Cohabitation Discovery via Spatial and Temporal Clustering

LIU RUIZHE
liurz05@vanke.com

Outline

01 Introduction

02 Problems

03 Solution

04 Experiments

05 Conclusion

Introduction

- The relationships among different people are not explored much.
- In a scenario of residential entries, knowing the relationships could
  a) prevent tailgaters;
  b) identify unregistered strangers;
  c) prevent the disease from spreading during the Cov-19 period.
- There are very limited work on such relationship discovery problems using cameras. This may be due to the lack of public datasets on long-term video records.
- The figures captured from cameras lack identity information due to privacy concern.
- Contingency: two people entering or exiting the same entry at the same time by chance does not mean they know each other.
- A long-term observation can increase the confidence in the relationships of two persons showing together.

Introduction – Related Works

Single Entity Recognition

- Using person features; no accompanies.
- Comparing face similarity; requiring proper camera setting.

Relationship Detection

- Detecting the relationship using ranking function; not human figures but static objects.

Relationship Modeling

- Modeling the relationship using a graph; static images of objects rather than persons.

Problems

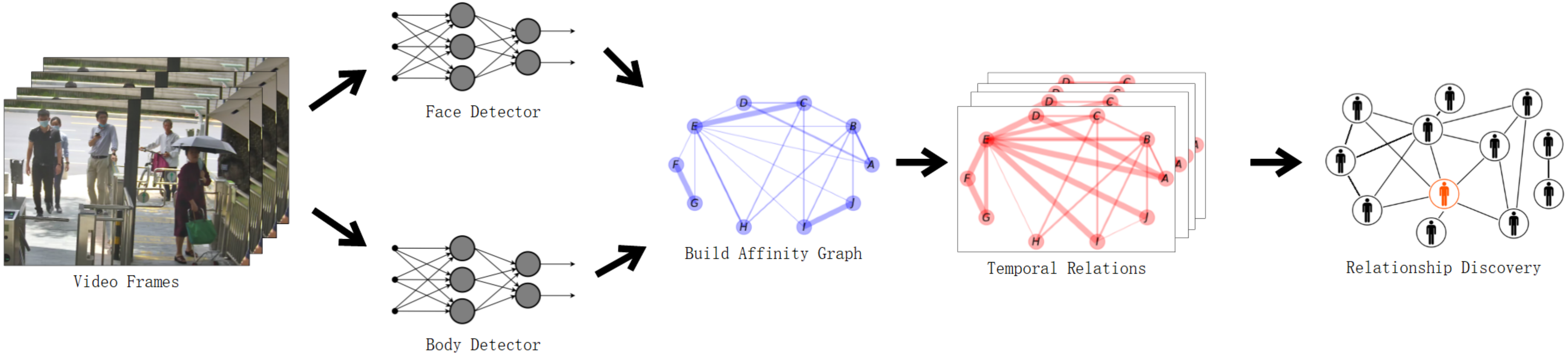
Assumptions	Frequent co-occurrence at a residence’s entry in a long timespan indicates cohabitation
Constraints	Faces are masked
	No pre-recording faces
	Occlusions always exist
	Computation power is limited

“Long-term co-occurrence relationship mining using unsupervised learning on residence entry cameras"

