Human-centered AI for Intelligent Vehicles

C. Papaioannids, D. R. Papadam, I. Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 2.3

Artificial Intelligence & Information Analysis Lab
Funded by the European Union

Contents

- **Human-centered AI overview**
- Human detection
- Human segmentation
- Human pose estimation
- Human action recognition
- Human gesture recognition
- Applications

Human-centered AI overview

- **Autonomous vehicles** (e.g., self-driving cars, UAVs) are increasingly being employed in real-world applications.
  - Autonomous transportation.
  - Infrastructure inspection.
  - Disaster management.
- **Human-Vehicle interaction**: Autonomous vehicles should understand humans and interact with them effectively.
  - Special case of Human-Robot Interaction (HRI).

Human-centered AI overview

- Autonomous vehicles need to be equipped with **visual and auditory perception** systems and **AI algorithms**.
- These systems and AI algorithms have to demonstrate:
  - High **perception accuracy**.
  - **Robustness** to input data variations.
  - Produce **quick state estimations** to ensure **safety** and **timely actions**.

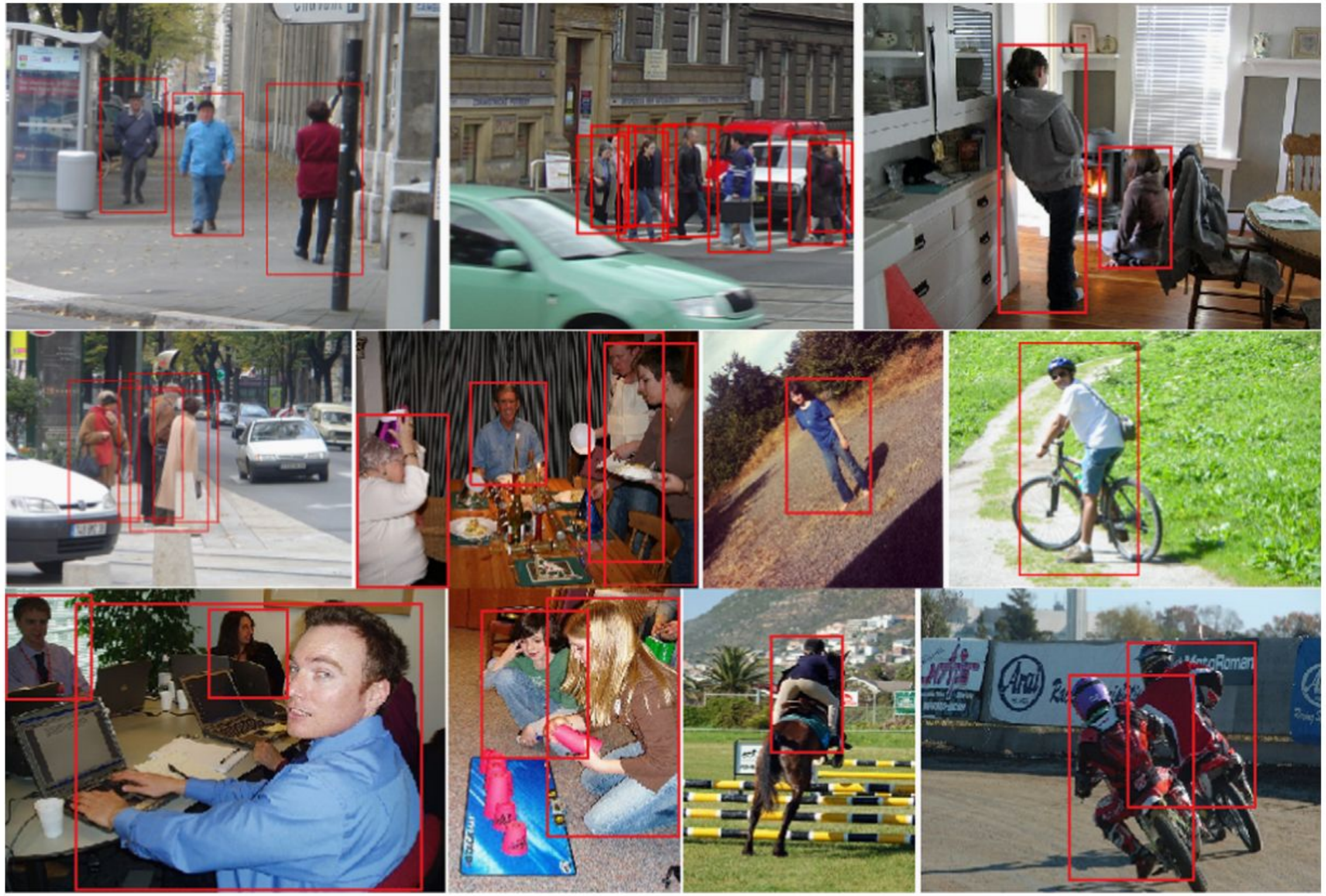
Human-centered AI overview

- **Deep Neural Networks** (DNNs) are actively being used to build such advanced systems.
  - **Convolutional Neural Networks** (CNNs).
  - **Transformer networks**.
- Main tasks:
  - Human detection.
  - Human segmentation.
  - Human pose/posture estimation.
  - Human action/activity recognition.
  - Human gesture recognition.

Contents

- Human-centered AI overview
- **Human detection**
- Human segmentation
- Human pose estimation
- Human action recognition
- Human gesture recognition
- Applications

Human Detection



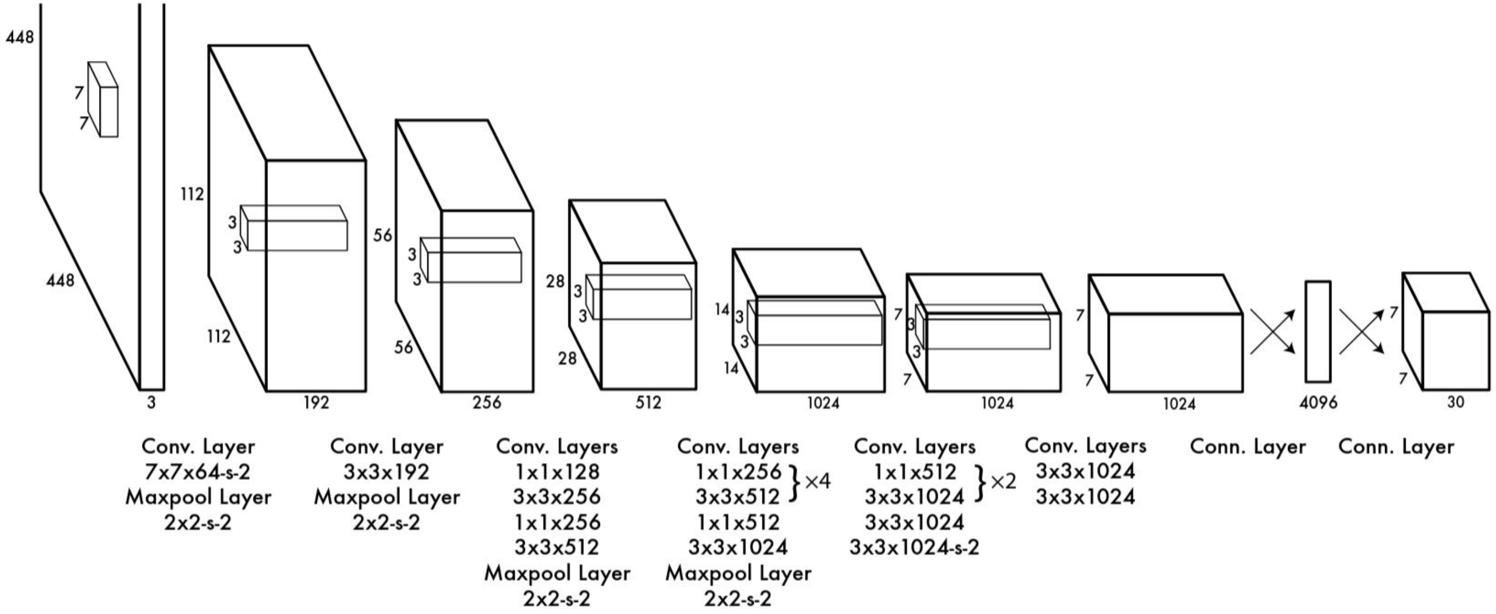
Examples of human detection results [NGU2016].

Human Detection

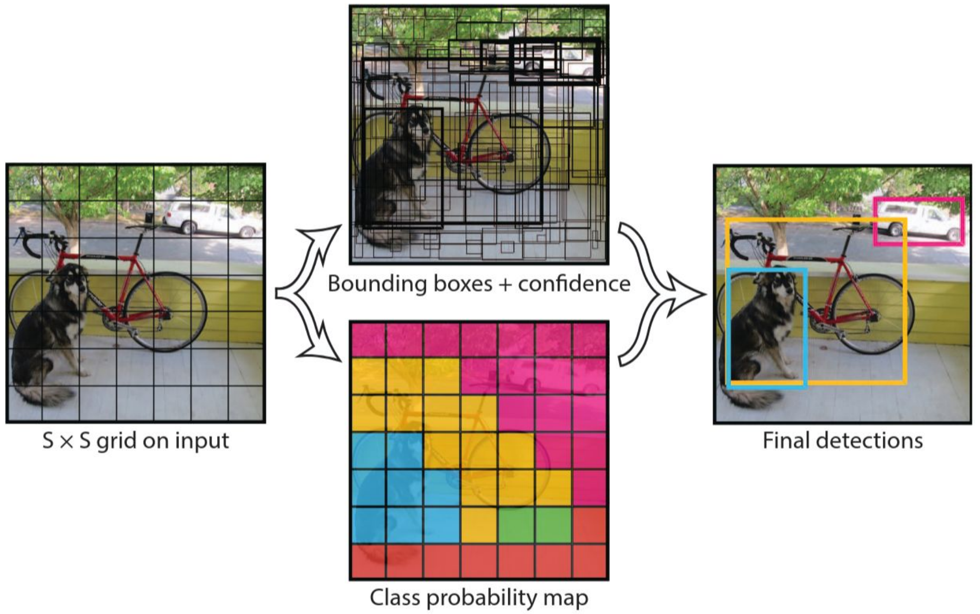
- Object detection mathematical formulation:
  - We are given:
    - RGB image $\mathbf{I} \in \mathbb{R}^{3 \times H \times W}$, where H is the height and W the width.
    - Ground truth $\mathbf{Y}_\mathbf{I} \in \mathbb{R}^{K \times 5}$, where K is the number of bounding boxes.
    - $\mathbf{Y}_{\mathbf{I},k} = [c_k, x_k, y_k, w_k, h_k], \ \forall \ k \in \{1, 2, \dots, K\}$, where:
      - c_k is the bounding box class.
      - x_k and y_k are the coordinates of the bounding box center.
      - w_k and h_k are the width the height of the bounding box respectively.
  - We predict:
    - $\hat{\mathbf{Y}}_\mathbf{I} \approx \mathbf{Y}_\mathbf{I}$, for all images $\mathbf{I}$.
    - We use a neural network $f(\mathbf{I}; \boldsymbol{\theta})$, where $f: \mathbf{I} \rightarrow \hat{\mathbf{Y}}_\mathbf{I}$.
    - The neural network learns parameters $\boldsymbol{\theta}$ during training.
    - SOTA object detection models have **tens of millions** of **parameters**.

Human Detection

- You Only Look Once (YOLO) algorithm.



YOLO model architecture [RED2016].

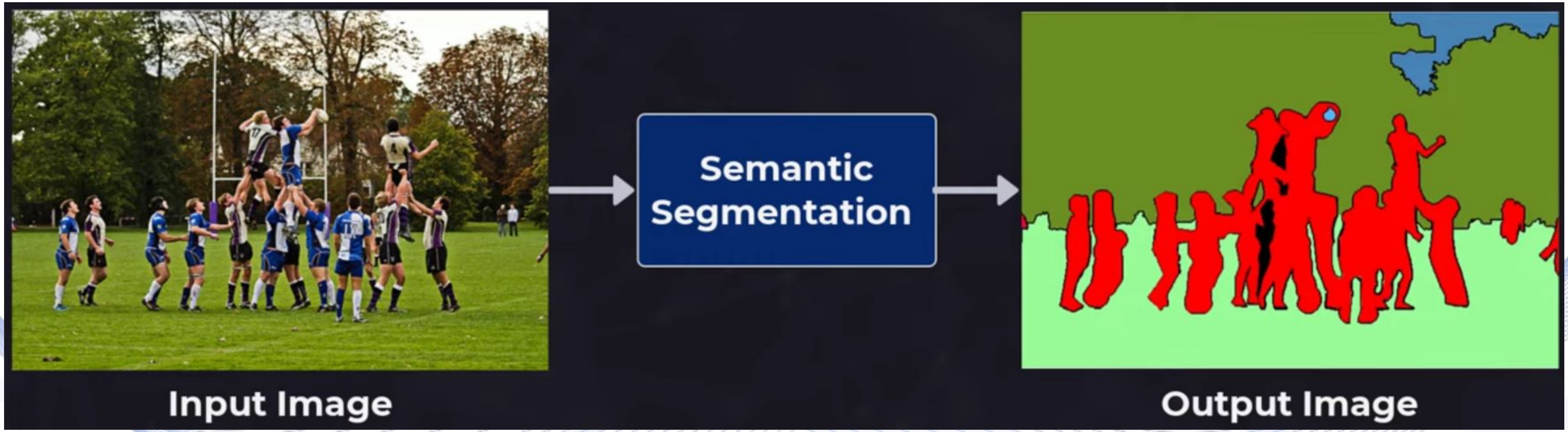


YOLO algorithm [RED2016].

