



Beyond accuracy: Tracking more like Human via Visual Search

Dailing Zhang^{1,2} Shiyu Hu⁵ Xiaokun Feng^{1,2}

Xuchen Li^{1,2} Meiqi Wu³ Jing Zhang² Kaiqi Huang^{1,2,4}

¹School of Artificial Intelligence, University of Chinese Academy of Sciences

²Institute of Automation, Chinese Academy of Sciences

³School of Computer Science and Technology, University of Chinese Academy of Sciences

⁴Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences

⁵School of Physical and Mathematical Sciences, Nanyang Technological University

zhangdailing2023@ia.ac.cn, shiyu.hu@ntu.edu.sg, fengxiaokun2022@ia.ac.cn,
wumeiqi18@mails.ucas.ac.cn, {lixuchen2024, jing_zhang, kqhuang}@ia.ac.cn

Background Introduction

- Task Definition
 - {Short-term → **Long-term** → **Global Instance**} Tracking



**Short-term tracking
(STT)**

- Short sequence
- Remain visible
- **Perceptual level**



**Long-Term Tracking
(LTT)**

- Long sequence
- Temporary absent
- **Partial cognitive level**

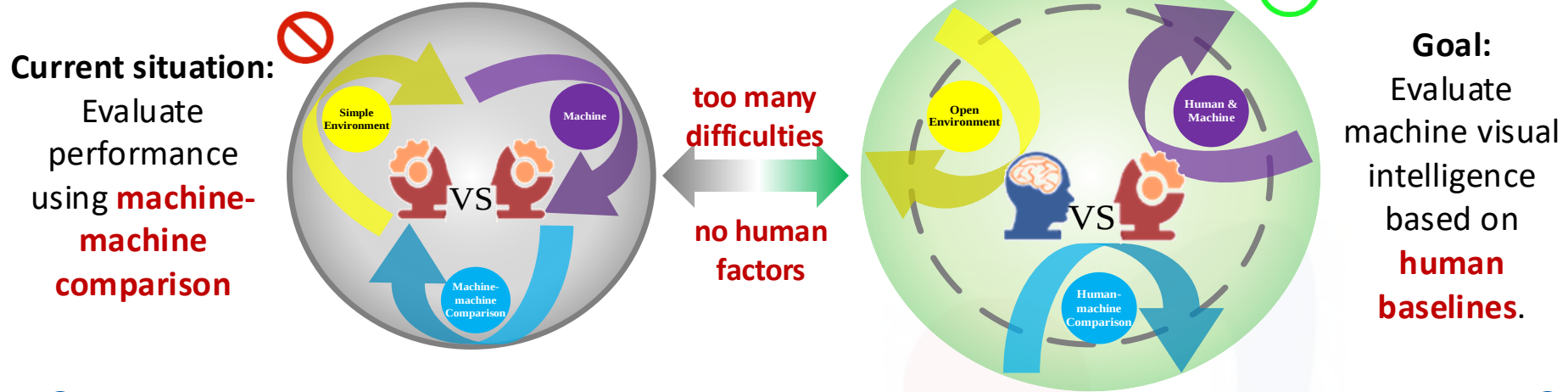


**Global Instance Tracking
(GIT)**

- Longer Sequence
- shot-cut and absent
- **Cognitive level**

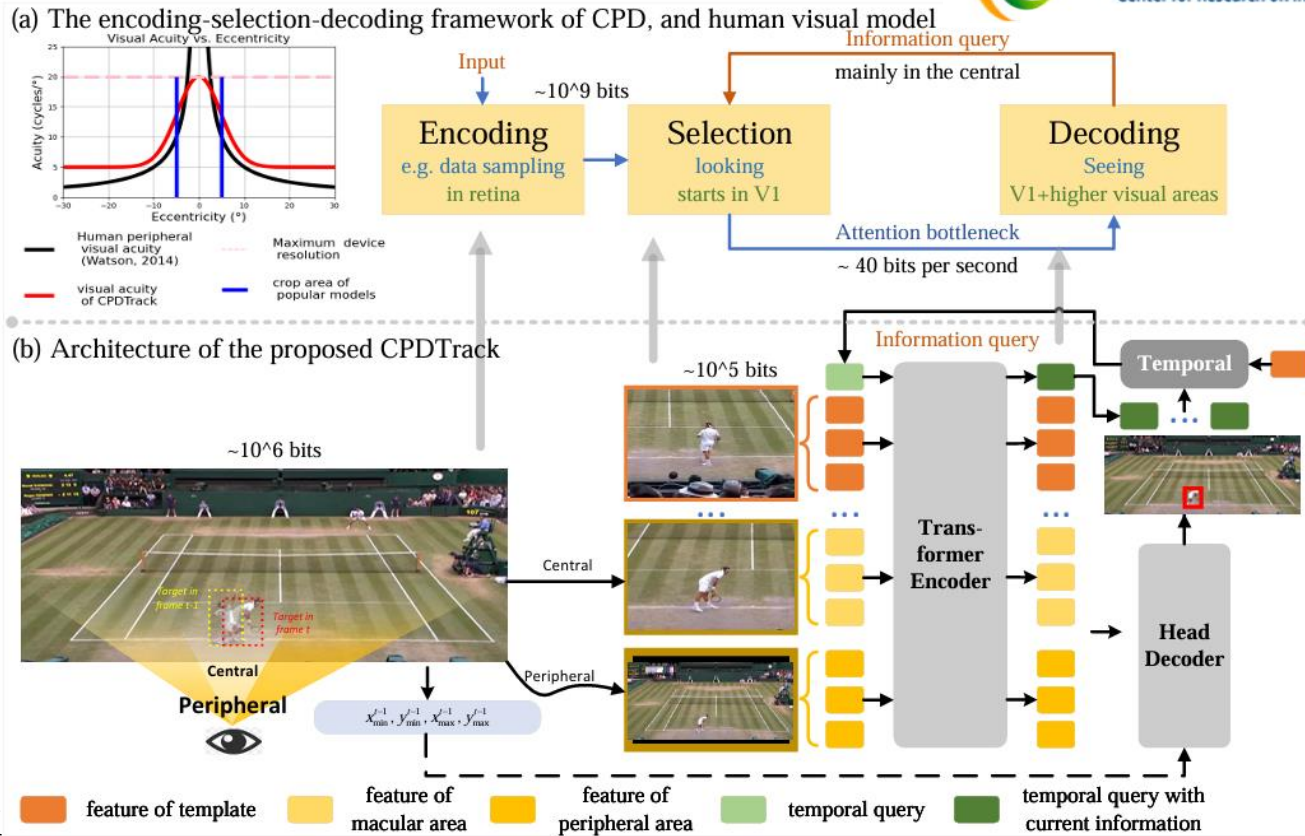
- LTT (featuring “absent”) and GIT (featuring “shot-cut”) disrupt the **continuity** of trajectory and apparent information of the target.
