NEURAL INFORMATION PROCESSING SYSTEMS

中国科学院自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES

RISE 智能系统与工程研究中心 Center for Research on Intelligent System and Engineering

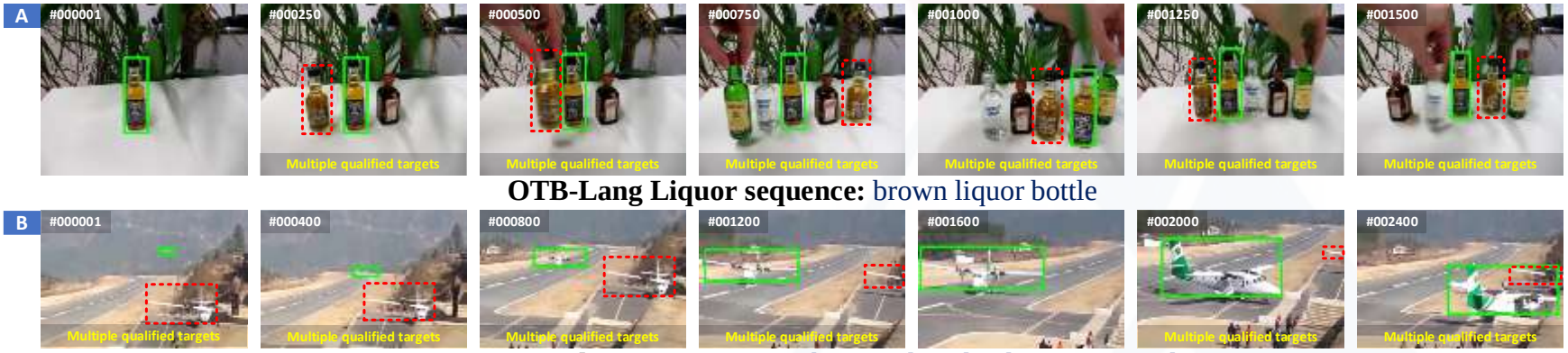
A Multi-modal Global Instance Tracking Benchmark (MGIT): Better Locating Target in Complex Spatio-temporal and Causal Relationship

Shiyu Hu[1,2] **Dailing Zhang**[1,2] **Meiqi Wu**[3] **Xiaokun Feng**[1,2]
Xuchen Li[4] **Xin Zhao**[1,2] **Kaiqi Huang**[1,2,5]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Institute of Automation, Chinese Academy of Sciences
[3]School of Computer Science and Technology, University of Chinese Academy of Sciences
[4]School of Computer Science, Beijing University of Posts and Telecommunications
[5]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
{hushiyu2019, zhangdailing2023, fengxiaokun2022}@ia.ac.cn, wumeiqi18@mails.ucas.ac.cn,
xuchenli@bupt.edu.cn, {xzhao,kqhuang}@ia.ac.cn

Motivation

- The single object tracking algorithm performs poorly in **complex scenes (long sequences & complex spatio-temporal causal relationships).**
- Some recent researches have considered studying from a multi-modal perspective:
  - **Limitations 1. Short sequence** (from hundreds of frames to thousands of frames)
    → **Simple narrative content**
  - **Limitations 2. Inaccurate semantic annotation** (describing only the information of the first frame, and there may be multiple objects in the scene that fit the description)
    → **Misguide algorithms**