Contents

Human segmentation

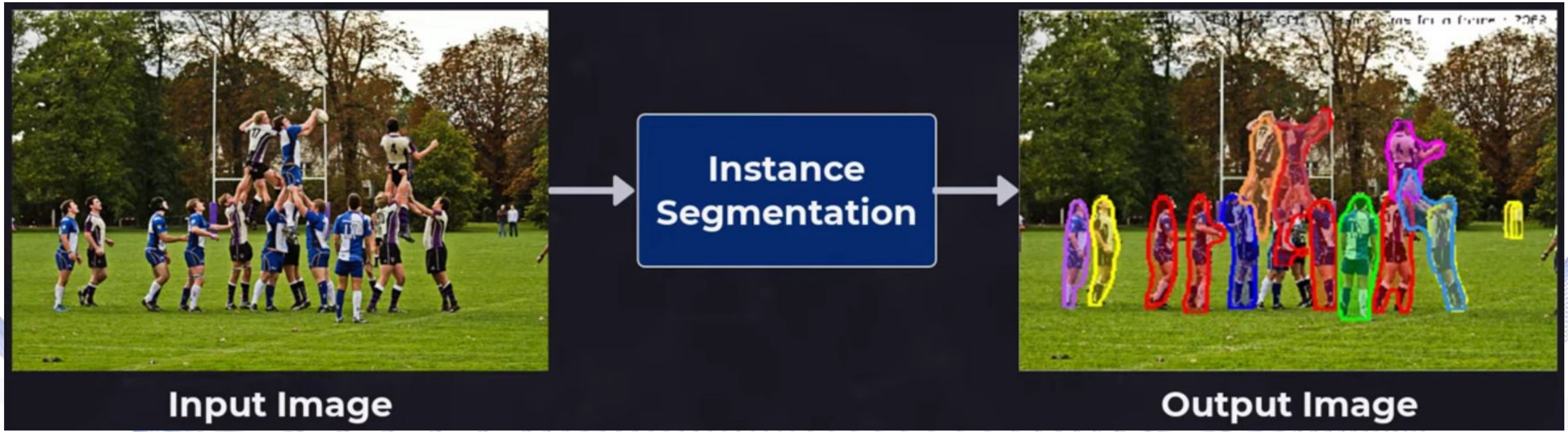
- Semantic segmentation



Semantic segmentation example [LEA2022].

Human segmentation

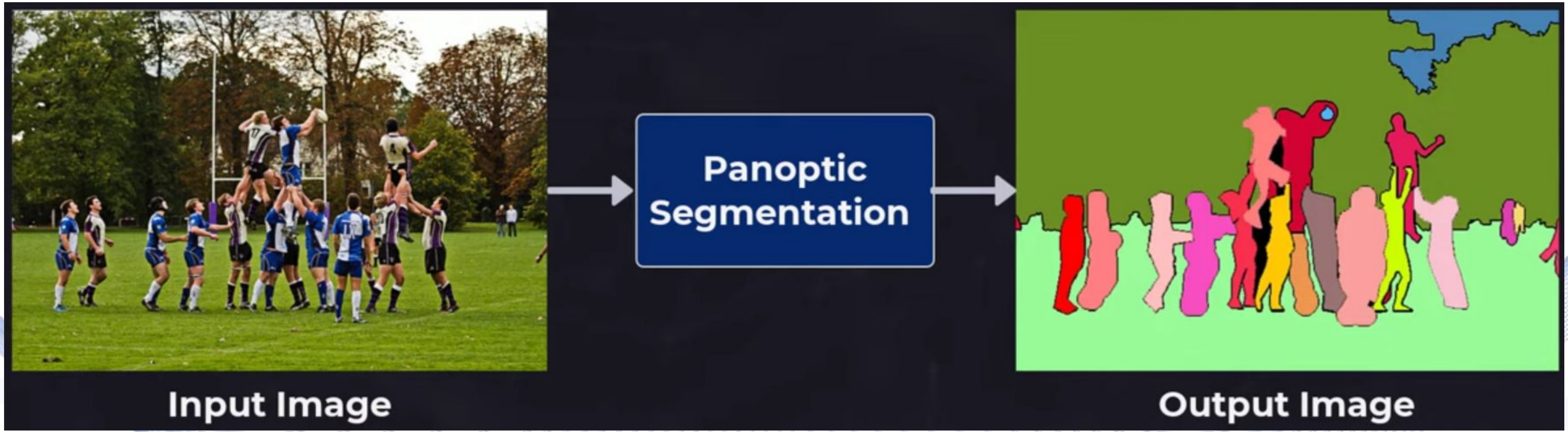
- Instance segmentation



Instance segmentation example [LEA2022].

Human segmentation

- Panoptic segmentation



Panoptic segmentation example [LEA2022].

Human segmentation

Person instance segmentation.

Scene semantic segmentation [COR2016].

Human segmentation

Crowd detection via image segmentation.

- Avoid detected crowds to ensure safety.

Human segmentation

- Image segmentation partitions the image domain $\mathcal{I}$ into the subsets $\mathcal{R}_i$, $i = 1, \ldots, N$, having the following properties:

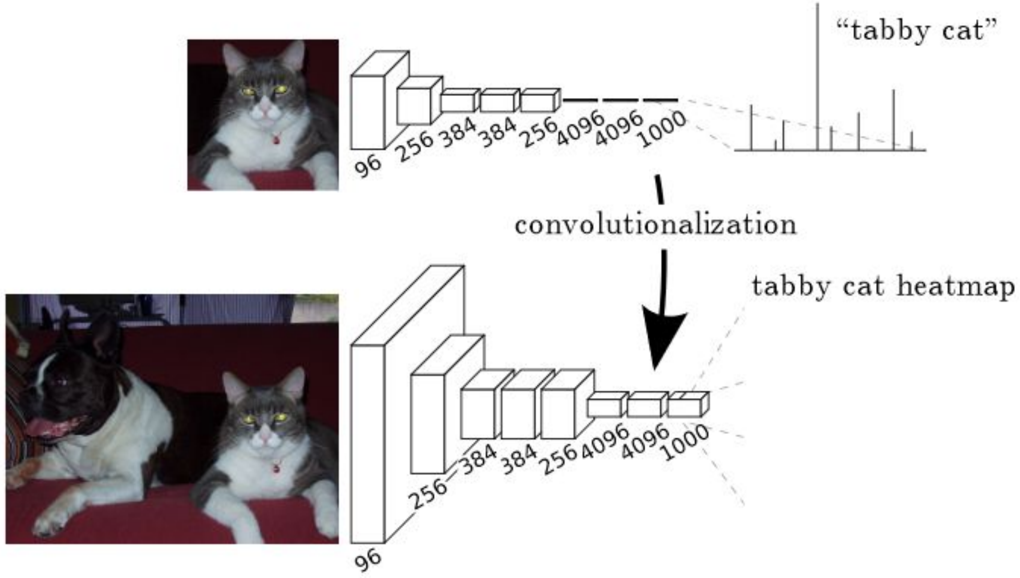
$$\mathcal{I} = \bigcup_{i=1}^{N} \mathcal{R}_i,$$

$$\mathcal{R}_i \cap \mathcal{R}_j = \emptyset, \qquad \text{for } i \neq j,$$

- **Semantic segmentation**: Classifies each pixel into a category (e.g., road, car, person).
- **Instance segmentation**: It also separates different objects of the same class, but considers all non-objects as background.
- **Panoptic segmentation**: Combines semantic segmentation and instance segmentation.

Human segmentation

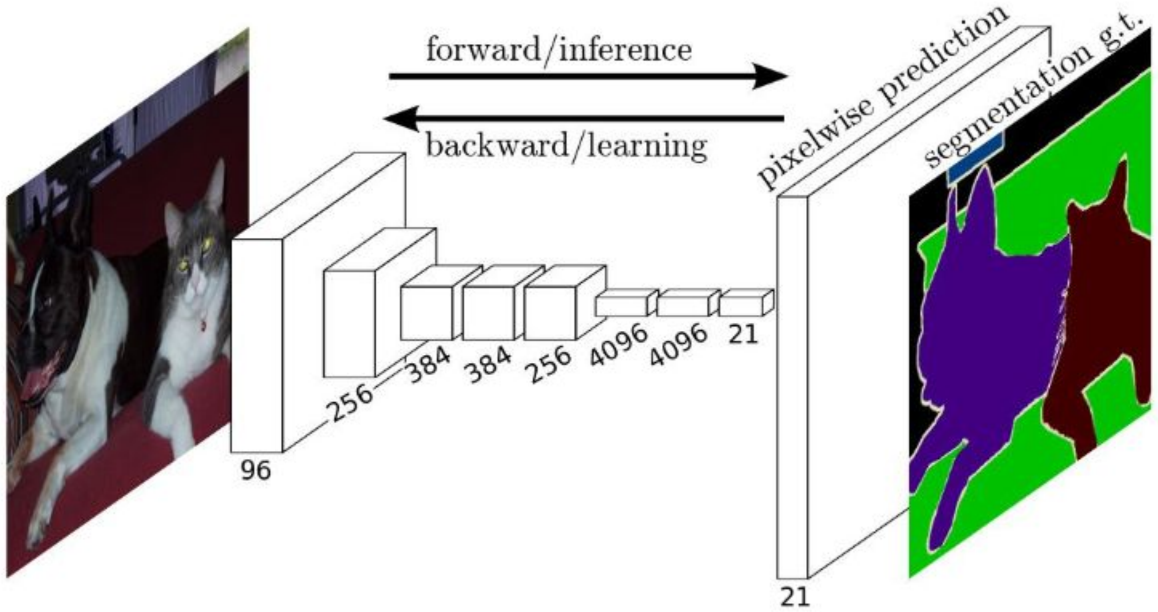
- **Convolutionalization**: Transformation of the fully connected layers of image classification networks (e.g., AlexNet) into convolution layers.
- End-to-end dense learning is possible.



Convolutionalization [LON2015].

Human segmentation

- **Fully convolutional networks** (FCNs) for image semantic segmentation.
- This FCN architecture modifies a pre-trained AlexNet.

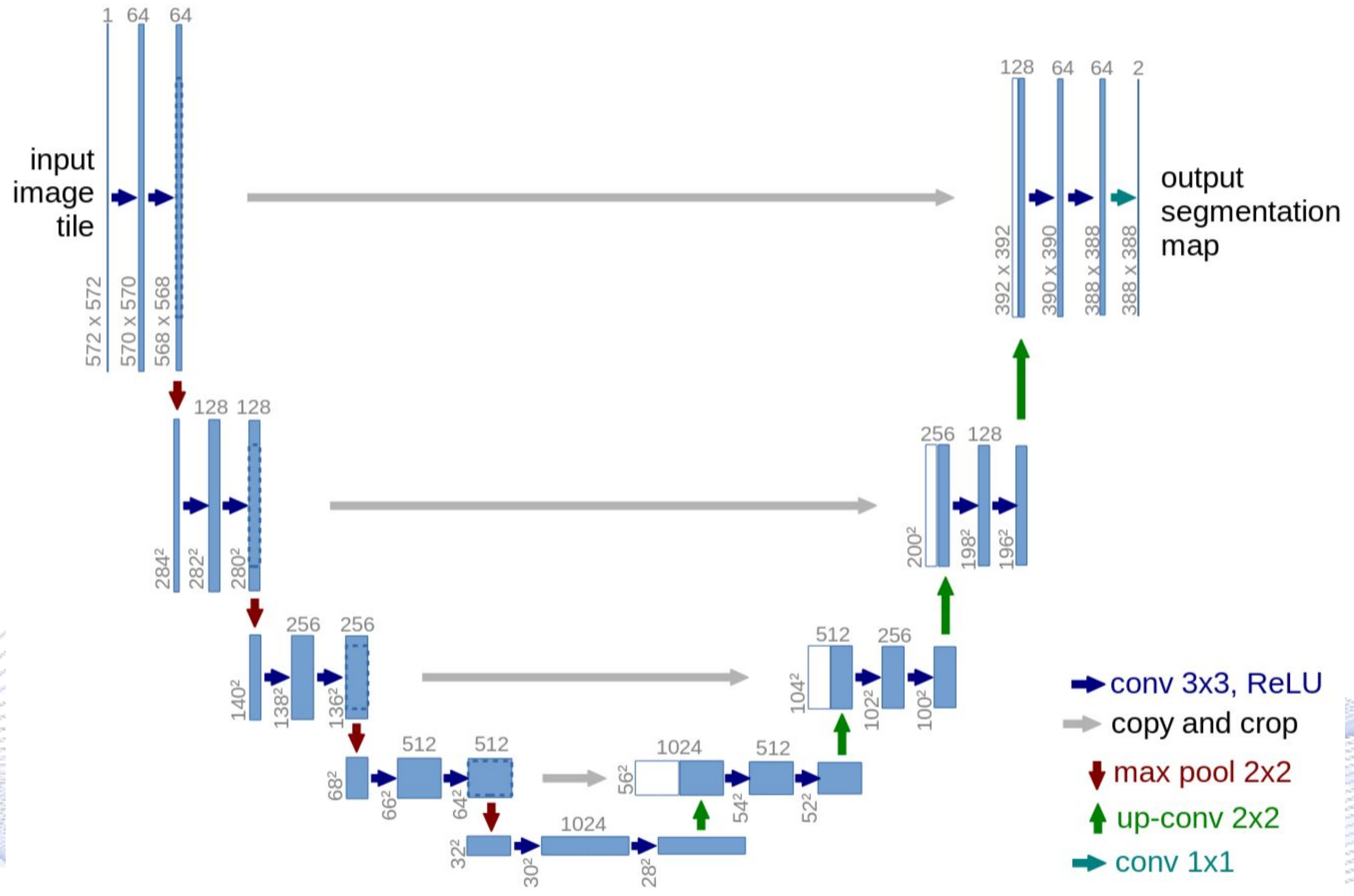


FCN for 21-class semantic segmentation [LON2015].

