Human-Centered AI for Autonomous Vehicles

C. Papaioannids, I. Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 2.0

Artificial Intelligence & Information Analysis Lab

Contents

- **Human-centered AI**
- Human pose/posture estimation
- Human action/activity recognition
- Human gesture recognition
- Semantic image segmentation
- Applications

Human-centered AI

- ***Autonomous vehicles*** (self-driving cars, UAVs) have been increasingly employed to assist humans in real-world applications.
  - Autonomous transportation.
  - Infrastructure inspection.
  - Natural disaster management.
- ***Human-Vehicle Interaction***: Autonomous vehicles should understand and interact with humans.
  - Special case of human-robot interaction.

Human-centered AI

- To that end, autonomous vehicles must be equipped with advanced ***visual and aural perception*** systems and human-centered AI algorithms.
- These systems/algorithms should demonstrate:
  - increased ***perception accuracy***,
  - ***robustness*** to input data variations and ***attacks***,
  - produce ***timely HRI state and action estimations*** to ensure ***safety***.

Human-centered AI

Deep Neural Networks (DNNs) in particular:

- ***Convolutional Neural Networks*** (CNNs) and
- ***Attention/transformer networks***

have been widely used to build such advanced systems.

- Main tasks:
  - Human pose/posture estimation.
  - Human action/activity recognition.
  - Human gesture recognition.
  - Contextual (in-cabin and exterior scene) ***human understanding***.

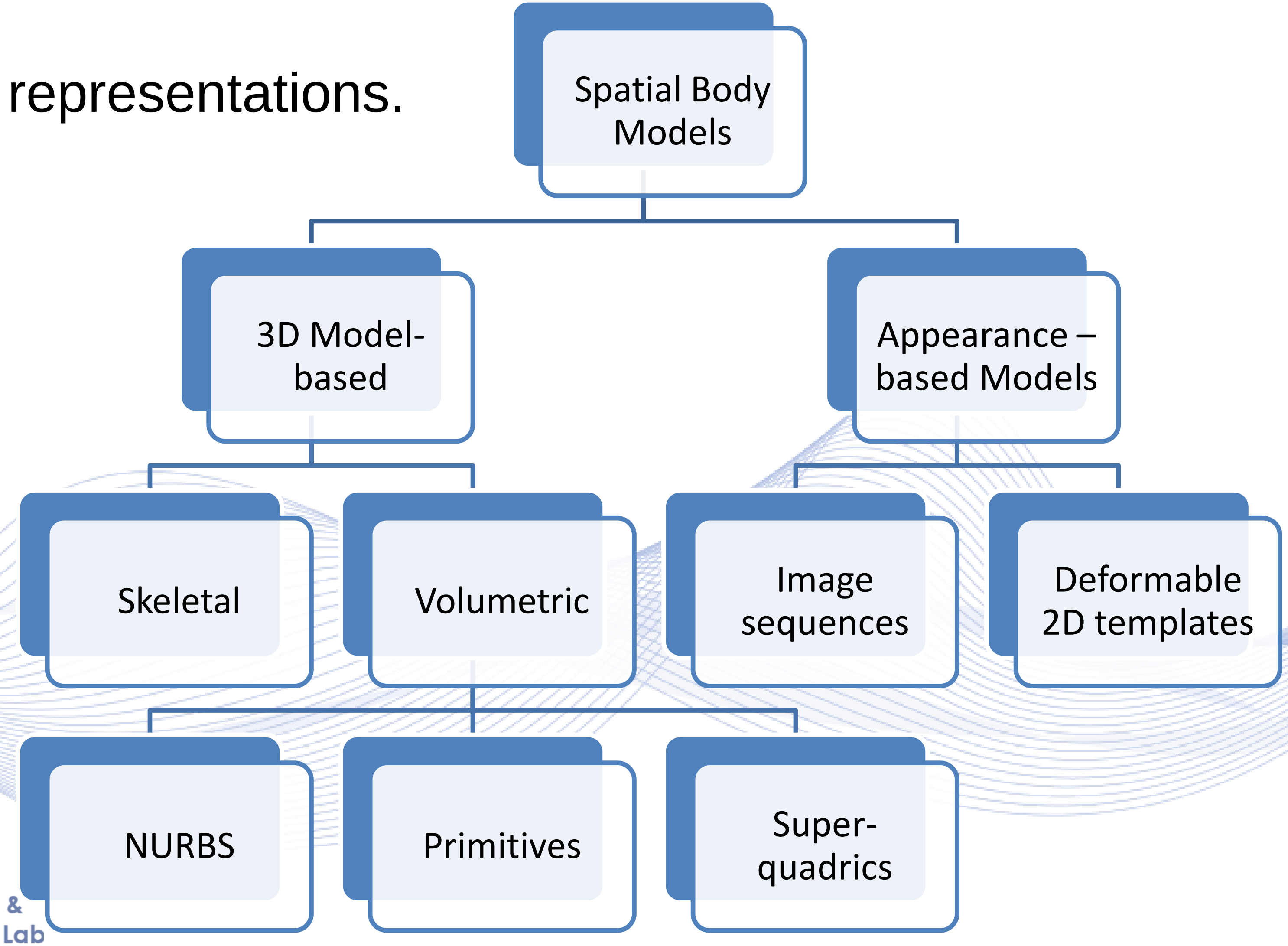
Human-centered AI

Human body representations.

- ***Appearance-based***: features are obtained directly from images or videos.
  - ***Video-based***: Analyze a video frame sequence to recognize the depicted human gestures.
- ***3D model-based***: human body represented by a human model, e.g., 3D mesh model: a list of vertices and lines.
  - ***Skeletal-based***: human body represented by 2D/3D human skeletons → more compact representation than 3D mesh.

Human-centered AI

Human body representations.



Artificial Intelligence & Information Analysis Lab

Human-centered AI

The human body anatomic/kinematic modeling allows its representation of 2D and 3D human poses as graphs:

- The body joints and bones are the graph nodes and the edges.
- ***Human body graph***: $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of K body joints/nodes and $\mathcal{E}$ is a set of B bones/edges.
- Human body graph can have various detail levels.

Human-centered AI

RGB data

- ***Are easy to obtain, massive datasets available.***
- Can be used as basis for neural feature extraction.
- Depth can be regressed from monocular video.

RGB data

- ***Do not protect user privacy.***
- Skeletons extracted from color images are of lower accuracy.

Human-centered AI

Depth data

- Depth images/videos
- ***Protect user privacy.***
- Highly accurate 3D skeletons can be extracted.

Depth data

- ***Difficult to obtain***, depth cameras are more expensive / difficult for outdoor environments.
- SoA CNNs mostly use RGB data.

Human-centered AI

Wearable sensor data

- Passive body joint/part tagging
- Magnetic field trackers,
- body suits,
- Instrumented gloves (active or passive).
- ***Good for skeleton-based analysis.***

Wearable sensor data

- ***Difficult to obtain.***
- Wearable sensors are intrusive and may obstruct body motion.

Contents

- Human-centered AI
- **Human pose/posture estimation**
- Human action/activity recognition
- Human gesture recognition
- Semantic image segmentation
- Applications

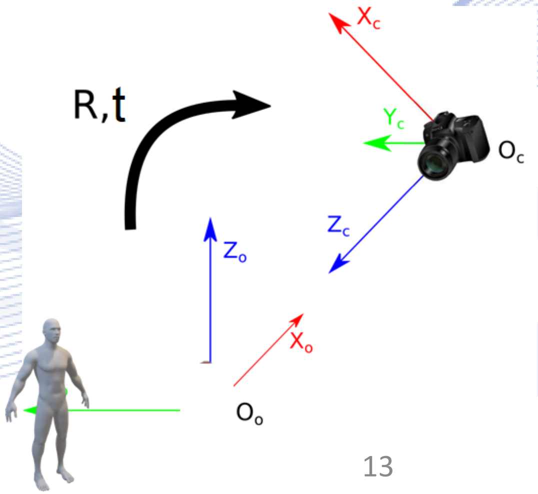
Human pose estimation

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 **R** and

& translation **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

Artificial Intelligence & Information Analysis Lab

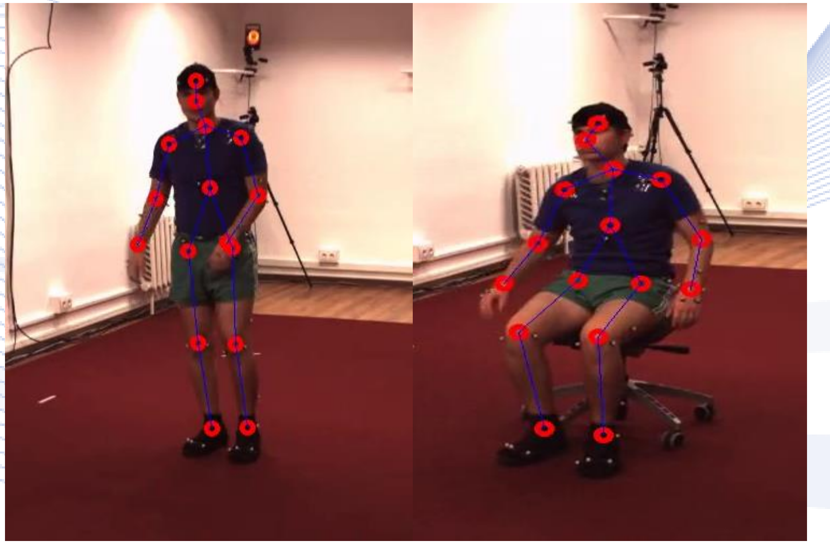
Human pose estimation

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 $\mathbf{c}$:

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

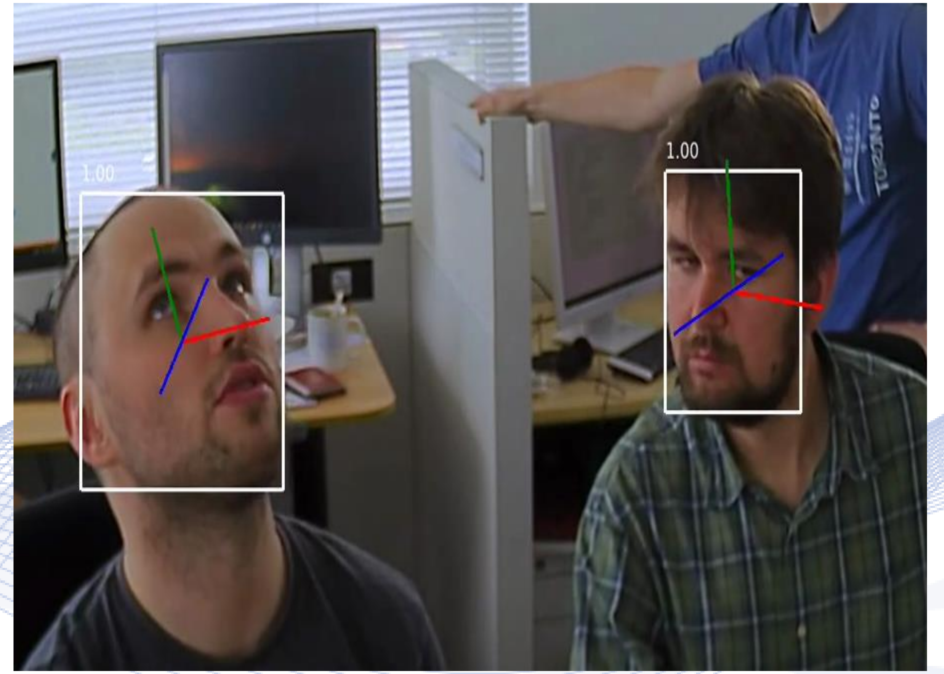
- Applications:
  - human-robot interaction (HRI),
  - sign language communication,
  - physical and rehabilitation training.



