UniGe | MaLGa

On the multi-level nature of human motion analysis

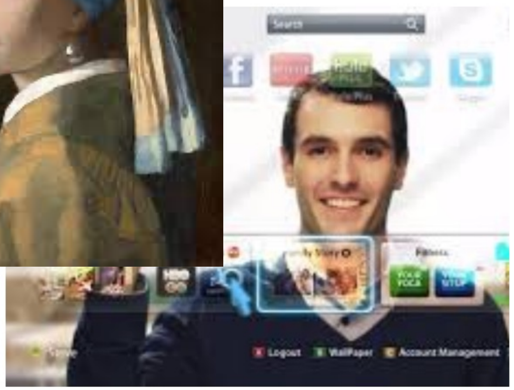
Francesca Odone

Joint work with P. Alfano, M. Casadio, E. De Vito, F. Figari Tomenotti, G. Goyal, M. Moro, I. Mouawad, E. Nicora, N. Noceti, A. Sciutti, A. Vignolo …

SSVM 2023

the complexity
of human motion

head



hands

upper body

full body

Introduction

- **Human motion analysis** touches on aspects that have an intrinsic **multi-level** nature.

- At a *low level*, there is a need to develop algorithms for estimating the flow field and for detecting features with specific dynamic and semantic characteristics.
- At a *medium level*, it is necessary to conceive models to integrate information on wider time intervals and spatial regions: semantic segmentation, feature tracking and motion primitives detection
- As we rise to a *high level* we face tasks of action/activity recognition and anticipation.

A multidisciplinary research area

As a research field human motion analysis is highly multidisciplinary as it involves

- (multi-resolution) signal processing
- Computer vision
- Machine learning

But also

- Cognitive science / developmental learning
- Biomedical engineering and motor learning

Humans perceiving human motion

Developmental science tells us about our ability in perceiving human motion since the early stages of development

Devising computational models to emulate/imitate this ability would provide scientists with

- further means of understanding
- artificial perception tools

Kinematic primitives in action similarity judgments: A human-centered computational model
V Air et al IEEE TCDS 2023

Insights from cognitive science

Cognitive science

- **Early after birth**: Perception of the presence of potential interacting agents
- **Some months**: Understanding of motion properties
- **Around third year**: Full action recognition and reaction, if needed

Computational science

- Understanding the presence of biological motion
- Understanding motion qualities, properties, affinities
- Classical action recognition

Overview of our research

Multi-level or multi-scale video analysis?

- We are interested in deriving ”semantic” information, as we are focusing on a specific class of dynamic events: *human motion with its kinematic rules*

- So far, we have addressed the different levels individually, inspired by biological motivations
- End-to-end models are a possible direction, as long as we read in-between results useful in several applications

Challenges we are addressing

But also a table of contents for this presentation

LOW LEVEL

HIGH LEVEL

- Detect **space-time keypoints** and compute **low-level motion patterns**
- Analyse their **evolution over time** and **detect motion primitives** (e.g, gait)
- Integrate them in space (and time) with **graph-based representations**
- **Distil higher-level information** in heading estimation, action recognition

- In the process we need to deal with limited resources (training/test time, few data, few labels)

Low level image and video analysis

Low level analysis

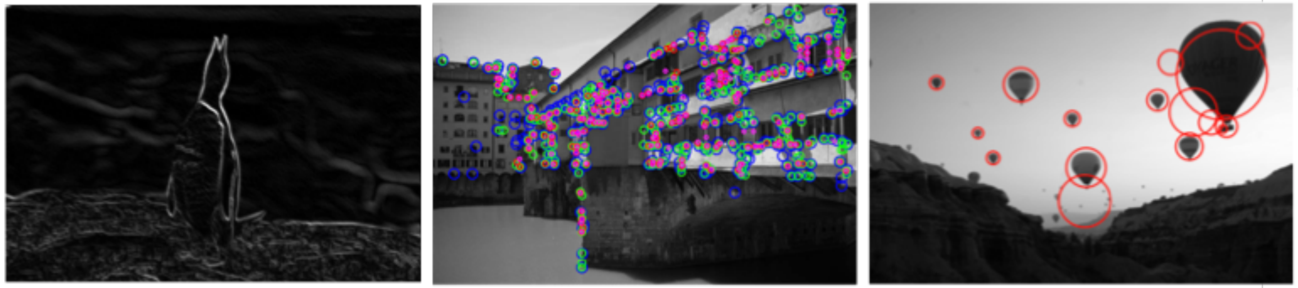
Videos are intrinsically high dimensional data. For some dynamic events we need minutes of videos translating into thousands of image frames

For this reason a common treat is to enhance meaningful regions and/or detect features to reduce the data redundancy

Not to forget, several applications call for efficient methods

Reducing image redundancy

Low-level Shearlet-based feature detection



As we are interested in enhancing information at multiple-scales we consider Shearlets.

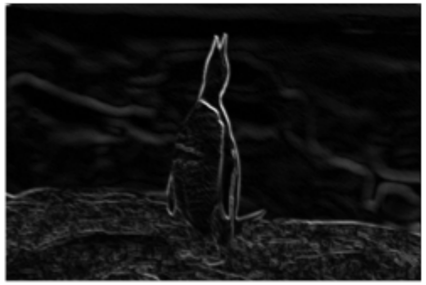
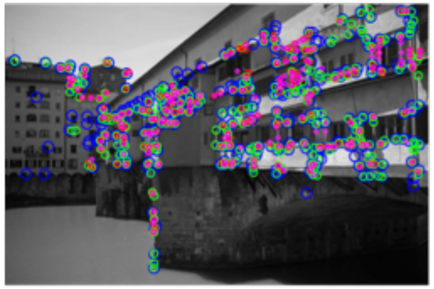
Shearlets are a multiscale framework which allows efficient encoding of anisotropic features in multivariate signals.

MA Duval-Poo, et al "Scale invariant and noise robust interest points with shearlets" IEEE Transactions on Image Processing 2017

M. Duval Poo, et al "Edges and corners with Shearlets" IEEE Transactions on Image Processing, 2015

Reducing image redundancy

Low-level Shearlet-based feature detection

As we are interested in enhancing information at multiple-scales we consider Shearlets.

Shearlets are a multiscale framework which allows efficient encoding of anisotropic features in multivariate signals.

