Autonomous Systems and Biomechatronics Lab
Mechanical and Industrial Engineering
UNIVERSITY OF TORONTO

Robots Autonomously Detecting People: A Multimodal Deep Contrastive Learning Method Robust to Intraclass Variations

Authors
Angus Fung (Student Member, IEEE)
Beno Benhabib
Goldie Nejat (Member, IEEE)

Goal

Develop a deep learning person detection architecture for mobile robots in human-centered environments which is robust under intraclass variations (*e.g.*, occlusions, varying lighting, deformation).

Motivation

Robots need to autonomous detect multiple, dynamic people in various human-centered environments to engage in human-robot interactions, for example:

- In grocery stores, service robots can be used to find and guide customers to products,
- In long-term care homes, they can search and locate residents to perform daily activities,
- In hospitals, they can provide directional guidance,
- In urban search and rescue (USAR), they can search for victims.

Challenges

- Human-centered environments are typically crowded with people, or cluttered with objects (results in occlusions)
- Human-centered environments have variable lighting due to natural and artificial lighting sources (or lack thereof)
- Human body is articulated, and the presence of different types of clothing can result in deformation


Occlusions


Varying lighting


Deformation

Novel Contributions

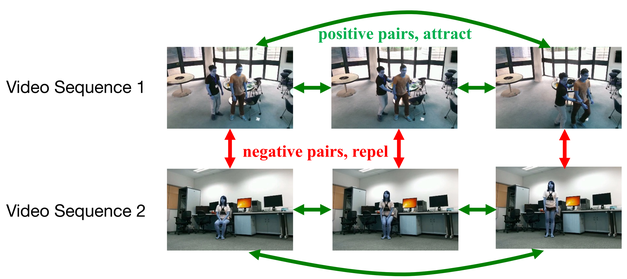
We present a two-stage multimodal DL person detection architecture for mobile robots consisting of:

1) a unique pretraining method Temporal Invariant Multimodal Contrastive Learning (*TimCLR*)
2) a Multimodal YOLOv4 (*MYOLOv4*) detector for finetuning.

Our main contributions include:

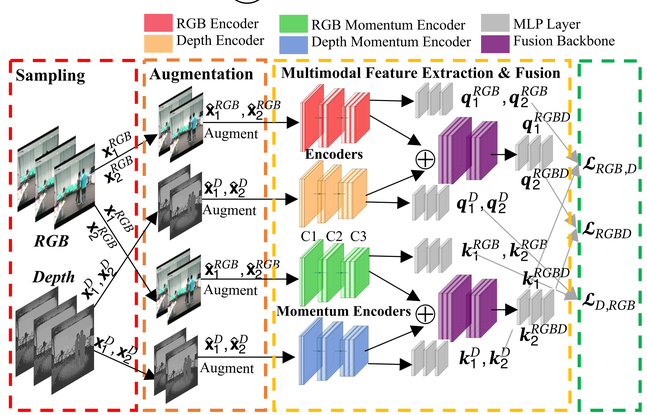
1) a new pretraining method, *TimCLR* which uniquely incorporates intraclass variations by generating multimodal image pairs from sampling video frames within a short temporal interval, thereby capturing natural variations in occlusion, pose and lighting
2) we uniquely incorporate contrastive learning within a fusion backbone network to contrast image features thereby capturing cross-modal invariances.

TimCLR Pair Generation

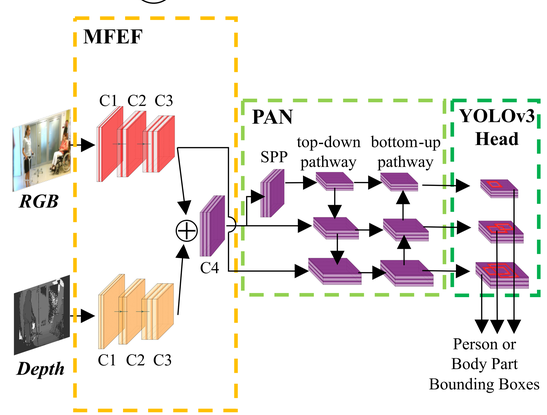


- **Contrastive learning** is an unsupervised framework which learns through positive pairs (attraction) and negative pairs (repulsion)
- Video frames within a short time interval capture natural variations in occlusion, lighting, and pose deformation

1 TimCLR Network



2 MYOLOv4 Detection Network



Experiments

Datasets:

- **Mobility Aids (MA):** contains multiple dynamic people in crowded hospital environment
- **USAR**: contains multiple static people/body parts in cluttered real-world environment
- **InOutDoorPeople (IOD)**: multiple dynamic people in in/outdoor campus environments

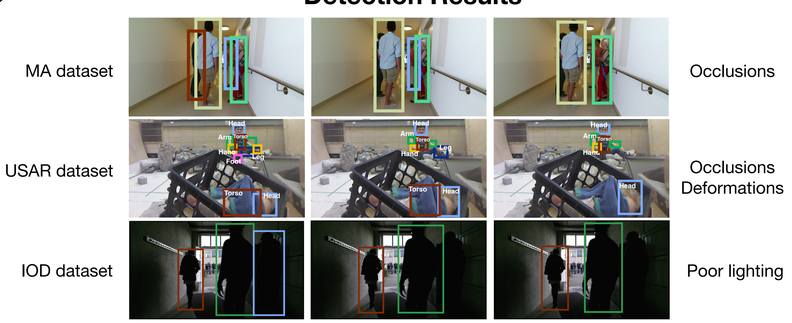
Metrics:

- **AP:** the mean average precision (mAP) where predictions with an Intersection over Union (IoU) > 0.5 are considered true positives
- **AP_{50}:** the averaged mAP over IoU = {0.5, 0.55, …, 0.95}

Comparison Results:

Methods	MA		USAR		IOD		Mem	FPS
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}		
Comparison of Detection Methods								
TimCLR + MYOLOv4	49.20	75.30	20.20	44.40	63.40	96.70	1.4	41
MYOLOv4	45.20	71.40	18.30	40.60	60.80	92.90	1.4	41
MEfficientDet	45.71	74.54	17.80	40.45	60.68	92.06	2.0	24
Comparison of Contrastive Pretraining Methods								
TimCLR + MYOLOv4	49.20	75.30	20.20	44.40	63.40	96.70	-	-
SimCLR + YOLOv4	43.40	70.20	14.50	28.30	58.30	92.40	-	-
Barlow + YOLOv4	43.60	70.30	13.70	32.10	60.50	92.10	-	-
MoCo v3 + YOLOv4	44.40	70.11	15.10	34.20	60.50	91.90		

Detection Results



- In the first and third scene, only our method identified all people
- In the second scene, only our method detected the partially occluded right foot

Conclusion

- Our multimodal person detection architecture for mobile robots learns person features which are invariant to natural variations in the environment such as person/body part occlusions, pose deformations, and varying lighting.
- Our new pretaining method *TimCLR + MYOLOv4* outperforms the existing detection methods in finding people in crowded and/or with varying lighting hospitals, university campuses, and cluttered USAR environments.
- Future work includes integrating our detection architecture within a mobile robot for real-time person detection in varying human-centered environments.