From Certain to Uncertain: Toward Optimal Solution for Offline Multiple Object Tracking

Kaikai Zhao, Takashi Imaseki, Hiroshi Mouri

Department of Mechanical Systems Engineering
Tokyo University of Agriculture and Technology

Einoshin Suzuki, Tetsu Matsukawa

Department of Informatics, ISEE
Kyushu University

Offline multiple object tracking

same or different objects? $\longrightarrow$ affinity measure

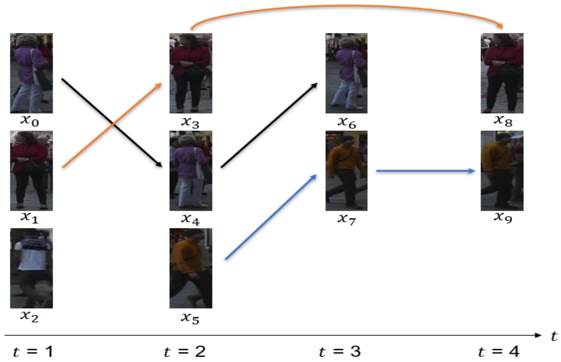


Figure 1: An example of object tracking. The task is to assign identities for detected objects across a series of frames.

Uncertain region and early mistakes

1. imperfect affinity measure $\longrightarrow$ uncertain region $\longrightarrow$ threshold θ $\longrightarrow$ mistakes
2. sequential tracking with pre-decided θ $\longrightarrow$ early mistakes

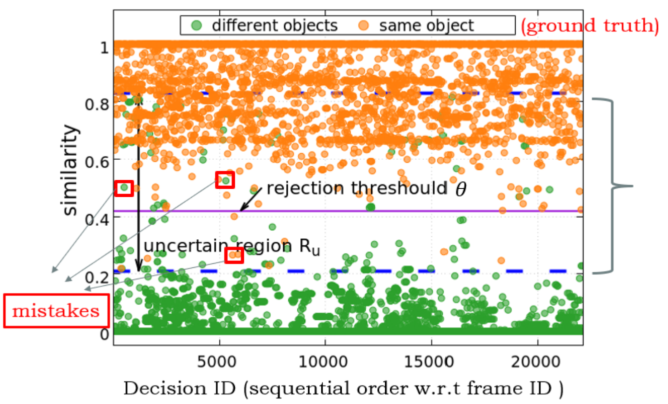
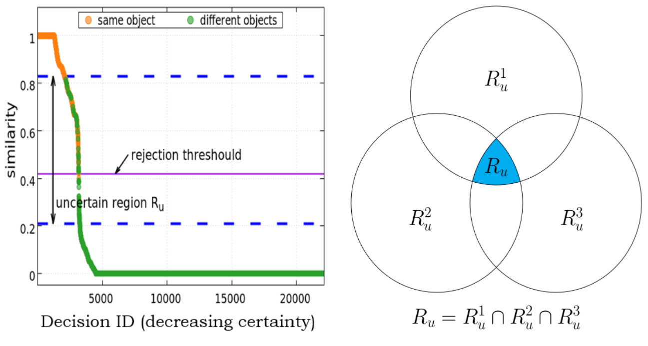


Figure 2: Two typical issues for previous offline object tracking.

Ideas to tackle the two issues



(a) tracking from certain to uncertain (b) ensemble multiple tracking experts

Figure 3: Ideas to tackle uncertain region and early mistake issues.