UniGe | MaLGa

MA Duval-Poo, et al "Scale invariant and noise robust interest points with shearlets" IEEE Transactions on Image Processing 2017

M. Duval Poo, et al "Edges and corners with Shearlets" IEEE Transactions on Image Processing, 2015

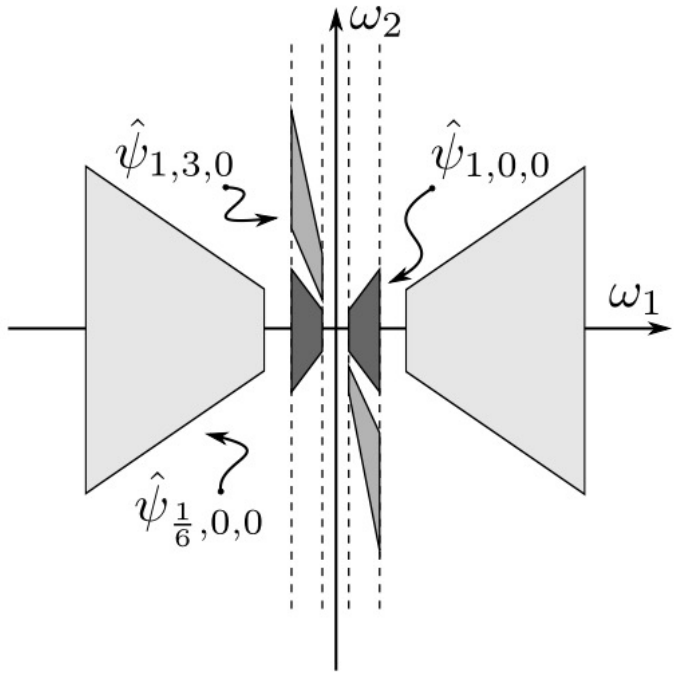
Shearlets

A shearlet ψ is generated by the dilation, shearing and translation of a mother shearlet function

$$\psi_{a,s,t}(x) = a^{-3/4}\psi\left(\begin{pmatrix} \frac{1}{a} & -\frac{s}{a} \\ 0 & \frac{1}{\sqrt{a}} \end{pmatrix}(x-t)\right)$$

Classical mother shearlet

$$\widehat{\psi}(\xi_1, \xi') = \underbrace{\widehat{\psi_1}(\xi_1)}_{\text{1D-wavelet}} \quad \underbrace{\widehat{\psi_2}(\xi'/\xi_1)}_{\text{bump function}}$$



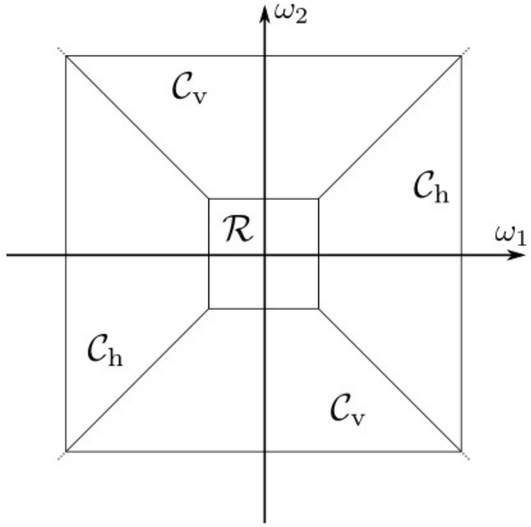
Shearlets support $\hat{\psi}_{a,s,t}$ (frequency domain) for different a and s.

Shearlet Transform

The discrete shearlet transform of an image $\mathcal{I}$ is defined by

$$\mathcal{SH}(\mathcal{I})(j,k,m) = \begin{cases} \langle \mathcal{I}, \phi_m \rangle \\ \langle \mathcal{I}, \psi^{\mathrm{h}}_{j,k,m} \rangle \\ \langle \mathcal{I}, \psi^{\mathrm{v}}_{j,k,m} \rangle \end{cases}$$

where j, k, m are the discretized scale, shear and translation parameters.



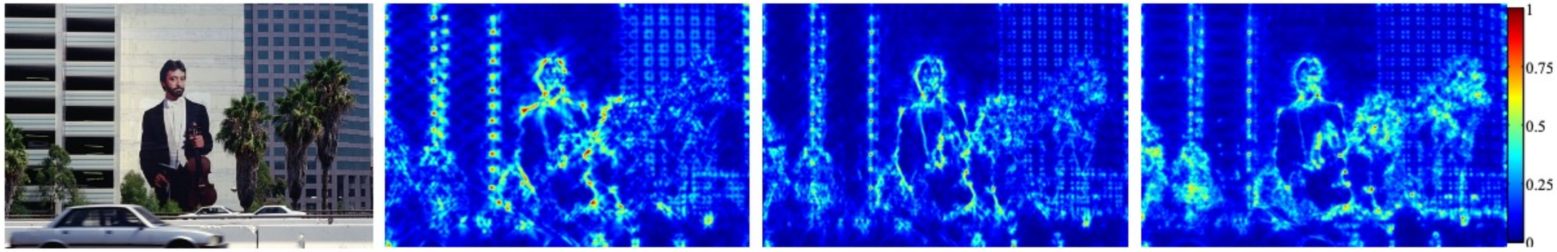
Decomposition of the frequency domain into cones

Reducing image redundancy

Shearlet-based Corner detection

- The **shearlet cornerness measure**[1] for a point $m \in \mathcal{I}$ and a fixed scale j is estimated as

$$\mathcal{C}(m,j) = \sum_{u \in W(m)} \sum_{k} |\mathcal{SH}(\mathcal{I})(j,k,u)| \sin(|\theta_k - \theta_{k_{\max}}|)$$



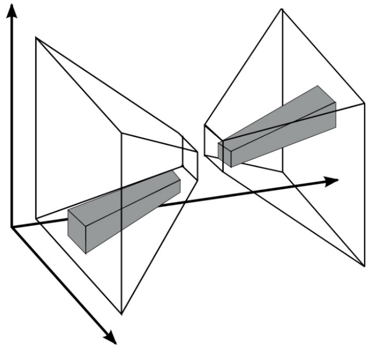
Original image | $j=0$ | $j=1$ | $j=2$

UniGe | MaLGa

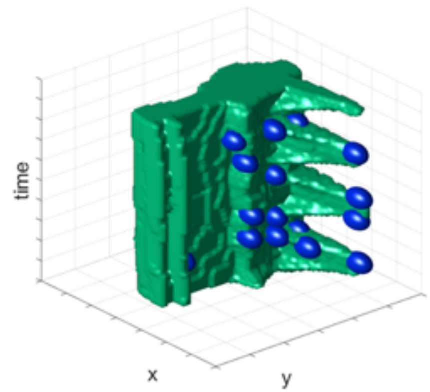
[1] Duval, Odone, De Vito. Edges and Corners With Shearlets (2015) IEEE Transactions on Image Processing.