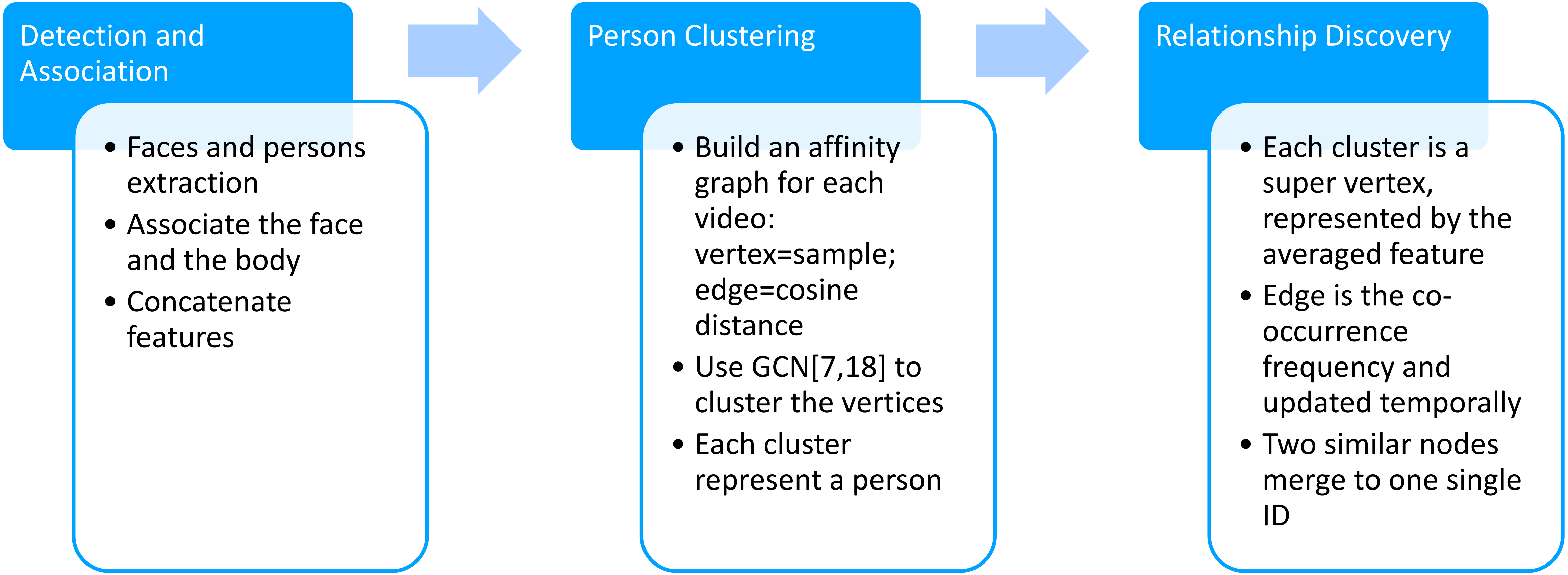
Solution

- Utilize the characters of “Residence”:

Features		Constraints		
Fixed Population	**Long-term Video Records**	**Mask**	**Privacy**	**Cost Efficient**
• People normally do not move frequently	• Similar background • Fixed position • Good error tolerance	• Facial masks are mandatory during the pandemic	• No pre-recorded data (e.g. face)	• Video must be computed in real time locally



Solution – Methodology



Detection and Association

- Faces and persons extraction
- Associate the face and the body
- Concatenate features

Person Clustering

- Build an affinity graph for each video: vertex=sample; edge=cosine distance
- Use GCN[7,18] to cluster the vertices
- Each cluster represent a person

Relationship Discovery

- Each cluster is a super vertex, represented by the averaged feature
- Edge is the co-occurrence frequency and updated temporally
- Two similar nodes merge to one single ID

[7] Thomas N. Kipf and Max Welling, “Semi-supervised classification with graph convolutional networks,” in Proceedings of international conference on learning representations, 2017.

[18] Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, and Dahua Lin, “Learning to cluster faces on an affinity graph,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 2298–2306.

Experiments – Implementation

Data Collection

- Entry security camera videos
- 24*7 operating
- Mask-wearing is mandatory
- 1080p and 30fps