Standing Sitting [ION2013].

Artificial Intelligence & Information Analysis Lab

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***.

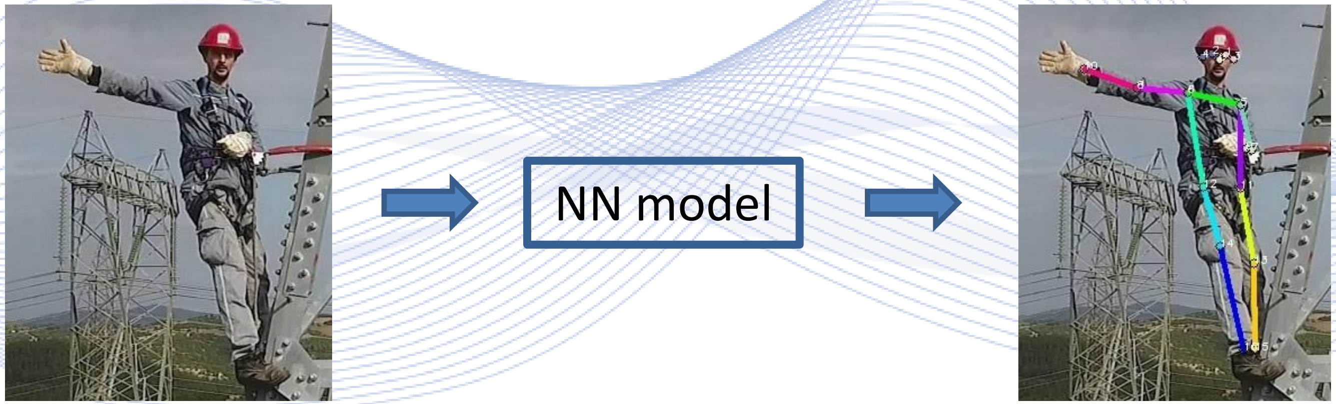
- Single-person 2D HPE:
  - direct regression methods,
  - heatmap-based methods.
- Multi-person 2D HPE:
  - top-down approach,
  - bottom-up approach.

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):

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

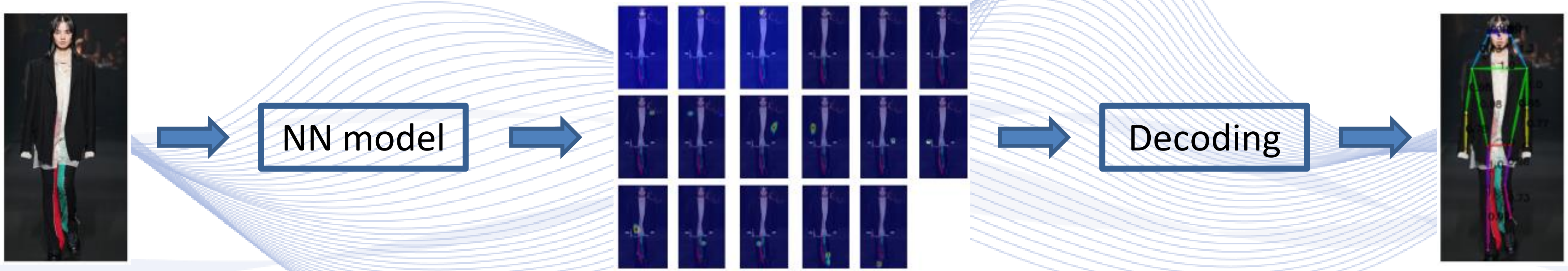
- $\mathbf{p} = [\mathbf{j}_1^T, \mathbf{j}_2^T, \dots, \mathbf{j}_K^T]^T$: pre-defined set of body joints that constitute the 2D human pose,
- K is the number of the body joints,
- $\mathbf{j}_k = [x_k, y_k]^T \in \mathbb{N}^2, k = 1, \dots, K$ human skeleton joint representation in pixel coordinates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

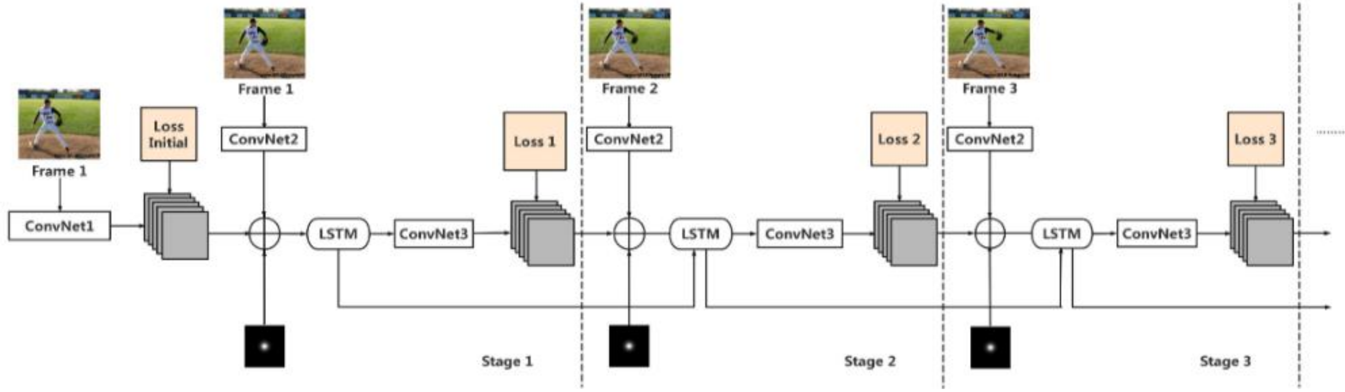
- Heatmaps provide richer supervision information, by preserving the spatial location information.
- Allow using the powerful ***Convolutional Neural Networks*** (CNNs).
- Facilitate DNN/CNN training.
- Used in state-of-the-art 2D HPE approaches.

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].

Artificial Intelligence & Information Analysis Lab

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.

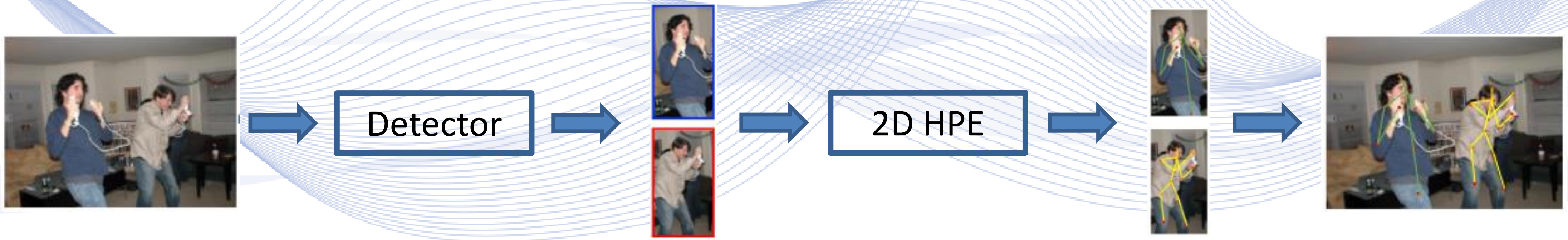
[CAO2017]

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.



[DAN2019]

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]

Artificial Intelligence & Information Analysis Lab

3D human pose estimation

- Predicts the body joint locations in 3D space.
- Provides 3D structure information related to human body.
- It remains a challenging task.
- ***3D pose annotation*** for ground-truth creation is costly and time-consuming.
- ***Limited availability of datasets***:
  - Generalization issues.
  - Problems in real-world applications.

3D human pose estimation

3D HPE from monocular images/videos

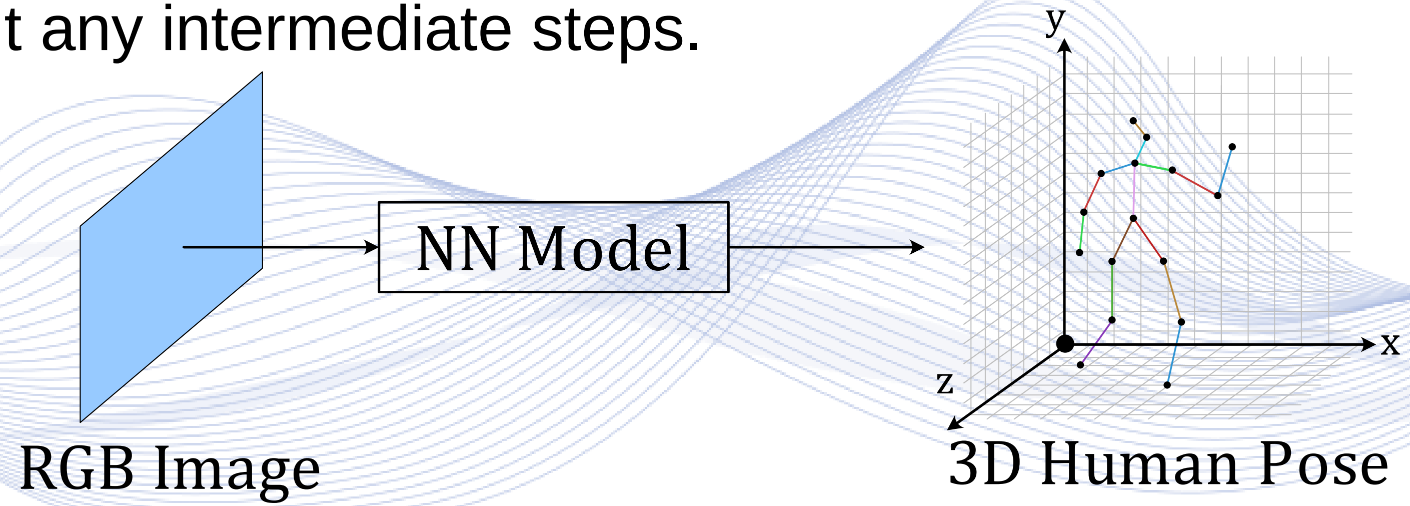
- 3D HPE from monocular images/videos is the most popular approach.
- One monocular RGB camera is required.
- 3D HPE in this setting is very challenging due to:
  - occlusions,
  - depth ambiguities,
  - insufficient data,
  - ***different 3D human poses can be projected to similar 2D poses***.