Our proposal

Agglomerative Hierarchical Clustering with Ensemble of Tracking Experts (AHC_ETE)

Notations

S: image sequence; D: detection set; N: number of detections; x_i: i^{th} detection; T_k: k^{th} track; e: a tracking expert (method)

Adapting AHC for object tracking

- memory complexity: $\mathcal{O}(N^2) \longrightarrow$ dividing S into $S_1, ..., S_n$, reduced to $\mathcal{O}(N_i^2)$;
- spatio-temporal constraint: detections in the same image should not belong to the same track $\longrightarrow$ building **cannot-link constraints**

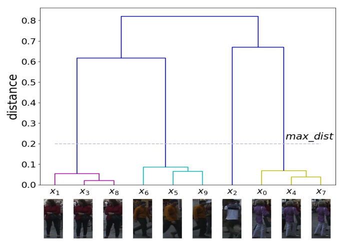


Figure 4: AHC based tracking

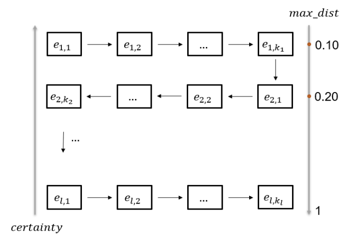


Figure 5: AHC_ETE framework

Defined distance measures

Appearance distance

$$\mathbf{dist}_{appe}(x_i, x_j) = 1 - \frac{a_i^T a_j}{||a_i||\ ||a_j||} \qquad (1)$$

a_i: extracted CNN feature vector of x_i

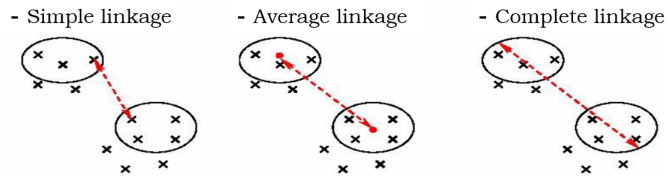


Figure 6: Linkages for the distance between two clusters

Motion (Kalman Filter) distance

state of object: $(u, v, \gamma, h, \dot{u}, \dot{v}, \dot{\gamma}, \dot{h})$ [Wojke et al., 17], centers, aspect ratio, height of a bbox

$$\mathbf{dist}_{kf}^{'}(T, x) = \sqrt{(y - \hat{y})^T \mathbf{\Sigma}^{-1} (y - \hat{y})} \qquad (2)$$

y: detection, $\hat{y}$: prediction of Kalman Filter

cyoukaikai@gmail.com

Temporal distance

Γ_k: set of frame IDs for detections in T_k

$$dist_{temp}(T_u, T_v) = \begin{cases} |\Gamma_u \cap \Gamma_v| - |\Gamma_u \cup \Gamma_v| & \text{if } \Gamma_u \cap \Gamma_v \neq \emptyset \\ min(\Gamma_v) - max(\Gamma_u) & \text{elseif } max(\Gamma_u) < min(\Gamma_v) \\ min(\Gamma_u) - max(\Gamma_v) & \text{elseif } max(\Gamma_v) < min(\Gamma_u) \\ 0 & \text{else} \end{cases} \quad (4)$$

frame IDs overlap $\rightarrow$ negative value; one track appears later than another $\rightarrow$ closest frame gap; no overlap & not earlier, later tracks $\rightarrow$ 0

Integrated distance

$$\text{dist}(T_u, T_v) = dist_{major}(T_u, T_v) * F_1(\cdot) * F_2(\cdot) * \ldots \quad (5)$$

We use appearance distance $dist_{appe}$ as $dist_{major}$, $F(\cdot)$ to filter other distances.

$$F(\nu, condition) = \begin{cases} 1 & \text{if } \nu \text{ satisfies } condition, \\ inf & \text{else.} \end{cases} \quad (6)$$

Defined tracking experts

1. Preprocessing: build T_{fp} for detections with $score \leq 0.3$ or suppressed by NMS with threshold 0.1; impose cannot-links for T_{fp}, i.e., for any track T_k, $dist(T_k, T_{fp}) = inf$.
2. Connecting detections to tracks: track with *complete* linkage (e_1), then *single* linkages (e_2 and e_3) $\rightarrow$ remove cannot-links on T_{fp} and track with weak constraints (e_4 and e_5)
3. Post-processing: remove T_k if $|T_k| < 3$

Table 1: Settings of our defined experts. Here ***appe***, ***temp*** and ***kf*** represent the settings for the appearance, temporal and Kalman Filter distance, respectively.