OTB-Lang Liquor sequence: brown liquor bottle

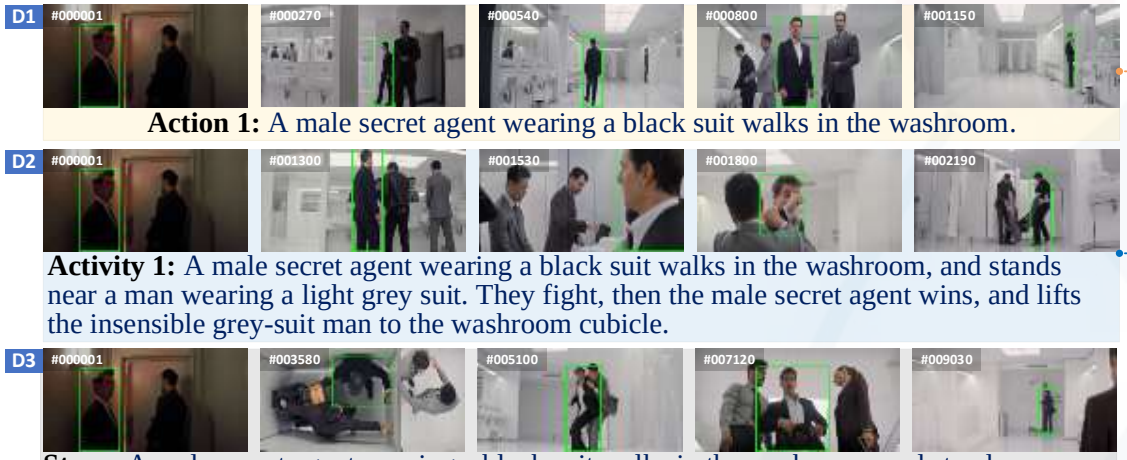
LaSOT airplane-1 sequence: white airplane landing on ground

TNL2k Arrow_Video_ZZ04_done sequence: the second arrow from left to right

Limitations of existing works

Motivation

- Some recent researches have considered studying from a multi-modal perspective:
  - **Limitations 1. Short sequence** (from hundreds of frames to thousands of frames)
    - → **Simple narrative content**
    - → **Using longer sequences with more complex narratives**
  - **Limitations 2. Inaccurate semantic annotation** (describing only the information of the first frame, and there may be multiple objects in the scene that fit the description)
    - → **Misguide algorithms**
    - → **Design a multi-granular annotation strategy to portray long videos**



An example of our methods

Methods and Contributions

Contribution 1 : Multi-modal Global Instance Tracking Dataset (MGIT)

- **Rich subject matter** → Sufficiently covers the complex spatio-temporal causal relationships of long videos

Story: A pink cartoon pig wearing red clothes talks to her family members on the grassland. Today, the red-clothes pig and her family aim to visit a castle. They go to the castle in a red car, and the red-clothes pig sits in the back. They stop the vehicle nearby the foothills and walk to the castle. At the entrance of the castle, they meet a white cartoon pig wearing gray armor. The red-clothes pig first talks with the gray-armor pig, then they are invited to visit the castle. The red-clothes pig walks with her family into the castle and sits beside a blue-clothes pig on the chair. After that, they have a meal in the castle's living room, and the red-clothes pink pig gets a gift from a yellow-clothes pig after the meal. Finally, the red-clothes pig walks with her family members on the stairway, and then stands at the top of the tower.

Story: A black gorilla holding a lady in white crouches on a gray building, and some airplanes attack them. He then walks and climbs to the top of the grey building. After that, he stands atop the grey building, hits an airplane, fights with a gray soldier in the other airplane, and finally crouches on the gray building.

Story: A black motorcycle is checked by a man with orange and white clothes in the yard; then, the man rides this black motorcycle in the yard. As an obstacle race, the black motorcycle first bounces across obstacles in the playground, then bounces across obstacles in the street. After that, it bounces across obstacles near the pool and across obstacles in the stream. After a brief break, the black motorcycle bounces across obstacles in the playground, then across obstacles near the pool, and finally across obstacles in the stream.

Story: A small basketball is played by a boy with a grey t-shirt and black shorts, and then inflated by a man with a red t-shirt and black pants in the skatepark. After that, the basketball is played by the boy, and then played by the man. After they practice, the basketball is holden by the boy from the skatepark to outdoors; then it is played by the boy outdoors. Finally, The basketball then is carried away by the boy.

Story: A brown cello is played by a man with white shirt and black pants in the room.