Data Pre-processing

- “Silent” clips removed
- Number of bounding box < 2

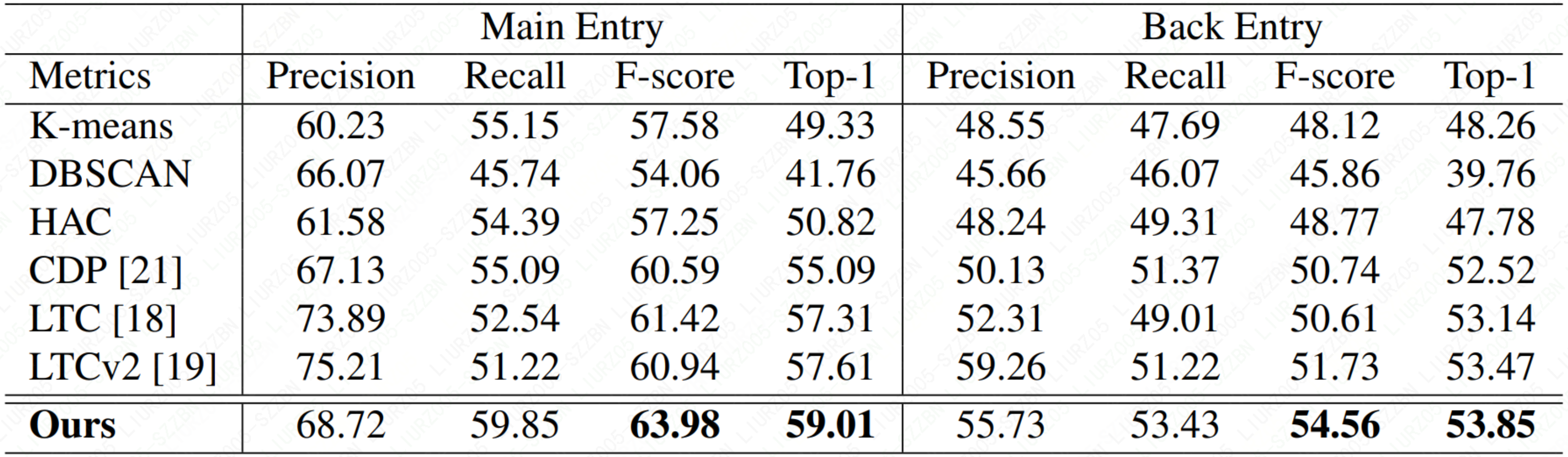
Implementation Details

- Using edge-computing to save the network load
- A light model is required
- ”Smart” gateway with Intel VEGA-300 series

Implementation Strategy

- Sacrifice accuracy while winning the computation efficiency
- The confidence and accuracy will be gradually improved by observing long-term data

Experiments – Person Clustering



	Main Entry				Back Entry			
Metrics	Precision	Recall	F-score	Top-1	Precision	Recall	F-score	Top-1
K-means	60.23	55.15	57.58	49.33	48.55	47.69	48.12	48.26
DBSCAN	66.07	45.74	54.06	41.76	45.66	46.07	45.86	39.76
HAC	61.58	54.39	57.25	50.82	48.24	49.31	48.77	47.78
CDP [21]	67.13	55.09	60.59	55.09	50.13	51.37	50.74	52.52
LTC [18]	73.89	52.54	61.42	57.31	52.31	49.01	50.61	53.14
LTCv2 [19]	75.21	51.22	60.94	57.61	59.26	51.22	51.73	53.47
Ours	68.72	59.85	**63.98**	**59.01**	55.73	53.43	**54.56**	**53.85**

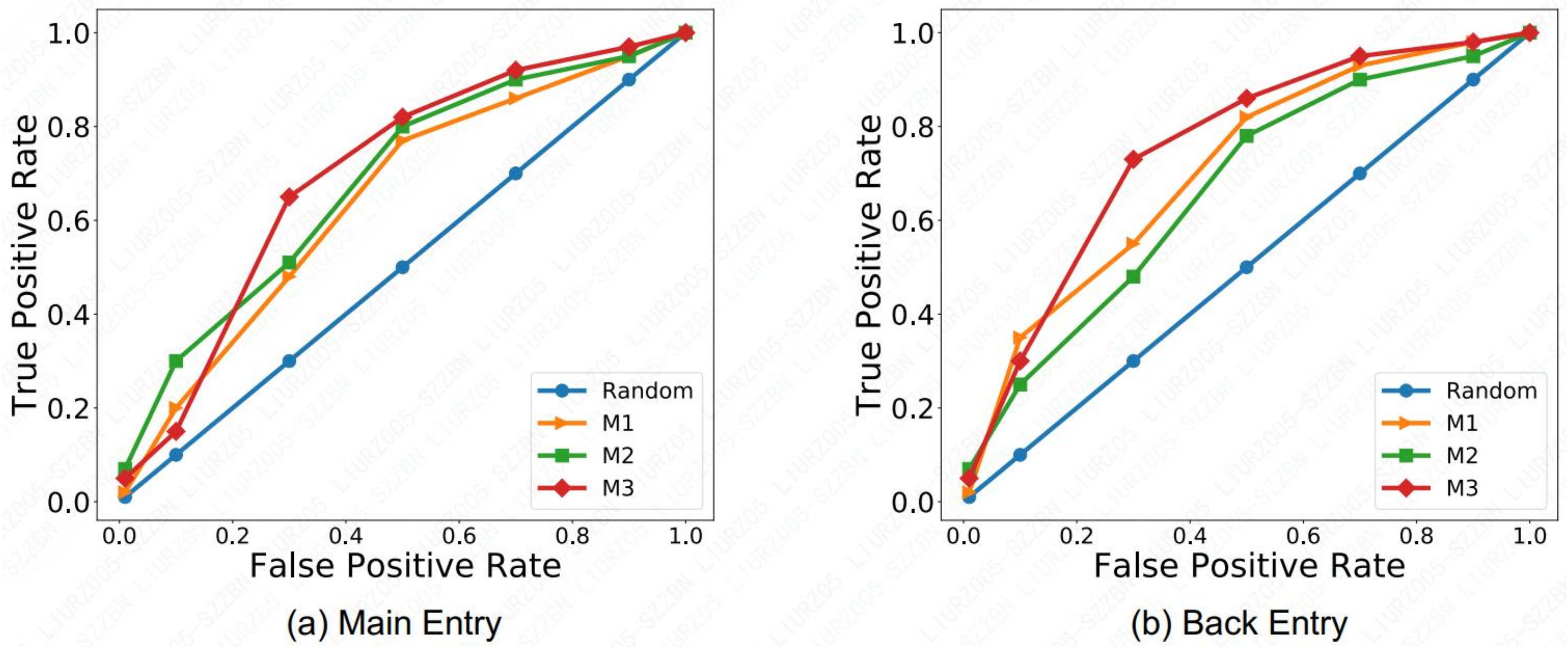
- F-score slightly better, due to utilizing both **SPATIAL** and **TEMPORAL** information.
- K-means: requires number clusters predefined, while we do not have.
- HAC, DBSCAN: require distribution assumption, while we cannot assume in real scenario.
- CDP, LTC, and LTCv2: use graph convolution techniques, but the lack of generalizability does not give good performance.
- Main entry performs better because much more people use main entry than back one.

