Dynamic Group Detection using VLM-augmented Temporal Groupness Graph

Kaname Yokoyama[1], Chihiro Nakatani[1], Norimichi Ukita[1]

[1] Toyota Technological Institute, Japan

Task

Detecting the **groups** of people in videos

Inputs: Video, Bounding Boxes

outputs: G1, G2

Limitations of Previous methods

Limitation 1： Ignore global context

Without Global Context / With Global Context

Still image

Video

t_1 t_2 Time

Limitation 2： Cannot detect the temporal changes of groups

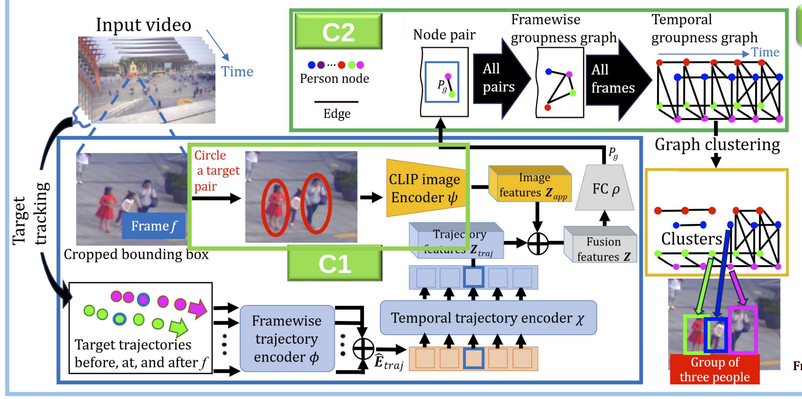
These two people join up to form a group.

Correct

NOT Correct

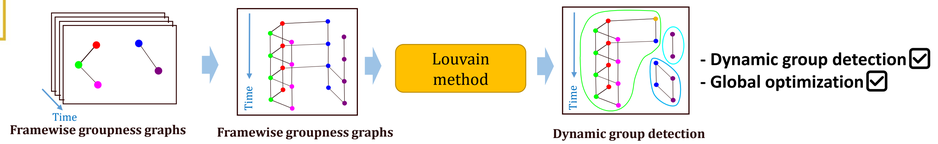
Note: People enclosed in bounding boxes of the same color are identified as being in the same group.

Proposed Method



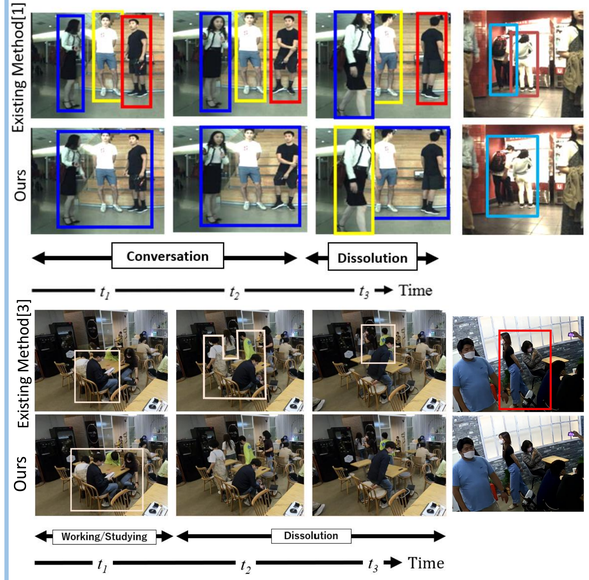
Contribution1: Capture local and global context with CLIP and red-circle annotations

Contribution2: Clustering for dynamic group detection



Experiments

Qualitative results



Quantitative results

Dynamic group detection on JRDB

Method	Precision	Recall	F1
S3R2 [3]	0.522	0.662	0.584
P-HAR [2]	0.568	0.697	0.626
GroupTrans [1]	0.514	0.534	0.524
Ours	**0.724**	**0.820**	**0.769**

Dynamic group detection on Cafe

Method	Precision	Recall	F1
S3R2 [3]	0.576	0.700	0.631
GroupTrans [1]	0.263	0.305	0.283
Ours	**0.681**	**0.904**	**0.776**

Component Analysis: Encoder

Using other image encoders (Trained only on images)

Dataset (Task)	Model	Group detection Precision	Recall	F1
JRDB (dynamic)	ResNet50	0.686	0.609	0.645
	ViT-L/16	0.652	0.666	0.659
	DINOv2	0.612	0.600	0.644
	Ours	**0.724**	**0.820**	**0.769**

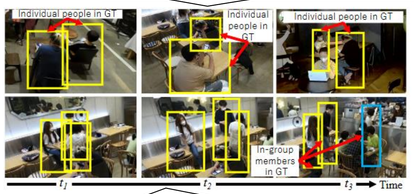
Component Analysis: Visual prompts

Using other visual prompts

Dataset (Task)	Model	Group detection Precision	Recall	F1
JRDB (dynamic)	No prompt	0.700	0.689	0.695
	Mask	0.584	0.750	0.657
	A circle	0.640	0.666	0.653
	Two circles	**0.724**	**0.820**	**0.769**

Limitations

Stronger positional cues



Lack of temporal appearance features

References

1. J. Zhang et al. Toward Grouping in Large Scenes With Occlusion-Aware Spatio-Temporal Transformers. TCSVT2024.
2. R. Han et al. Panoramic Human Activity Recognition. ECCV2022.
3. J. Li et al. Self-supervised Social Relation Representation for Human Group Detection. ECCV2022.