- The requirements for machine capabilities gradually rise from the **perception** level to the **cognition** level. Therefore, it is considered a more realistic simulation of the real world
- We combine these challenges into the “Spatial-Temporal Discontinuity challenge” (**STDChallenge**) of the target.

Motivation



- We try to address the STDChallenge from the perspectives of **environment-executor-evaluation** mechanism.
 - **Environment:** To address the lack of **STDChallenge** representation in current datasets, we developed a dedicated video environment.
 - **Executor:** To overcome the limitations of *motion consistency assumption* in mainstream algorithms, we designed a tracker inspired by human visual search that integrates both local and global perspectives.
 - **Evaluation:** To improve the accuracy of intelligence assessment, we introduced human participants in the benchmark and applied **Visual Turing Test** to precisely evaluate algorithmic intelligence.

Executors



CPDTrack

- **Central Vision**: Drawing from the **Central-Peripheral Dichotomy (CPD)** theory, the central vision is modeled as a Gaussian distribution. The region is cropped and resized to a specified dimension.

$$w_{t-1}^e = \frac{w_{t-1}}{sens_{x-1}} = \mathcal{S} \frac{w_{t-1}}{2\Phi(\frac{3w_{t-1}}{W}) - 1},$$

- **Peripheral Vision**: In line with the **CPD** theory's concept of peripheral vision, the current frame is resized to match the size of the central vision.