[18] Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, and Dahua Lin, “Learning to cluster faces on an affinity graph,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 2298–2306.

[19] Lei Yang, Dapeng Chen, Xiaohang Zhan, Rui Zhao, Chen Change Loy, and Dahua Lin, “Learning to cluster faces via confidence and connectivity estimation,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13369–13378.

[21] Xiaohang Zhan, Ziwei Liu, Junjie Yan, Dahua Lin, and Chen Change Loy, “Consensus-driven propagation in massive unlabeled data for face recognition,” in Proceedings of European conference on computer vision, 2018.

Experiments – Relationship Discovery



(a) Main Entry

(b) Back Entry

- Adopt the protocol similar to the evaluation of face verification. [20]
- M1 is a vertex-related method that assumes the relationship is connecting the densest vertex.
- M2 is an edge-related method that chooses the edge with maximum weight (highest similarity) as the relationship connection.
- M3 is a multi-hop method that partitions the graph into components and relationships exist in the components.
- We can see all methods outperforms the random guess. Specifically, M3 gives the best result overall.

[20] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

Experiments – Complexity

- Three parts: feature extraction, person clustering, and relationship discovery.

Feature Extraction

- Using ResNet-50 and it can be replaced by a lightweight backbone in practice

Person Clustering

- Similar to [18] and it has a more efficient alternative [19] for real-world applications

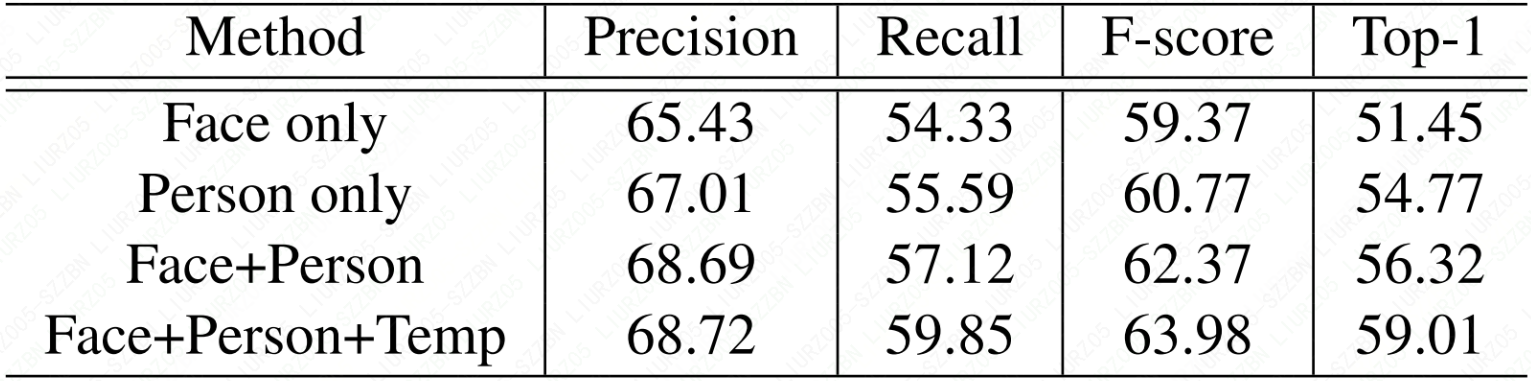
Relationship Discovery

- Only adds a little computation compared to person clustering, as the entire graph has been clustered into some large groups in the previous stage