3D human pose estimation

3D HPE from monocular images

Single-person

- Direct ***3D skeleton regression*** (estimation) from an RGB image: The 3D human pose is obtained directly from the input image without any intermediate steps.



3D human pose estimation

3D HPE from monocular images

Single-person

- Methods based on CNNs.
- If $\mathbf{I}$ is an input RGB image of resolution $M \times N$ and f is the 3D HPE CNN, direct 3D skeleton estimation methods aim to predict:

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

- $\mathbf{P} = [\mathbf{J}_1^T, \mathbf{J}_2^T, \dots, \mathbf{J}_K^T]^T$ is the set of 3D skeleton body joints,
- K is the number of the body joints
- $\mathbf{J}_k = [X_k, Y_k, Z_k]^T \in \mathbb{R}^3, k = 1, \dots, K$ represents the 3D coordinates of each 3D human body.

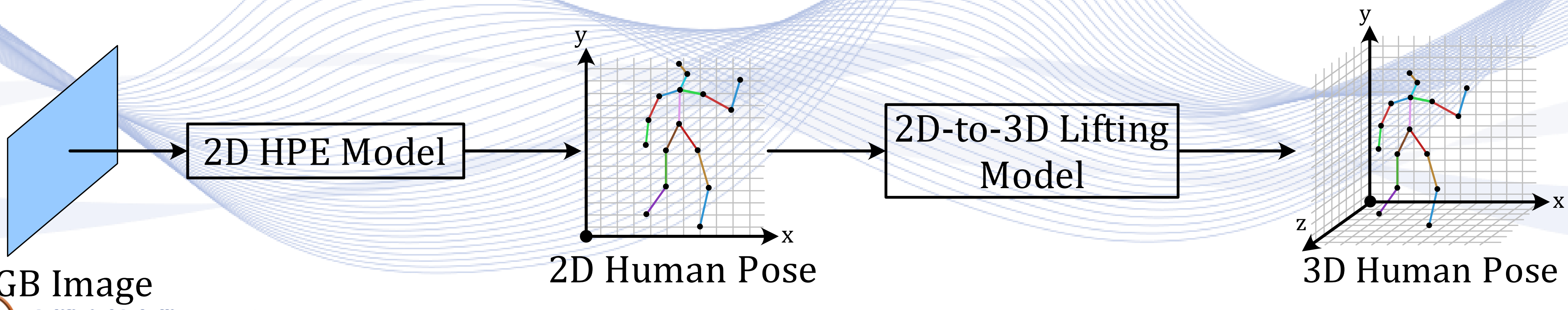
Artificial Intelligence & Information Analysis Lab

3D human pose estimation

3D HPE from monocular images

Single-person

- ***2D-to-3D lifting***: A 2D skeleton is first extracted from the input RGB image, which is then lifted to the corresponding 3D skeleton.
- 2D-to-3D lifting to be performed using ***Graph Convolutional Networks*** (GCNs).

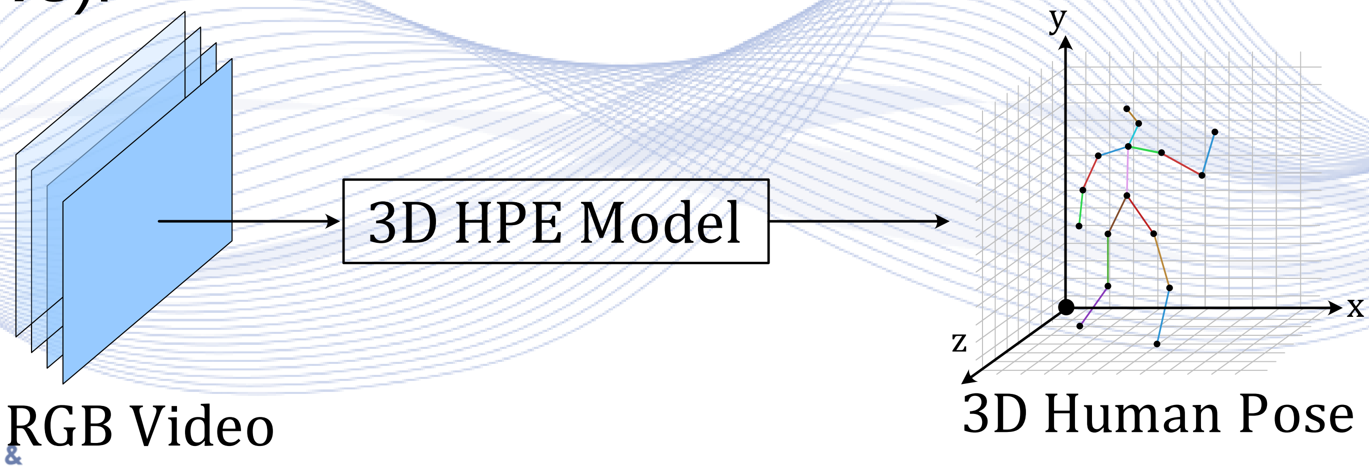


3D human pose estimation

3D HPE from monocular videos

Single-person

- Videos provide temporal information, which can improve the accuracy and the robustness of 3D HPE.
- Use of local temporal video frame neighborhood information (***3D tensors***).



3D human pose estimation

3D HPE from monocular videos

Single-person

- The temporal information of a video can be exploited by a model capable of handling sequential data, such as ***RNNs*** or ***LSTM network***.
- Occlusions or ambiguities on a single frame can be alleviated by additional information provided by neighbouring frames.
- Video-based approaches:
  - LSTM-based [HOS2018],
  - GCN-based [CAI2019],
  - Transformer-based [LI2022].

Artificial Intelligence &
Information Analysis Lab

3D human pose estimation

3D HPE from monocular images

Multi-person

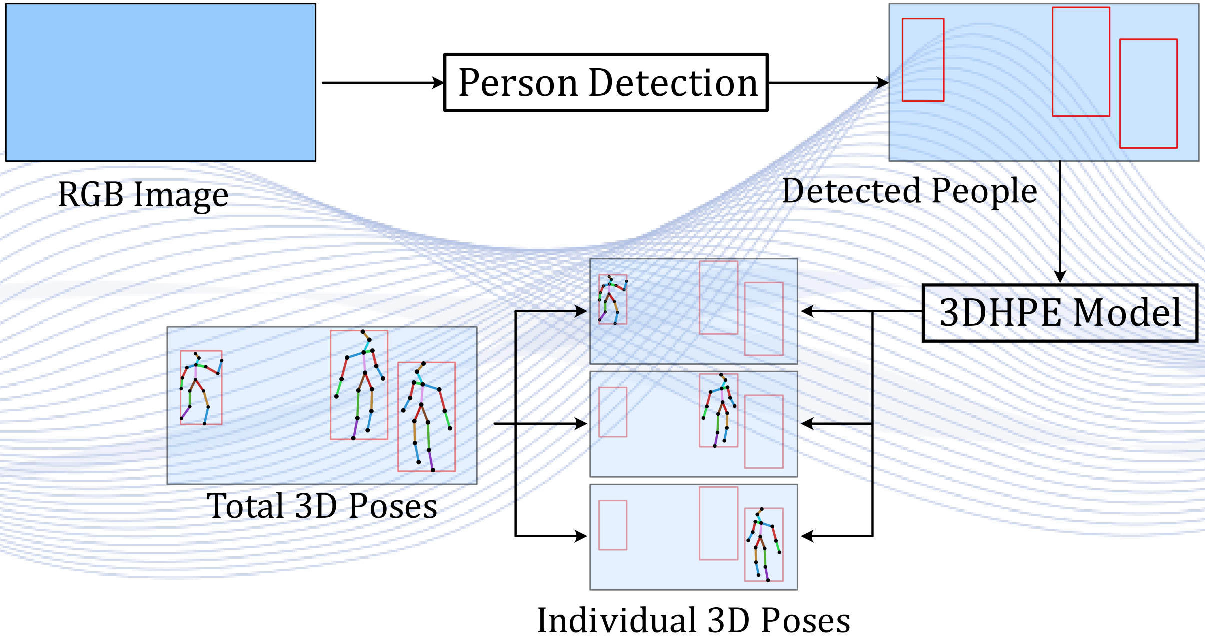
- Estimate the 3D skeletons of multiple persons in an input image.
- ***Top-down pipeline***: Similar to the 2D HPE case,
  - each person is first detected on the input image and
  - individual 3D skeletons are then estimated.
- ***Bottom-up pipeline***:
  - First predict all body joints and depth maps and then
  - group and associate all detected body parts to each person.

3D human pose estimation

3D HPE from monocular images

Multi-person

- Top-down pipeline.