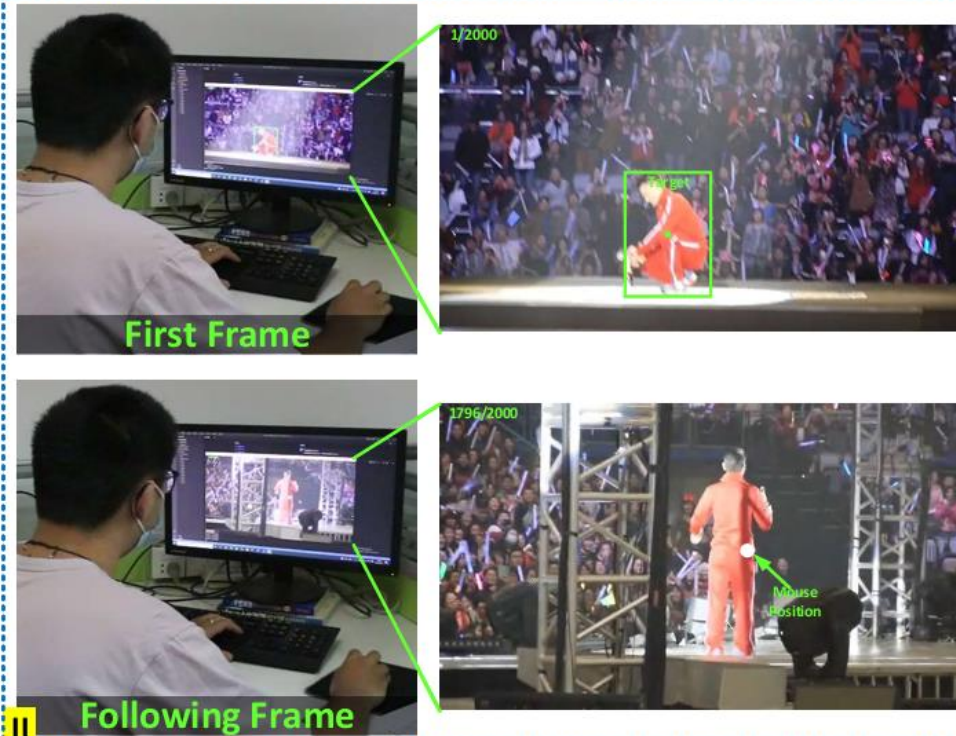
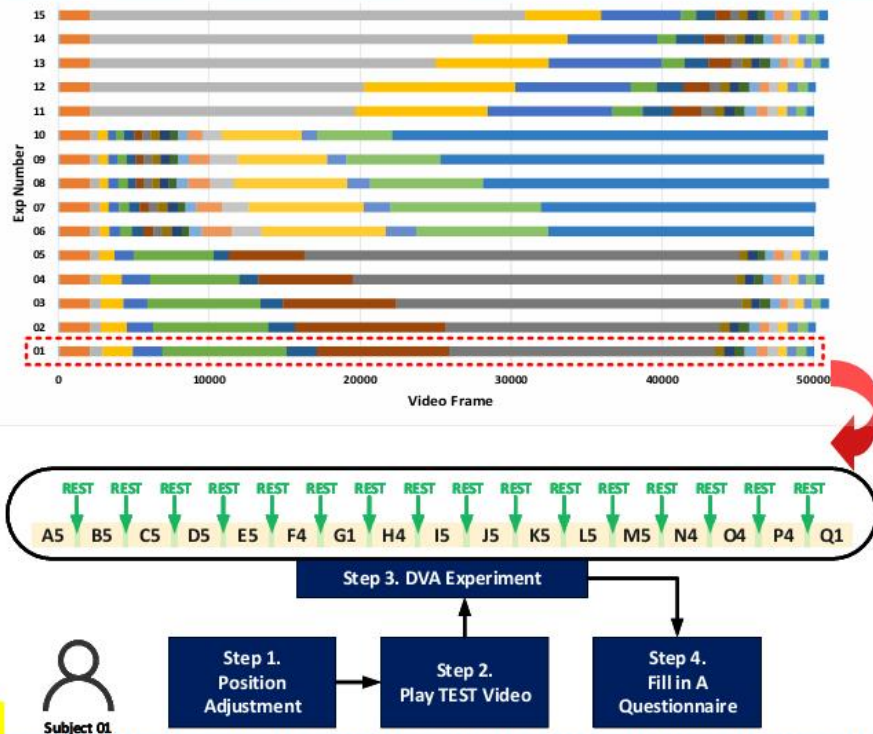
Table 3: Representative Benchmarks in STT, LTT, GIT and *STDChallenge Benchmark*

