{t-1})$$

Location **Viewpoint**

$$P(X_t \mid X_{t-1}, V_t) \propto \prod_{i=1}^{n} P(X_{it} \mid X_{i(t-1)}) \prod_{(i,j)} \Lambda(X_{it}, X_{jt}, V_t)$$

$$\Lambda(X_{it}, X_{jt}, V_t) = P(\Delta_t(x_i, x_j) \mid V_t) P(\Delta_t(y_i, y_j) \mid V_t)$$



❑ Multiview Particle Filtering Object Tracking

References

[1] B. Babenko, M.H. Yang, S. Belongie. Robust object tracking with online multiple instance learning. TPAMI, 2011.

[2] C. Bao, Y. Wu, H. Ling, H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In CVPR, 2012.

[3] Z. Kalal, K. Mikolajczyk, J. Matas. Tracking-learning-detection. TPAMI, 2012.

[4] S. Hare, A. Saari, P.H. Torr. Struck: Structured output tracking with kernels. In ICCV, 2011.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan. Object detection with discriminatively trained part-based models. TPAMI, 2010.

[6] A. Geiger, P. Lenz, R. Urtasun. Are we ready for autonomous driving? In CVPR, 2012.

[7] Y. Xiang, S. Savarese. Estimating the aspect layout of object categories. In CVPR, 2012.

Acknowledgement

We acknowledge the support of DARPA UPSIDE grant A13-0895-S002 and NSF CAREER grant N.1054127.

Experiments

❑ 2D Object Tracking

Video	MIL [1]	L1 [2]	TLD [3]	Struct [4]	DPM [5]+PF	Category Model	Full Model
YouTube	0.37	0.44	0.38	0.40	0.74	0.74	**0.75**
KITTI [6]	0.34	0.28	0.29	0.36	0.54	0.55	**0.58**
06_car [3]	0.19	0.52	**0.85**	0.48	0.70	0.67	0.70

Metric: mean bounding box overlap ratio

❑ Continuous Viewpoint Estimation

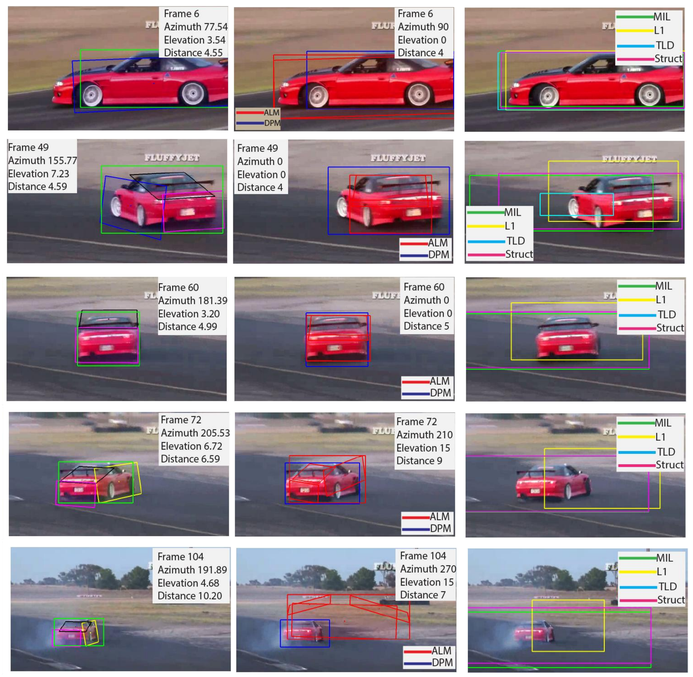
Video	Full Model	Category Model	ALM [7]
YouTube	**13.46°**	18.38°	47.24°
KITTI [6]	**14.66°**	23.20°	37.89°

Metric: mean absolute difference in azimuth angle

❑ 3D Aspect Part Localization

Video	Full Model	Category Model	ALM [7]
YouTube	**0.41**	0.40	0.30
KITTI [6]	**0.36**	0.30	0.26

Metric: mean overlap ratio of part shape



Ours Object Detection Online Tracking