Reducing image sequences redundancy

Space-time Shearlet local features



We look for few spatio-temporal keypoints on the basis of the relation they have with their neighbourhood



D. Malafronte, E. De Vito, F. Odone "Space-Time Signal Analysis and the 3D Shearlet Transform Journal of Mathematical Imaging and Vision" 2018

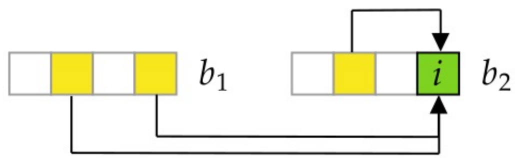
Dense image sequences analysis

Grey Code Kernels for efficient dense low-level analysis

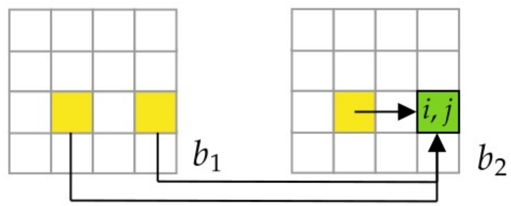
Family of filter kernels that, under specific circumstances, can be used as an **highly efficient filtering scheme.**

Successive convolutions of an image with a set of such filters require **only two operations per pixel** for each filter, regardless of size or dimension of the filter.

$$b_2(i) = b_1(i) \pm b_1(i-\Delta) \pm b_2(i-\Delta)$$



$$b_2(i,j) = b_1(i,j) \pm b_1(i,j-\Delta) \pm b_2(i,j-\Delta)$$



Given the result b_1 of the application of the first kernel v_1 to an image I, we can obtain the result b_2 of filtering with the second kernel v_2 with just two summation per pixel

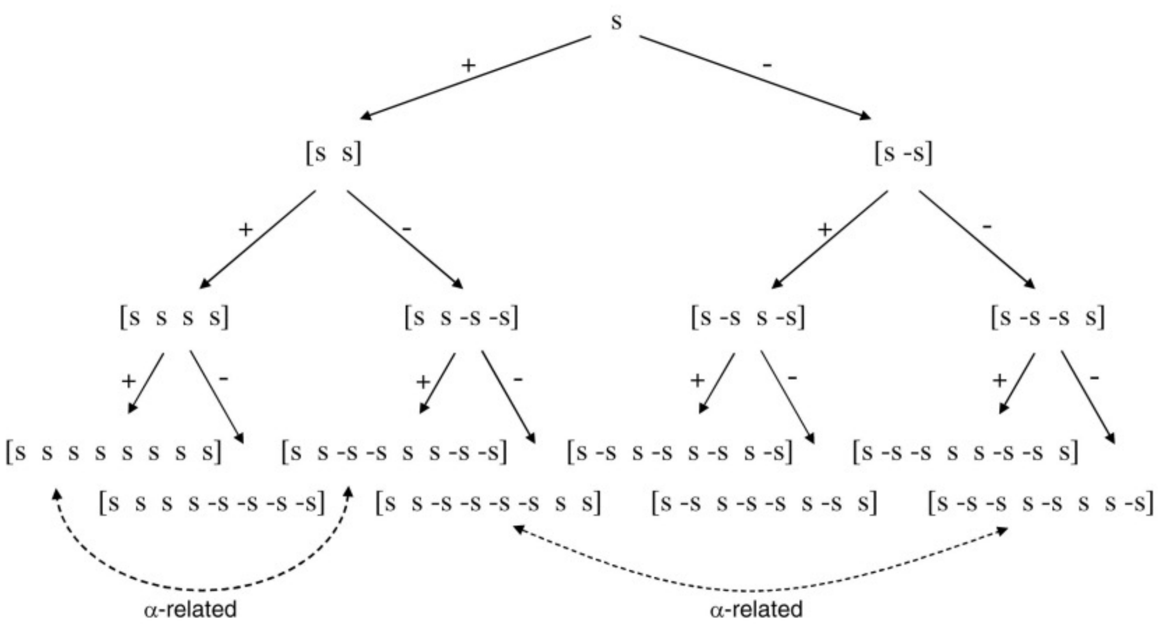
Ben-Artzi et al. "The Gray-Code Filter Kernels" IEEE transactions on pattern analysis and machine intelligence (2007), pp.382–393

GCK Sketch in 1D

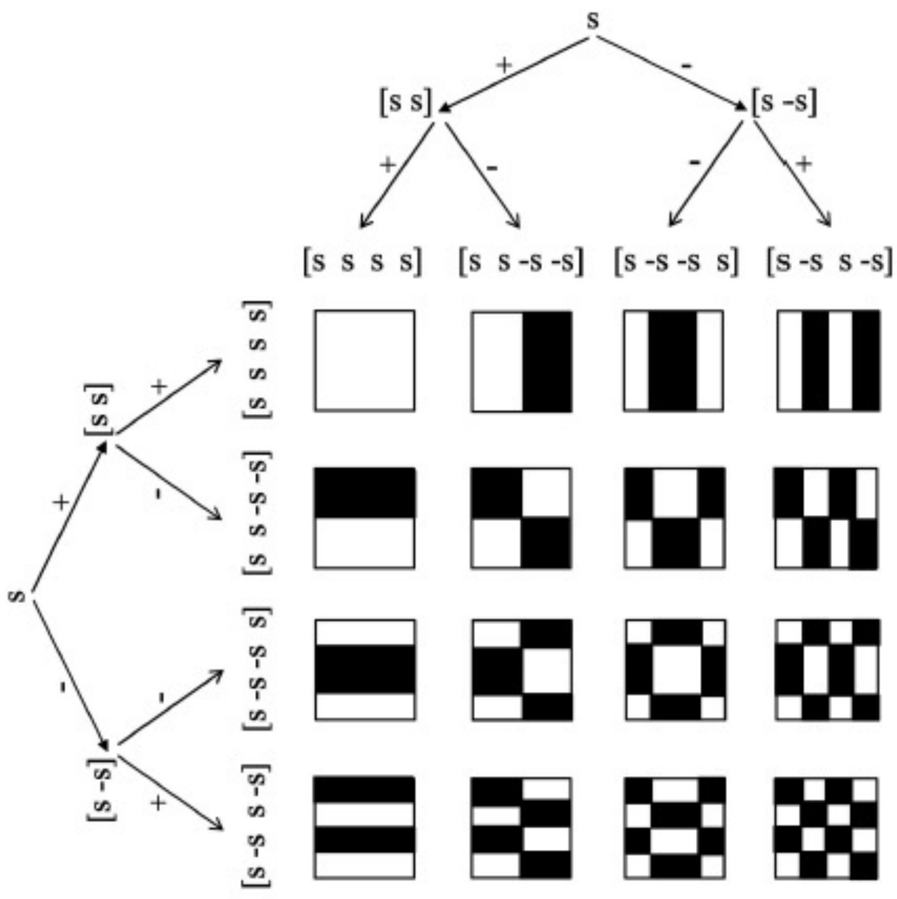
We consider a recursive definition of 1D filter kernels expanded from an initial seed vector s as follows:

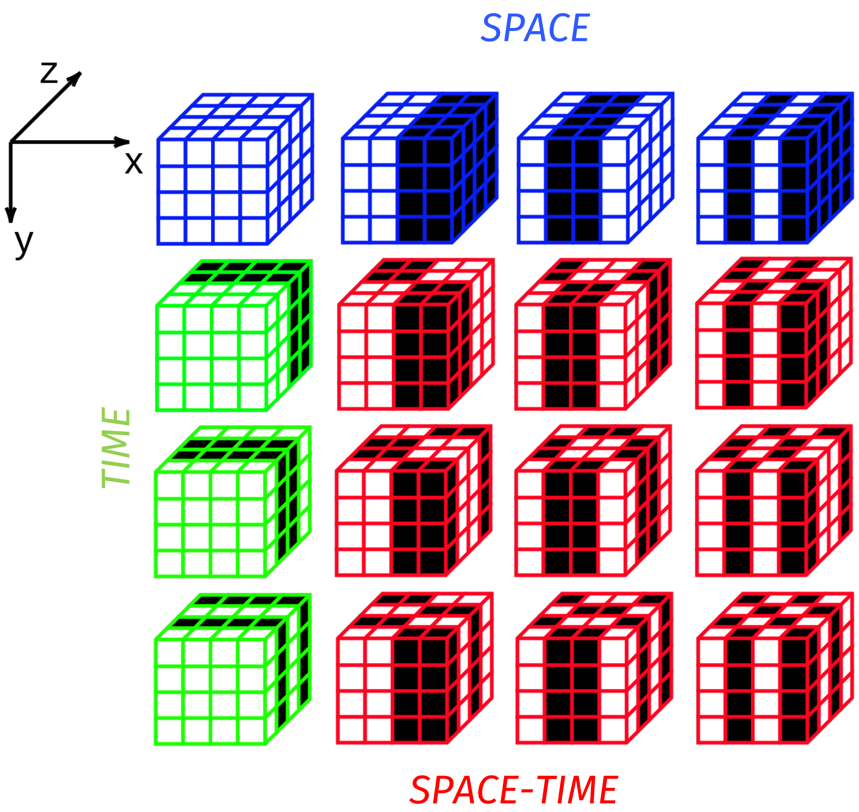
$$
\begin{aligned}
V_s^{(0)} &= \mathbf{s}, \\
V_s^{(k)} &= \{[\mathbf{v}_s^{(k-1)} \alpha_k \mathbf{v}_s^{(k-1)}]\} \quad s.t. \quad \mathbf{v}_s^{(k-1)} \in V_s^{(k-1)}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\quad \alpha_k \in \{+1, -1\},
\end{aligned}
$$