Human segmentation

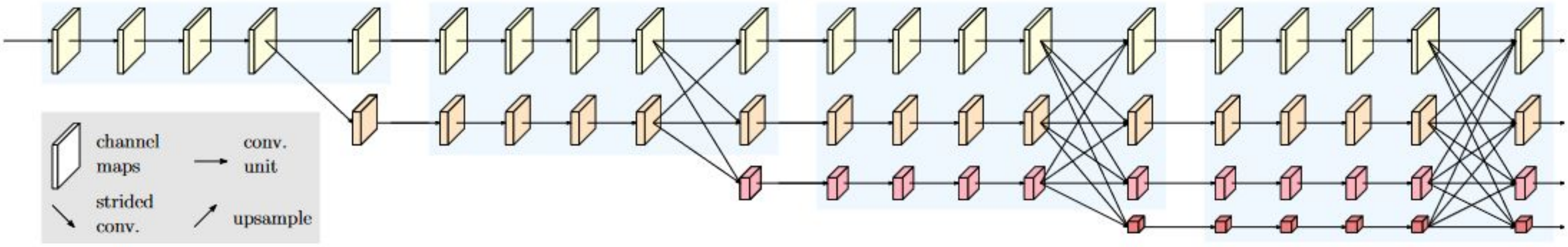
- Input resolution is radically reduced → hard to produce fine-grained segmentations.
- Improvements:
  - Skip connections.
  - U-shaped network architecture (e.g., U-Net [RON2015]).
    - Multiple skip connections to maintain information from high-resolution feature maps.
  - High-resolution networks (e.g., HR-Net [WAN2020]).
    - Maintain high-resolution feature maps throughout the forward pass process.

Human segmentation



U-Net network architecture [RON2015].

Human segmentation



High-resolution image segmentation networks [WAN2020].

Human segmentation

- Similar DNN approaches can also be used for ***monocular depth estimation***.
  - Goal is to ***regress depth maps*** that correspond to input images.

[ZHE2019]

[GEI2013]

Contents

- Human-centered AI overview
- Human detection
- Human segmentation
- **Human pose estimation**
- Human action recognition
- Human gesture recognition
- Applications

Human pose estimation

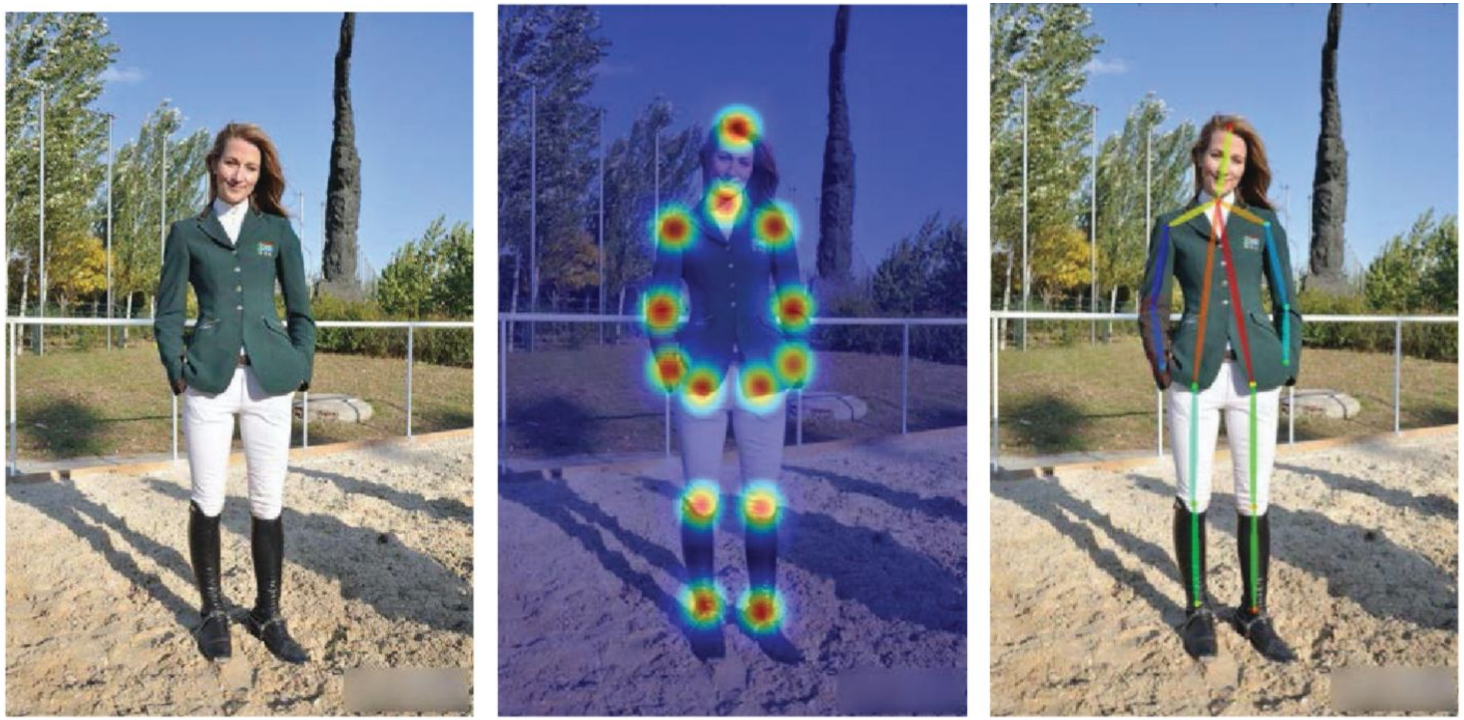
2D body pose.

Human pose estimation

- **Human pose estimation** (HPE) algorithms describe the configuration of human body parts.
- Input:
  - RGB images.
  - Depth maps.
  - Multi-view cameras.
- Output:
  - Set of 2D keypoint coordinates: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$.
  - Set of 3D keypoint coordinates: $\{(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_n, y_n, z_n)\}$.
  - Set of confidence scores for each keypoint: $\{c_1, c_2, \dots, c_n\}$.

Human pose estimation

- **Heatmap-based** methods for **HPE**.



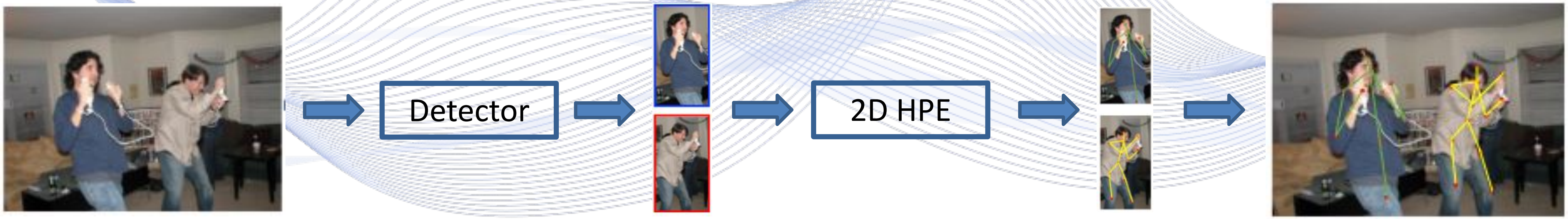
[DAN2019]

Human pose estimation

Multi-person 2D HPE

Top-down pipeline

- Each person is detected on the input image (2D bounding boxes) using off-the-shelf person detectors [NGU2016].
- Single-person HPE is performed to each person bounding box.
- Inference speed increases linearly with the number of persons.



[DAN2019]

Contents

- Human-centered AI overview
- Human detection
- Human segmentation
- Human pose estimation
- **Human action recognition**
- Human gesture recognition
- Applications

Human action recognition


run


walk


jump f.


jump p.


bend

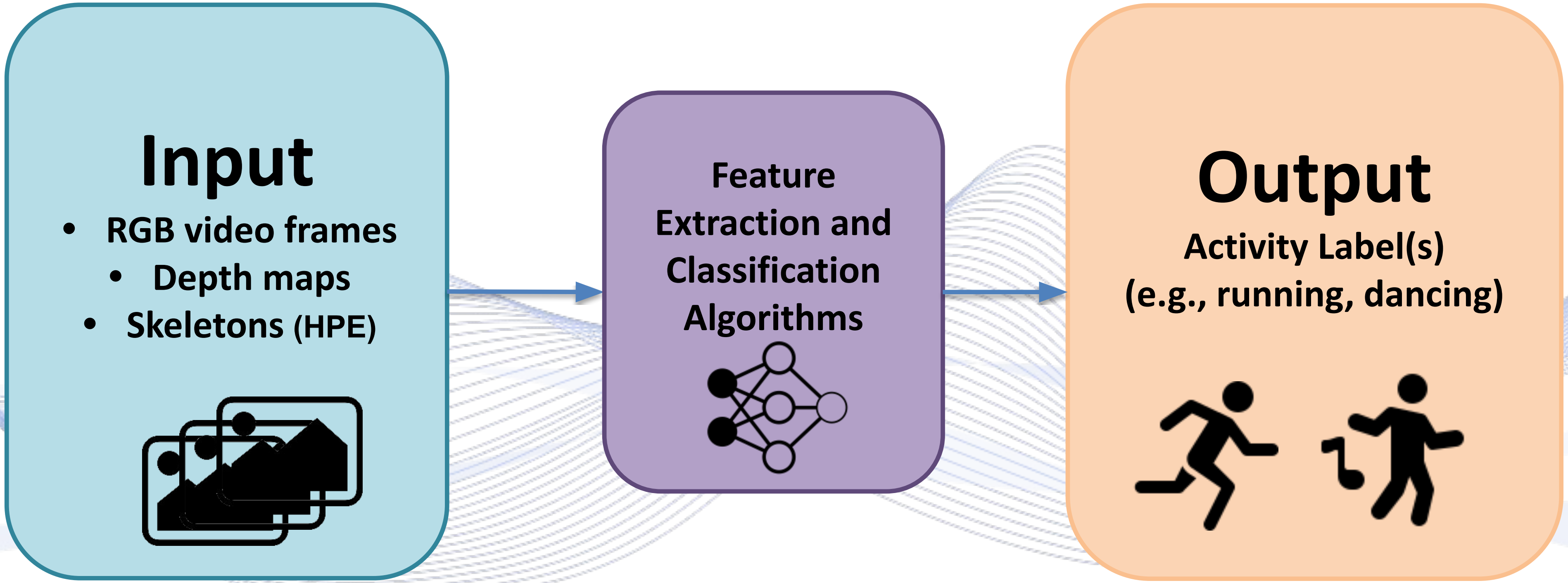

sit


wave


fall

Human action recognition

- **Human Action Recognition** (HAR) aims at automatically recognizing the actions of persons given a sequence of input data.



Human action recognition

- **3D CNNs** employ 3D convolution between kernels and data to produce feature tensors.
- Can be applied on spatio-temporal (video) or volumetric data analysis (e.g., medical imaging).
- Can learn ***spatio-temporal neural features*** from raw frame sequences, without complex hand-crafted features or multi-stream DNN architectures.