3D human pose estimation

3D HPE from monocular images

Multi-person

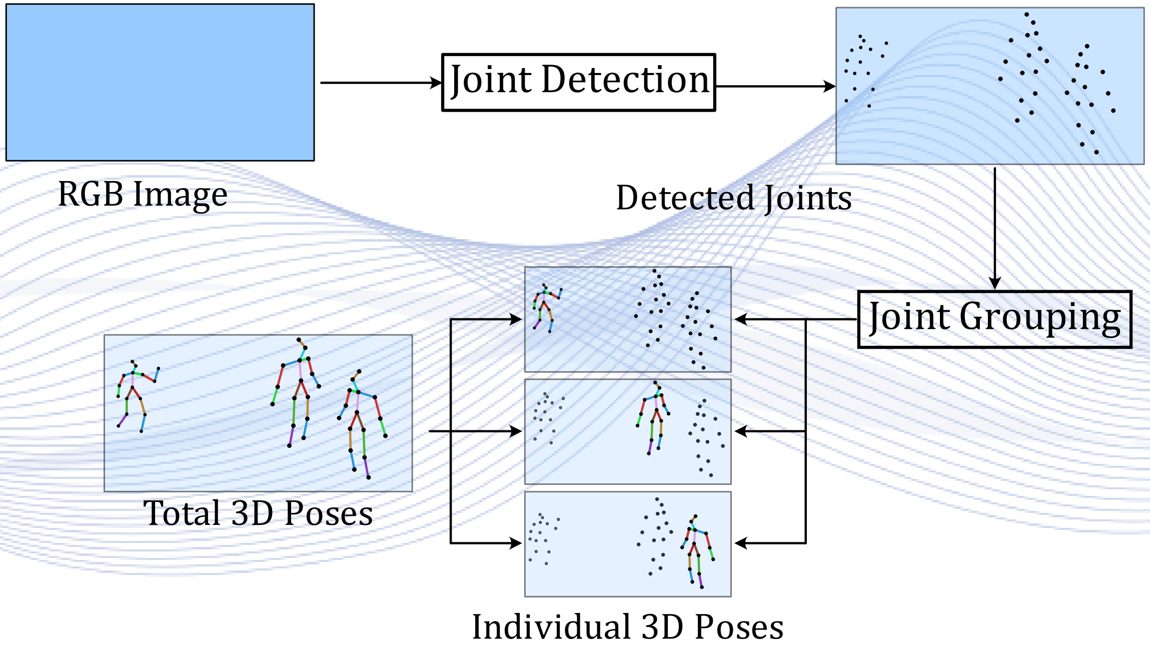
- ***Top-down pipeline***:
  - It achieves promising results.
  - Human mesh reconstruction is straightforward.
  - ***Computations increase linearly with the person number.***
  - ***Global scene information is lost***, since a detection step is first applied.
  - Popular approaches:
    - LCR-Net [ROG2017], LCR-Net++ [ROG2019], PandaNet [BEN2020].

3D human pose estimation

3D HPE from monocular images

Multi-person

- Bottom-up pipeline.



3D human pose estimation

3D HPE from monocular images

Multi-person

- ***Bottom-up pipeline***:
  - ***Faster execution speed.***
  - Human mesh reconstruction is not straightforward.
  - ***Body joint grouping is challenging.***
  - Occlusions can cause inaccurate predictions.
  - Popular approaches:
    - Single-stage multi-person Pose Machine [NIE2019],
    - Occlusion-Robust Pose-Maps (ORPM) [MEH2018].

Contents

- Human-centered AI
- Human pose/posture estimation
- **Human action/activity recognition**
- Human gesture recognition
- Semantic image segmentation
- 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

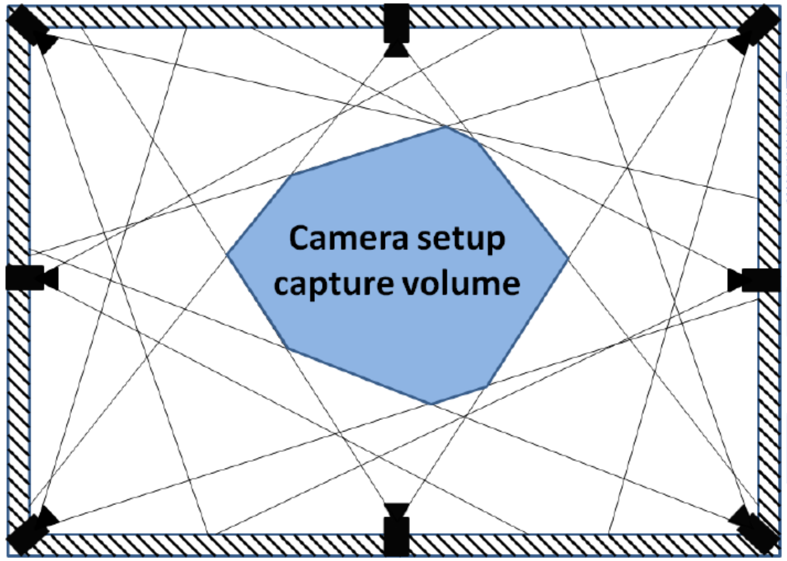

sit


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 setup ($N_C = 8$).

Neural HAR

- 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.

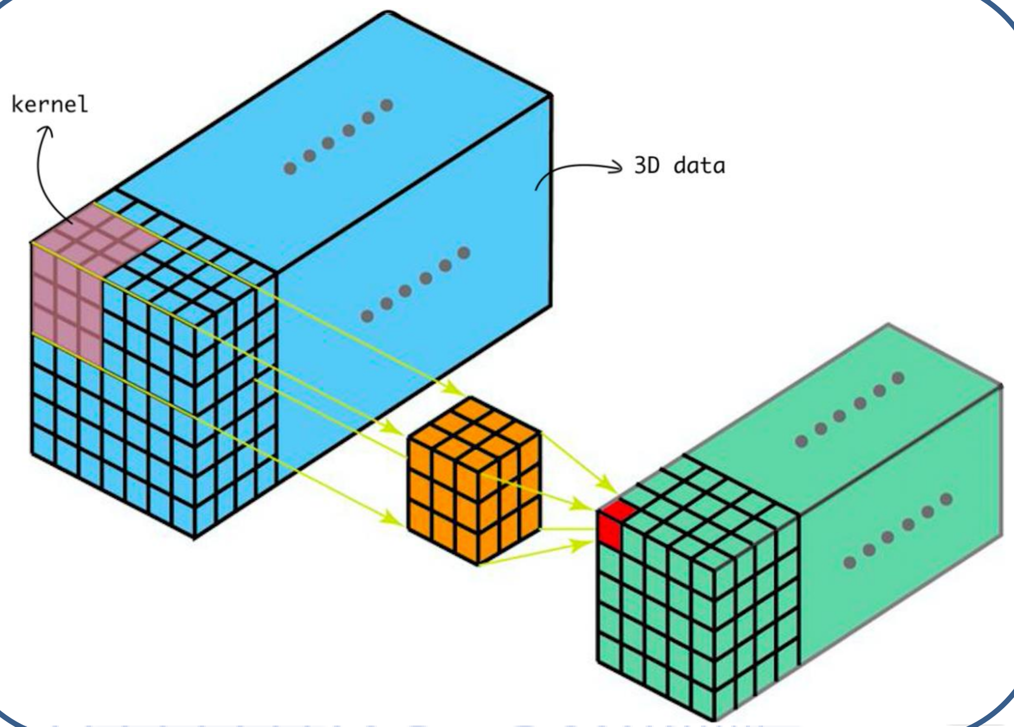


image from https://towardsdatascience.com/understanding-1d-and-3d-convolution-neural-network-keras-9d8f76e29610

HAR with 3D CNNs

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

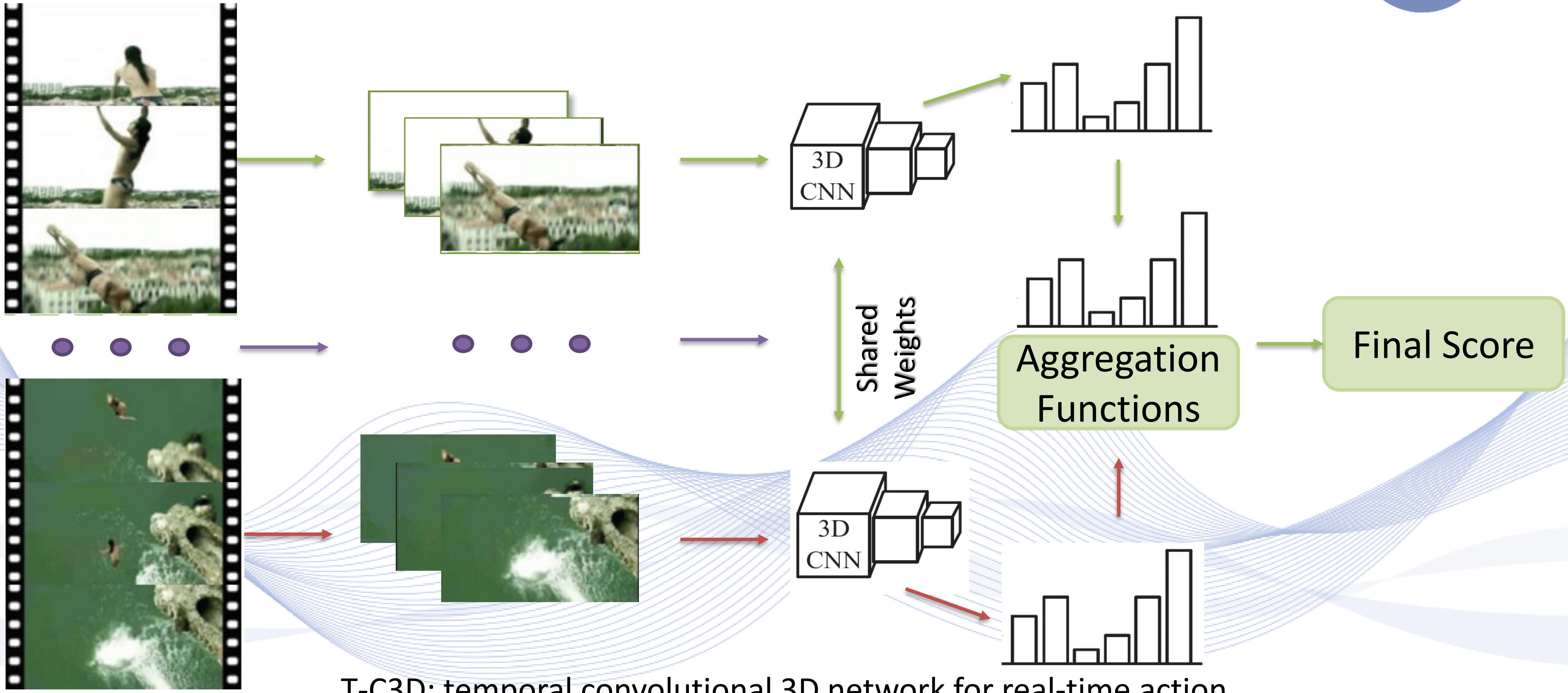
Objective:

- 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].