Subtask	Benchmark	Videos	Min frame	Mean frame	Max frame	Total frame	<i>absent</i>	<i>shotcut</i>
STT	OTB2015[5]	100	71	590	3872	59K	✗	✗
	VOT2016[54]	60	41	357	1500	21K	✗	✗
	VOT2018[55]	60	41	356	1500	21K	✗	✗
	VOT2019[44]	60	41	332	1500	20K	✗	✗
	GOT-10k[17]	10000	29	149	1418	1.45M	✗	✗
LTT	VOTLT2019[44]	50	1389	4305	29700	215K	✓	✗
	LaSOT[7]	1400	1000	2502	11397	3.5M	✓	✗
GIT	VideoCube[3]	500	4008	14920	29834	7.46M	✓	✓
LTT+GIT	<i>STDChallenge Benchamrk</i>	252	1000	5192	29700	1.3M	✓	✓

- We extracted sequences containing the *STDChallenge* from the LTT and GIT tasks to create the *STDChallenge Benchmark*, reducing bias from any single dataset.
- Additionally, we quantified the difficulty of the *STDChallenge*, taking into account the challenges of “absent” and “shot-cut” within the sequences.

$$STD = \frac{(n_a + n_s) \cdot l_a}{l^2},$$

Evaluation



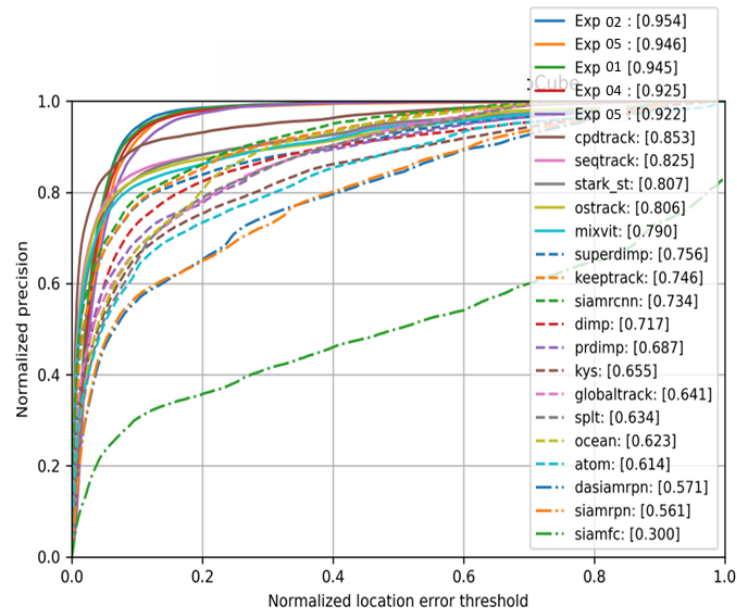
- Equipment Adjustment
- Testing and Training
- Target Tracking
- Questionnaire Completion

During the experiment, participants are permitted a limited number of pauses to adjust their state. This approach aims to eliminate cumulative errors and enhance hand-eye coordination.

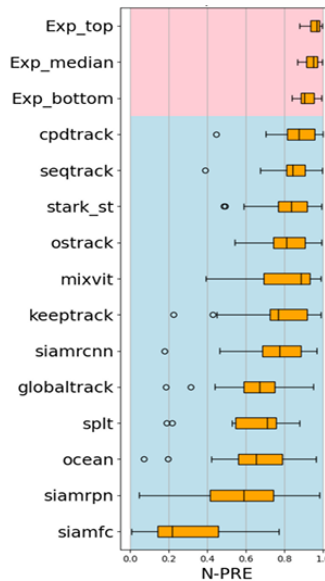
Results

Visual Turing

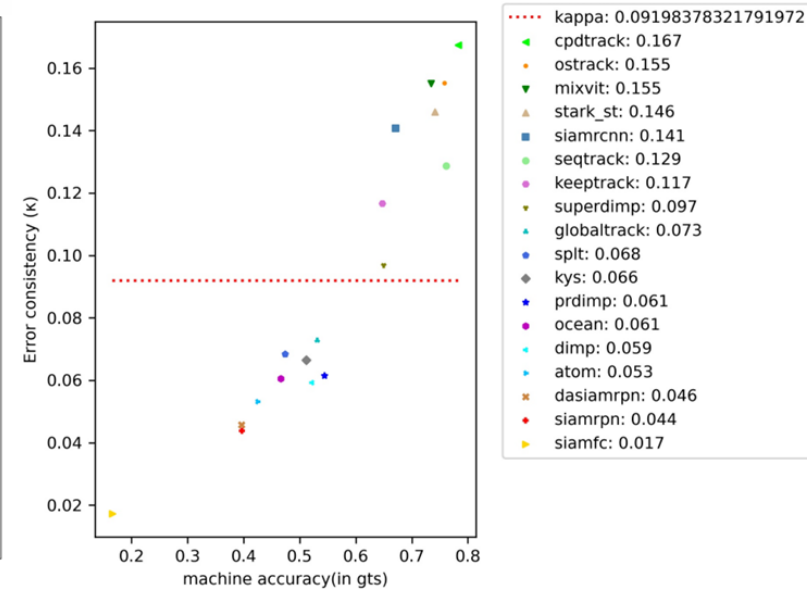
(a) N-PRE Score in *STDChallenge-Turing*



