Elderly Assistance and Monitoring Using a Robotic Platform

Prof. PhD Eng. Irina Mocanu

irina.mocanu@upub.ro,
http://aimas.cs.pub.ro/people/irina.mocanu/

FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE

Computer Science & Engineering Department

AI-MAS Group

Introduction

Robots are being used

in a wide variety of domains:

- industry
- social
- entertainment or health care
- monitoring activities of daily living

TurtleBot

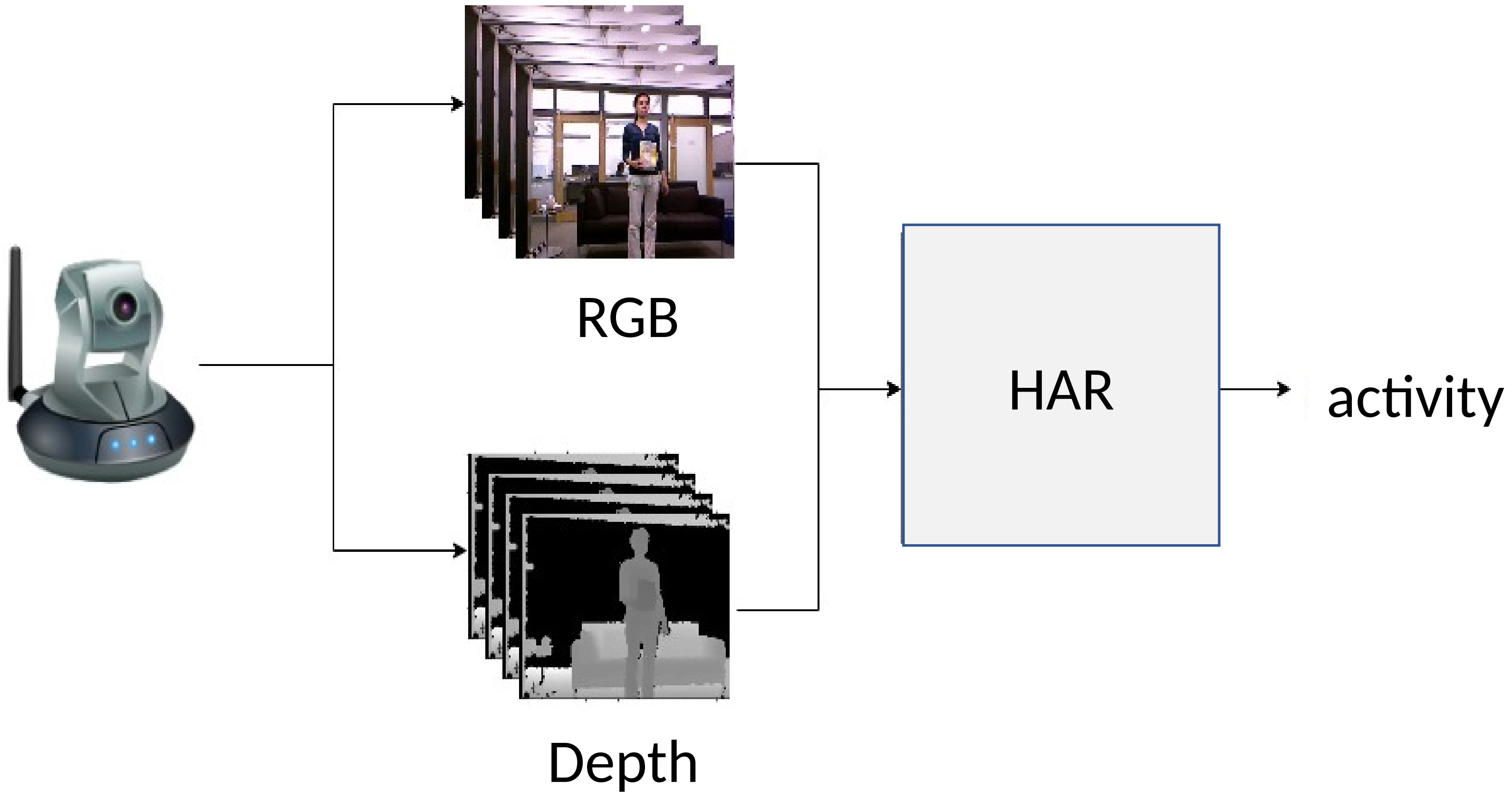
TIAGo

Robots provided by IT Center for Science and Technology

Solution

- Design a robotic framework for indoor monitoring and assistance
- Implement modules suitable for the robotics framework:
  - human activity recognition
  - object recognition
  - gesture recognition for human-robot interaction
  - abnormal object detection

Human Activity Recognition

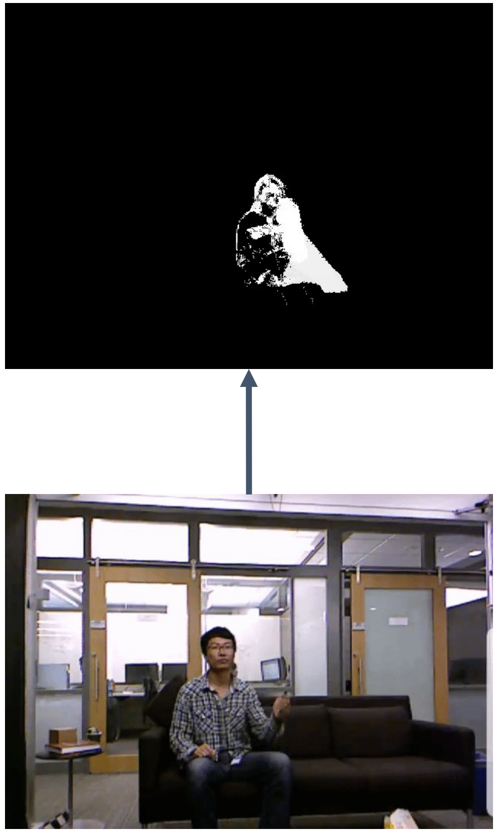


Fusion Mechanisms

Data fusion

- Motion history images -static description of a temporal action.
- Encode how an object or a subject move through time

Motion history images

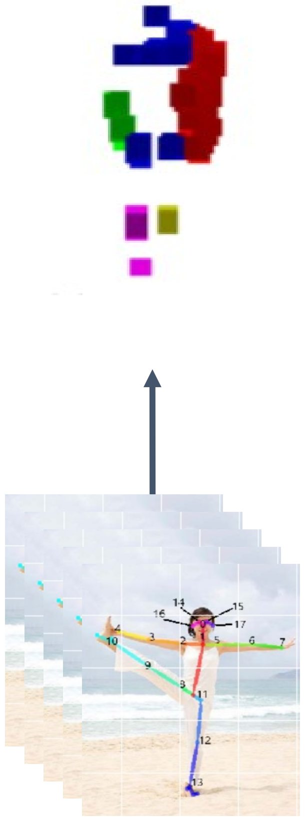


Fusion Mechanisms

Data fusion

Skeleton images represent a visual summary of skeleton information present throughout a video.

Skeleton images

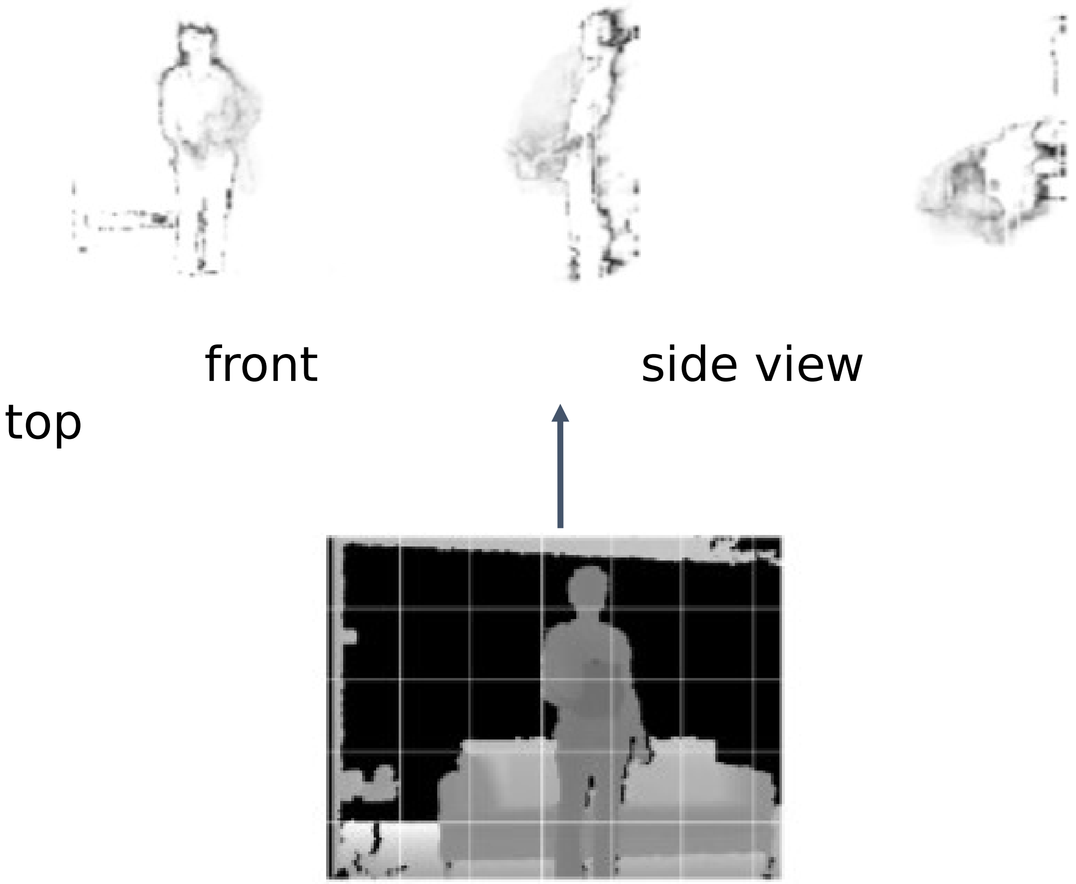


Fusion Mechanisms

Data fusion

Depth motion maps encapsulate motion information extracted from depth data.