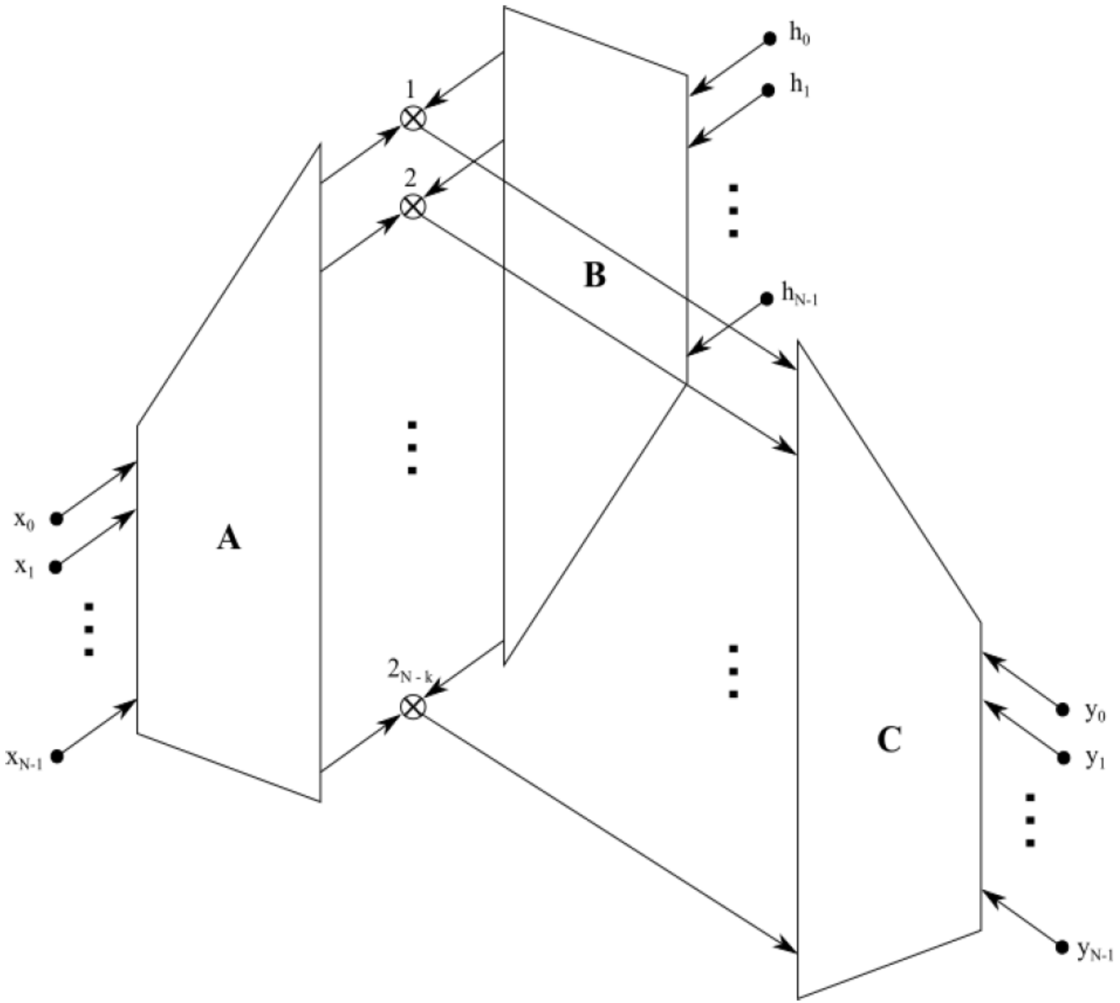
HAR with 3D CNNs

3D convolutions are notoriously computationally expensive.

- ***Fast 3D convolution algorithms***:

$$\mathbf{y} = \mathbf{C}(\mathbf{Ax} \otimes \mathbf{Bh}).$$

- GEneral Matrix Multiplication (GEMM) BLAS or cuBLAS routines can be used.



HAR with multi-stream DNNs

- Multi-stream networks are implemented using model architectures (e.g., CNNs for image classification tasks) which are trained separately.
- Their softmax scores are combined by late fusion considering different fusion methods, such as averaging or training multi-class classifiers (e.g. SVM) on stacked L_2-normalized softmax scores as features.

Human visual cortex contains two pathways:

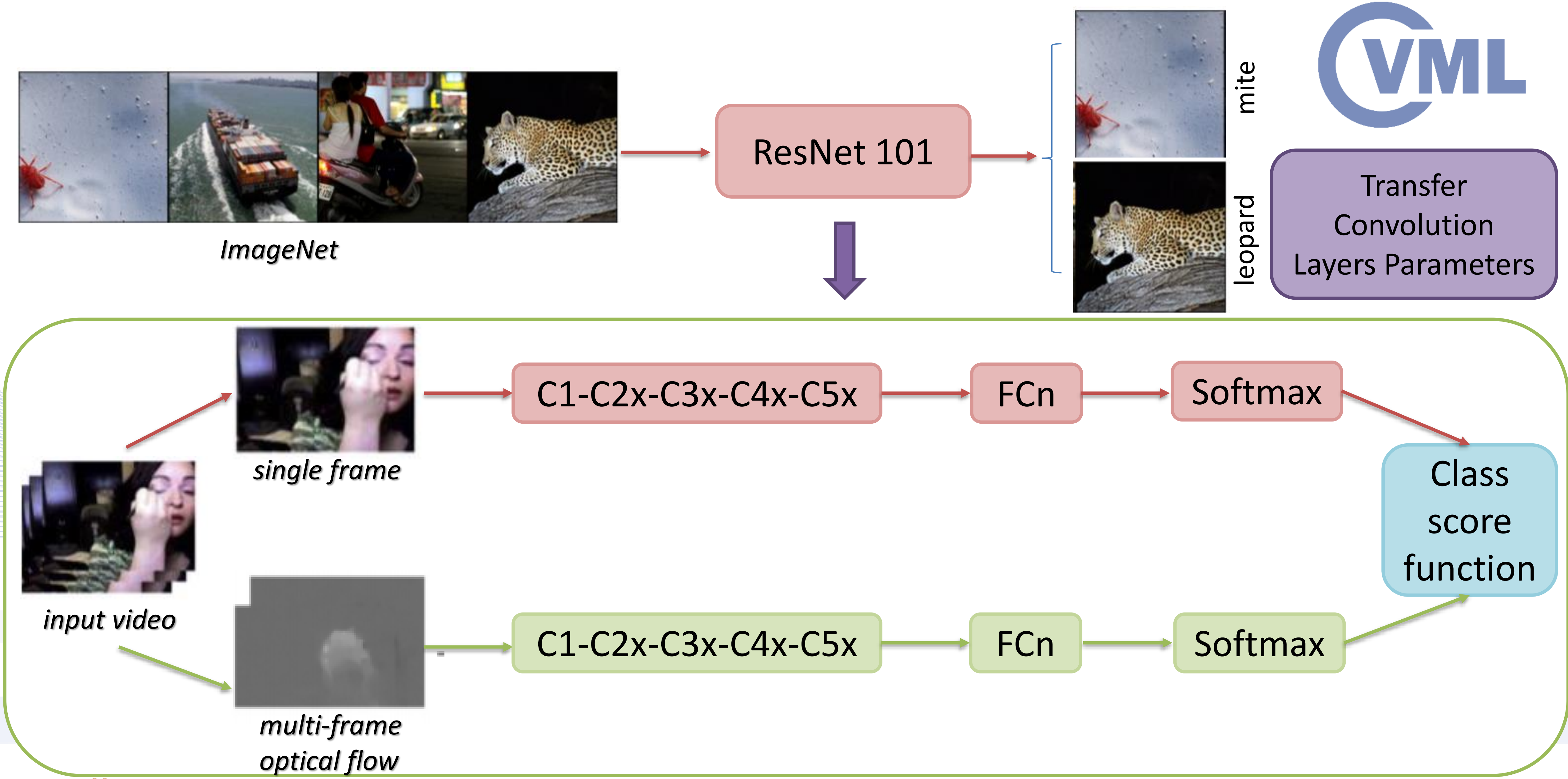
1. the ***ventral stream*** (which performs object recognition),
2. the ***dorsal stream*** (which recognizes motion).

First stream: ***spatial stream*** performs object recognition on still images.

Second stream: ***temporal stream*** conveys motion information using features like optical flow.

HAR with multi-stream DNNs

- A ***two-stream network*** architecture is capable to manage both spatial and temporal information [HAN2018].
  - Pretraining on ImageNet dataset to overcome over-fitting.
  - Deeper CNN architectures can model challenging datasets more efficiently.
- ***Experiments & Accuracy***
  - Increased accuracy on publicly available datasets (93-95% accuracy).



ImageNet
ResNet 101
mite
leopard
Transfer
Convolution
Layers Parameters
single frame
input video
multi-frame
optical flow
C1-C2x-C3x-C4x-C5x
FCn
Softmax
C1-C2x-C3x-C4x-C5x
FCn
Softmax
Class
score
function

[HAN2018].

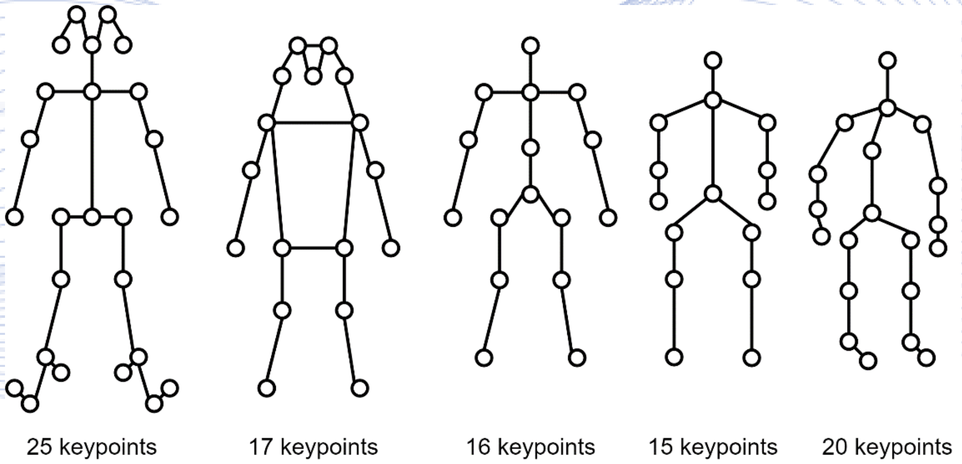
Skeleton-based HAR

Methodology

- Use a ***human pose estimator*** to extract 2D/3D skeletons of humans in each frame.
- Collect extracted 2D/3D skeletons to form features volumes.
- This fixed-size representation for an entire video clip is suitable to classify actions using shallow networks (DNNs, CNNs, LSTMs, GCNs, Transformers).