(b) Boxplot of Human and Machine



(c) Error consistency between humans and machines



- Current algorithms **still exhibit significant disparities** compared to human abilities in dynamic visual tasks; however, the gap between machines and human performance is **gradually narrowing**.
- When subjected to **STDChallenge**, machines performance tends to fluctuate considerably, whereas humans maintain relatively stable and robust tracking abilities.
- CPDTrack not only maintains SOTA performance but also achieves the highest error consistency (16.7), demonstrating the effectiveness of **human-like modeling**.

Results

Motion Model	Method	STDChallenge			VideoCube			VideoCube R-OPE		
		N-PRE	PRE	SUC	N-PRE	PRE	SUC	N-PRE	SUC	Robust
CPD	CPDTrack	84.2	73.3	65.9	82.9	67.1	70.4	89.5	75.6	75.3
	SeqTrack [11]	81.9	71.9	66.8	76.8	54.0	63.5	88.3	72.5	74.6
	OSTrack [10]	79.1	68.9	64.6	73.7	50.7	61.8	85.8	71.3	74.4
	MixViT [9]	82.5	71.6	66.7	76.9	52.2	63.1	88.5	72.7	74.7
	STARK [45]	80.7	68.2	64.5	76.3	49.4	62.1	86.8	70.4	74.5
	KeepTrack [13]	80.4	64.3	62.8	73.0	37.9	54.3	83.0	64.4	73.8
Local Crop	Ocean [46]	57.1	39.9	40.7	53.9	19.5	34.2	74.8	51.2	73.7
	SuperDiMP [47]	72.6	56.7	56.5	64.6	31.4	47.4	80.1	61.2	74.3
	PrDiMP [47]	70.3	51.7	52.7	65.4	28.6	44.5	79.6	58.3	74.3
	DiMP [48]	65.9	47.0	48.6	54.6	18.7	37.1	77.2	56.0	74.0
	SiamRPN [30]	53.4	35.6	37.3	46.7	15.0	29.0	72.6	50.3	73.6
	ATOM [49]	57.8	39.8	40.8	43.6	14.0	26.7	75.2	53.1	73.8
Local-Global	KYS [50]	60.1	42.6	44.5	49.3	17.1	33.7	80.1	59.4	73.3
	SiamFC [14]	33.6	21.2	20.6	15.8	3.6	7.4	52.1	35.6	72.7
Global	SPLT [33]	60.9	38.2	40.3	56.5	15.7	33.7	72.4	47.6	73.5
	DaSiamRPN [34]	53.4	35.4	37.1	46.3	14.4	29.1	72.2	50.4	73.6
Global	SiamRCNN [36]	75.3	62.8	60.7	72.6	47.9	58.8	80.5	65.8	74.5
	GlobalTrack [35]	65.5	49.5	49.5	64.3	29.6	46.1	72.7	53.7	74.3



• Visual Turing Test

- Human performance **does not always indicate correctness**, but humans can quickly relocate the target after STDChallenge.
- Humans can **recognize environmental factors** closely related to the target.
- Even when the target is absent, humans **are not distracted by the background**.
- humans show **strong robustness against occlusion**.

• Benchmark Evaluation

- CPDTrack outperformed existing trackers in STDChallenge. Specifically, N-PRE and PRE improved by 1.7 and 1.4, respectively, with notable performance gains on challenging datasets (e.g. VideoCube).

Summary

- Inspired by the **CPD theory**, we propose a new tracker named **CPDTrack** to achieve **human-like visual search ability**.
- To further evaluate and analyze **STDChallenge**, we create **the STDChallenge Benchmark**.
- Additionally, by introducing human subjects, we conduct a detailed assessment of the algorithm's intelligence by comparing its performance to human responses under the **STDChallenge**.
- Our extensive experiments demonstrate that the proposed CPDTrack not only achieves SOTA performance in this challenge but also narrows the behavioral differences with humans.
- In summary, our research underscores the importance of human-like modeling and offers strategic insights for advancing intelligent visual target tracking