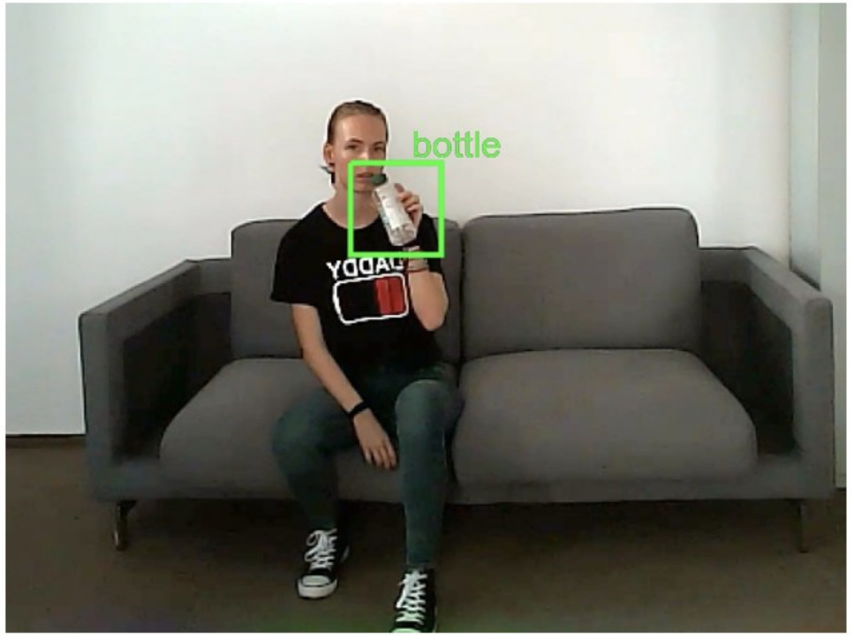
Depth motion maps

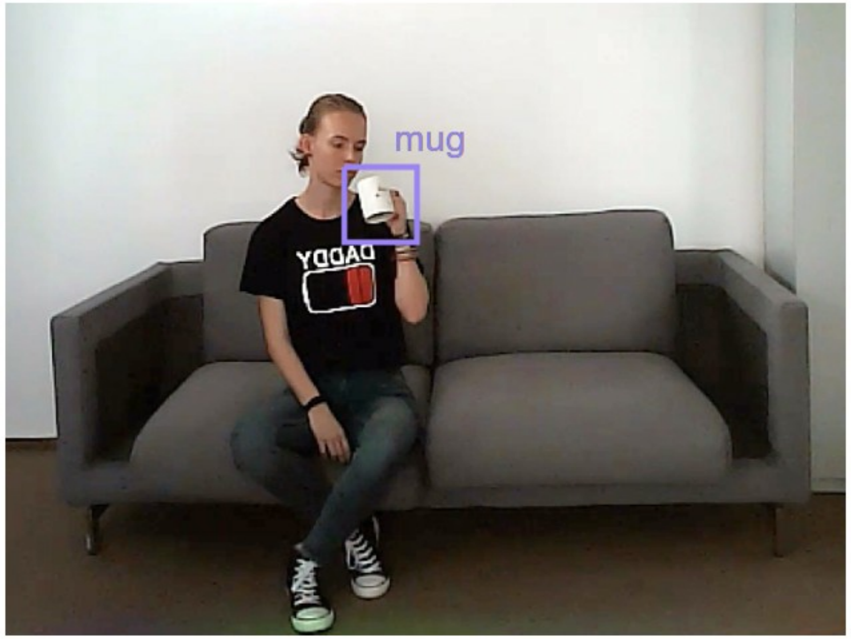


Fusion Mechanisms

Data fusion

Context

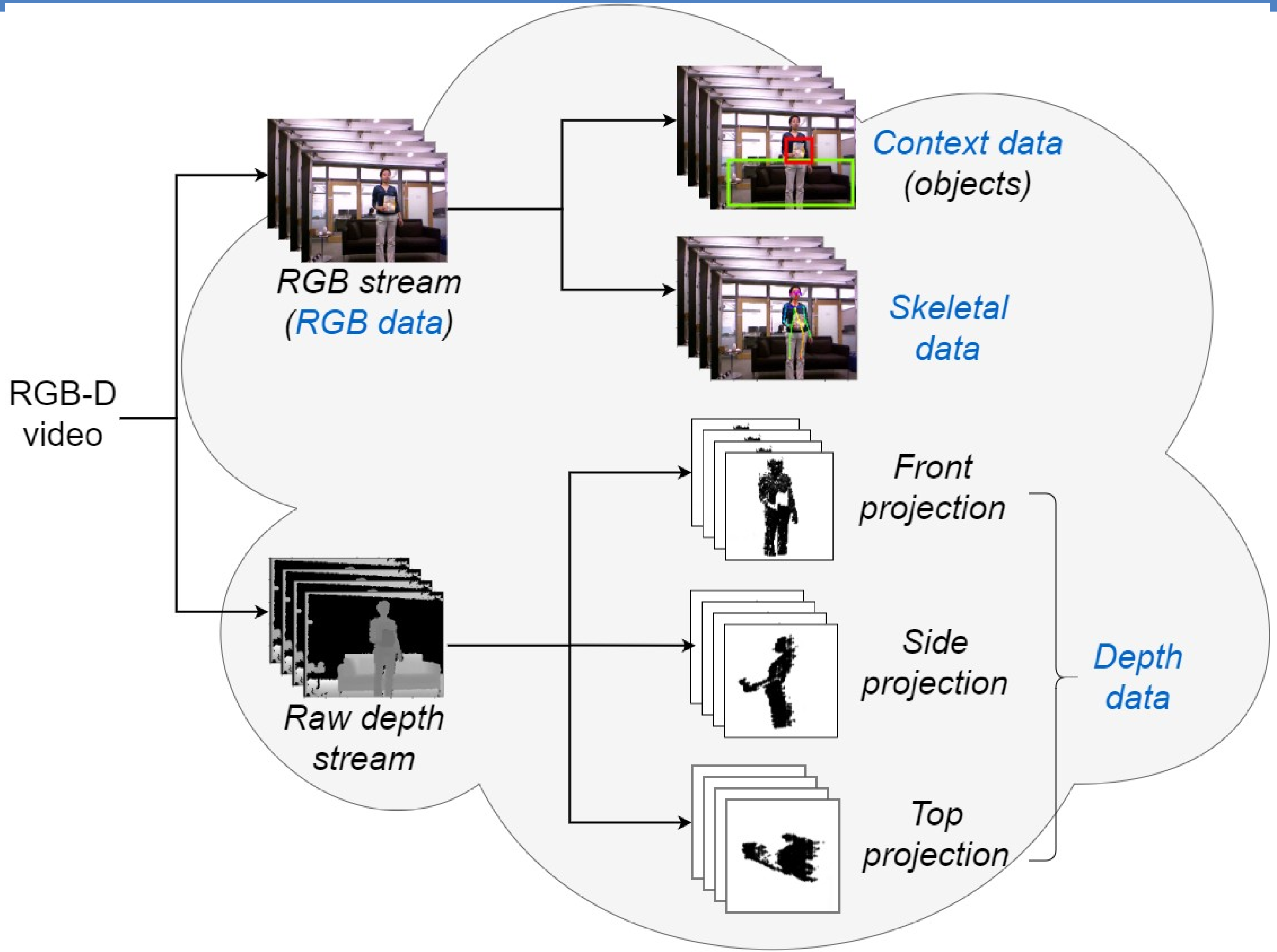




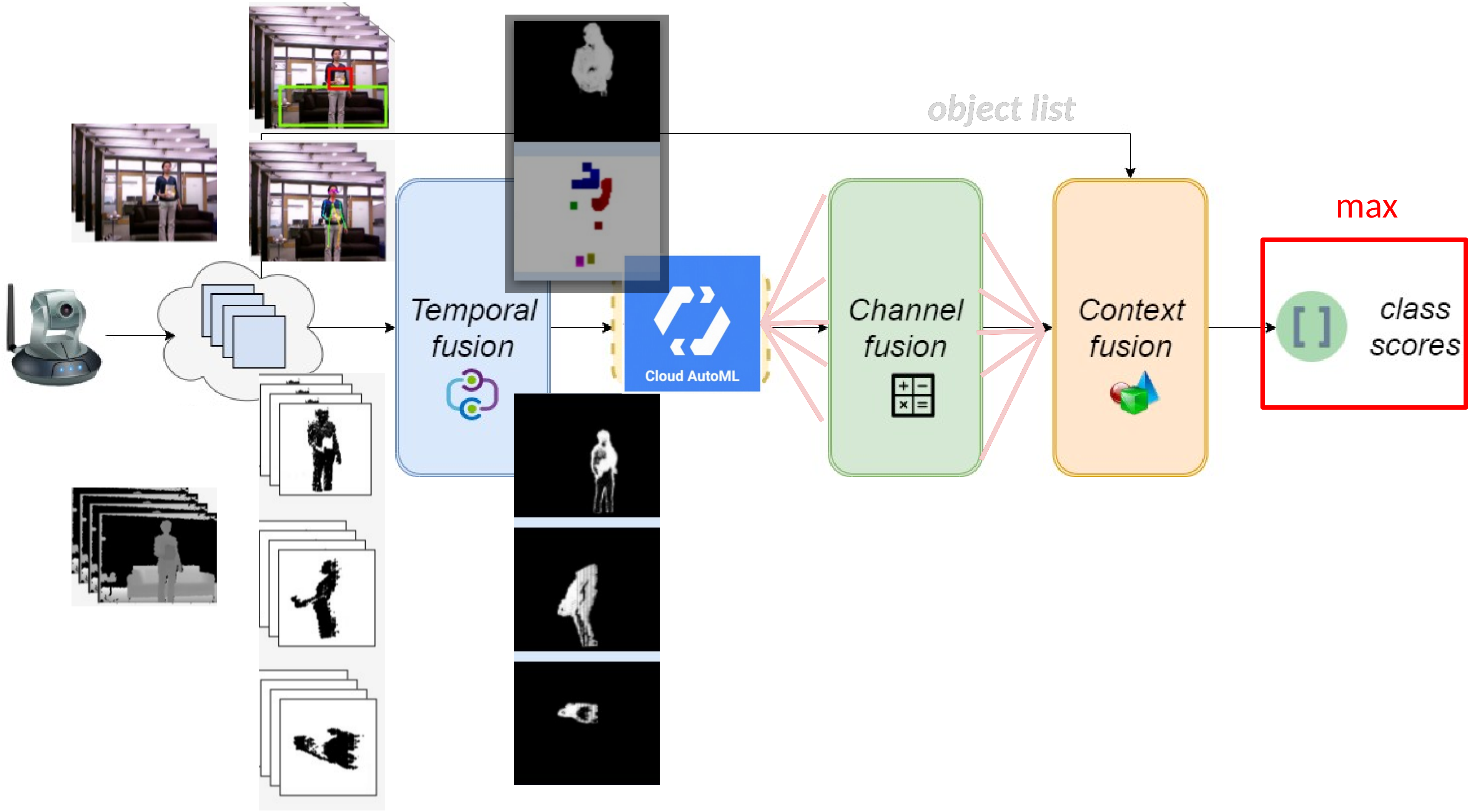
Human Activity Recognition

Data fusion:

- RGB images
- Skeleton
- Depth information
- Context



Human Activity Recognition

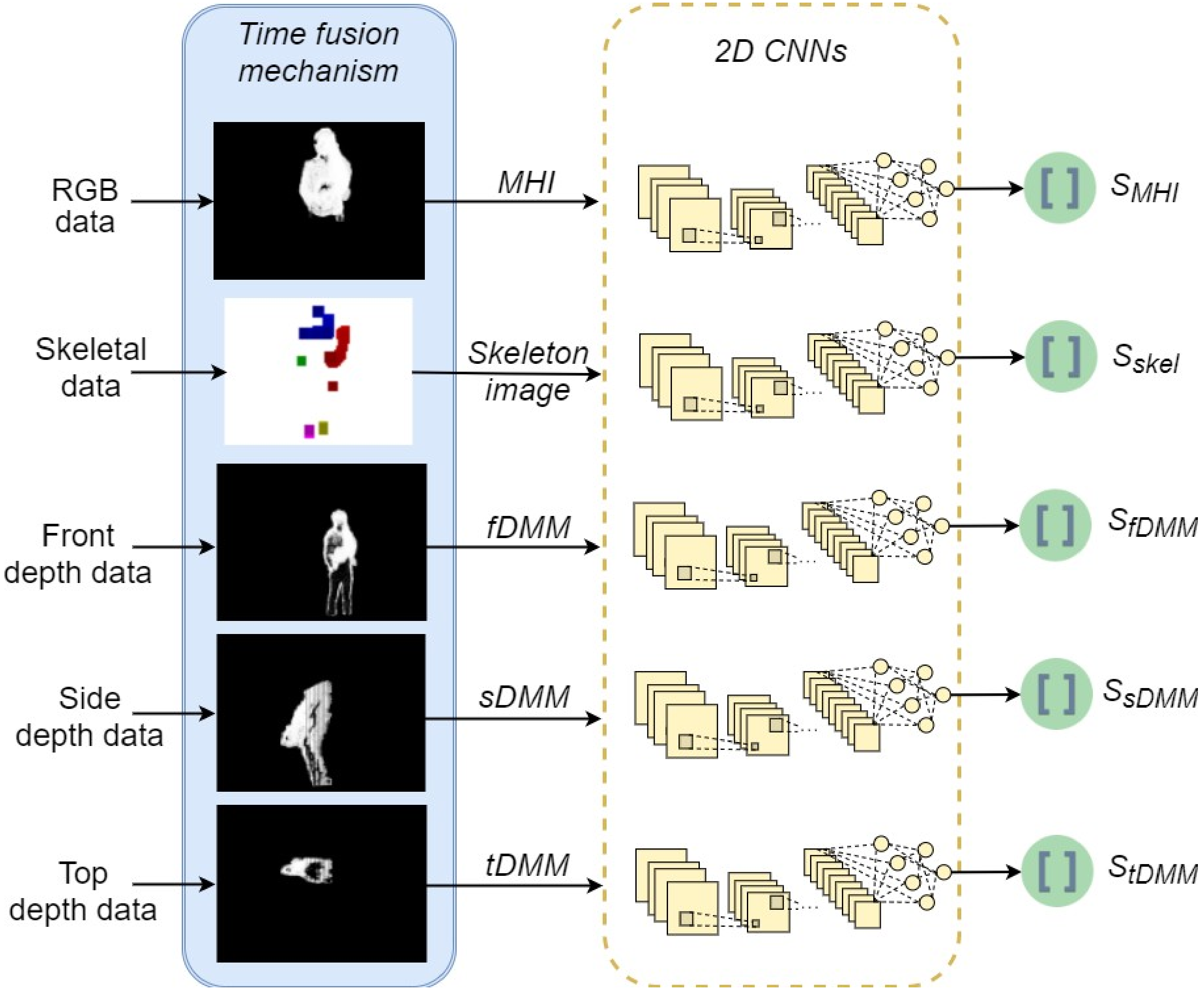


Time Fusioning

Motion history images

Skeleton images

Depth motion maps

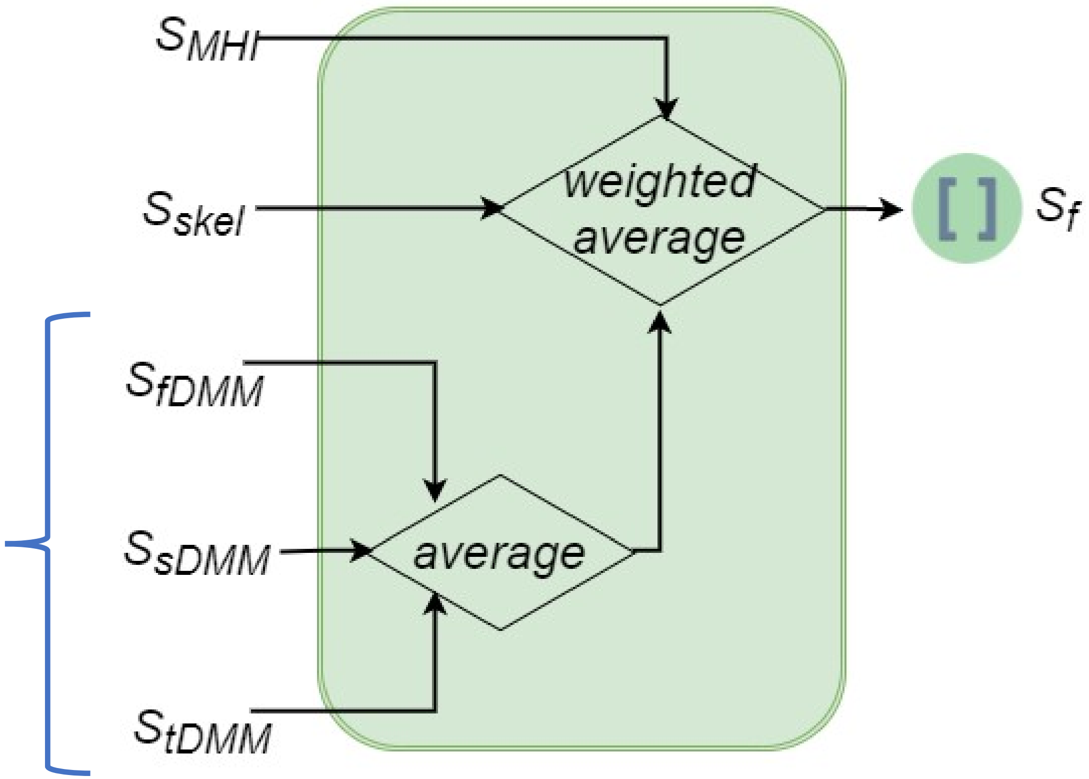


Channel Fusioning

Motion history images

Skeleton images

Depth motion maps



$$\varphi(c) = \frac{1}{3} \cdot (s_c^{MHI} w_{MHI} + s_c^{skel\_img} w_{skel\_img} + \frac{s_c^{fDMM} + s_c^{sDMM} + s_c^{tDMM}}{3} w_{DMM})$$