Efficiency depends on the ordering in which they are applied to an image



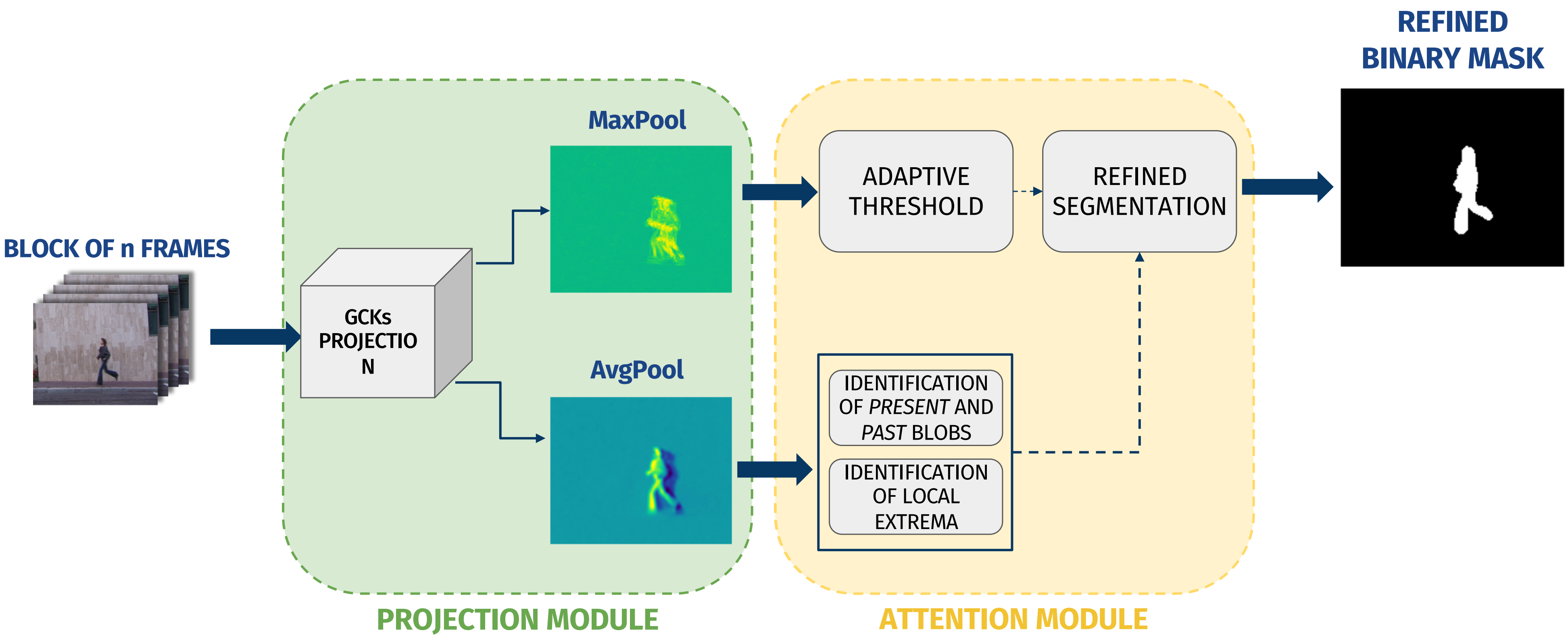
GCK Extensions to higher dimensions





From low to mid-level analysis

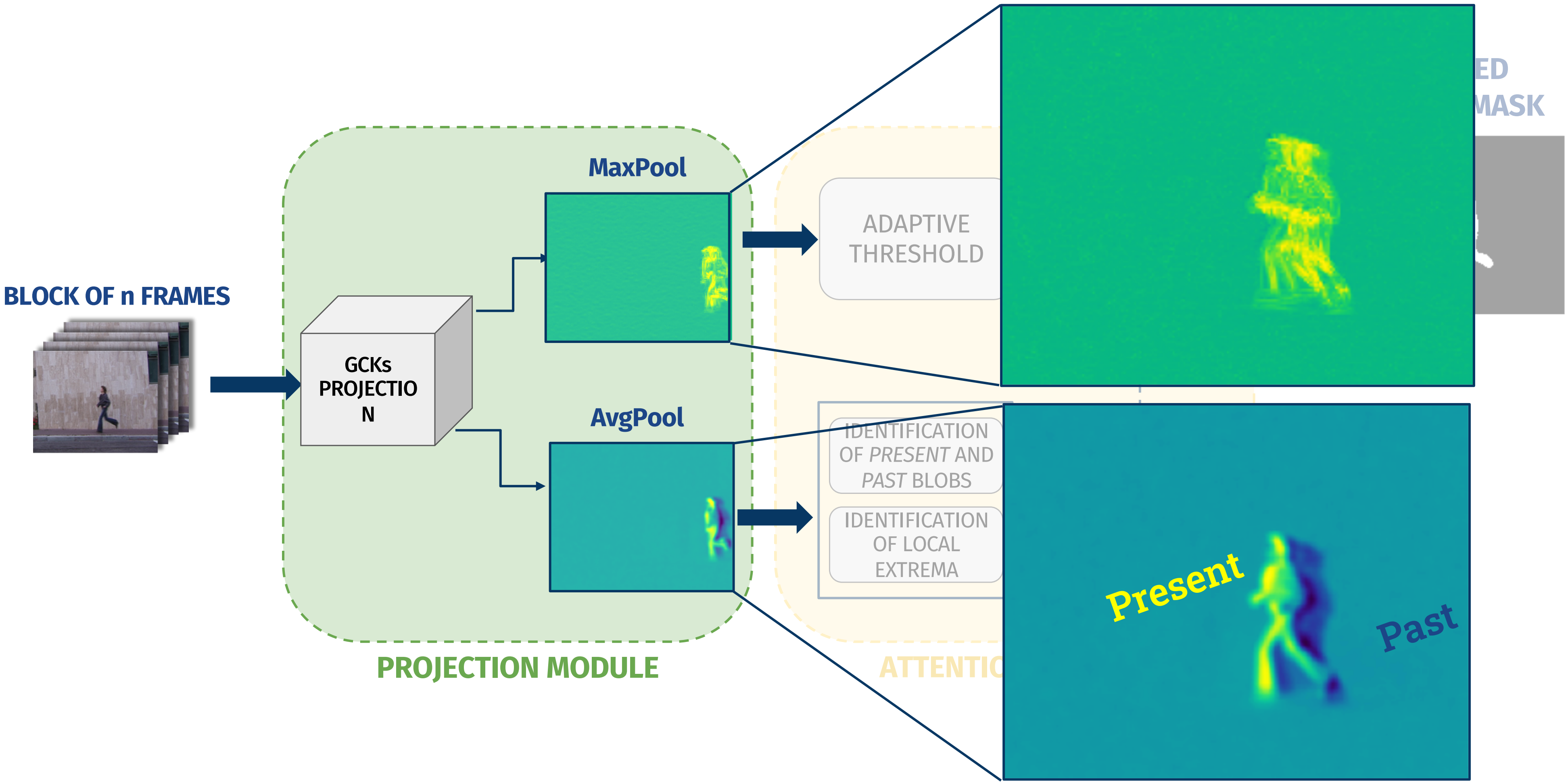
GCK for efficient motion segmentation



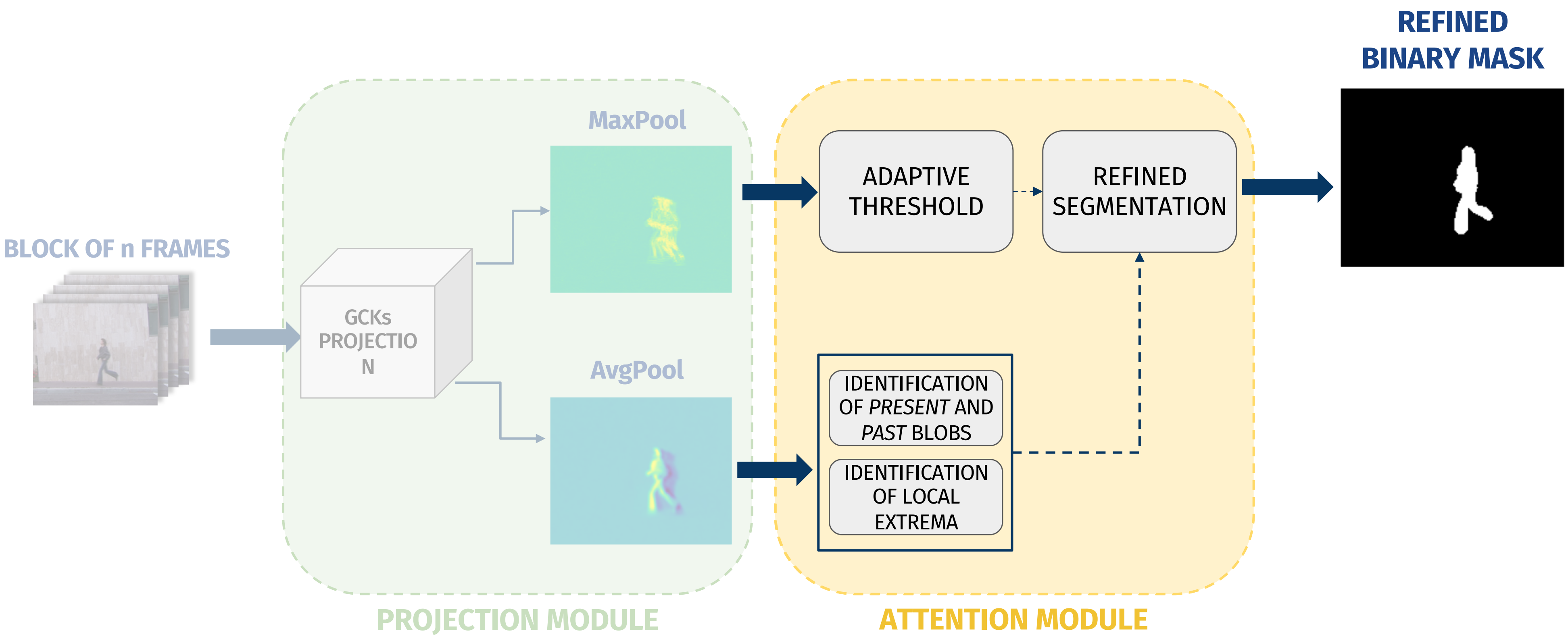
UniGe | MaLGa

On the Use of Efficient Projection Kernels for Motion-Based Visual Saliency Estimation, *E. Nicora, Frontiers Computer Science 2022*