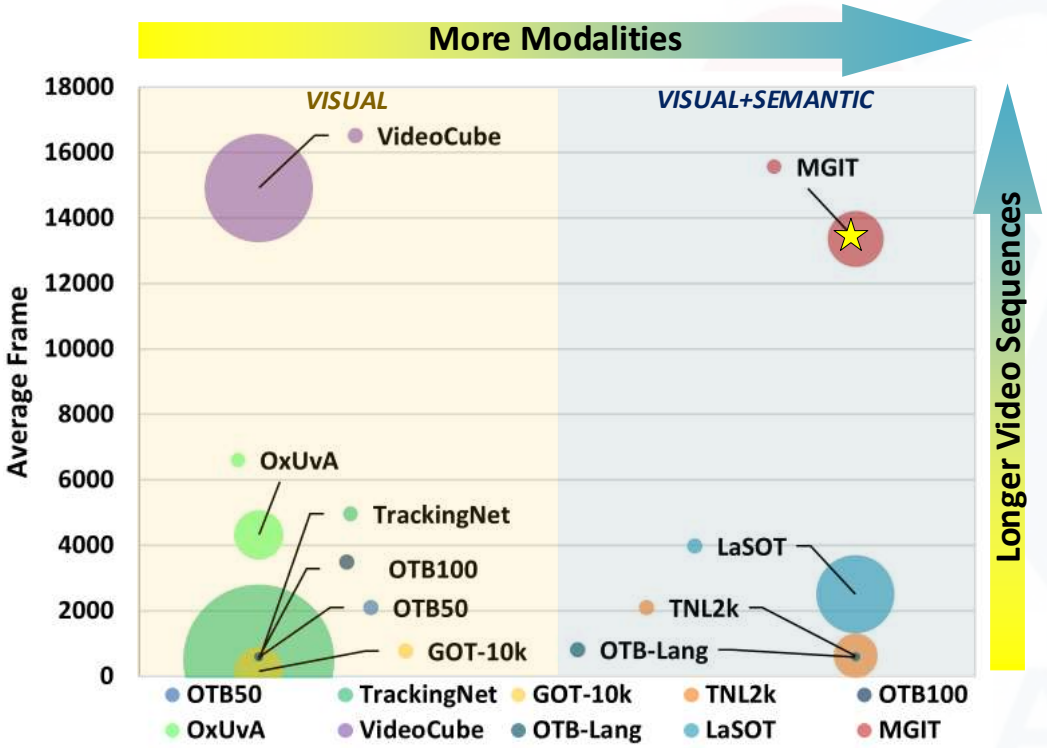
Story: A red cap is worn by a man with a gray t-shirt on the soccer court.

Several examples of MGIT

Methods and Contributions

Contribution 1 : Multi-modal Global Instance Tracking Dataset (MGIT)

- **Large-scale, Multi-modal**
  - 150 long videos
  - **2.03 million frames**
  - The average length of a single video is **13,500 frames**



Comparison between MGIT with other SOT benchmarks

Methods and Contributions

Contribution 2: Hierarchical Multi-Granularity Semantic Annotation Strategy

- Applying **hierarchical structure inspired by human cognition** for multi-granular annotation
  - **Action :** Determining annotation dimensions from both **natural language grammar structure** and **video narrative content**
    - **Natural Language Grammar Structure :** Subject, Predicate, Object, Adverbial of time, Adverbial of place
    - **Video Narrative Content :** Time, Location, Character, Event

Methods and Contributions

Contribution 2: Hierarchical Multi-Granularity Semantic Annotation Strategy

- Applying **hierarchical structure inspired by human cognition** for multi-granularity annotation
  - **Action :** Determining annotation dimensions from both **natural language grammar structure** and **video narrative content**
  - **Activity :** Using **causality** as a basis for classification

Action 4 : more suitable as the Result of activity 1

Action 5 : more suitable as the Cause for activity 2

Cause → **Result**

Methods and Contributions

Contribution 2: Hierarchical Multi-Granularity Semantic Annotation Strategy

- Applying **hierarchical structure inspired by human cognition** for multi-granularity annotation
  - **Action :** Determining annotation dimensions from both **natural language grammar structure** and **video narrative content**
  - **Activity :** Using **causality** as a basis for classification
  - **Story :** To enhance **temporal and causal relationships**, guiding words such as "first, then, after that, finally," can be used on the basis of actions and activities