Automated Machine Learning

Automated machine learning:

- Finding optimal architecture
- Fine tuning of parameters

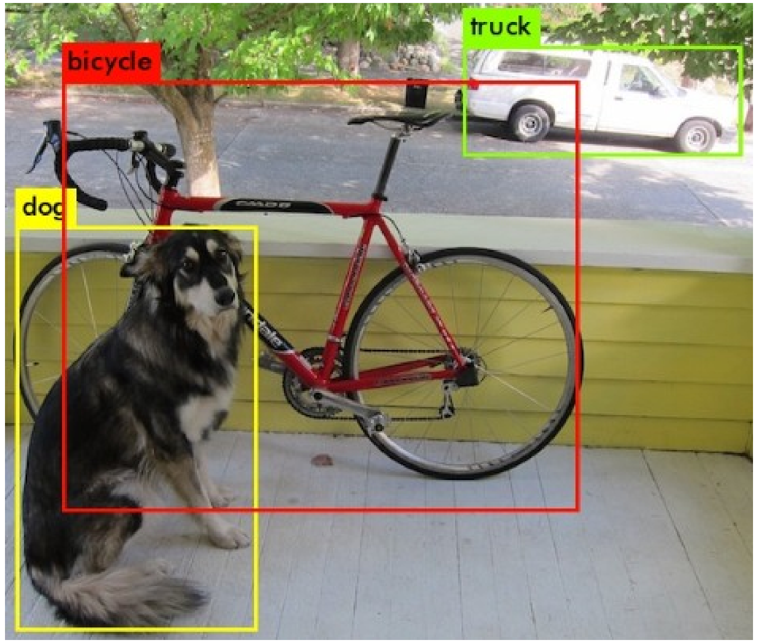
Google AutoML:

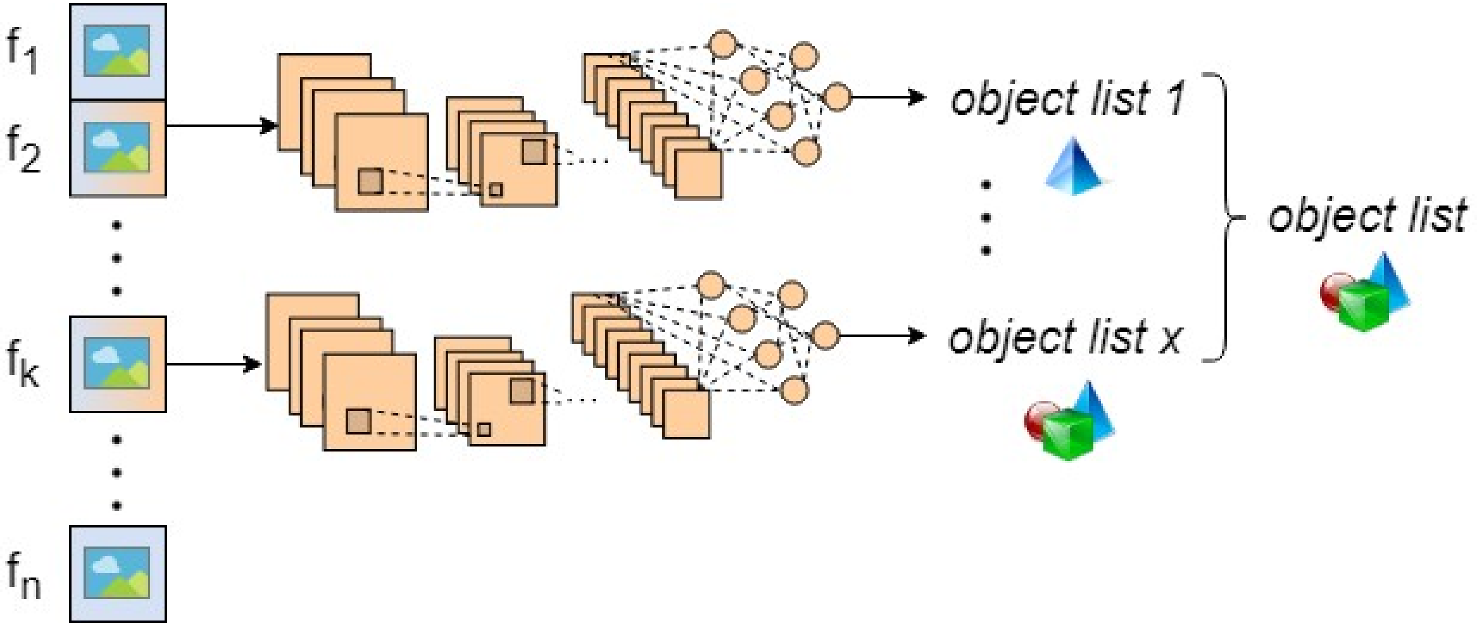
- Fast training
- Model is not provided (it can be used)

Cloud AutoML

Context Fusioning

Object recognition: YOLOv4





- 5 frames with objects are randomly selected
- Objects: person, bottle, cup, sofa, laptop, cell phone, book

Datasets

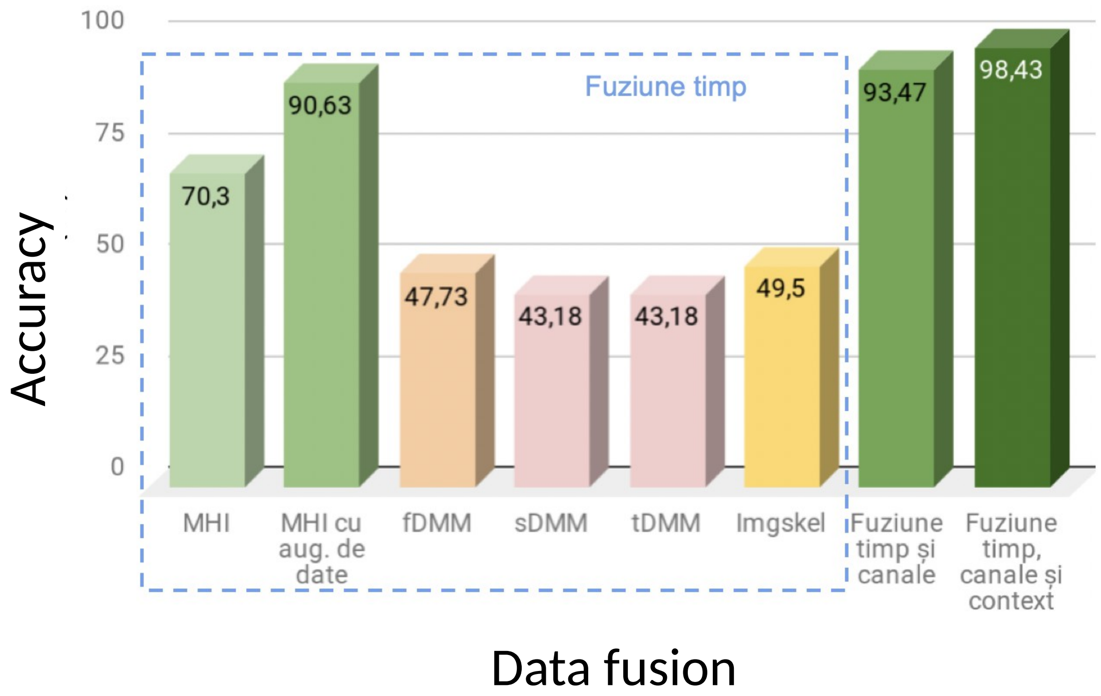
Public dataset: MSRDailyActivity3D (320 videos)

- 16 different activities: drink, eat, read book, call cell-phone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down,
- Each activity is performed twice (once standing up, once sitting down) by ten subjects, in an office.

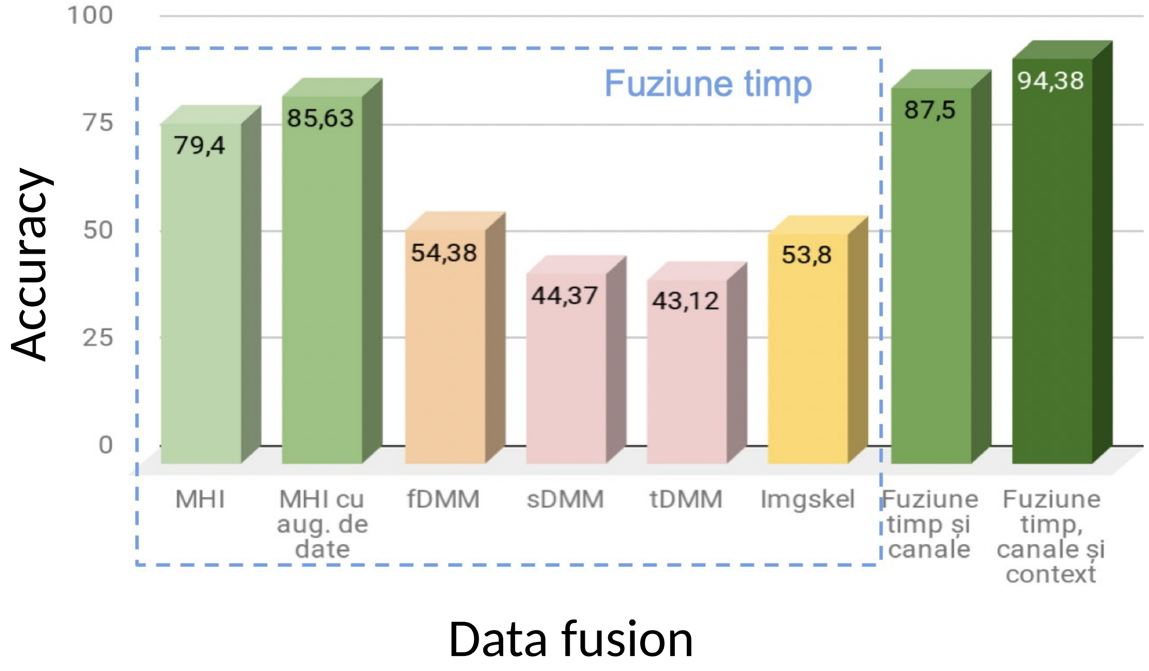
PRECIS HAR Dataset

- **PRECIS HAR**:
- : ***PRECIS HAR*** *(800 video)*
  - 50 subjects
  - 16 activities: stand up, sit down, sit still, read, write, cheer up, walk, throw paper, **drink from a bottle, drink from a mug**, move hands in front of the body, move hands close to the body, raise one hand up, raise one leg up, fall from bed and faint