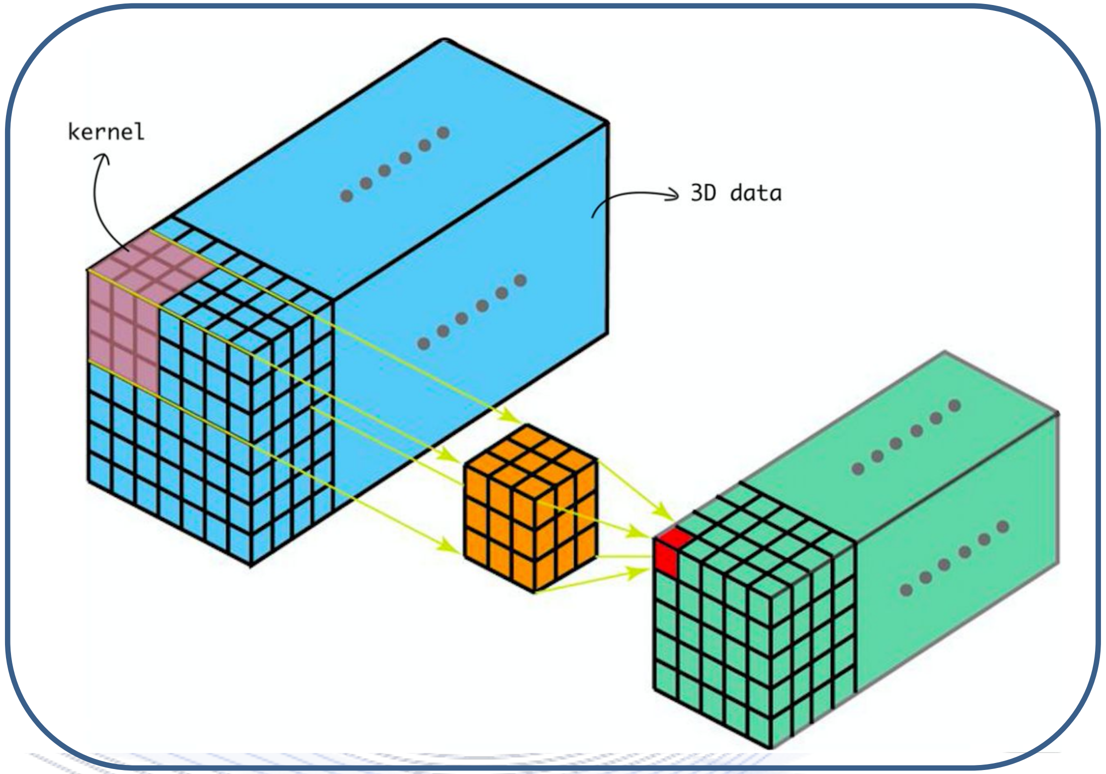


image from https://towardsdatascience.com/understanding-1d-and-3d-convolution-neural-network-keras-9d8f76e29610

Human action recognition

T-C3D: temporal convolutional 3D network for real-time action recognition [LIU2018].

Objective:

- Real-time recognition of the action performed in video sequences using 3D convolutions.

Methodology:

- Temporal info is extracted using the nature of 3D networks.
- A temporal encoding technique is used to model characteristics of the entire video.
- The overall process is end-to-end trainable.
- Good accuracy.

Contents

Human gesture recognition

- **Gesture** is an expressive meaningful body motion involving physical movement of head, body, hands etc.
- Intention:
  - Convey meaningful information
  - Interact with environment.
- Gestures can be:
  - **Static**: certain body posture or configuration.
  - **Dynamic**: prestrike, stroke and poststroke phases.

Human gesture recognition

- Gestures can be **culture-specific**.
- Gestures can be categorized based on the body part as:
  - **Hand gestures**:
    - hand poses, sign language etc.
  - **Head and face gestures**:
    - Shaking head.
    - Speaking by opening and closing the mouth.
    - Raising the eyebrows.
    - Emotions: surprise, anger, happiness, sadness.
  - **Body gestures**: full body motion.

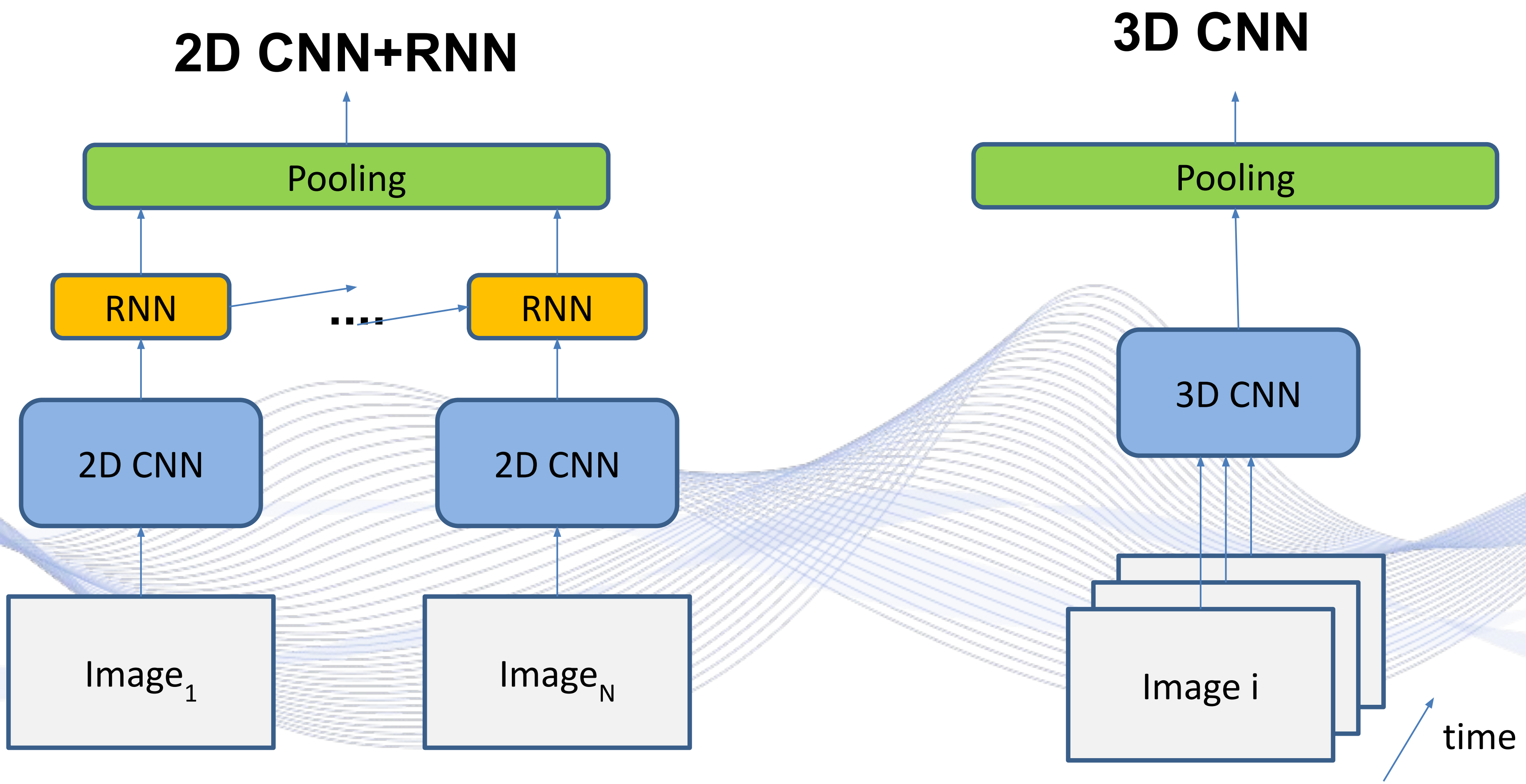
Human gesture recognition

- **Gesture recognition is similar to human action recognition.**
- Data sources:
  - Visual: RGB images, depth maps, thermal images.
  - Wearable: Magnetic field trackers, instrumented gloves (active or passive).
- Human gestures from visual data are analyzed by DNN algorithms.

Human gesture recognition

- Gesture recognition DNN architectures:
  - **2D CNN+RNN**:
    - RNNs are used to encode temporal information.
    - 2D CNNs are used to encode spatial information.
  - **3D CNN**: encodes both spatial and temporal relationships between the input frames.

Human gesture recognition



Contents

Artificial Intelligence & Information Analysis Lab

Applications

The presented algorithms have numerous applications on real-world scenarios that involve self-driving cars, UAVs, etc.

- **Pedestrian detection and intention recognition.**
- In-cabin human-vehicle interaction.
- Assessment and modeling of driver's behavior.
- Road scene understanding.
- **Gesture-based vehicle control.**

Applications

- Pedestrian intention (cross/no-cross) recognition.

Pedestrian intention recognition [PAP2022].

Applications

- Scene understanding.

[COR2016]

[GEI2013]

Road scene segmentation and depth estimation.

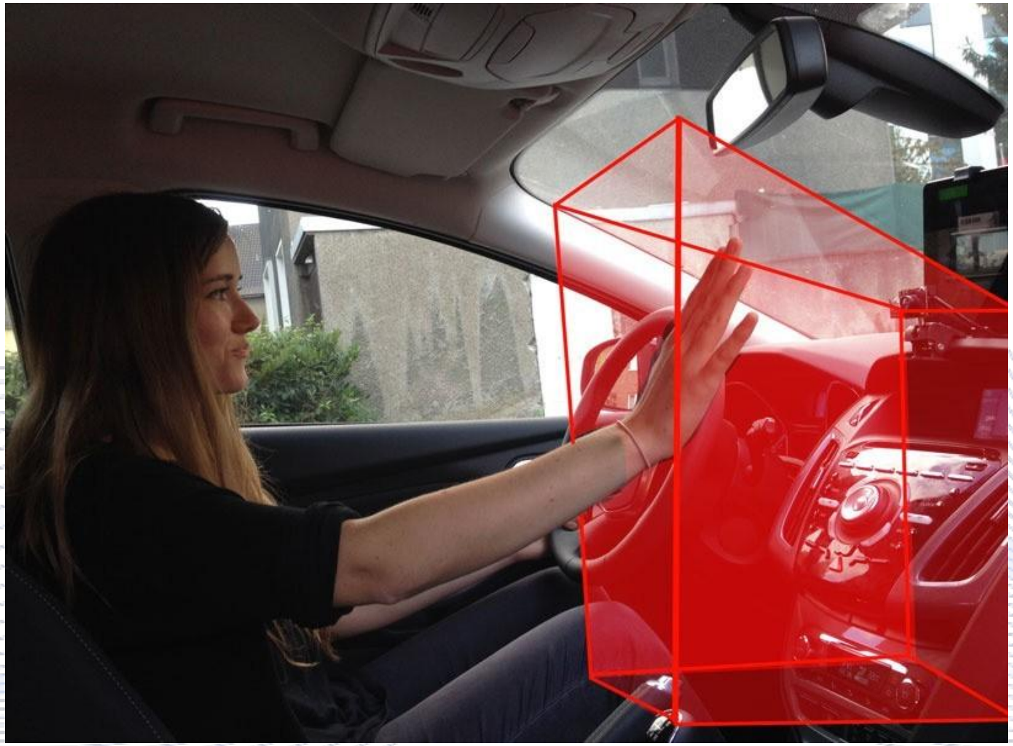
Applications

Human–vehicle interaction via gestures.

- Algorithms usually run onboard.
  - Estimation accuracy and execution speed of algorithms are crucial.
  - Specifically designed DNNs.
  - Software that translates DNN estimations to control commands.
- **Real-time gesture recognition.**

Applications

- Autonomous vehicle control.



Performing hand gesture detection in the range of the sensor of time-of-flight-ToF (area of detection in red) [ZEN2018].

Artificial Intelligence & Information Analysis Lab

Applications

- Autonomous vehicle control.