GCK for efficient motion segmentation



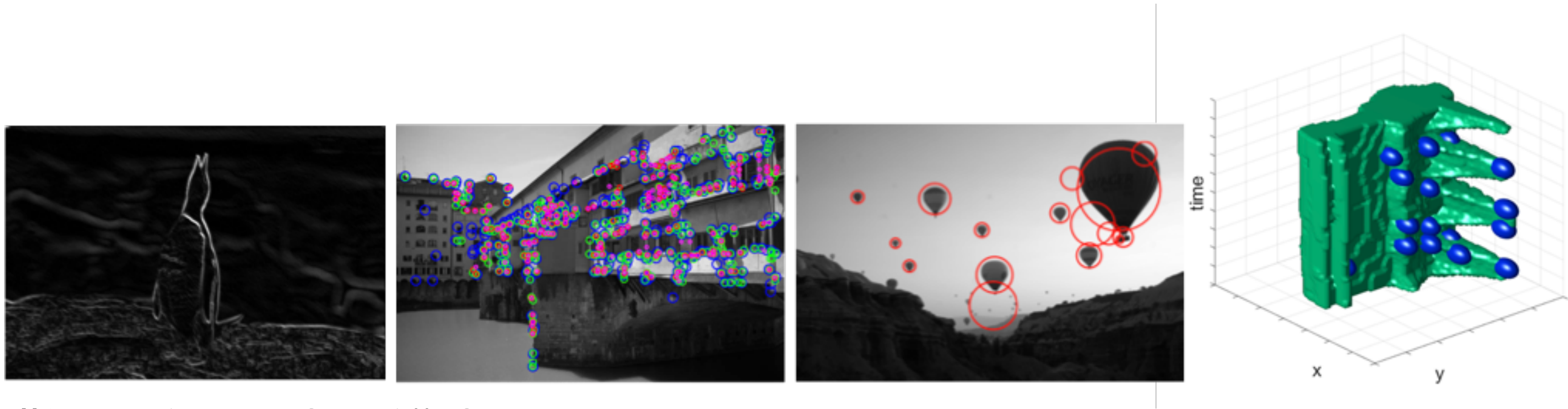
GCK for efficient motion segmentation



Mid level image and video analysis

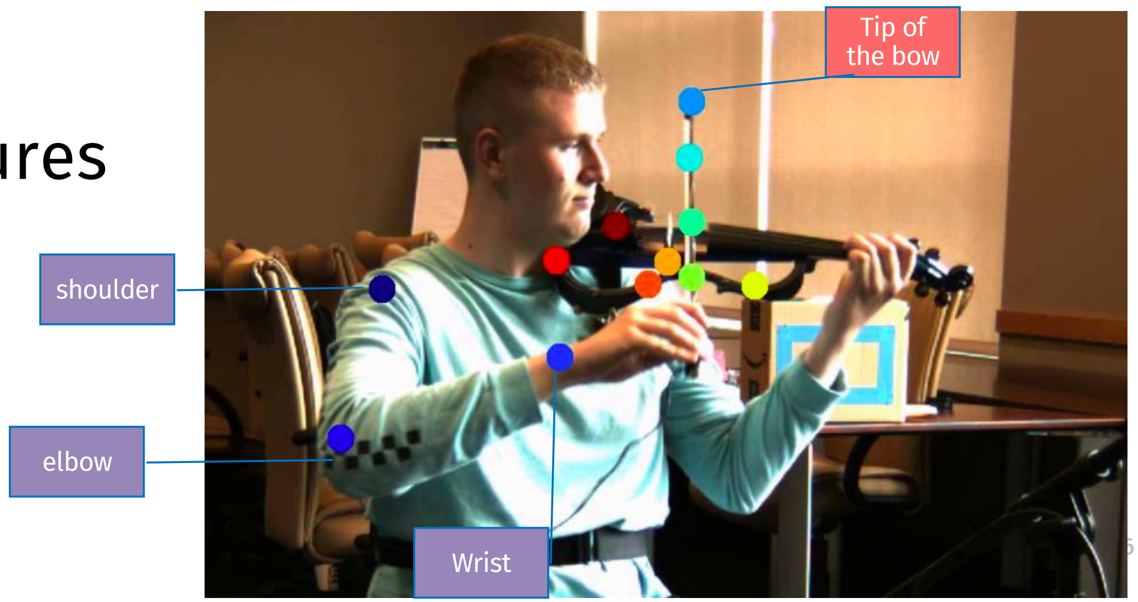
From low to mid-level analysis

Semantic features



“hand-crafted” features

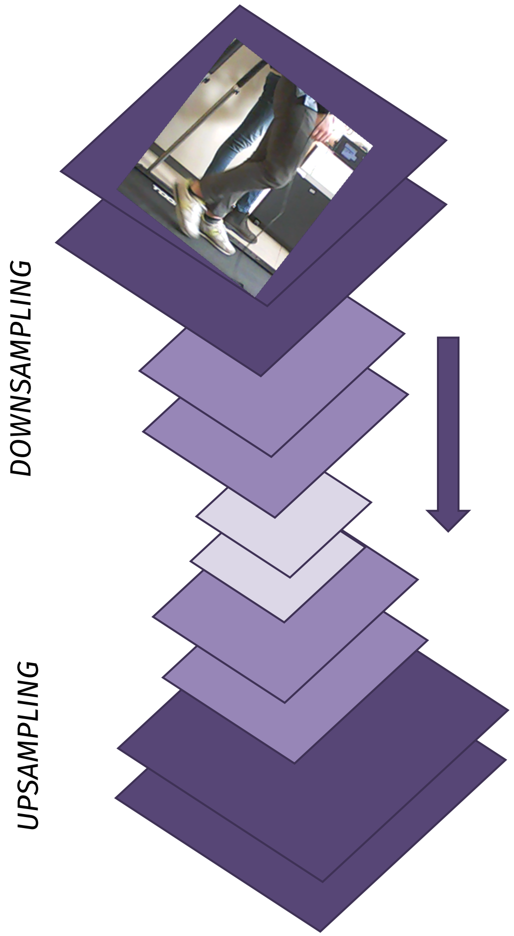
“data-driven” features



Low-level to high-level and back

Semantic features are image keypoints associated with a specific **appearance** and **semantic attributes**