Evaluation



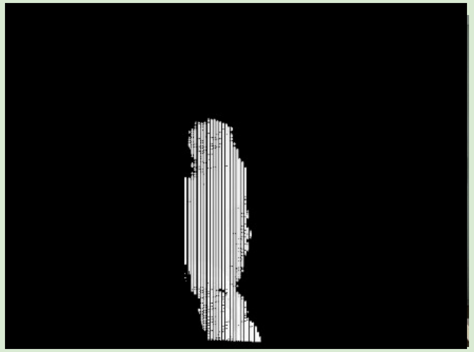
MSRDailyActivity3D



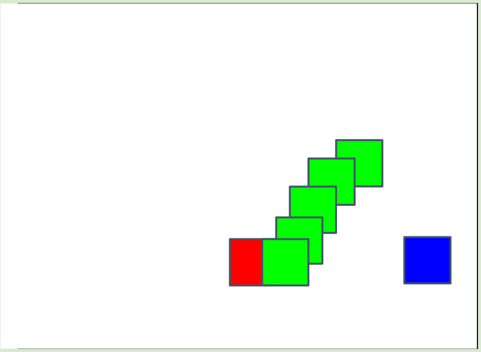
PRECIS HAR

Correct classifications

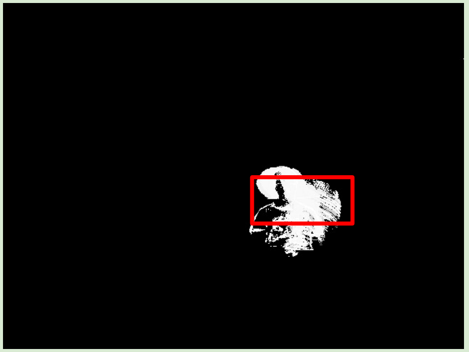
fall from bed



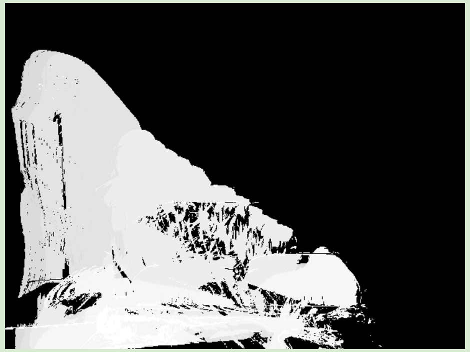
hands close to the body



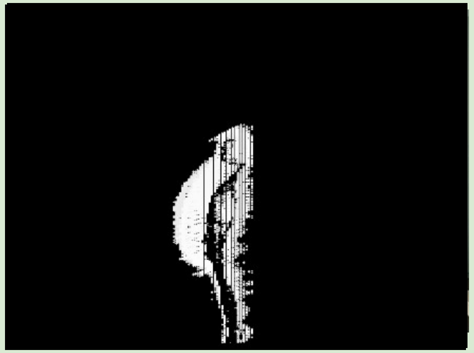
raise one hand up



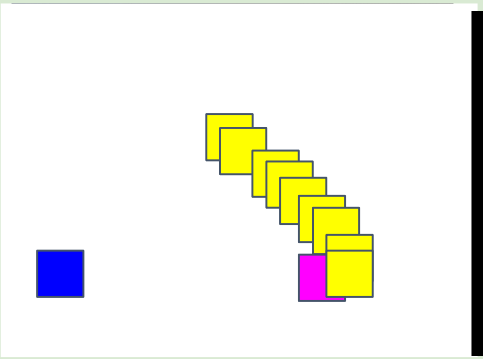
drink from a bottle



faint



hands in front of the body



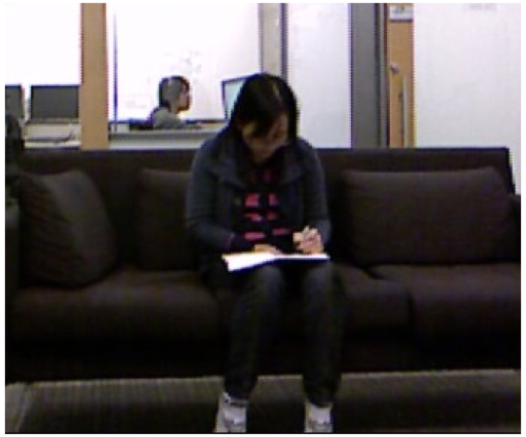
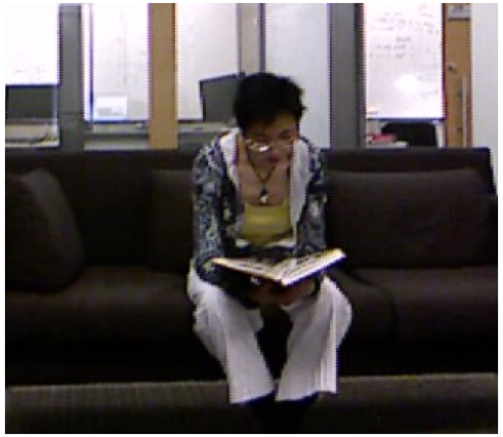
raise one leg up

drink from a mug

Difficult classifications

MSRDailyActivity3D:

- writing vs. reading vs. laptop

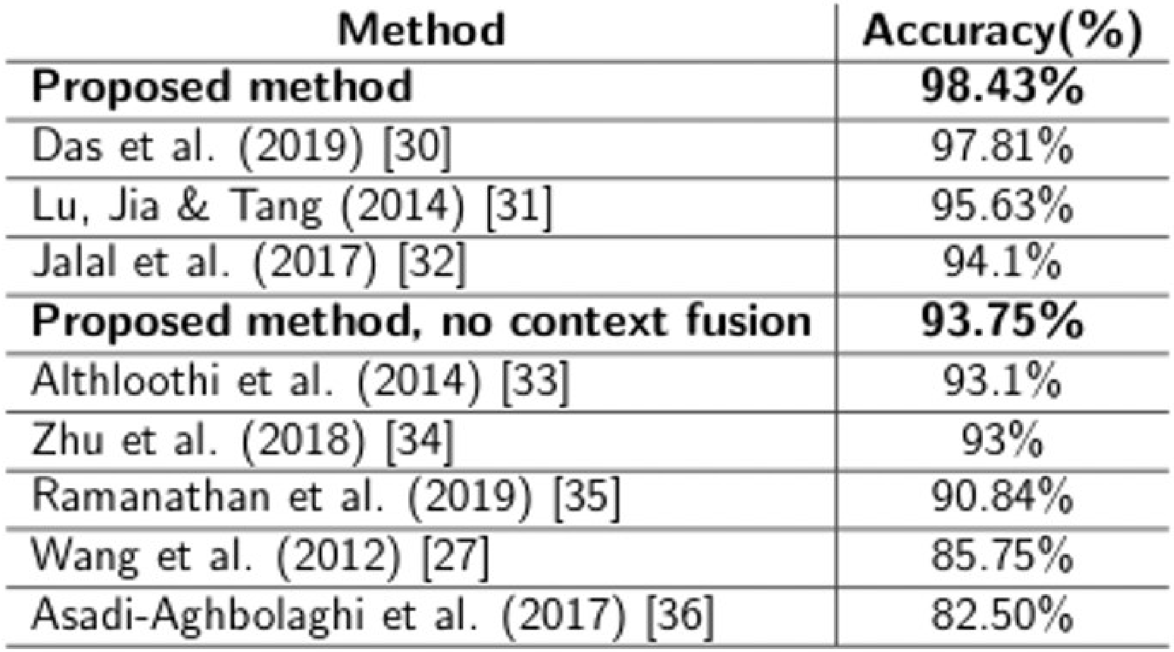



PRECIS HAR :

- writing vs. reading

Evaluation

MSRDailyActivity3D

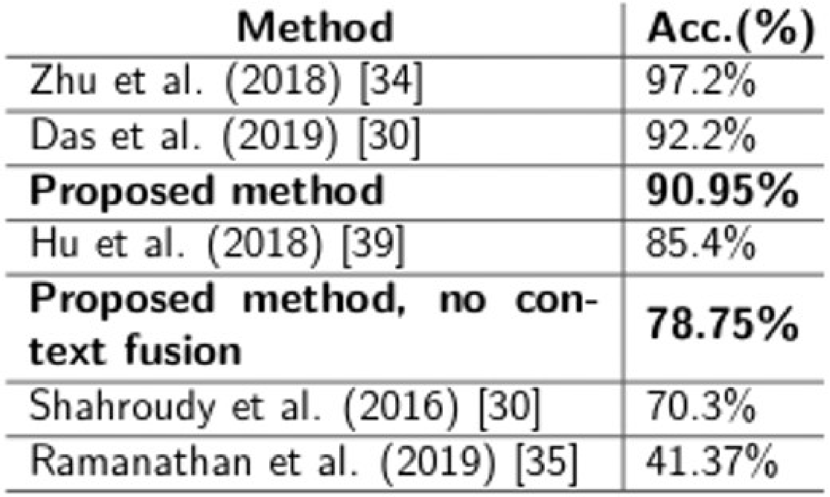


Method	Accuracy(%)
Proposed method	**98.43%**
Das et al. (2019) [30]	97.81%
Lu, Jia & Tang (2014) [31]	95.63%
Jalal et al. (2017) [32]	94.1%
Proposed method, no context fusion	**93.75%**
Althloothi et al. (2014) [33]	93.1%
Zhu et al. (2018) [34]	93%
Ramanathan et al. (2019) [35]	90.84%
Wang et al. (2012) [27]	85.75%
Asadi-Aghbolaghi et al. (2017) [36]	82.50%

- **Time**
  - training: 10 - 15 minute
  - classification: 0.1 - 1.8 sec

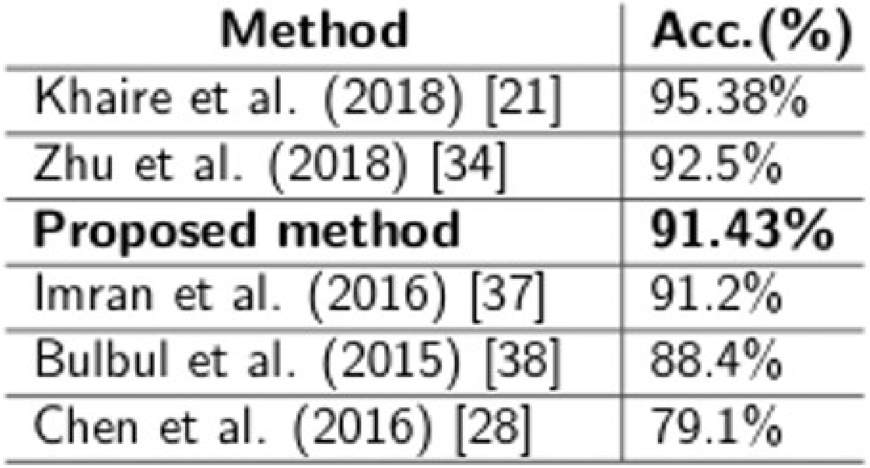
Evaluation

NTU RGB+D: 60 classes



Method	Acc.(%)
Zhu et al. (2018) [34]	97.2%
Das et al. (2019) [30]	92.2%
Proposed method	**90.95%**
Hu et al. (2018) [39]	85.4%
Proposed method, no context fusion	**78.75%**
Shahroudy et al. (2016) [30]	70.3%
Ramanathan et al. (2019) [35]	41.37%

UTD-MHAD: 26 classes



Method	Acc.(%)
Khaire et al. (2018) [21]	95.38%
Zhu et al. (2018) [34]	92.5%
Proposed method	**91.43%**
Imran et al. (2016) [37]	91.2%
Bulbul et al. (2015) [38]	88.4%
Chen et al. (2016) [28]	79.1%

Context fusion is relevant for datasets that contains users that interacts with objects

Difficult classification

UTD-MHAD:

- catch - knock

NTU RGB+D:

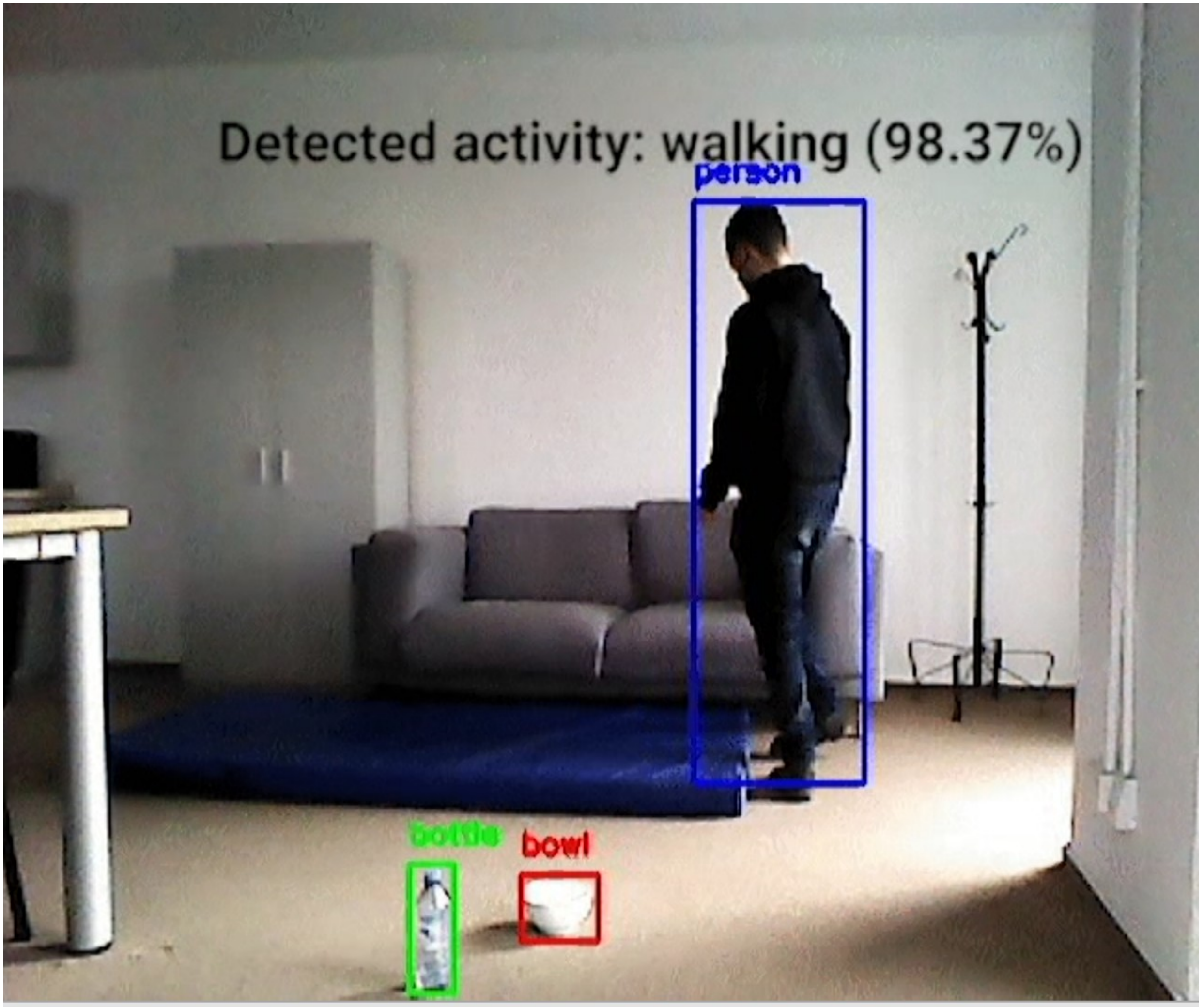
- salute - brush teeth

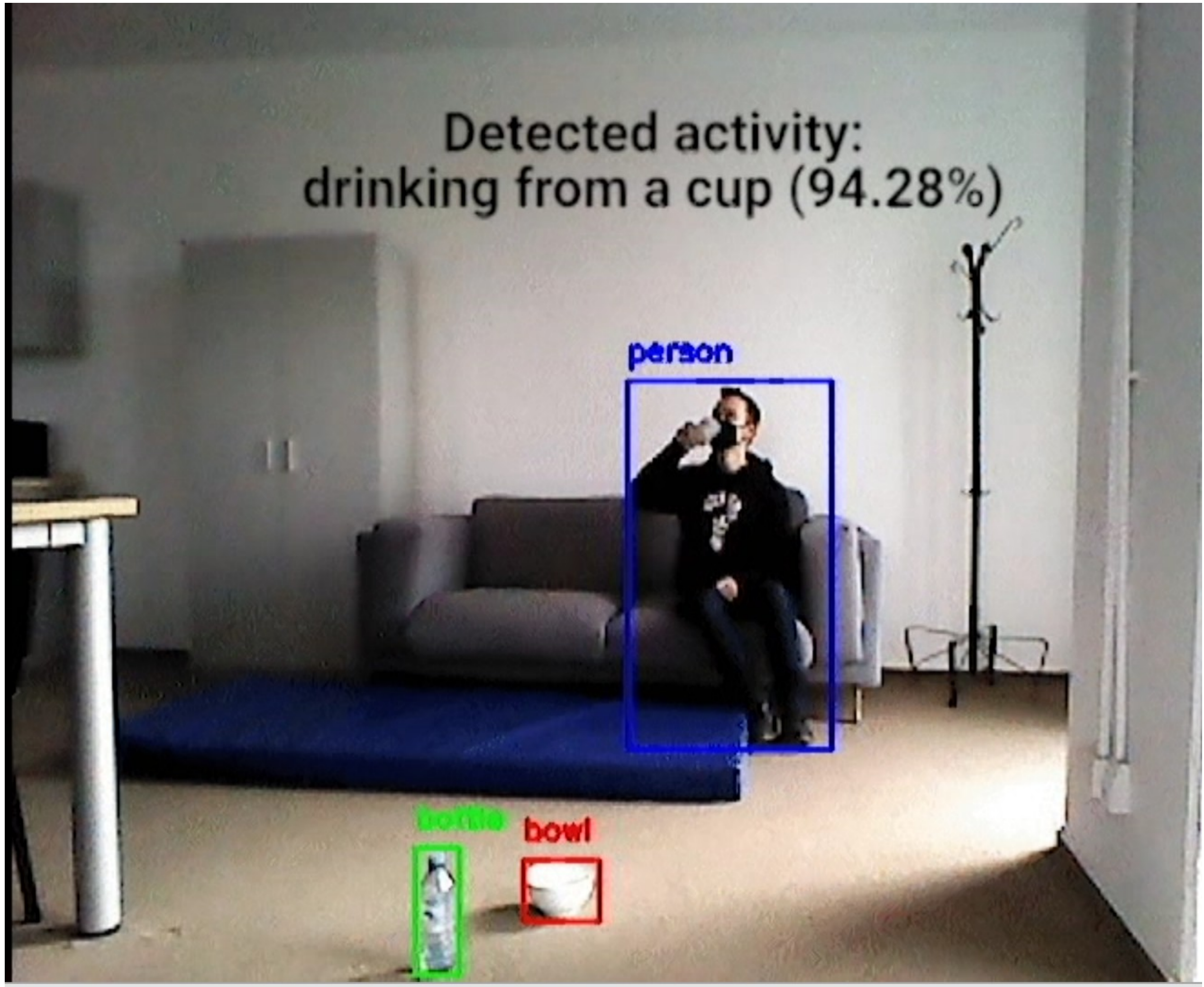
Testing with robotic platforms

TurtleBot

TIAGo

Activity Recognition





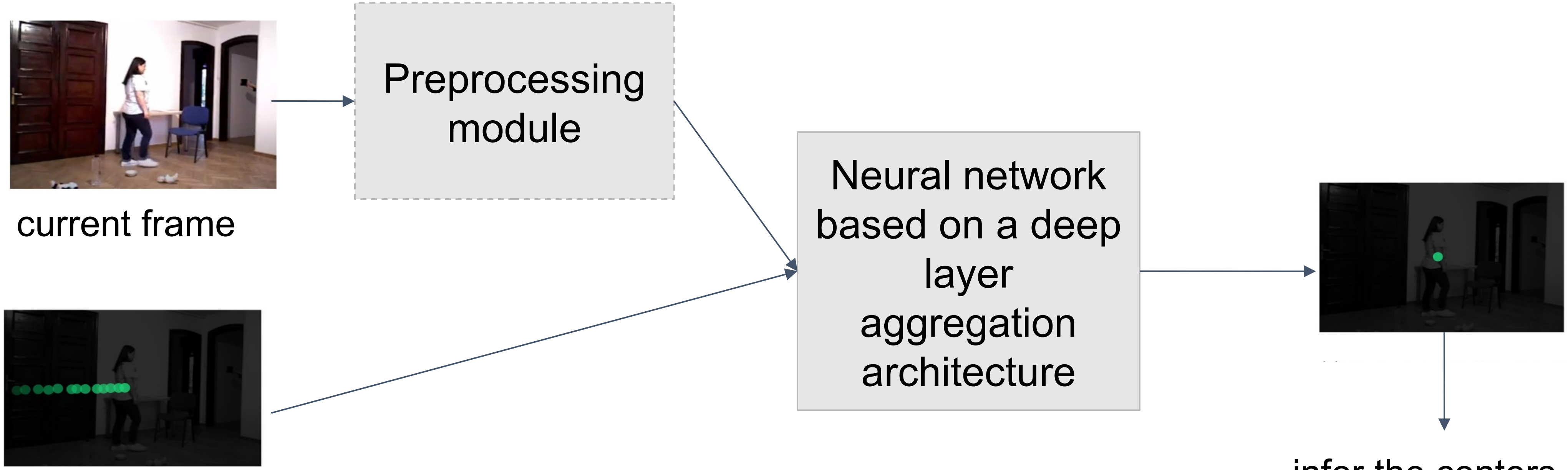
Object Detection

- Classical object detectors are the single-frame ones like YOLO
- Single-frame object detection has significant drawbacks:
  - it fails when objects are partially occluded/filmed from particular angles,
  - can lack either speed or accuracy when ported on real-life applications—that can lead to functional anomalies.
- Multi-frame object detection solves the previously stated drawbacks

Object Detection

- Center-based descriptor
  - Describe an object by using only the central point of its bounding box
  - Infer via regression other properties (size, location, orientation) from the keypoint feature in the center
- Train a fully connected network that generates heatmaps;
  - The peaks in the heatmap will represent the centers of the predicted objects.

Object Detection



current frame

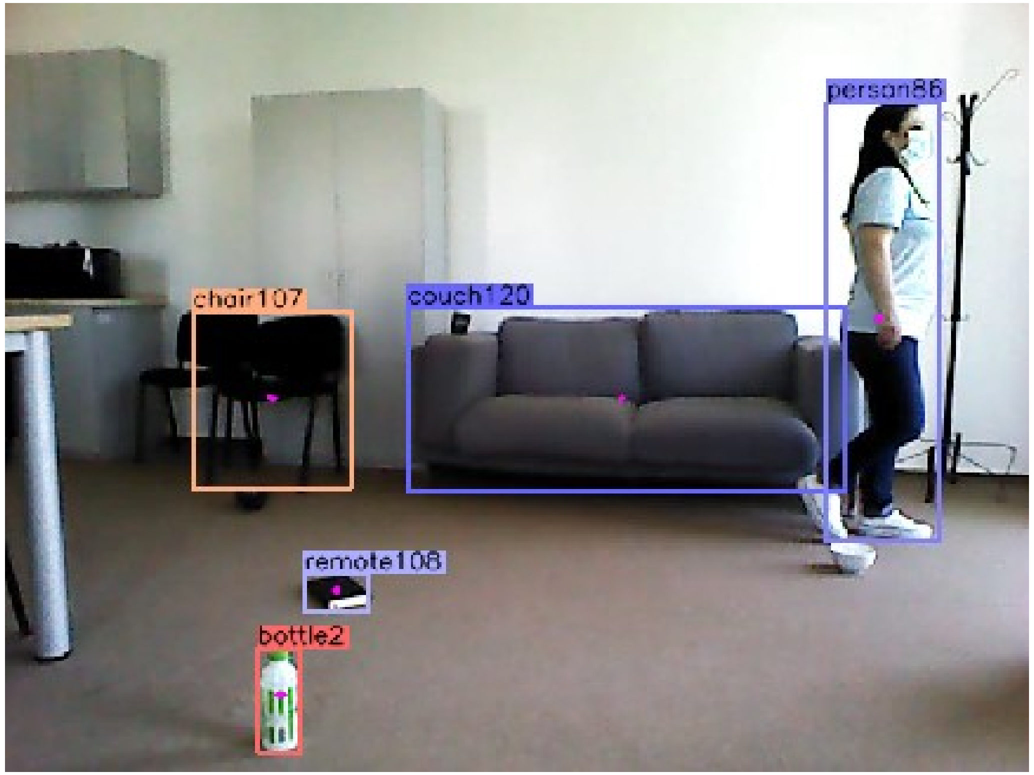
Preprocessing module

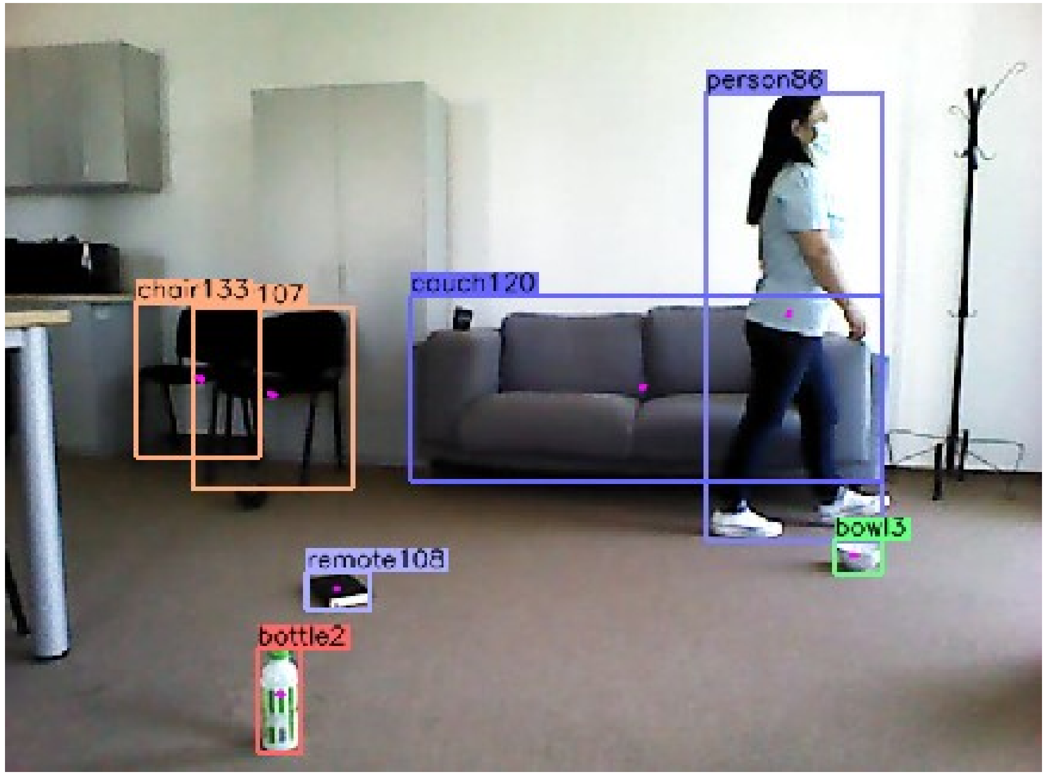
Neural network based on a deep layer aggregation architecture