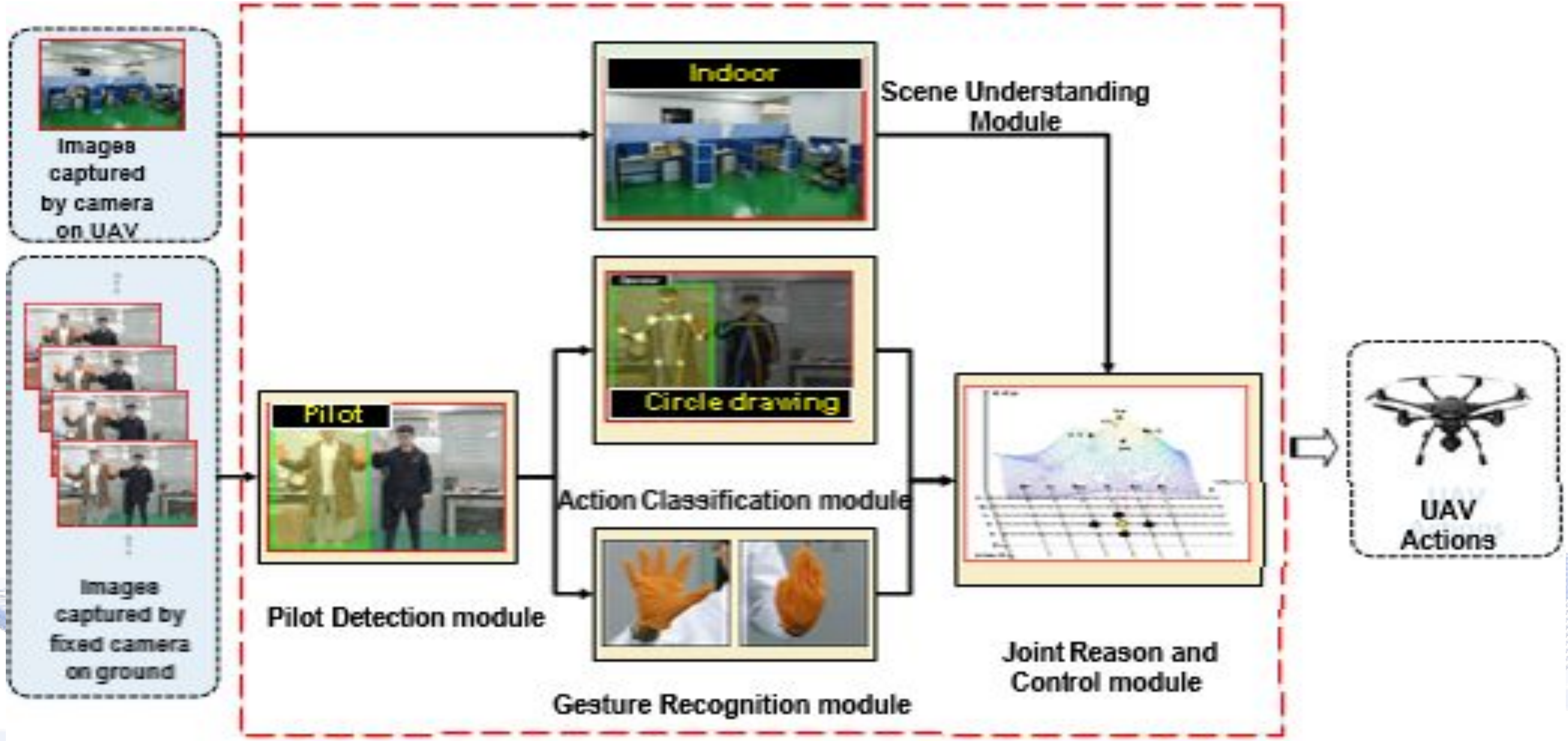
Lane change with gesture control [ZEN2018].

Applications

Gesture-controlled Drones

- Video stream is recorded through the camera and segmented into sequences of images.
- Each image is then recognized by a classification process.
- Typical commands:
  - Take off.
  - Land.
  - Move right or left.

Applications



Human-Drone Interaction model [HUA2019].

Applications

- Autonomous vehicle control.

Applications

- Crowd detection for autonomous UAV navigation.

[PAP2021].

Bibliography

[NGU2016] Nguyen, D. T., Li, W., & Ogunbona, P. O. (2016). Human detection from images and videos: A survey. *Pattern Recognition*, *51*, 148-175.

[RED2016] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[LEA2022] LearnOpenCV. Image Segmentation, Semantic Segmentation, Instance Segmentation, and Panoptic Segmentation. Available at: https://www.youtube.com/watch?v=5QU mlX Bb0MY.

[LON2015] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

[RON2015] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.

[WAN2020] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3349-3364.

[DAN2019] Dang, Q., Yin, J., Wang, B., & Zheng, W. (2019). Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, *24*(6), 663-676.

[LIU2018] Liu, K., Liu, W., Gan, C., Tan, M., & Ma, H. (2018, April). T-C3D: Temporal convolutional 3D network for real-time action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Bibliography

[PAP2022] Papaioannidis, C., Mademlis, I., & Pitas, I. (2022). Fast CNN-based single-person 2D human pose estimation for autonomous systems. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(3), 1262-1275.

[COR2016] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[GEI2013] Geiger, Andreas, et al. “Vision meets robotics: The KITTI dataset,” The International Journal of Robotics Research 32, 11, pp. 1231-1237, 2013.

[ZEN2018] Nico Zengeler , Thomas Kopinski and Uwe Handmann “Hand Gesture Recognition in Automotive Human–Machine Interaction Using Depth Cameras”.

[HUA2019] Bo Chen, Chunsheng Hua, Decai Li, Yuqing He and Jianda Han “Intelligent Human–UAV Interaction System with Joint Cross-Validation over Action–Gesture Recognition and Scene Understanding”.

[PAP2021] C. Papaioannidis, I. Mademlis and I. Pitas, "Autonomous UAV Safety by Visual Human Crowd Detection Using Multi-Task Deep Neural Networks," 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr

Artificial Intelligence & Information Analysis Lab

Funded by the European Union

Contents

- Human-centered AI overview
- Human detection
- Human segmentation
- **Human pose/posture estimation**
- Human action/activity recognition
- Human gesture recognition
- Applications

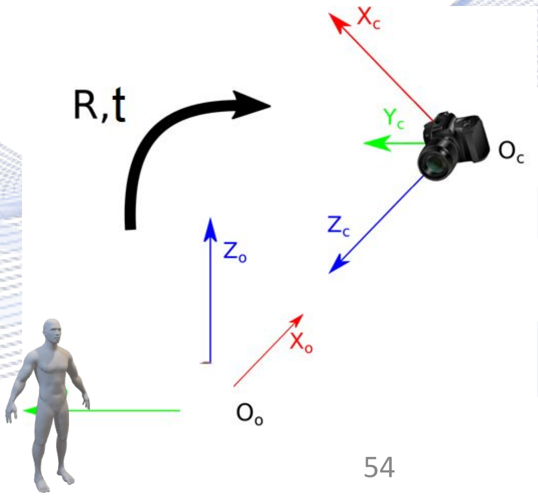
Human pose estimation

Human body pose describes the configuration of human body parts.

- Human body can be described by a graph of its parts.
- Graph nodes contain body joint descriptions:
  - 2D or 3D rotation angles
  - 2D or 3D joint coordinates.
- Confused with ***camera pose***:
- Camera 3D rotation $\mathbf{R}$ and

& translation $\mathbf{t}$ parameters.

2D body pose.



Camera pose.

Human pose estimation

Human Pose Estimation (HPE) estimates the configuration of human body parts from input data captured by sensors:

- usually images and videos.
- Provides geometric/motion information of the human body.
- ***Regression*** of human body parameters $\mathbf{p}$:

$$\mathbf{p} = \boldsymbol{f}(\mathbf{I}).$$

- Wide range of applications:
  - human-robot interaction (HRI),
  - motion analysis, AR/VR, healthcare.

2D HPE

3D HPE

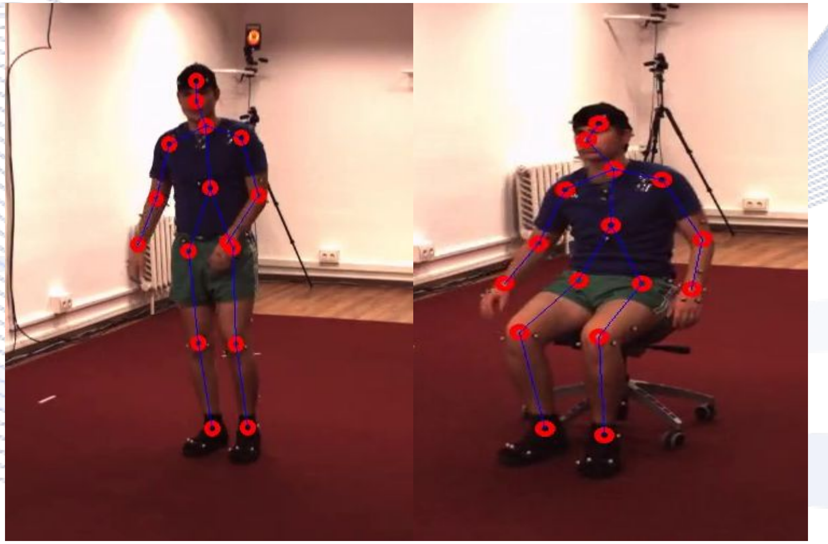
Human pose estimation

Human body posture is a specific body state, i.e., a ***labeled configuration of the body joints***: standing, sitting, lying, etc.

- Human postures are static,
- Human actions are dynamic.
- ***Classification problem*** of posture class $\mathbf{c}$:

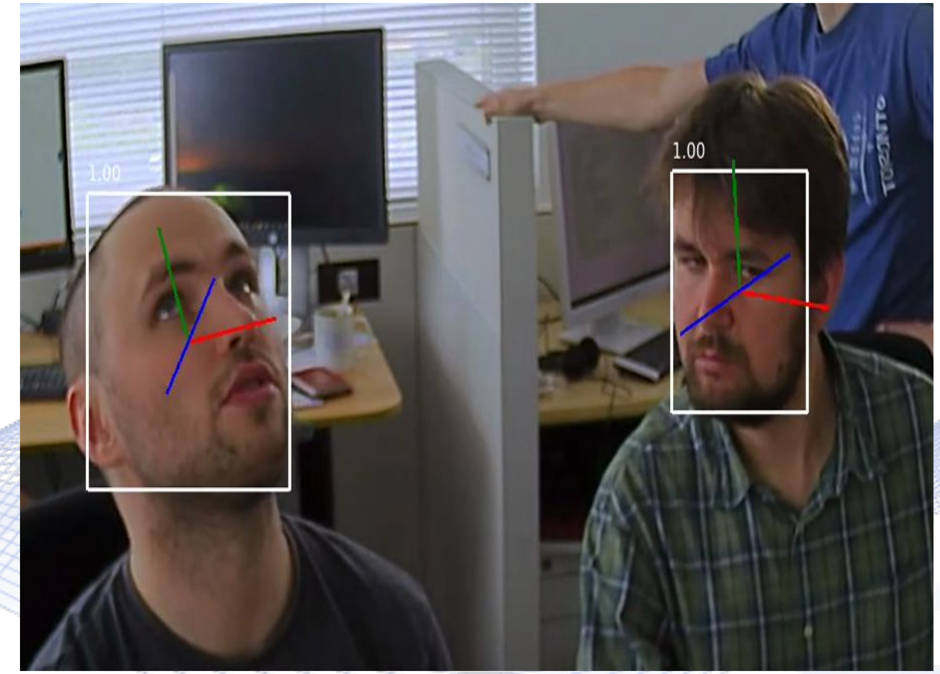
$$\mathbf{c} = \boldsymbol{f}(\mathbf{I}).$$

- Applications:
  - human-robot interaction (HRI),
  - sign language communication,
  - physical and rehabilitation training.



Standing Sitting [ION2013].

Human pose estimation



Camera pose estimation in facial images.

Human pose estimation

- ***Deep Neural Networks*** (DNNs) have achieved remarkable results in HPE.
- DNN-based approaches have outperformed classical computer vision methods.
- HPE challenges:
  - human body part ***occlusion***,
  - training data availability,
  - depth information availability, form and ambiguity.

2D human pose estimation

- Prediction of the 2D spatial location of human body key-points/joints from images or videos.
- Joint description in the ***image plane***.

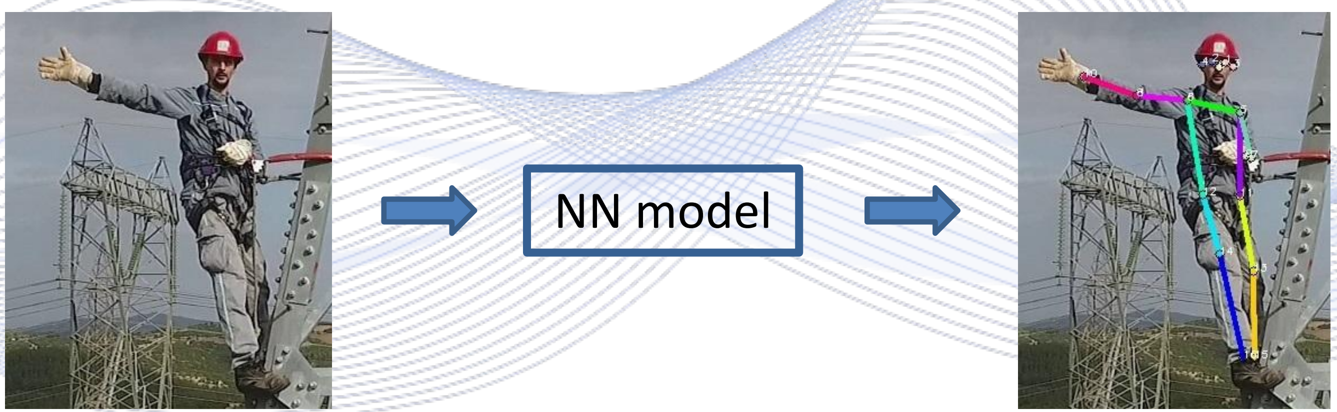
- Single-person 2D HPE:
  - direct regression methods,
  - heatmap-based methods.
- Multi-person 2D HPE:
  - top-down approach,
  - bottom-up approach.