Methods and Contributions

Contribution 2: Hierarchical Multi-Granularity Semantic Annotation Strategy

- **High-quality annotation**
  - **Accuracy :** 77,652 words
  - **Generality :** 921 non-repetitive vocabulary

Word cloud visualization of semantic annotations in the MGIT benchmark

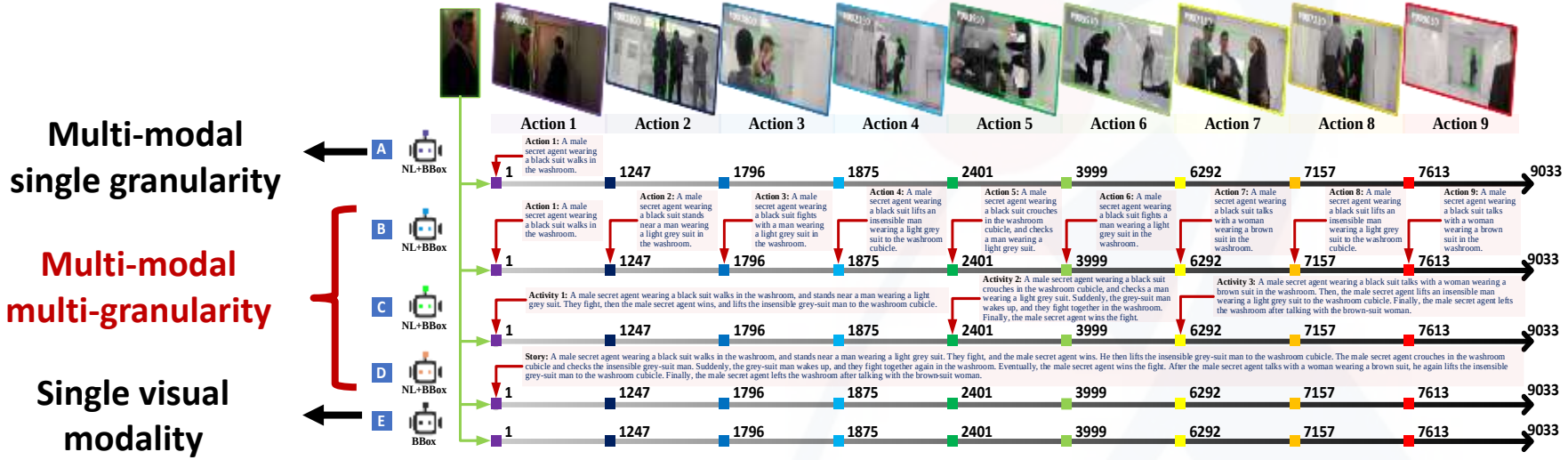
Word cloud visualization of actions in the MGIT benchmark

Word cloud visualization of scenes in the MGIT benchmark

Methods and Contributions

Contribution 3: Evaluation mechanism for multimodal tracking tasks

- Expand the evaluation mechanism by conducting experiments under both **traditional evaluation mechanisms (multi-modal single granularity, single visual modality)** and **evaluation mechanisms adapted to this work (multi-modal multi-granularity)**.