motion history
map of the center-based descriptors
from the previous n
frames

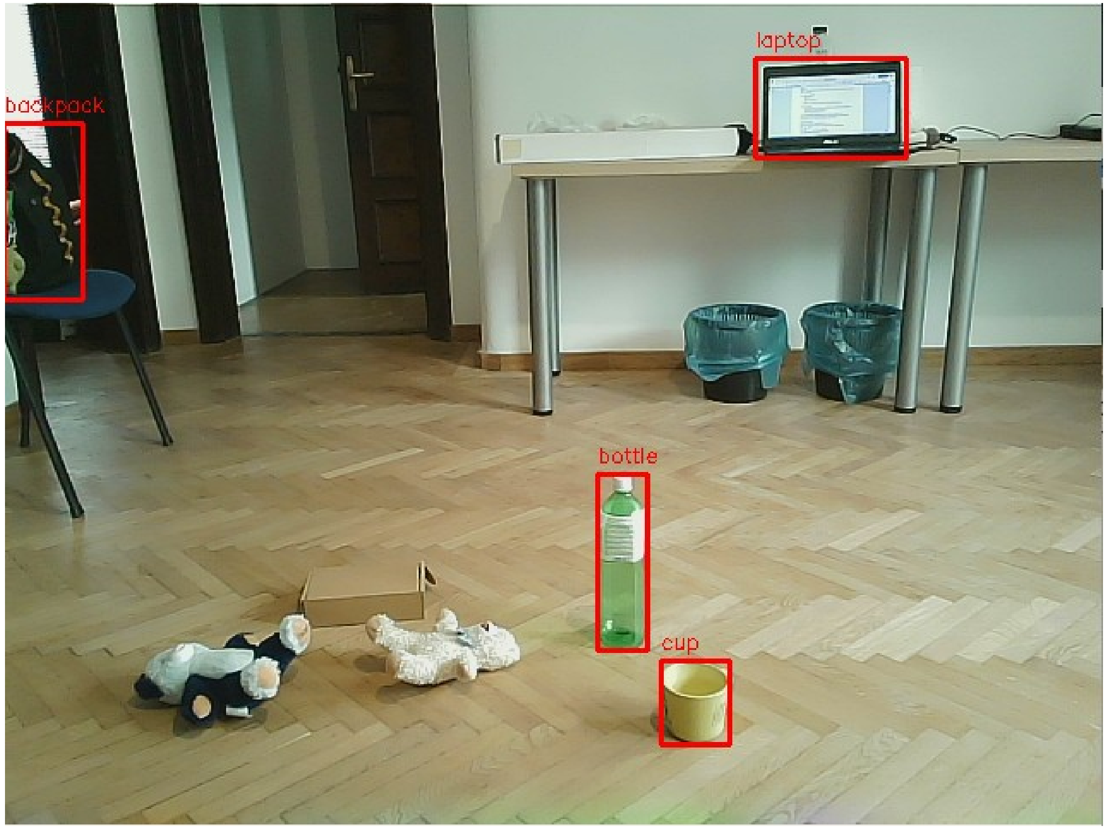
infer the centers of the objects, the sizes of the bounding boxes, as well as their orientation.

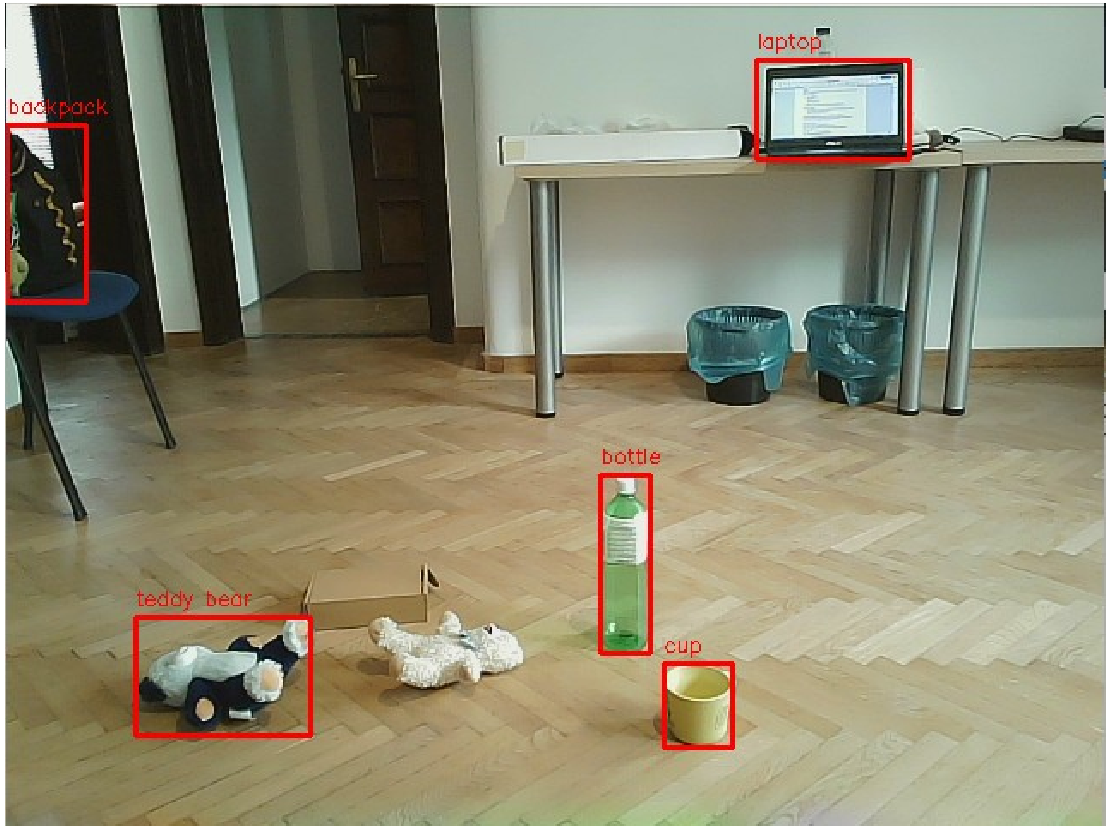
Object Detection





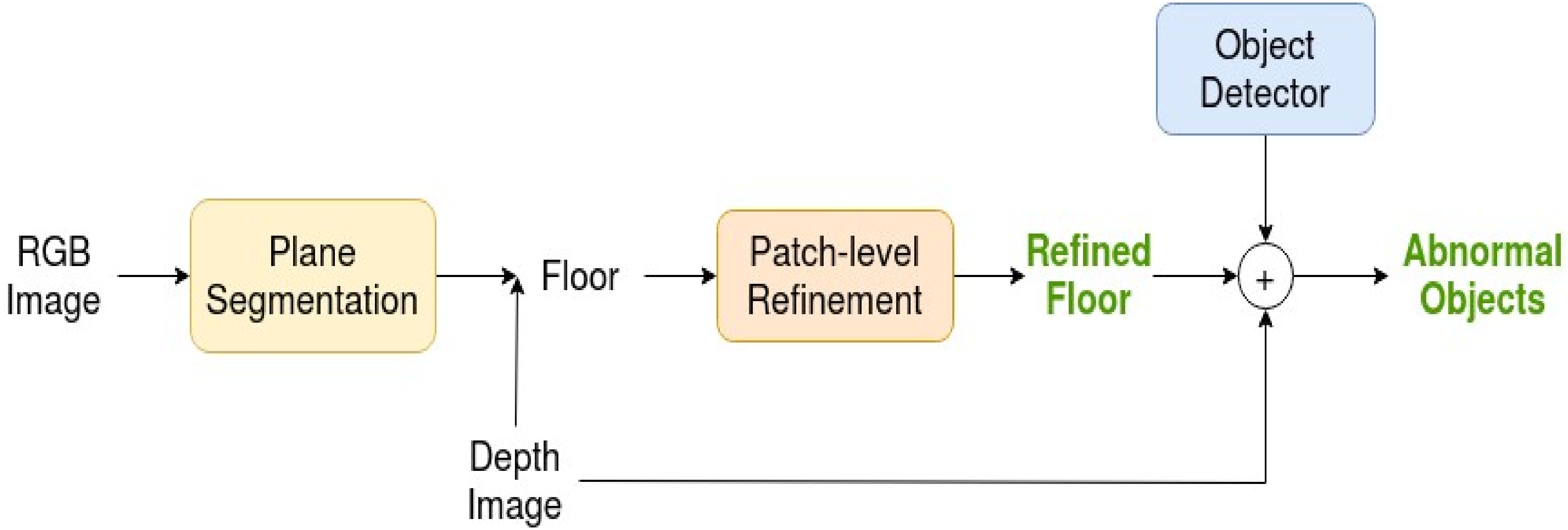
Object Detection





Abnormal Object Detection

Two-phase complex system for plane segmentation

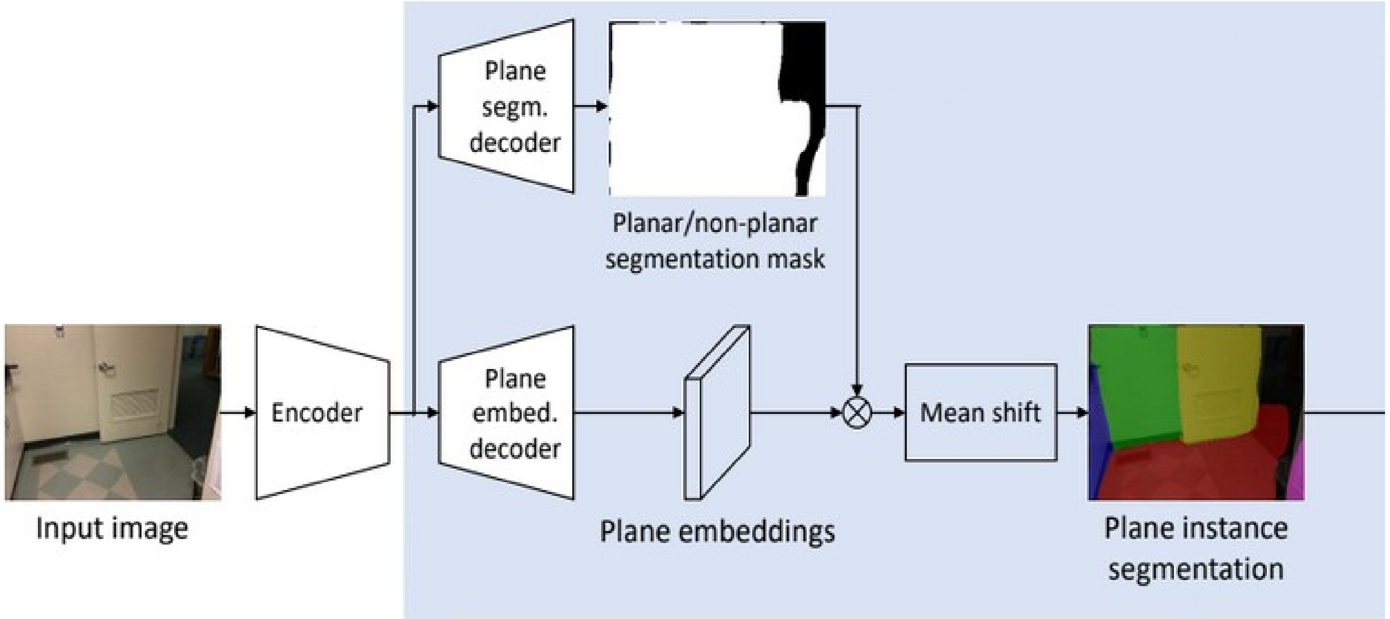


Abnormal Object Detection

Plane segmentation:

- Encoder-decoder architecture that is capable of distinguishing between regions that are planar and that are not.
- Encoder: an extended version of ResNet-101 Feature Pyramid Network, pre-trained on ImageNet
- Classification in plane regions is made using the mean shift clustering

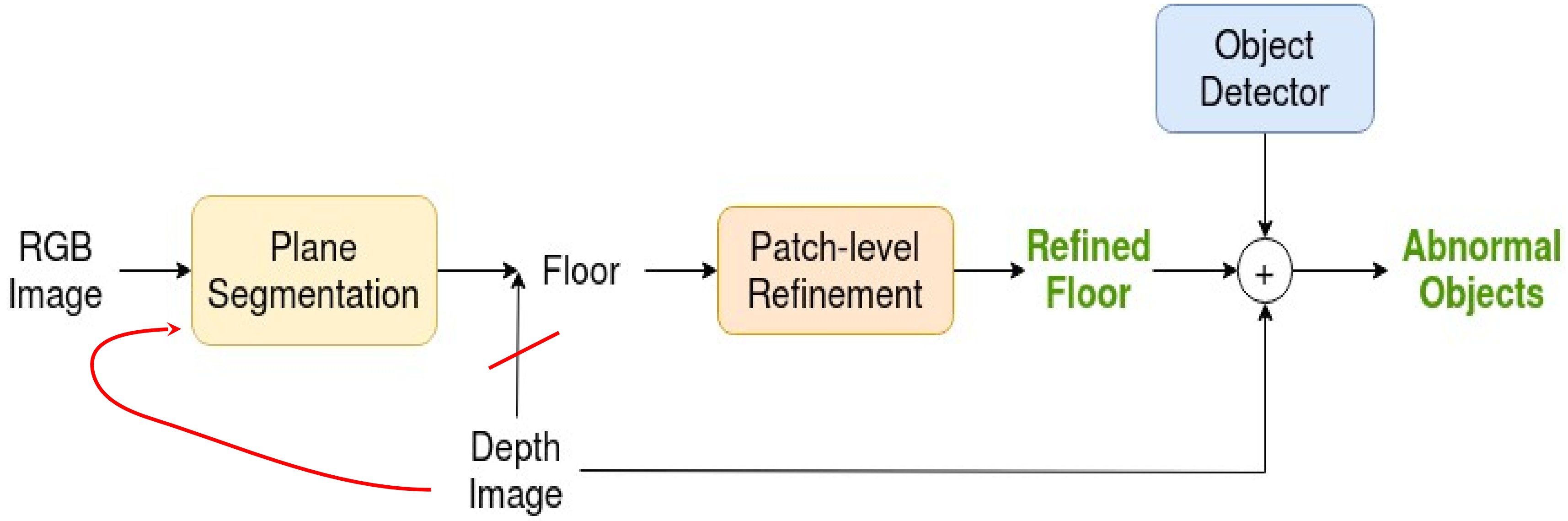
- Using depth data - the ground should be a plan that is lower than the other plans, on which the robot sits and that is large enough in the picture.