[18] Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, and Dahua Lin, "Learning to cluster faces on an affinity graph," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 2298–2306.

[19] Lei Yang, Dapeng Chen, Xiaohang Zhan, Rui Zhao, Chen Change Loy, and Dahua Lin, "Learning to cluster faces via confidence and connectivity estimation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13369–13378.

Experiments – Ablation Study



Method	Precision	Recall	F-score	Top-1
Face only	65.43	54.33	59.37	51.45
Person only	67.01	55.59	60.77	54.77
Face+Person	68.69	57.12	62.37	56.32
Face+Person+Temp	68.72	59.85	63.98	59.01

- It is noticed that by gradually adding in components, the performance increases through all metrics.
- For example, the F-score using face only is 59.37. It increases to 62.37 after considering personal features. It reaches 63.98 by adding in temporal information.
- Comparing the state-of-art algorithms on person clustering tasks, the performance has a significant drop, which shows the problem is challenging when most faces are covered by masks

Experiments – Knowledge

Person E

- Much more connections than others in a relation graph
- Low similarity with the majority of people
- May be the security guy

Person C

- Only guy with a similar appearance with person E
- No co-occurrences with others except E
- Might be also a security guy from another entry

Persons F and G

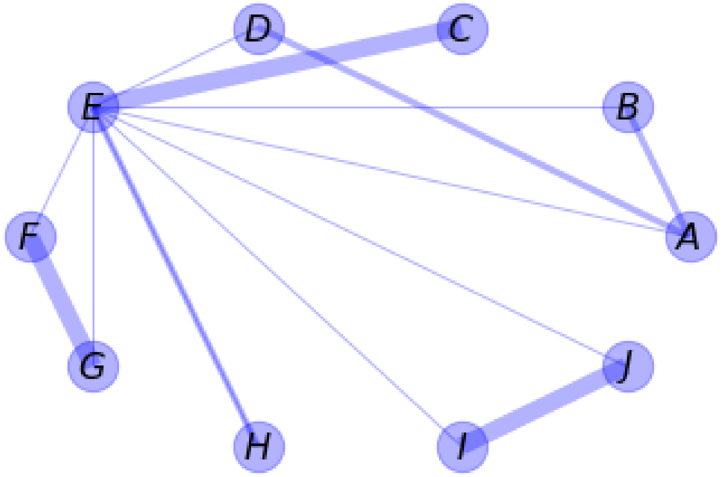
- Strong connections in both graphs
- Probably related to each other and stay in the same place
- Person F is recognized as male, mid-age and person G is recognized as female, teenager. They are likely father and daughter

Person H

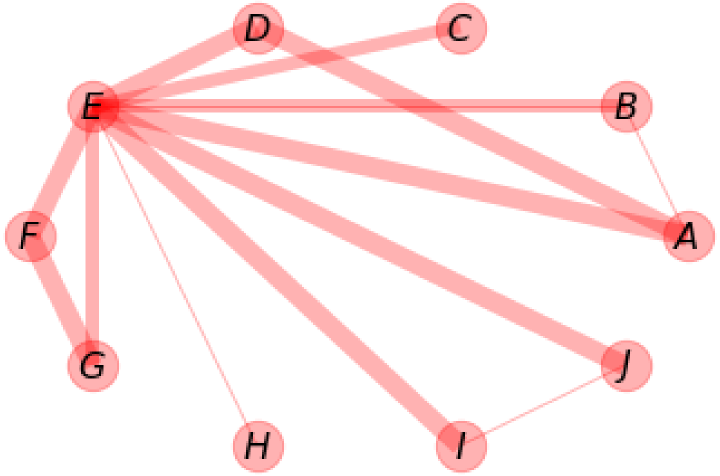
- Weak connections in both graphs
- Might be visitors or a “silent neighbor”
- We can further investigate the frequency of the person appearing in the long-term video records