E	$dist_{appe}$	$F_1(temp)$	$F_2(kf)$	$F_3(appe)$	max_dist
e_1	complete	≥ 0	complete: < 9.5	-	0.10
e_2	single	≥ 0	-	-	0.05
e_3	single	≥ 0	complete: < 9.5	-	0.10
e_4	single	≥ 0	complete: < 9.5	-	0.10
e_5	single	≥ 0	average: < 9.5	complete: ≤ 0.30	0.20

Design of experiments

Dataset: MOT15, MOT16 [Milan et al., 16] training sequences

Evaluation metrics: multiple object tracking accuracy (MOTA [Bernardin and Stiefelhagen, 08]), identification precision (IDP), recall (IDR), corresponding F_1 score (IDF$_1$ [Ristani et al., 16])

Benchmark: Deep Sort [Wojke et al., 17] (same features, appearance and motion distances)

Result: effects of merging order

our method generally outperforms Deep Sort [Wojke et al., 17]; IDF$_1$s, IDPs, IDRs and MOTAs generally increase as more experts integrated

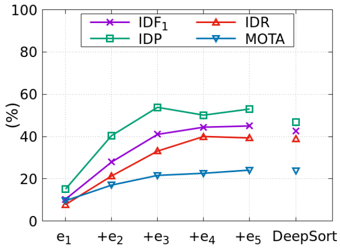


Figure 1: MOT15

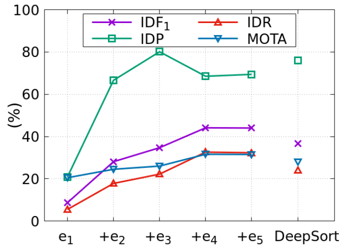


Figure 2: MOT16

Result: effects of different linkages

standard AHC [Day and Edelsbrunner, 84] based tracking

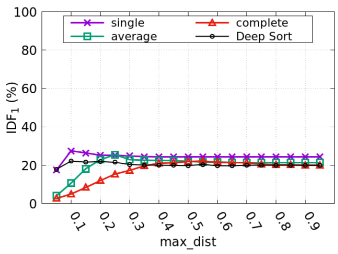


Figure 3: IDF1

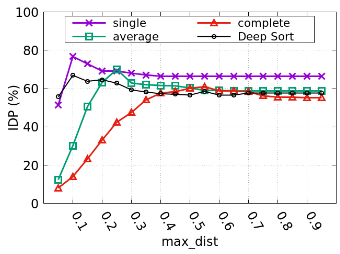


Figure 4: IDP

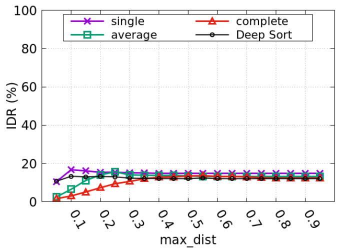


Figure 5: IDR

Test data: MOT16-02

best IDF$_1$s, IDPs and IDRs differ; certain region of *single* $<$ *average* $<$ *complete* linkage

Conclusion

- Tackling two typical issues for object tracking: 1) uncertain region, 2) early mistakes
- Proposed AHC_ETE: tracking from certain to uncertain, ensemble multiple tracking experts (a general framework for various distance measures and tracking experts)

Limitations and future work

- accepted all the progress of earlier experts $\rightarrow$ sensitive to the ordering of experts
- further experiments comparing with the state-of-the-art methods needed

Acknowledgments

This research was supported by the SAKURA project initiative of Japan Automotive Research Institute, Japan Ministry of Economy, Trade and Industry.

Code: https://github.com/cyoukaikai/ahc_ete