Abnormal Object Detection

Improved the plane segmentation by including depth data as direct input in the encoder-decoder of the architecture.

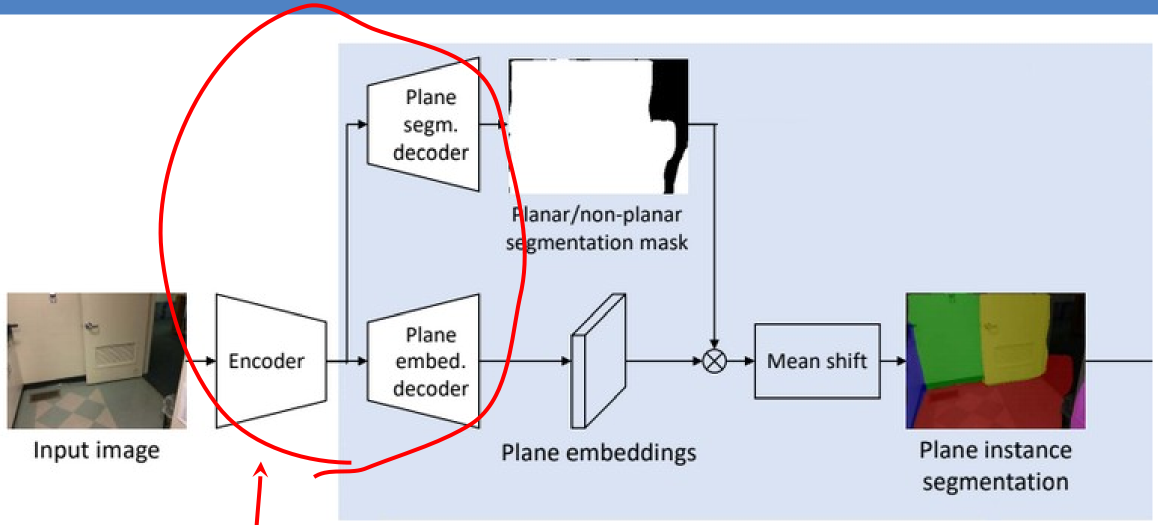
The depth information is used in an earlier stage, by encoding it in the encoder component of the plane segmentation component.

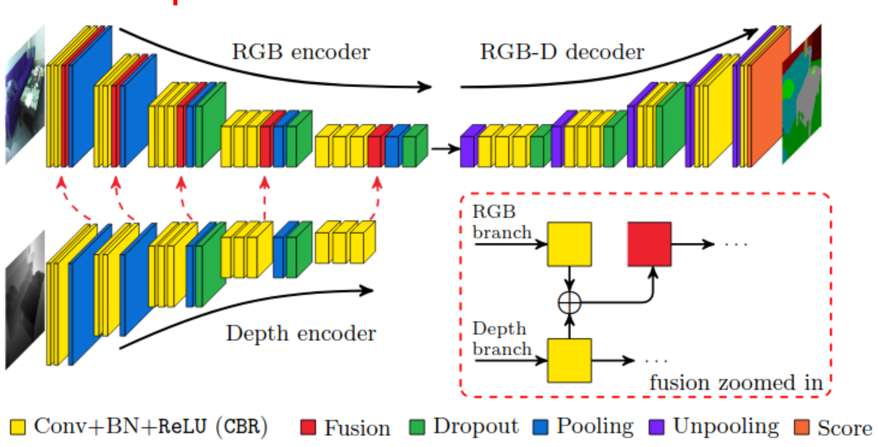


Process of transforming the simple architecture of the plane detection mechanism into a more robust one

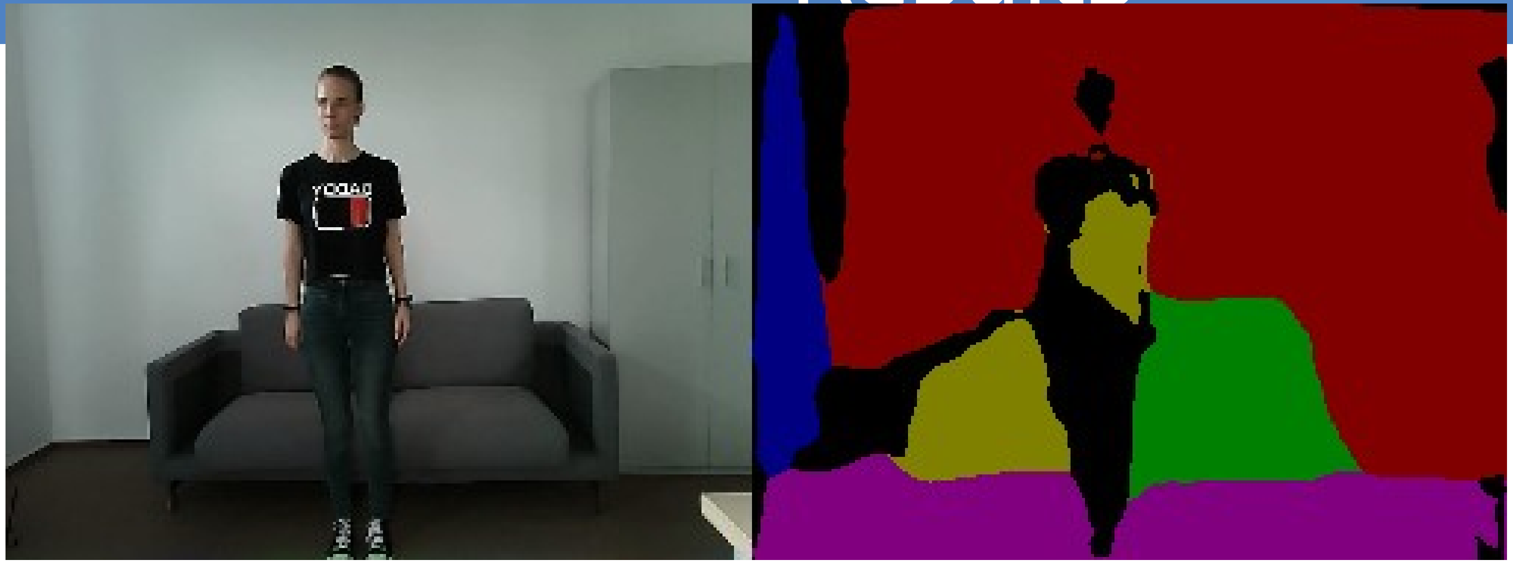
Abnormal Object Detection

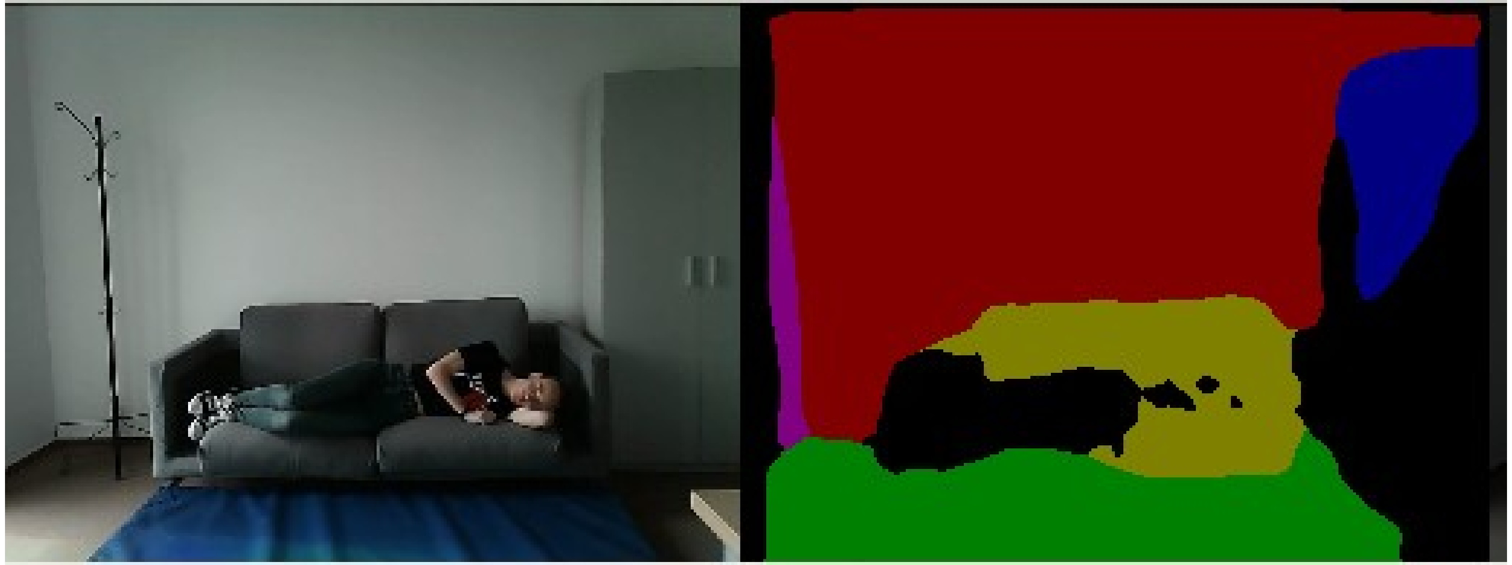
- FuseNet - the depth encoded information is progressively merged in the RGB encoded one.
- The network has two branches, each of them being specialized in extracting features from a type of information stream—RGB or depth data.
- The feature maps extracted from the depth stream are progressively merged into the ones extracted from the RGB data.





Abnormal Object Detection- Qualitative Results

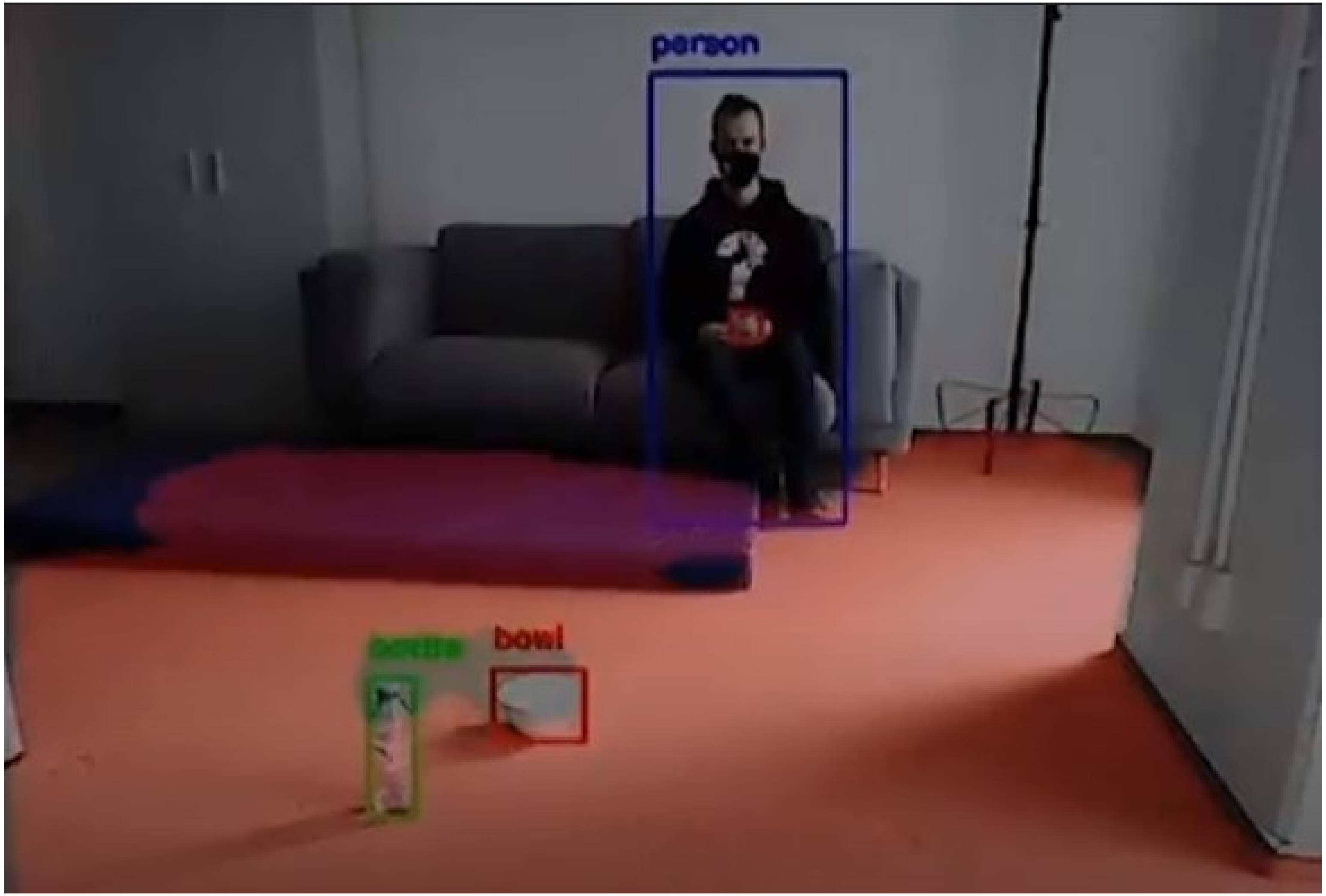




Not every plane is fully identified, but the floor plan is correctly marked

Abnormal Object Detection- Qualitative Results

Most of the qualitative experiments were successful, proving the fact that the proposed ground segmentation and object detection mechanism are suitable for being used in abnormal object detection