2D human pose estimation

Single-person 2D HPE

Direct regression methods

- End-to-end framework.
- Regress (learn) a mapping from the input image to body joints or parameters of human body models.



2D human pose estimation

Single-person 2D HPE

Direct regression methods

- If $\mathbf{I}$ is an input RGB image of resolution $M \times N$ and $\boldsymbol{f}$ is the 2D HPE DNN, direct regression methods aim to directly predict (estimate):

$$\mathbf{p} = \boldsymbol{f}(\mathbf{I}),$$

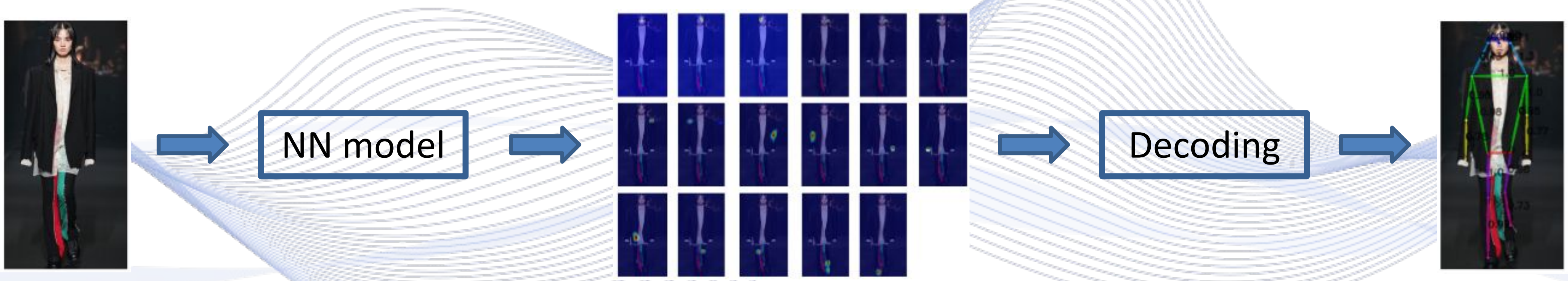
- $\mathbf{p} = [\mathbf{j}_1^T, \mathbf{j}_2^T, \dots, \mathbf{j}_K^T]^T$: pre-defined set of body joints that constitute the 2D human pose,
- K is the number of the body joints,
- $\mathbf{j}_k = [x_k, y_k]^T \in \mathbb{N}^2, k = 1, \dots, K$ human skeleton joint representation in pixel coordinates ***on the image plane***.

2D human pose estimation

Single-person 2D HPE

Heatmap-based methods

- Train a body part detector to predict the position of body joints.
- Estimate ***joint heatmap images*** that represent the joint locations.



[DAN2019]

2D human pose estimation

Single-person 2D HPE

Heatmap-based methods

- Instead of directly predicting $\{\mathbf{j}_1, \mathbf{j}_2, \dots, \mathbf{j}_K\}$, $\boldsymbol{f}$ predicts 2D body joint heatmaps $\{\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_K\}$ of resolution $M \times N$ (one for each joint):
$$\{\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_K\} = \boldsymbol{f}(\mathbf{I}).$$
- Each heatmap $\mathbf{H}_k \in \mathbb{R}^{M \times N}$ encodes the 2D location of the corresponding body joint by using a 2D Gaussian function centered at the 2D position of the body joint in the input image.
- 2D pixel coordinates of each body joint can be obtained by choosing the $\mathbf{j}_k = [x_k, y_k]^T$ pairs with the ***highest heat value***.

2D human pose estimation

Single-person 2D HPE

Heatmap-based methods

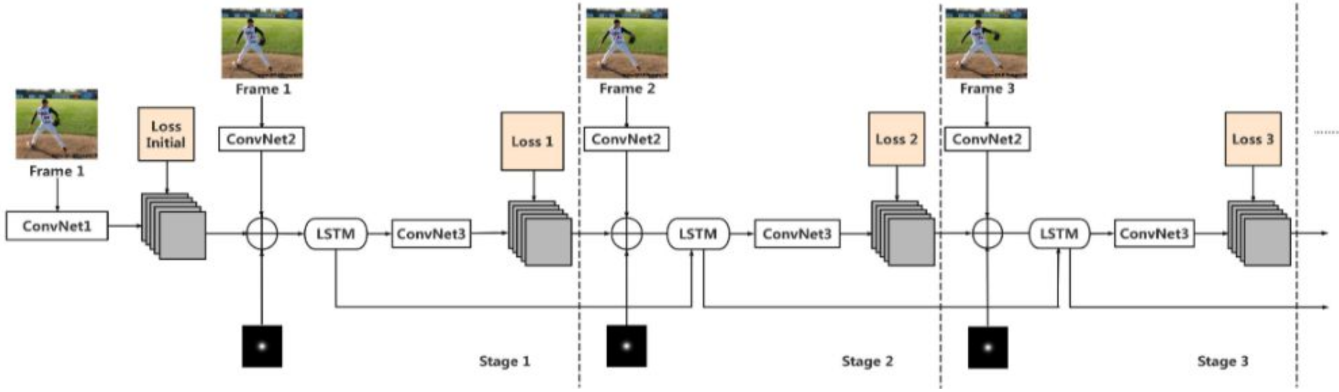
- Heatmaps provide richer supervision information, by preserving the spatial location information.
- Allow using the powerful ***Convolutional Neural Networks*** (CNNs).
- Facilitate DNN/CNN training.
- Used in state-of-the-art 2D HPE approaches.

2D human pose estimation

Single-person 2D HPE

2D HPE in video sequences

- Video sequences are spatio-temporal (3D) signals.
- Temporal information → model that can handle sequential data:
  - ***Recurrent Neural Networks*** (RNN), or
  - ***Long Shot-Term Memory*** (LSTM) networks.



[LUO2018].

2D human pose estimation

Multi-person 2D HPE

- Estimate the 2D skeletons of multiple persons that appear in the input image.
  - All persons must be localized.
  - Detected body keypoints must be grouped for different persons.

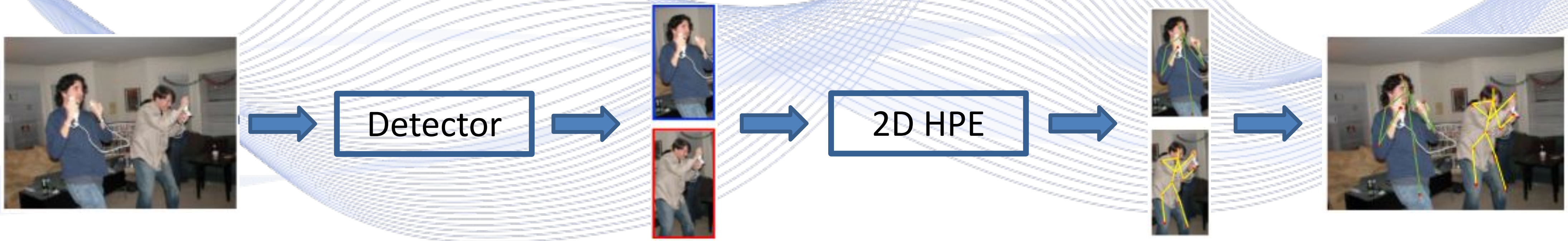
[CAO2017]

2D human pose estimation

Multi-person 2D HPE

Top-down pipeline

- Each person is detected on the input image (2D bounding boxes) using off-the-shelf person detectors [REN2015].
- Single-person HPE is performed to each person bounding box.
- Inference speed increases linearly with the number of persons.



AN2019]

2D human pose estimation

Multi-person 2D HPE

Bottom-up pipeline

- Localize all the body joints in the input image.
- Group the detected body joints to the corresponding persons.
- ***Increased inference speed*** compared to top-down approaches, since body joints for all persons are estimated simultaneously.
- Grouping of estimated body joints is required.

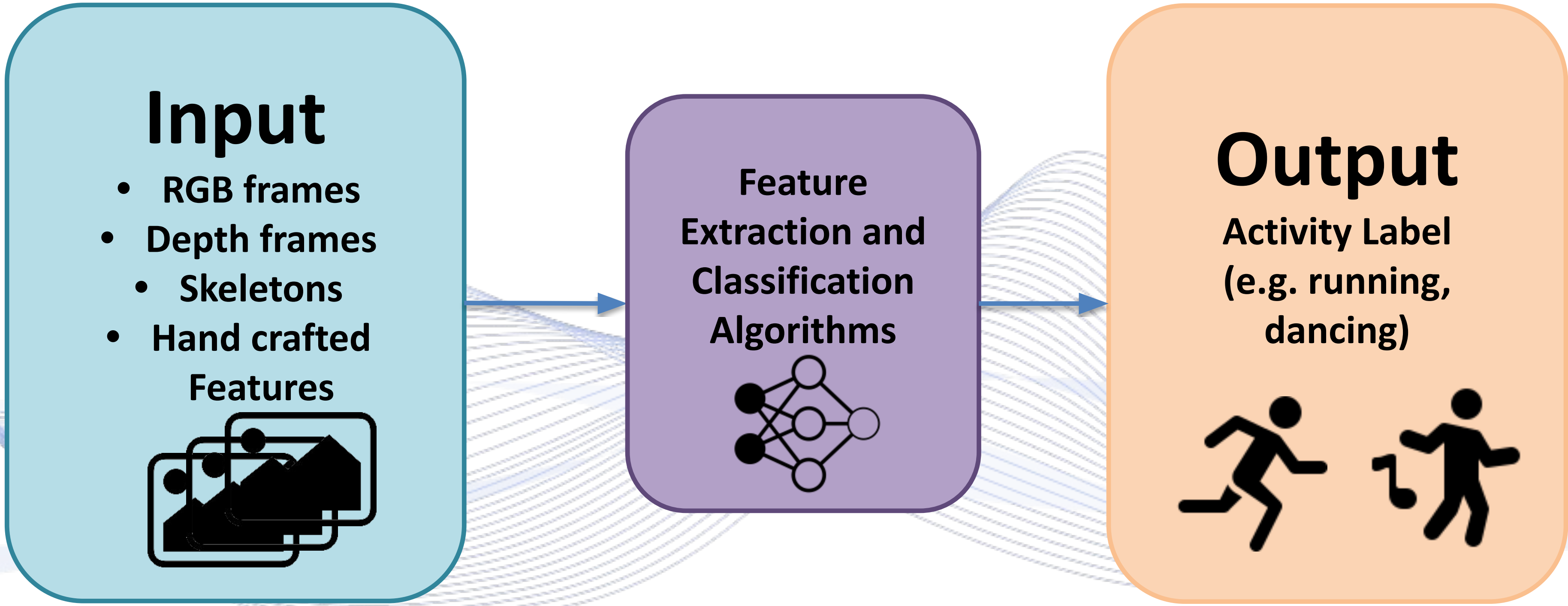
[DAN2019]

Contents

- Human-centered AI overview
- Human detection
- Human segmentation
- Human pose/posture estimation
- **Human action/activity recognition**
- Human gesture recognition
- Applications

Human action/activity recognition

- ***Human Activity/Action Recognition*** (HAR) aims to automatically recognize the actions of persons given a sequence of input data.



Human action/activity recognition

Human Activity/Action Recognition (***HAR***):

- To identify the action of a person.
- ***Action*** is an elementary ***human activity***.

Classification problem:

- ***Input***: a single-view or multi-view video or a sequence of 3D human body models (or point clouds).
- ***Output***: An action label belonging to a set of N_A action classes (e.g., walk, run) for each frame or for the entire sequence

Human action/activity recognition


run


walk


jump f.


jump p.


bend

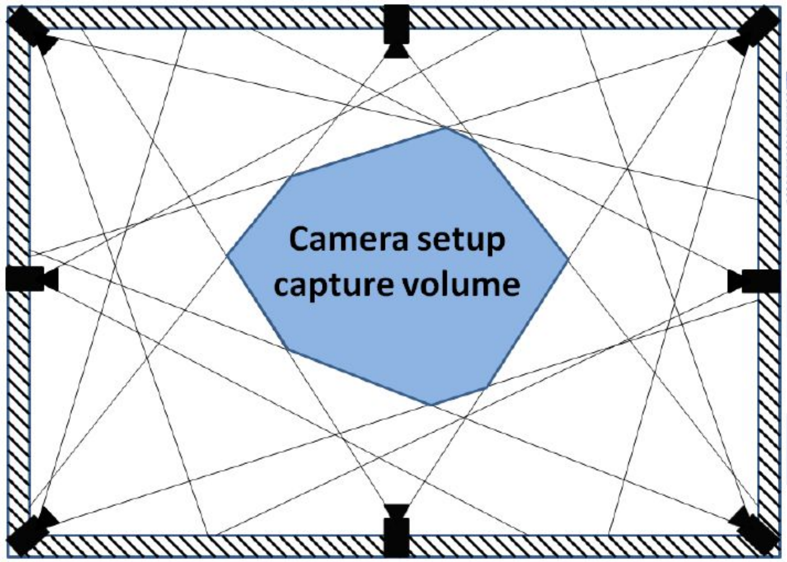

sit


wave


fall

Human action/activity recognition

- ***Single-view***: methods utilizing one camera:
  - special cases of multi-view ones, i.e., for $N_C = 1$.
- ***Multi-view***: methods utilizing multiple cameras forming a multi-camera setup.