Skeleton-based HAR with GCNs

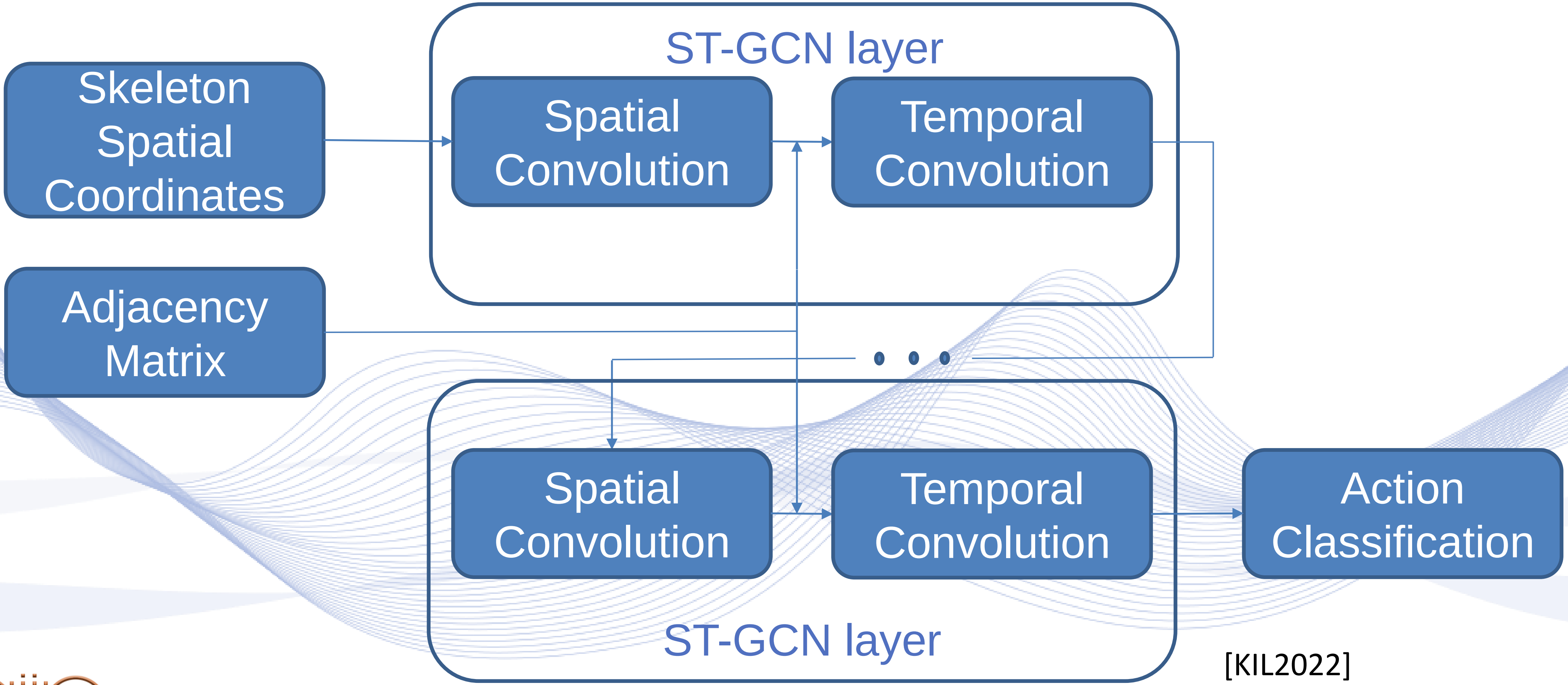
- Human skeleton:
  - Keypoints: Nodes in the Graph,
  - Connections: Edges in the Graph.
- Representation with 3D skeltongraphs:
  - ***Invariant to viewpoint and appearance***.



Skeleton-based HAR with GCNs

- ***Spatial Convolution block***:
  - Sums the values of all channels and gives us a single value for each node.
  - Multiplication with adjacency matrix creates graphical connections for each frame.
- ***Temporal Convolution block***:
  - Uses a temporal kernel $[t_1 \times 1]$ over the frames and extracts the temporal features for each node.
- These two blocks compose the ***ST-GCN layer***.
- Several ST-GCN layers compose the ***ST-GCN model***.

Skeleton-based HAR with GCNs



[KIL2022]

Contents

- Human-centered AI
- Human pose/posture estimation
- Human action/activity recognition
- **Human gesture recognition**
- Semantic image segmentation
- 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.