Gesture Recognition

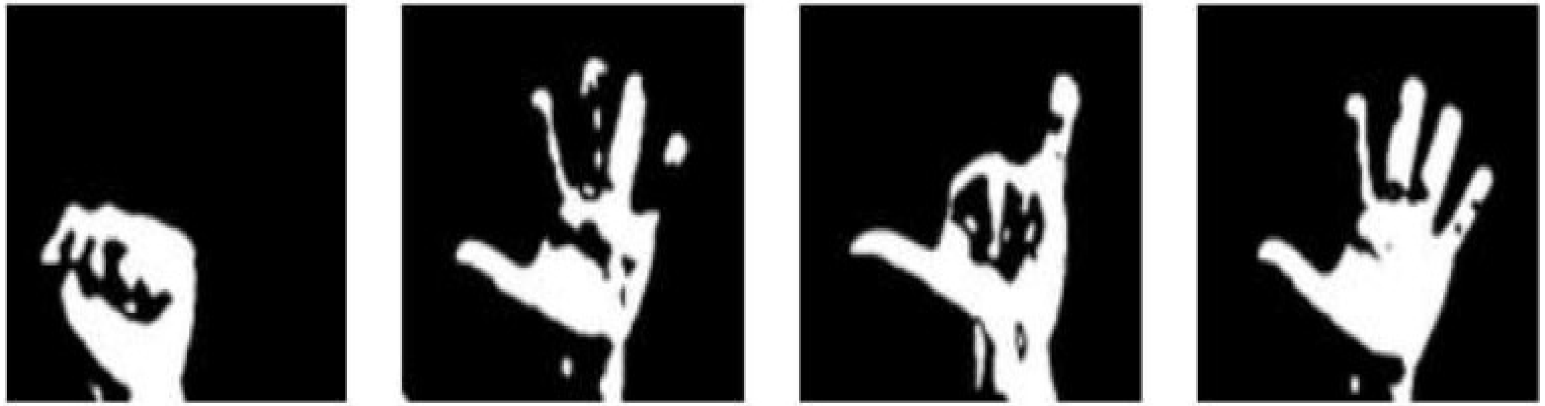
- Aimed to be an accessible interface between the robot and the human
- Hand area detection based on skeleton and depth information
- Simple CNN architecture, operating on binary images

Simple CNN architecture contains:

- 7 hidden layers (2D convolutions followed by 2D max pooling) with ReLU as activation function, and dropout and one fully connected layer with softmax activation in the end.
- The optimizer is Adam and the loss functions is cross-entropy.

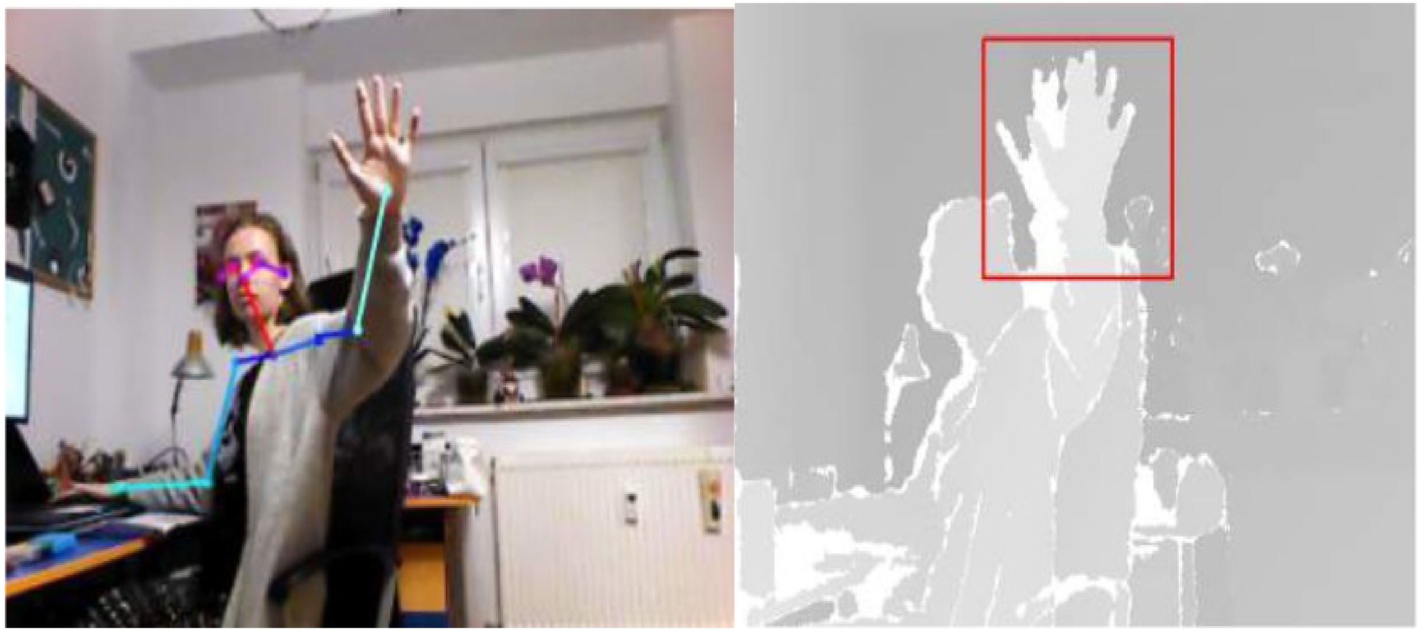
Gesture Recognition

The CNN gets as input binary images of reduced size (e.g. 100×100 pixels) representing shapes of the hand.



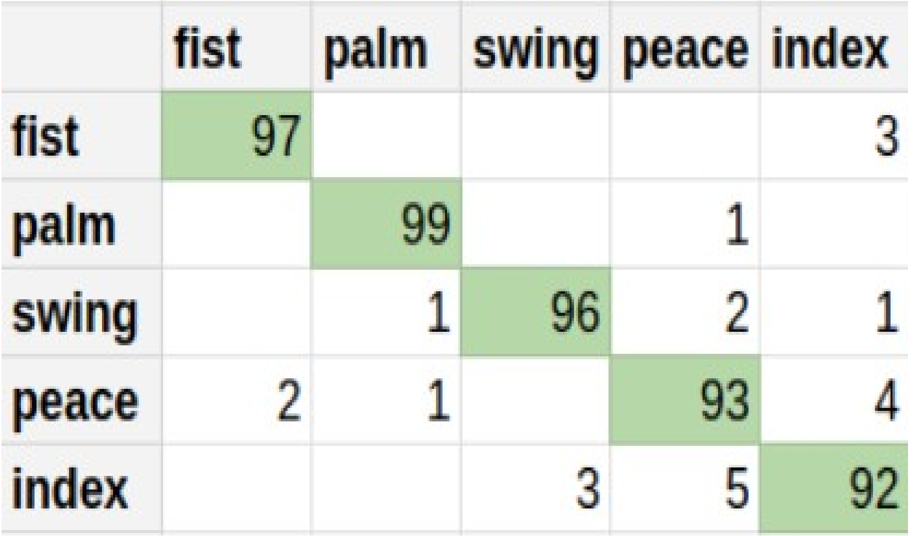
Gesture Recognition

- The detection of the hand area currently is based on the depth information and skeleton.
- The person points towards the robot when performing an interaction gesture:
  - they are facing the robot
  - the closest points in the point cloud to the robot correspond to the hand.



Gesture Recognition

- Confusion matrix

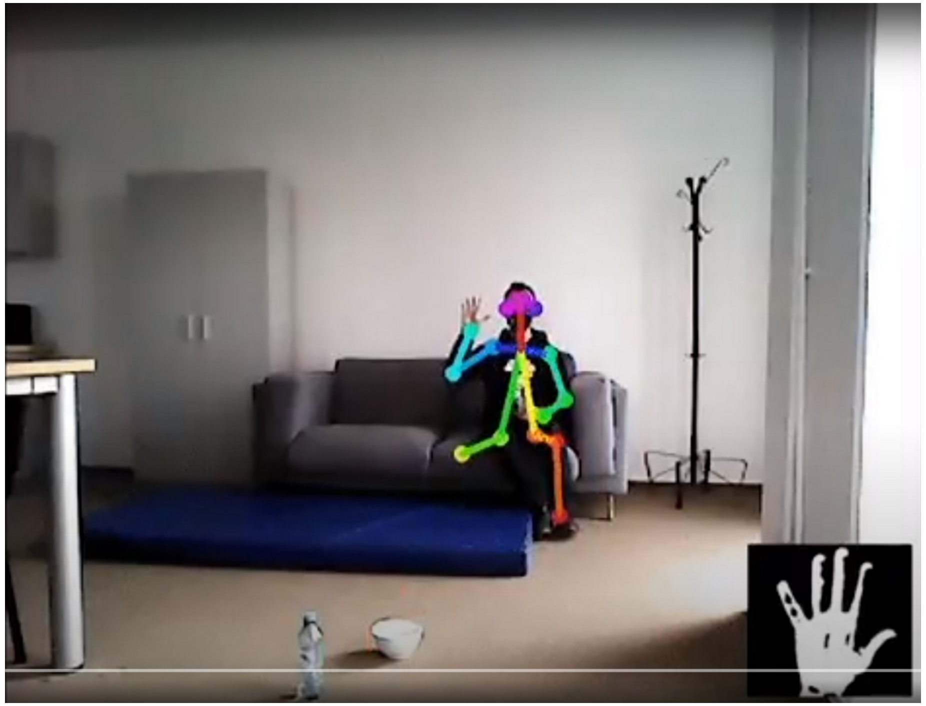


	fist	palm	swing	peace	index
fist	97				3
palm		99		1	
swing		1	96	2	1
peace	2	1		93	4
index			3	5	92

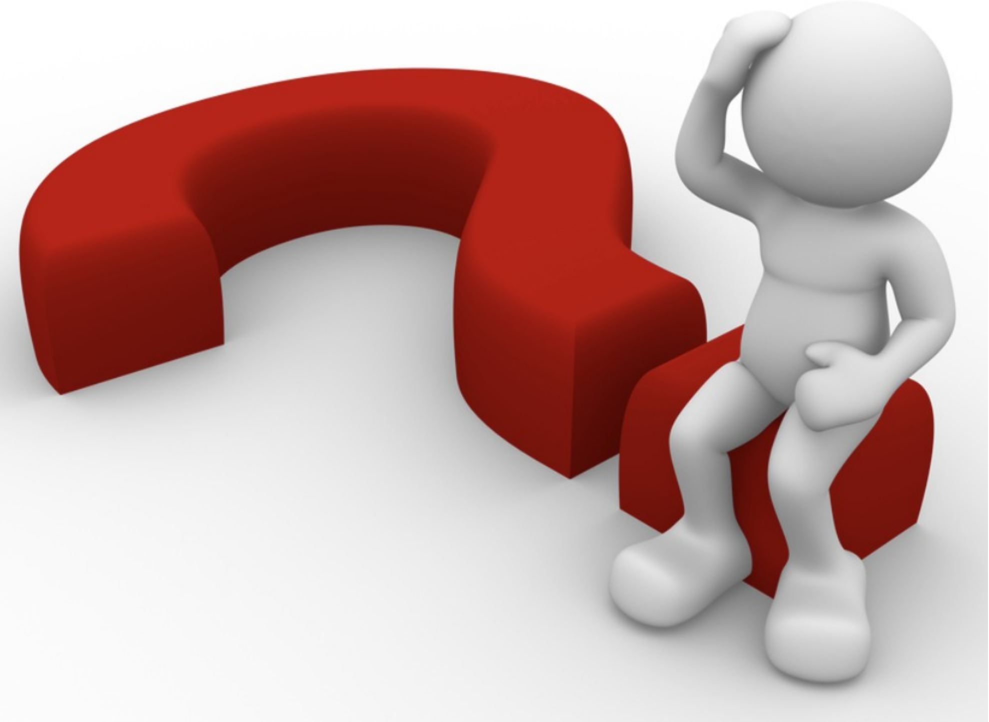
95.4% accuracy on own dataset with 1000 samples per class.

Gesture Recognition

Gesture recognition: palm identification with the help of the skeleton and gesture identification as “palm



Thank you!



irina.mocanu@upb.ro