An eight-view camera

Neural HAR

VML

- Still images □ ***spatial*** information.
- Multiple video frames □ ***temporal*** information.

- ***3D CNNs***
- ***Multi-stream DNN networks.***
- They capture both ***temporal & spatial information.***

HAR with 3D CNNs

- ***3D CNNs*** employ 3D convolution between kernels and data to produce feature tensors.
- Can be applied where spatio-temporal (video) or volumetric data (e.g., Medical Imaging) analysis is important.
- Can learn ***spatio-temporal neural features*** from raw frame sequences, without complex hand-crafted features or multi-stream DNN architectures.

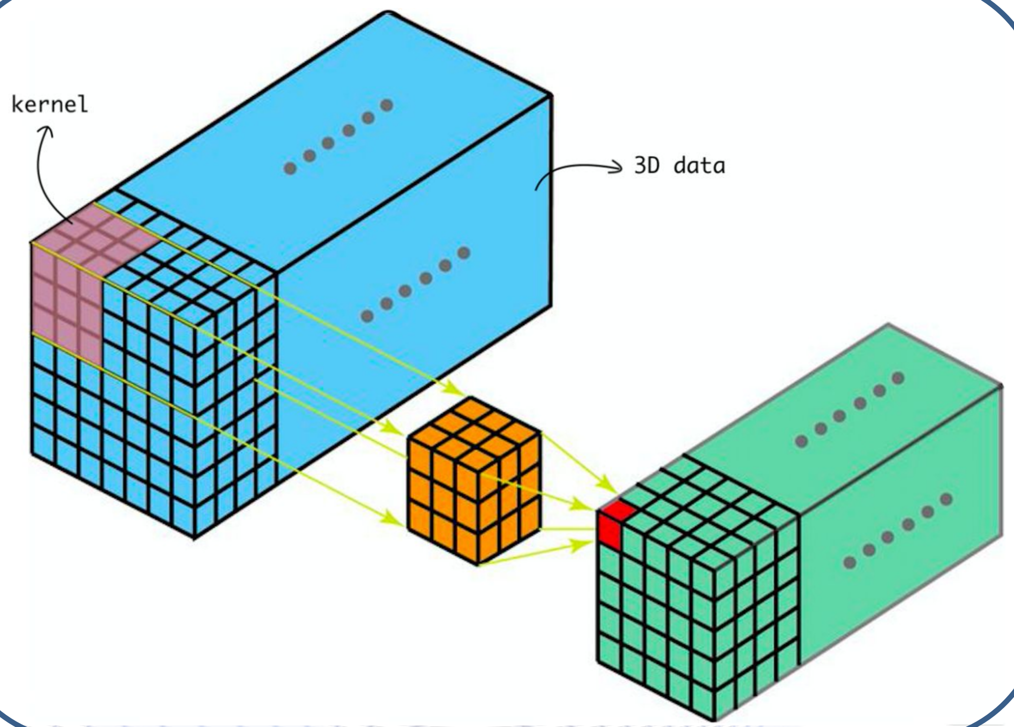


image from https://towardsdatascience.com/understanding-1d-and-3d-convolution-neural-network-keras-9d8f76e29610

HAR with 3D CNNs

T-C3D: temporal convolutional 3D network for real-time action recognition [LIU2018].

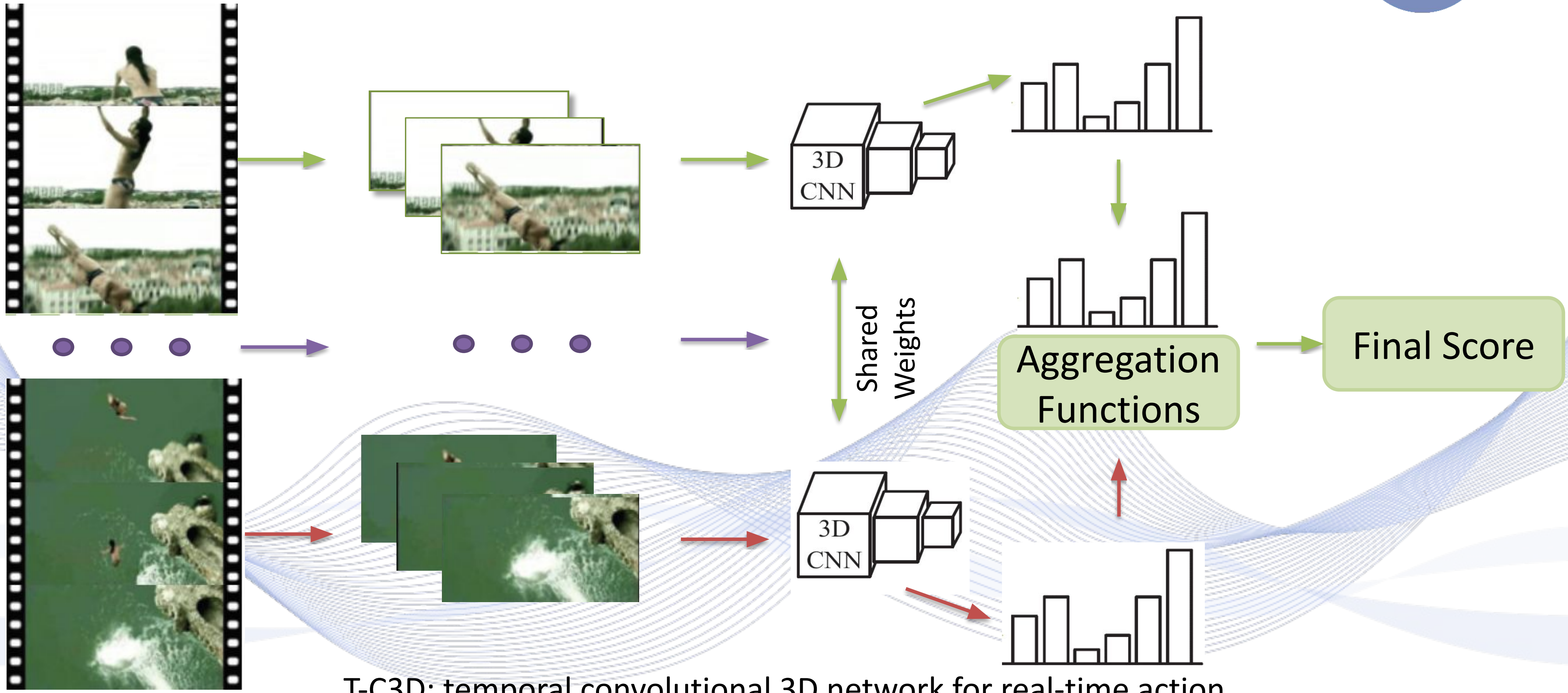
Objective:

- Real-time recognition of the action performed in video sequences using 3D convolutions.

Methodology:

- Temporal info is extracted using the nature of 3D networks.
- A temporal encoding technique is used to model characteristics of the entire video.
- The overall process is end-to-end trainable.
- Good accuracy.

HAR with 3D CNNs



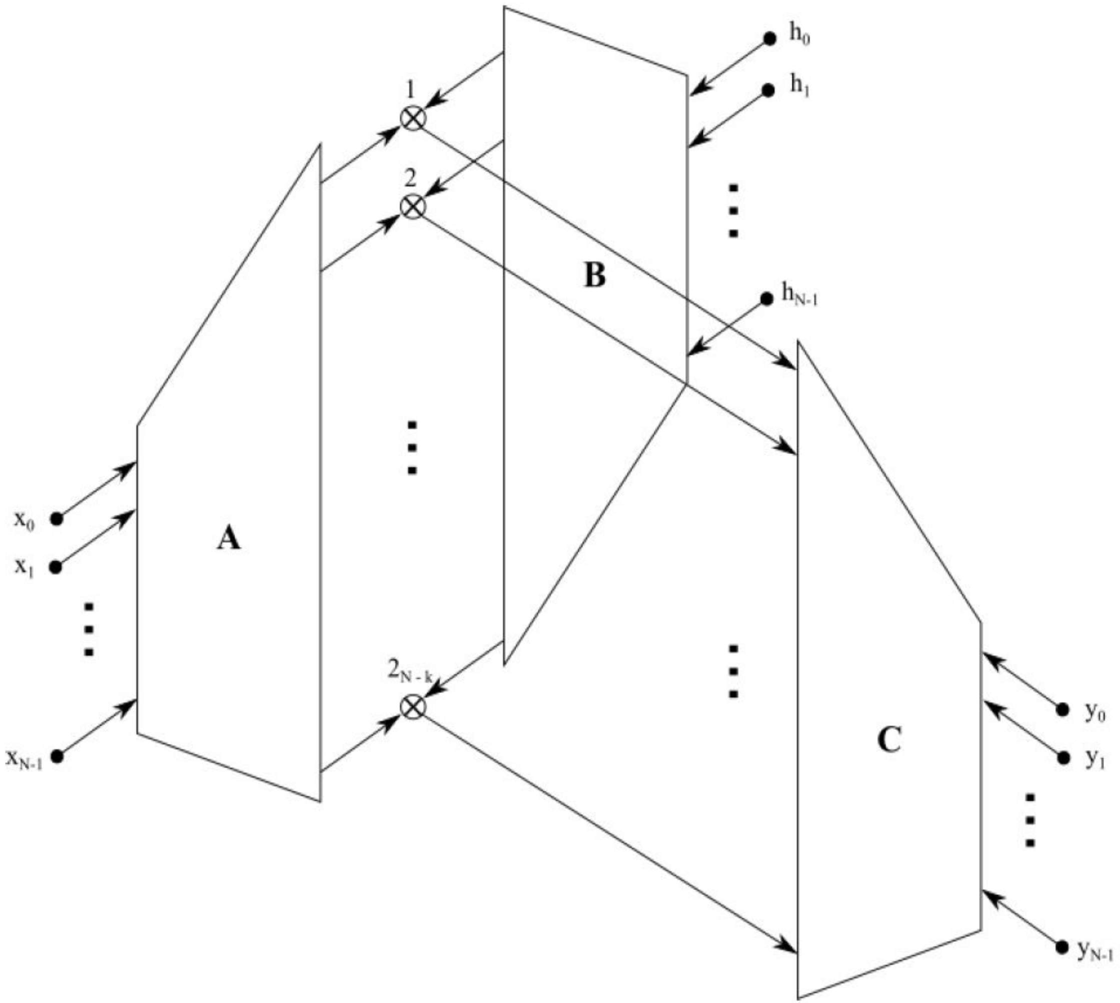
T-C3D: temporal convolutional 3D network for real-time action recognition [LIU2018].

HAR with 3D CNNs

- 3D convolutions are notoriously computationally expensive.
- ***Fast 3D convolution algorithms***:

$$\mathbf{y} = \mathbf{C}(\mathbf{A}\mathbf{x} \otimes \mathbf{B}\mathbf{h}).$$

- GEneral Matrix Multiplication (GEMM) BLAS or cuBLAS routines



Contents

- Human-centered AI overview
- Human detection
- Human segmentation
- Human pose/posture estimation
- Human action/activity recognition
- **Human gesture recognition**
- Applications

Gesture recognition

- ***Gesture*** is an expressive meaningful body motion involving physical movement of head, body, hands etc.
- Intention:
  - Convey meaningful information
  - Interact with environment.
- Gestures can be:
  - ***Static***: certain body posture or configuration.
  - ***Dynamic***: prestrike, stroke and poststroke phases.

Artificial Intelligence & Information Analysis Lab

Gesture recognition

- Gestures can be ***culture-specific***.
- Gestures can be categorized based on the body part as:
  - ***Hand gestures***:
    - hand poses, sign language etc.
  - ***Head and face gestures***:
    - Shaking head.
    - Speaking by opening and closing the mouth.
    - Raising the eyebrows.
    - Emotions: surprise, anger, happiness, sadness.
  - ***Body gestures***: full body motion.