Persons I and J

- Extremely high similarity
- Seldom appear in the same frame
- Might be the same person

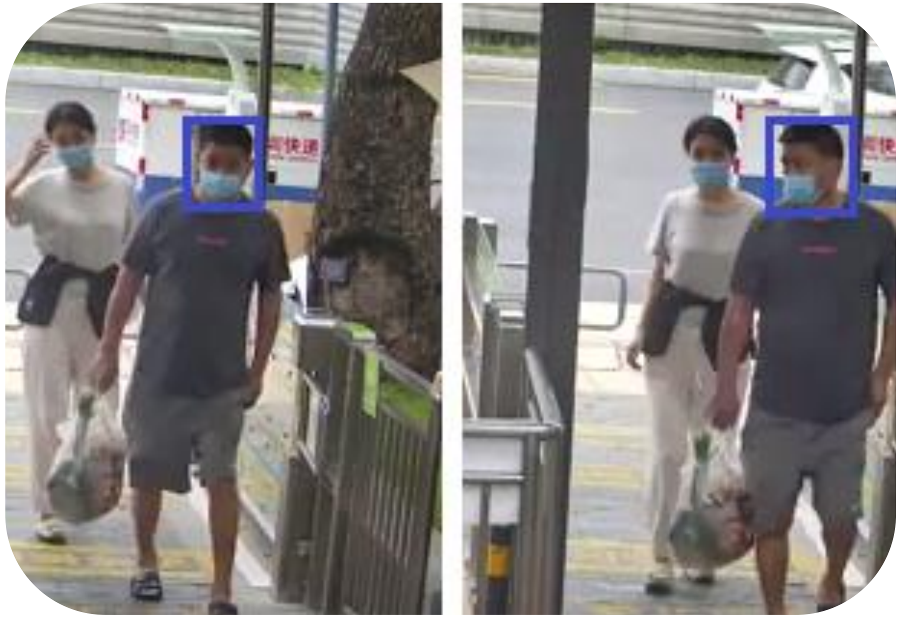


Similarity Graph



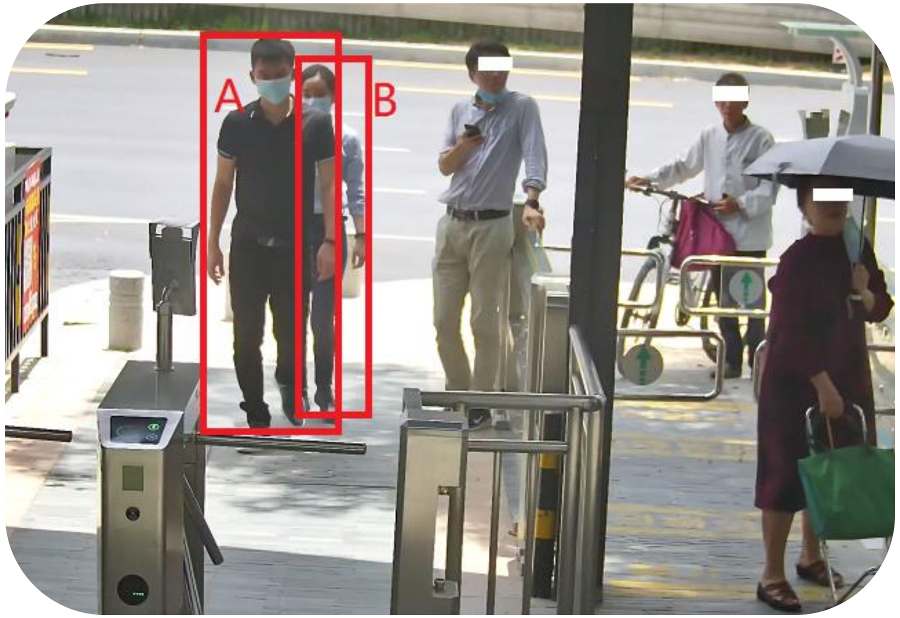
Frequency Graph

Experiments – Hard Cases



Feature Change

- The male person with the mask turns his head, leading to the change of face features



Occlusion

- Person B (female) hides behind person A (male)

Conclusion

- Use image features + spatial + temporal information to discover relationships
- Utilize features of “residence entry”.
- A long-term observation can increase the confidence in the relationships of two persons showing together.