Adapted evaluation process for different task settings

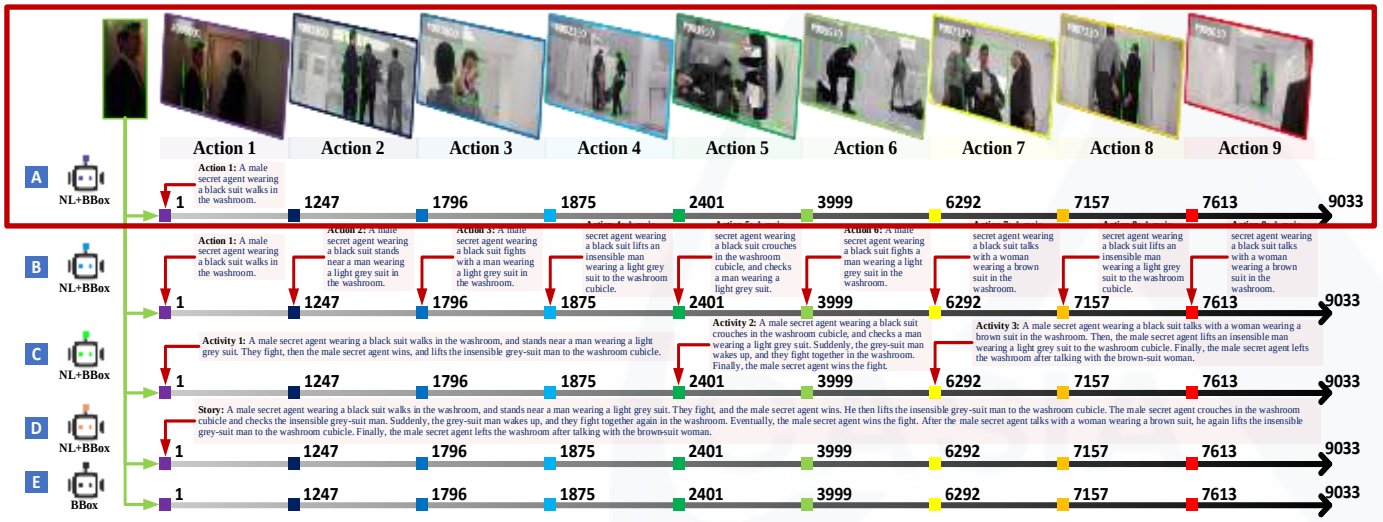
Experiments and Results

Experimental analysis: MGIT includes more challenges

- Compared to other multi-modal single object tracking benchmarks, the algorithm's performance significantly declines on MGIT (based on evaluation mechanism A)

Table 1: Results on different multi-modal benchmarks (based on mechanism A in Figure 6).

Tracker	OTB-Lang [1]		TNL2k [3]		LaSOT [2]		LaSOT$_{Ext}$ [17]		LaSOT$_{Sub}$		LaSOT$_{NLC}$		MGIT	
	PRE	SR	PRE	SR	PRE	SR	PRE	SR	PRE	SR	PRE	SR	PRE	SR
SNLT [46]	0.848	0.666	0.081	0.100	0.475	0.459	0.306	0.262	0.527	0.495	0.513	0.483	0.004	0.036
VLT_SCAR [42]	0.898	0.739	0.556	0.497	0.677	0.630	0.503	0.428	0.670	0.633	0.659	0.633	0.124	0.177
VLT_TT [42]	0.931	0.764	0.583	0.539	0.714	0.670	0.549	0.465	0.707	0.660	0.721	0.662	0.324	0.474
JointNLT [18]	0.856	0.653	0.598	0.552	0.640	0.607	0.457	0.398	0.624	0.583	0.707	0.651	0.433	0.603