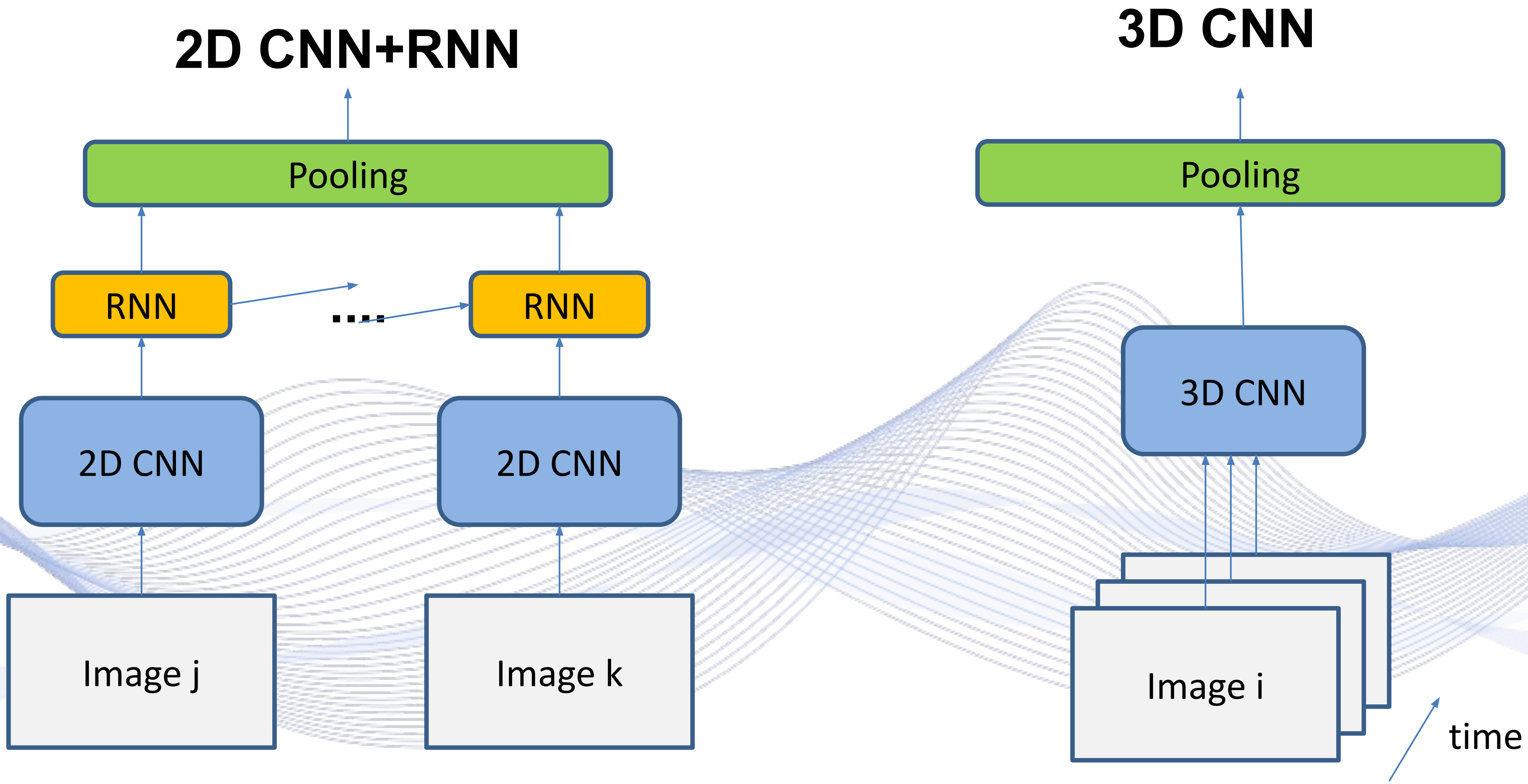
Gesture recognition

- ***Gesture recognition is similar to human action recognition.***
- Data sources:
  - Visual: RGB, depth, thermal images.
  - Wearable: Magnetic field trackers, body suits, instrumented gloves (active or passive).
- Human gestures from visual data are analyzed by DNN algorithms.
- Applications
  - ***Gesture-based vehicle control.***

DNN architectures for gesture recognition

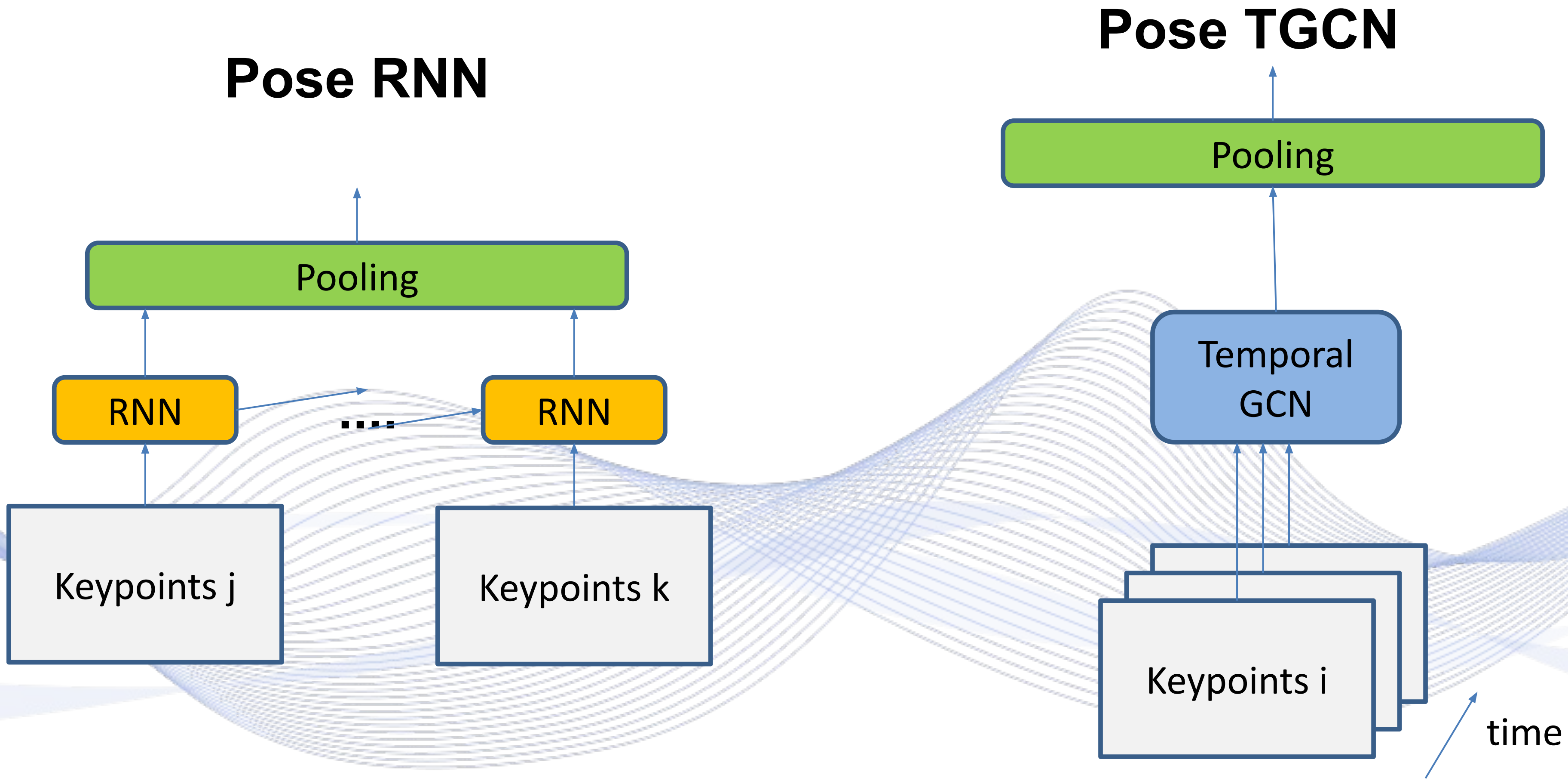
- Gesture recognition DNN architectures:
  - ***2D CNN+RNN***: RNNs are used to encode temporal information and 2D CNNs for spatial information from the input sequence.
  - ***3D CNN***: encodes spatial and temporal relationships between the input frames.
  - ***Skeleton-based models***: analyze input sequences of 2D/3D skeletons with RNNs/LSTMs to recognize gestures.
    - Spatio-temporal GCNs: model the spatio-temporal dependencies of the skeleton sequences.

DNN architectures for gesture recognition



Artificial Intelligence & Information Analysis Lab

DNN architectures for gesture recognition



Keypoints are the joints of human bodies.

Artificial Intelligence & Information Analysis Lab

Bibliography

[DAN2019] Dang, Qi, et al. "Deep learning based 2d human pose estimation: A survey." Tsinghua Science and Technology vol 24, no. 6, pp. 663-676, 2019.

[PAP2022] Papaioannidis, Christos, et al. “Fast CNN-based Single-Person 2D Human Pose Estimation for Autonomous Systems”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 3, pp. 1262-1275, 2022.

[LUO2018] Luo, Yue, et al. "Lstm pose machines." IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[CAO2017] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[REN2015] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in Neural Information Processing Systems, 2015.

[HOS2018] Hossain, Mir Rayat Imtiaz, and James J. Little. "Exploiting temporal information for 3d human pose estimation." European Conference on Computer Vision, 2018.

[CAI2019] Cai, Yujun, et al. "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks." IEEE International Conference on Computer Vision, 2019.

[LI2022] Li, Wenhao, et al. "Exploiting temporal contexts with strided transformer for 3d human pose estimation." IEEE Transactions on Multimedia, 2022.

[ROG2017] Rogez, Gregory, Philippe Weinzaepfel, and Cordelia Schmid. "Lcr-net: Localization-classification-regression for human pose." IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[ROG2019] Rogez, Gregory, Philippe Weinzaepfel, and Cordelia Schmid. "Lcr-net++: Multi-person 2d and 3d pose detection in natural images." IEEE Transactions on Pattern Analysis and Machine Intelligence vol.42 no. 5, pp. 1146-1161, 2019.

[BEN2020] Benzine, Abdallah, et al. "Pandanet: Anchor-based single-shot multi-person 3d pose estimation." IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[NIE2019] Nie, Xuecheng, et al. "Single-stage multi-person pose machines." IEEE International Conference on Computer Vision, 2019.

[MEH2018] Mehta, Dushyant, et al. "Single-shot multi-person 3d pose estimation from monocular rgb." IEEE International Conference on 3D Vision, 2018.

Bibliography

[HAN2018] Han Y, Zhang P, Zhuo T, Huang W, Zhang Y. Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recognition Letters. 2018 May 1;107:83-90.

[LIU2018] Liu K, Liu W, Gan C, Tan M, Ma H. T-C3D: temporal convolutional 3d network for real-time action recognition. InThirty-second AAAI conference on artificial intelligence 2018 Apr 27.

[PAP2021] C. Papaioannidis, D. Makrygiannis, I. Mademlis, and I. Pitas, “Learning Fast and Robust Gesture Recognition”, in Proccedings of the European Signal Processing Conference, 2021.

[ZEN2018] Nico Zengeler , Thomas Kopinski and Uwe Handmann “Hand Gesture Recognition in Automotive Human–Machine Interaction Using Depth Cameras”

[HUA2019] Bo Chen, Chunsheng Hua, Decai Li, Yuqing He and Jianda Han “Intelligent Human–UAV Interaction System with Joint Cross-Validation over Action–Gesture Recognition and Scene Understanding”

[RON2015] Ronneberger, Olaf and Fischer, Philipp and Brox, Thomas “U-net: Convolutional networks for biomedical image segmentation” in Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[WAN2020] Wang, Jingdong, et al. “Deep high-resolution representation learning for visual recognition” in IEEE transactions on pattern analysis and machine intelligence, 43, 10, pp. 3349-3364 2020.

[ION2013] Ionescu, Catalin, et al. "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 36, no. 7, pp. 1325-1339, 2013.

[KIL2022] N. Kilis, C. Papaioannidis, I. Mademlis and I. Pitas, "An Efficient Framework for Human Action Recognition Based on Graph Convolutional Networks," in Proceedings of the IEEE International Conference on Image Processing (ICIP), 2022.

[ZHE2019] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Learning the Depths of Moving People by Watching Frozen People,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[COR2016] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Bibliography

[GEI2013] Geiger, Andreas, et al. “Vision meets robotics: The KITTI dataset,” The International Journal of Robotics Research 32, 11, pp. 1231-1237, 2013.

[PAP2021b] C. Papaioannidis, I. Mademlis and I. Pitas, "Autonomous UAV Safety by Visual Human Crowd Detection Using Multi-Task Deep Neural Networks," 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021.

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.

[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).

[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.

[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.

[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

[1] I. Pitas, “Artificial Intelligence Science and Society Part A: Introduction to AI Science and Information Technology“, Amazon/Kindle Direct Publishing, 2022,

https://www.amazon.com/dp/9609156460?ref_=pe_3052080_397514860

[2] I. Pitas, “Artificial Intelligence Science and Society Part B: AI Science, Mind and Humans“, Amazon/Kindle Direct Publishing, 2022, https://www.amazon.com/dp/9609156479?ref_=pe_3052080_397514860

[3] I. Pitas, “Artificial Intelligence Science and Society Part C: AI Science and Society”, Amazon/Kindle Direct Publishing, 2022, https://www.amazon.com/dp/9609156487?ref_=pe_3052080_397514860

[4] I. Pitas, “Artificial Intelligence Science and Society Part D: AI Science and the Environment“, Amazon/Kindle Direct Publishing, 2022,

https://www.amazon.com/dp/9609156495?ref_=pe_3052080_397514860

Artificial Intelligence & Information Analysis Lab