Semantics is usually “inherited” by a more global understanding of the image content



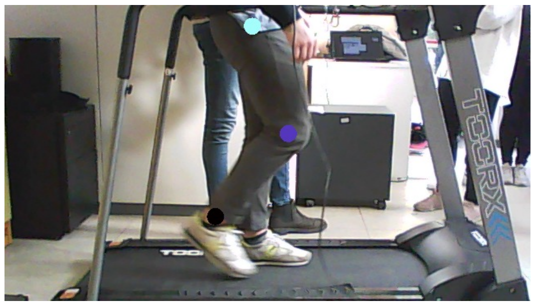
Semantic features are usually derived by semantic segmentation pipelines, usually based on encoding-decoding models, the goal of which is to classify individual pixels

Each obtained feature will be defined by its position on the image plane and its confidence level

$$f_i = (x_i, y_i, c_i)$$

Semantic features segmentation

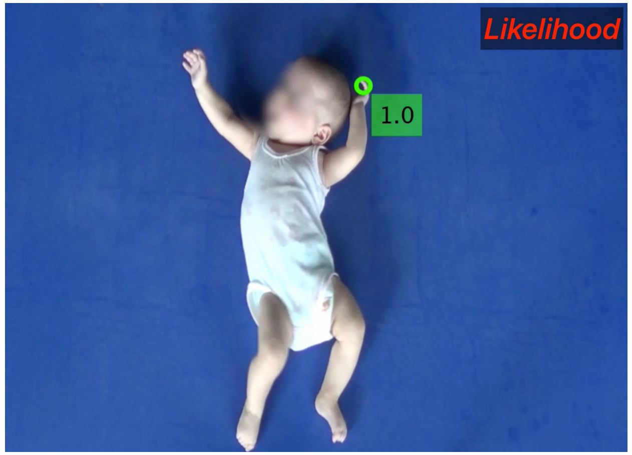
Feature detection

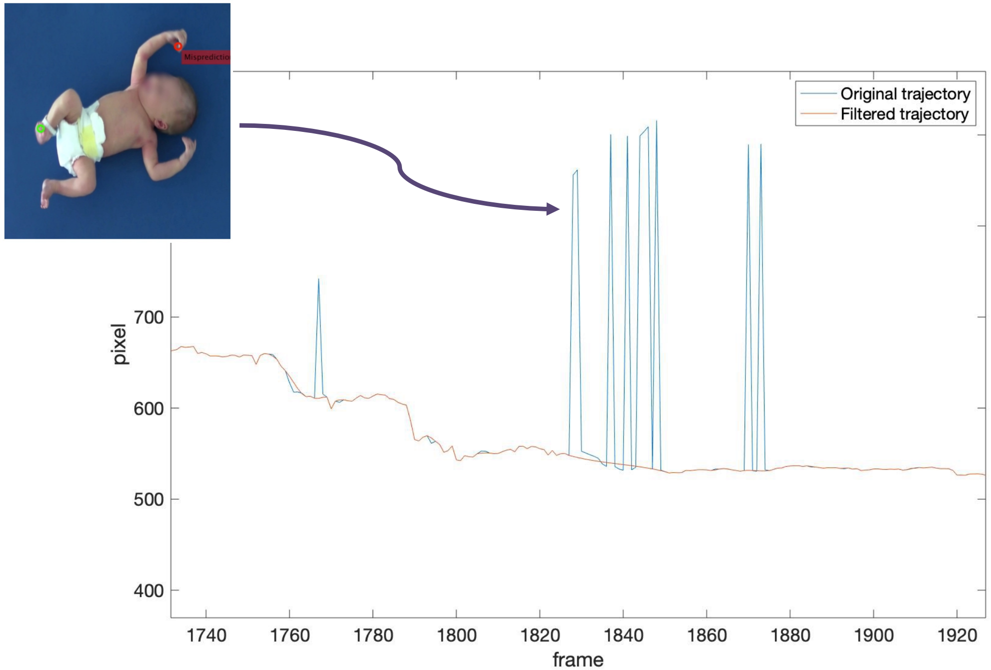


Example:
Analysing infants motion

Temporal continuity as a way to detect mispredictions and filter temporary occlusions

Confidence values and occlusions



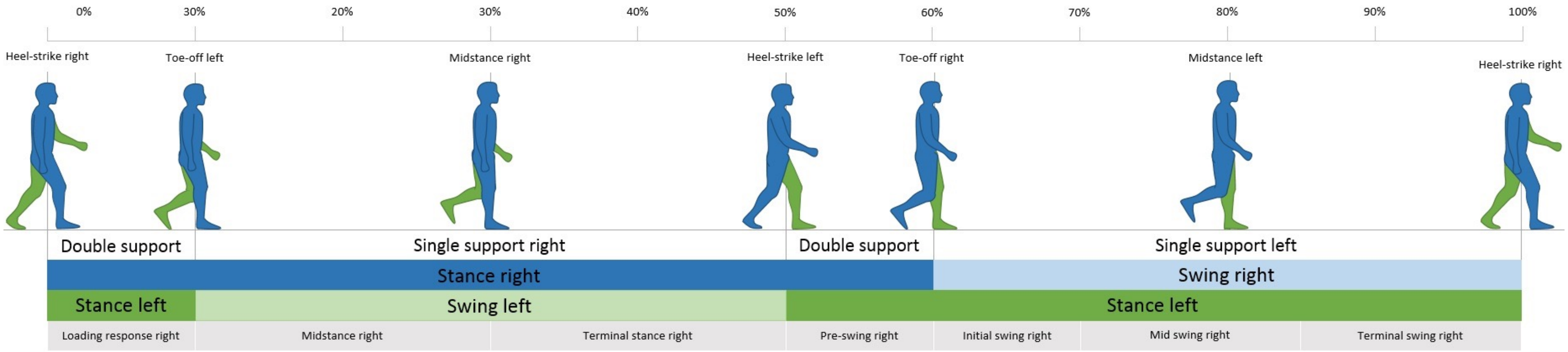


UniGe | MaLGa

A markerless pipeline to analyze spontaneous movements of preterm infants
M. Moro et al CMPB 2022