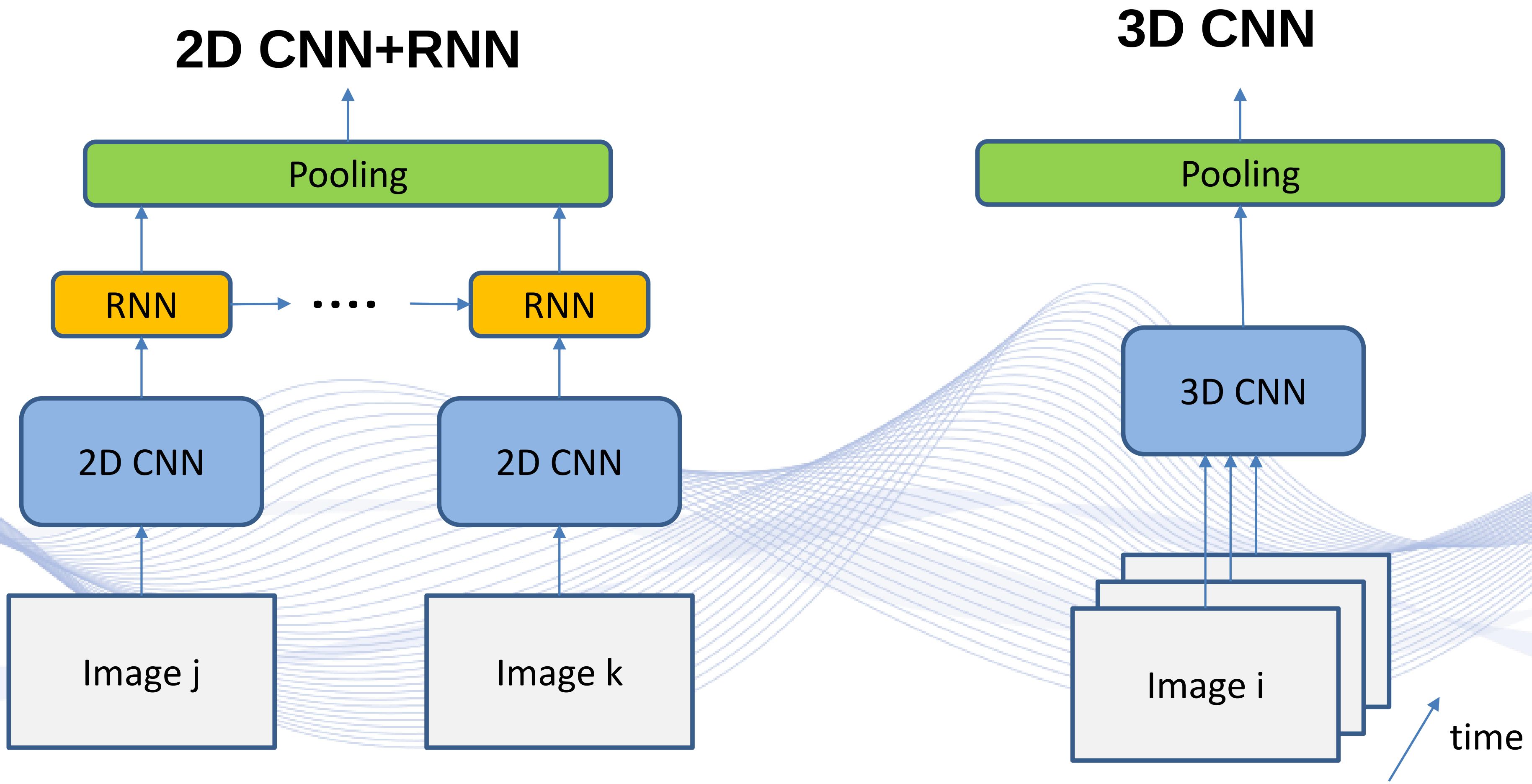
Gesture recognition

- ***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.***

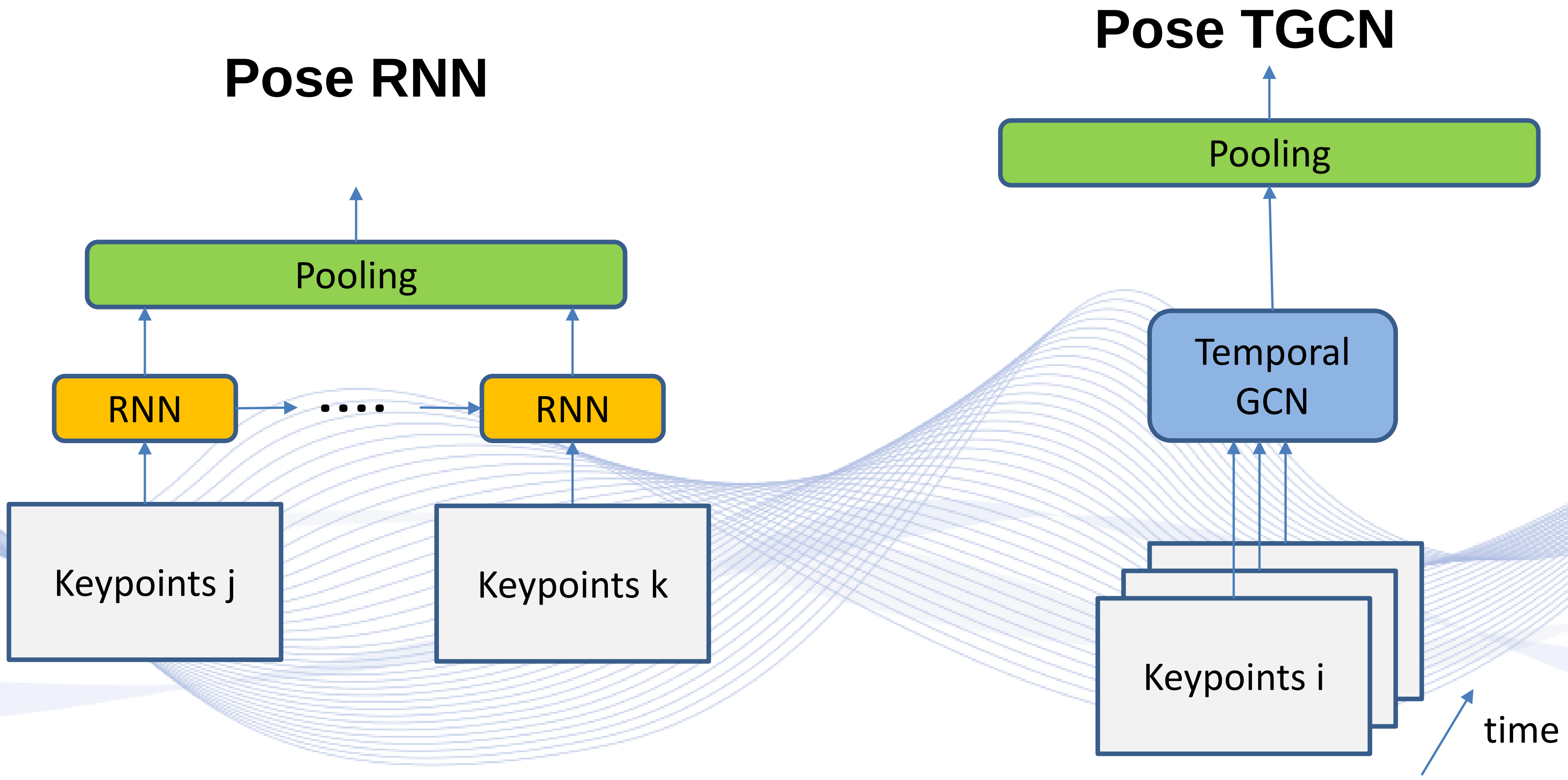
DNN architectures for gesture recognition

- 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



DNN architectures for gesture recognition



Keypoints are the joints of human bodies.

Contents

- Human-centered AI
- Human pose/posture estimation
- Human action/activity recognition
- Human gesture recognition
- **Semantic image segmentation**
- Applications

Semantic image segmentation

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

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

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

$$P(\mathcal{R}_i) = \text{TRUE}, \qquad \text{for } i = 1,2, \ldots, N,$$

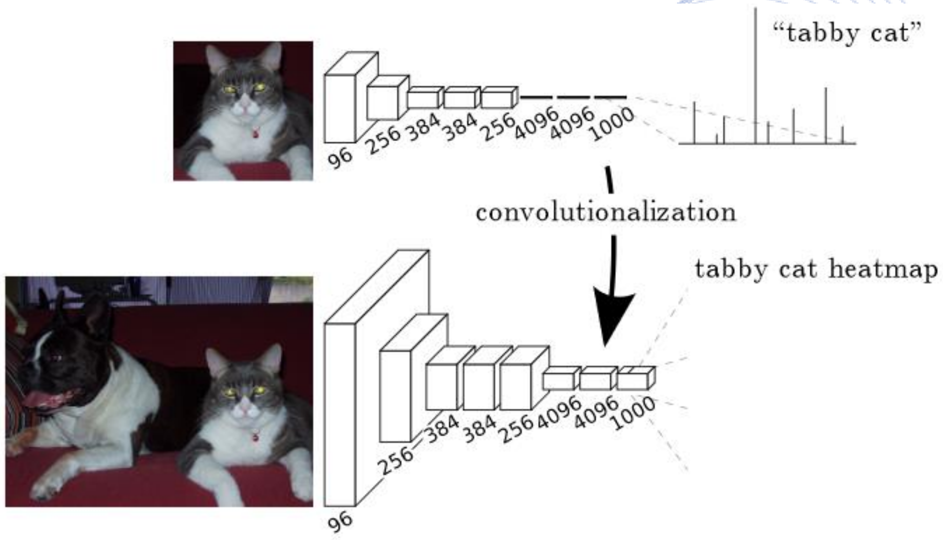
$$P\left(\mathcal{R}_i \cup \mathcal{R}_j\right) = \text{FALSE}, \qquad \text{for } i \neq j.$$

- $P(\mathcal{R})$: logical predicate.

Artificial Intelligence & Information Analysis Lab

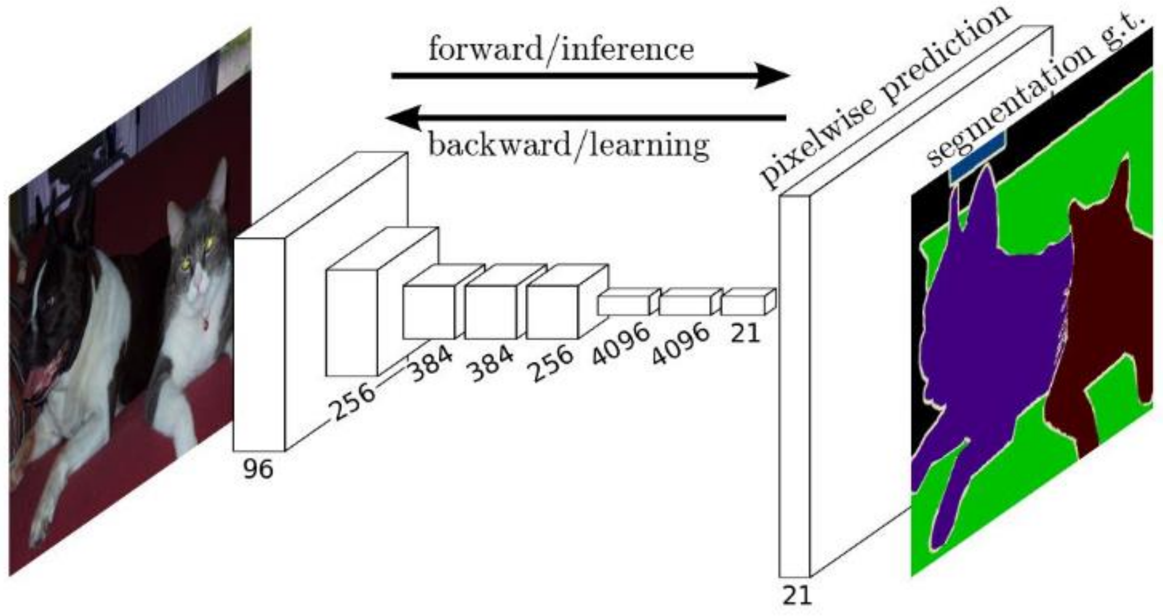
DNN-based semantic image segmentation

- Transforming the fully connected layers of image classification networks into convolution layers enables the transformed network to output heatmaps.
- End-to-end dense prediction learning is possible by adding extra layers.



DNN-based semantic image segmentation

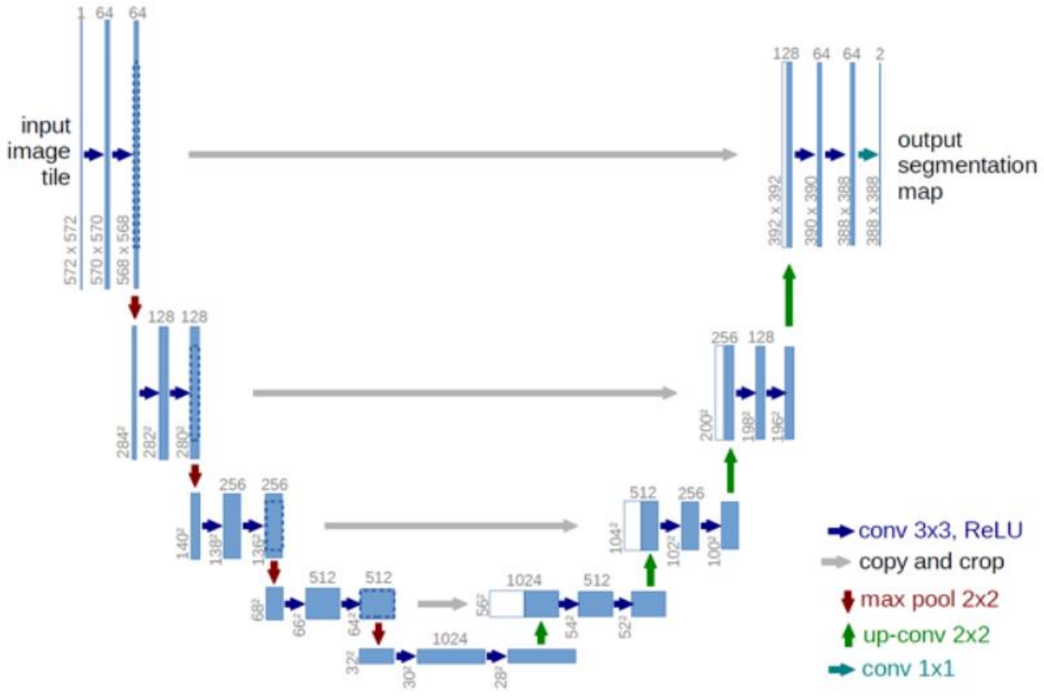
- Fully convolutional networks (FCNs) with encoder-decoder architecture for semantic image segmentation.



DNN-based semantic image segmentation

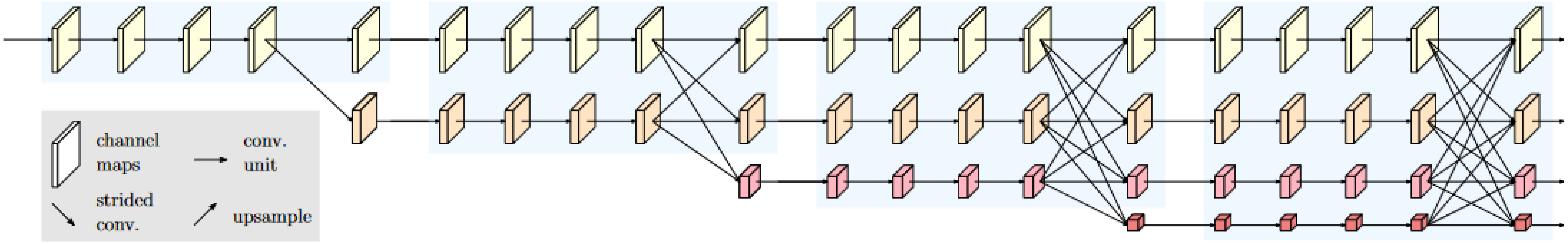
- Encoder radically reduces resolution inputs → decoder fails to produce fine-grained segmentations.
- Improvements:
  - Skip connections.
  - U-shaped network architecture (e.g., U-Net [RON2015]).
    - Multiple skip connections to maintain information from high-resolution feature maps.
  - High-resolution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

DNN-based semantic image segmentation



U-Net network architecture [RON2015].