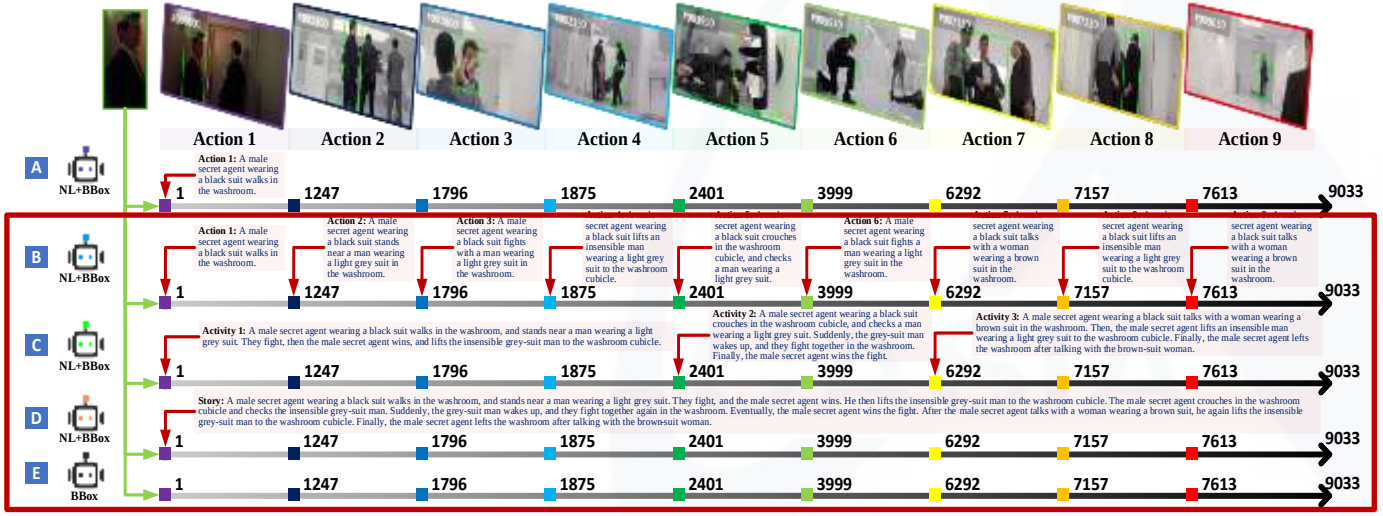
Experiments and Results

Experimental analysis: MGIT includes more challenges

- Multimodal algorithms still need to overcome performance bottlenecks (based on evaluation mechanisms B-E) :
  - How to **handle long semantic information**
  - How to **achieve alignment between visual modality and semantic modality**

Table 2: Results of different trackers on MGIT.

Tracker	Architecture	Initialize	Mechanism	PRE	N-PRE	SR
SiamCAR [11]	SNN	BBox	E	0.116	0.378	0.183
SiamRCNN [10]	SNN	BBox		0.512	0.707	0.591
PrDiMP [12]	SNN+CF	BBox		0.296	0.602	0.453
KeepTrack [13]	SNN+CF	BBox		0.373	0.695	0.519
TransT [39]	Transformer	BBox		0.447	0.670	0.539
MixFormer [14]	Transformer	BBox		0.526	0.775	0.629
OSTrack [15]	Transformer	BBox		0.476	0.706	0.583
GRM [16]	Transformer	BBox		0.500	0.718	0.597
SNLT [46]	SNN	NL&BBox	Action (B)	0.004	0.226	0.036
			Activity (C)	0.004	0.234	0.038
			Story (D)	0.005	0.230	0.040
VLT_SCAR [42]	SNN	NL&BBox	Action (B)	0.116	0.354	0.167
			Activity (C)	0.124	0.382	0.180
			Story (D)	0.127	0.403	0.184
VLT_TT [42]	Transformer	NL&BBox	Action (B)	0.318	0.602	0.468
			Activity (C)	0.325	0.627	0.485
			Story (D)	0.322	0.616	0.480
JointNLT [18]	Transformer	NL&BBox	Action (B)	0.445	0.786	0.610
			Activity (C)	0.441	0.780	0.605
			Story (D)	0.433	0.773	0.600



Summary

Summary of work

- Starting from improving the tracking capability of algorithms in complex spatio-temporal causal relationships:
  - Constructing a **multi-modal global instance tracking dataset**
  - Designing **hierarchical multi-granular semantic annotation strategy** based on human cognitive structures
  - Constructing **evaluation mechanisms** tailored for multi-modal single object tracking tasks to effectively analyze the performance bottlenecks of existing methods

Platform for more information

URL: http://videocube.aitestunion.com/

智能系统与工程研究中心
Center for Research on Intelligent System and Engineering

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

Thanks!

CRISE @ CASIA