DNN-based semantic image segmentation



High-resolution image segmentation networks [WAN2020].

DNN-based semantic image segmentation

Person instance segmentation.

Scene segmentation [COR2016].

Artificial Intelligence & Information Analysis Lab

DNN-based semantic image segmentation

Crowd detection via image segmentation.

- Avoid detected crowds to ensure safety.

Depth Estimation

- 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

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 and condition.
- Road scene understanding.
- ***Gesture-based vehicle control.***

Autonomous driving

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

Pedestrian intention recognition [PAP2022].

Autonomous driving

- Scene understanding.

[COR2016]

[GEI2013]

Road scene segmentation and depth estimation.

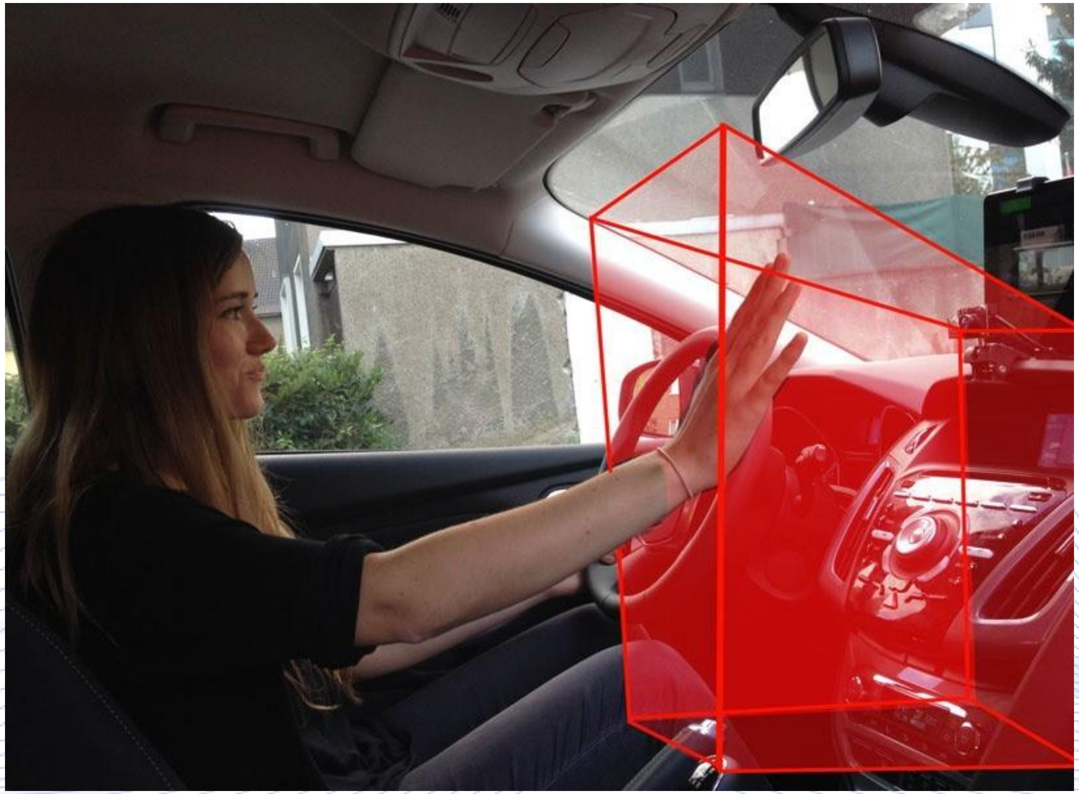
Autonomous vehicle control

Human–vehicle interaction via gestures.

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

- ***Gesture-based vehicle command language.***
  - Can interaction be done based on spontaneous gestures?

Autonomous vehicle control



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

Autonomous vehicle control

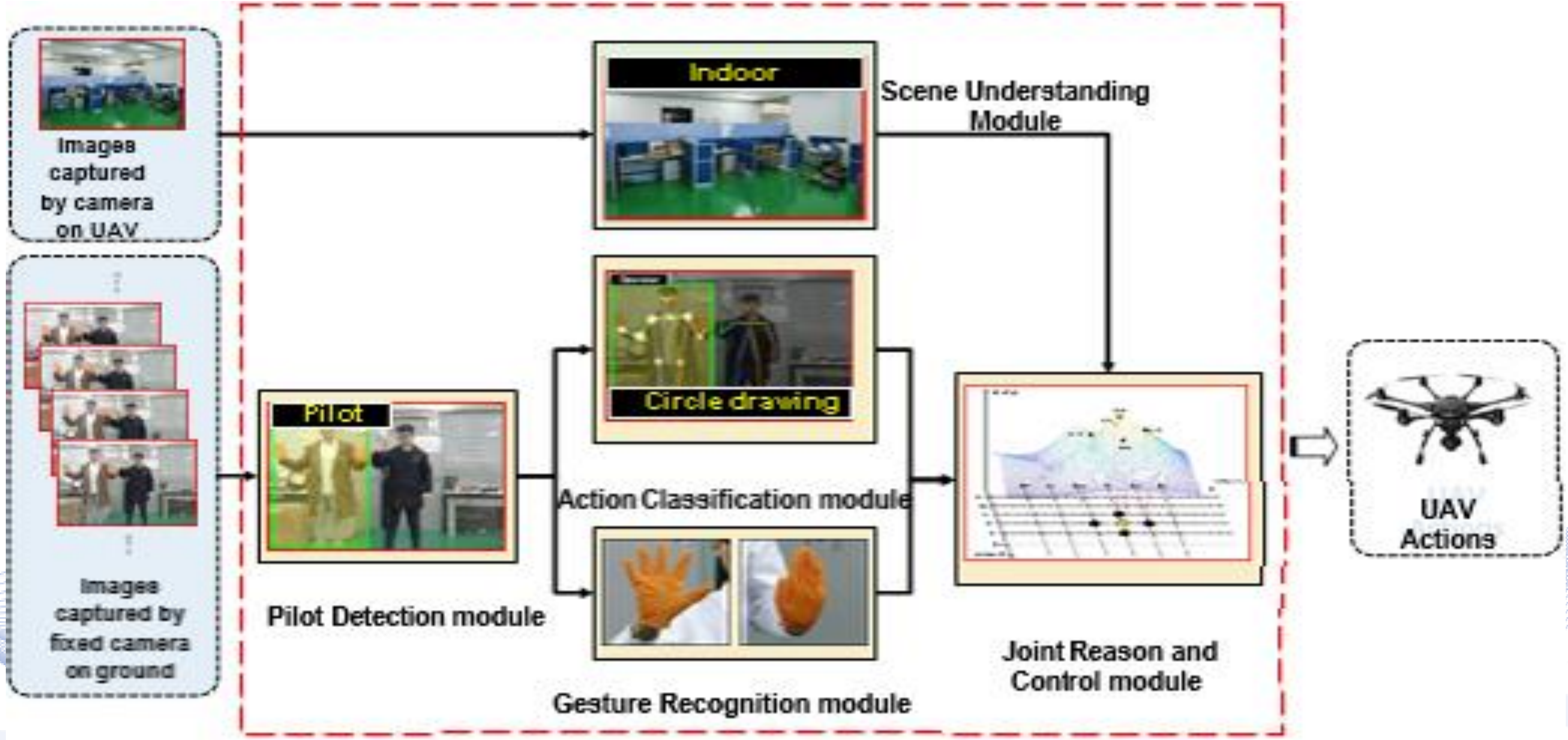
Lane change with gesture control [ZEN2018].

Autonomous vehicle control

Gesture-controlled Drones

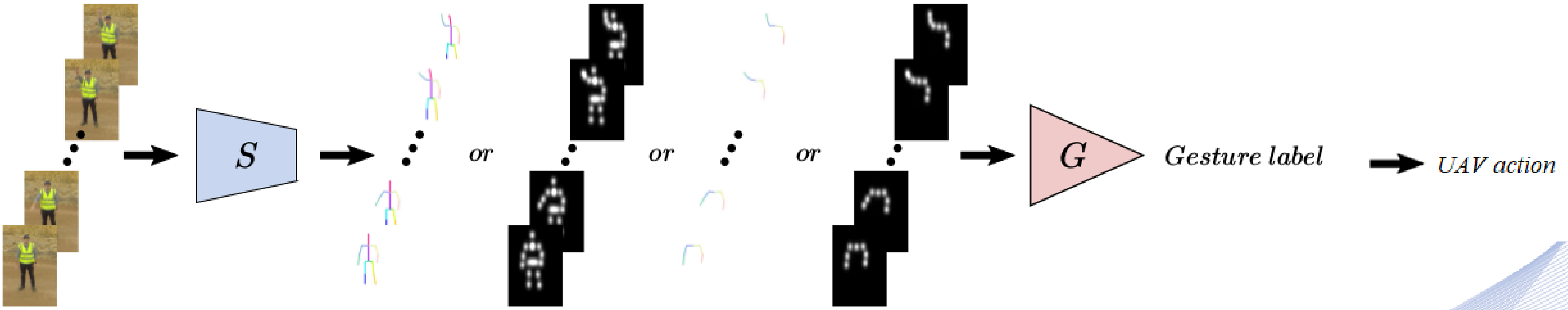
- Video stream is recorded through the camera and segmented into sequences of images.
- Each image is then recognized by a classification process.
- Typical commands:
  - Take off.
  - Land.
  - Move right or left.
- Finally, the action planner on the drone.

Autonomous vehicle control



Human-Drone Interaction model [HUA2019].

Autonomous vehicle control



Gesture recognition for Human-Drone Interaction [PAP2021].

Artificial Intelligence & Information Analysis Lab

Autonomous vehicle control

Crowd detection for autonomous UAV navigation

[PAP2021b].

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

Acknowledgements

- This lecture has received funding from the European Union's European Union Horizon 2020 research and innovation programme under grant agreement 871479 (AerialCore).

- This publication reflects only the authors' views. The European Commission is not responsible for any use that may be made of the information it contains.

Artificial Intelligence & Information Analysis Lab

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