Detecting and studying motion primitives

Gait cycle analysis



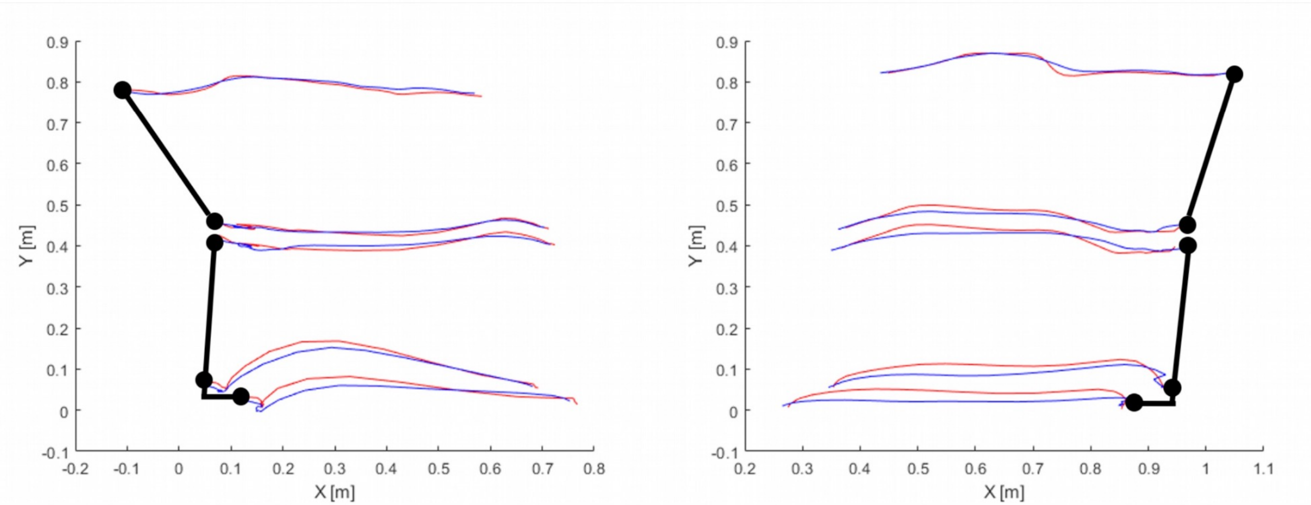
Gait cycle analysis is a common practice in rehabilitation and clinical applications

Gold standard techniques are marker-based systems

Video-based (marker-less) is less intrusive: semantic feature detection is becoming a valid alternative

Detecting and studying motion primitives

2D Marker-less gait analysis

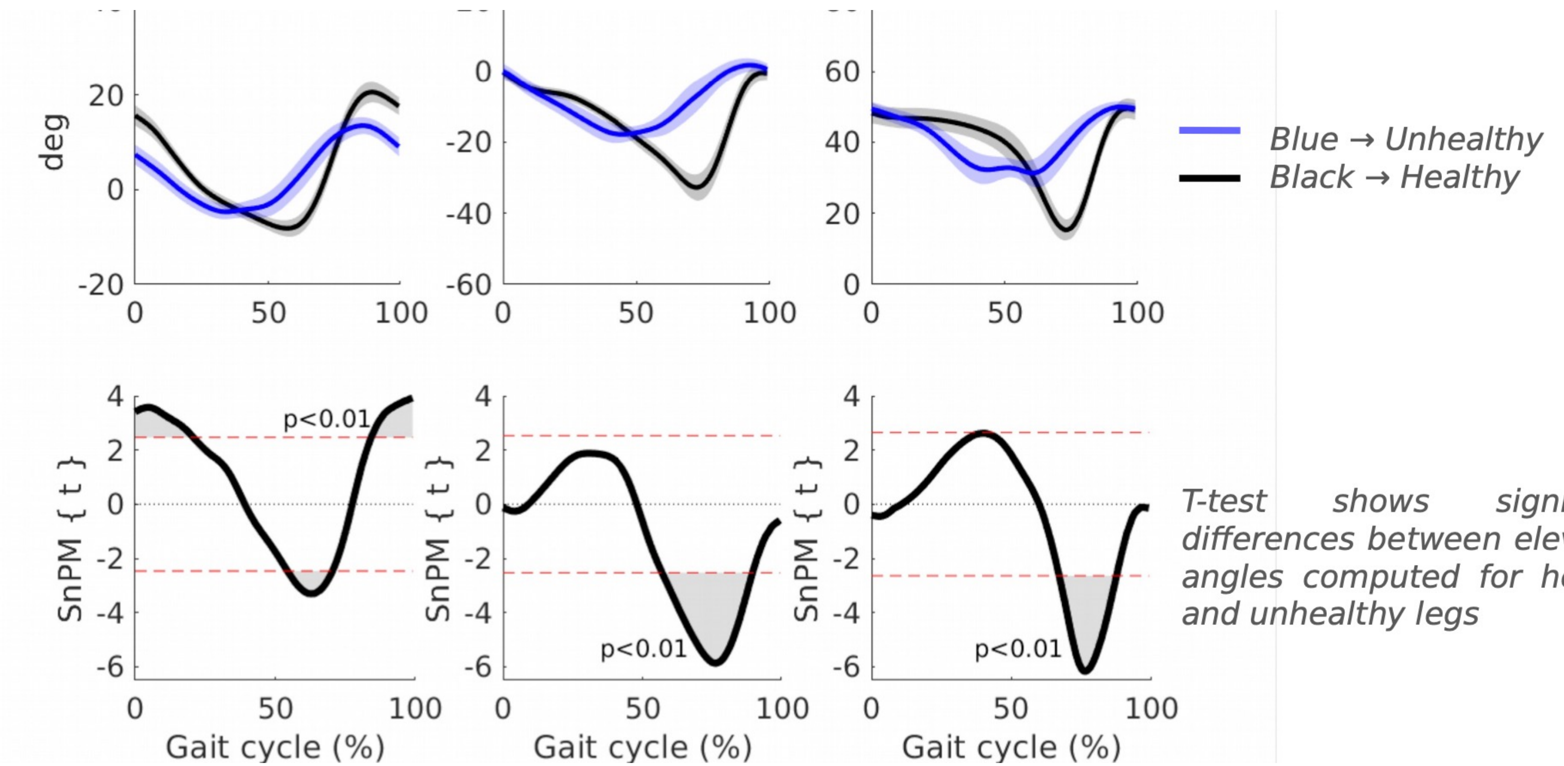


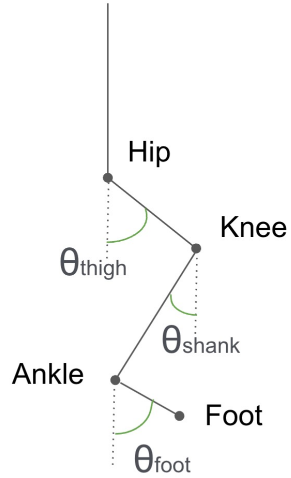
Red → Marker signal
Blue → Video signal

Moro Casadio Odone "Markerless gait analysis in stroke survivors based on computer vision and deep learning: a pilot study. ACM SAC 2020

Detecting and studying motion primitives

2D Marker-less gait analysis





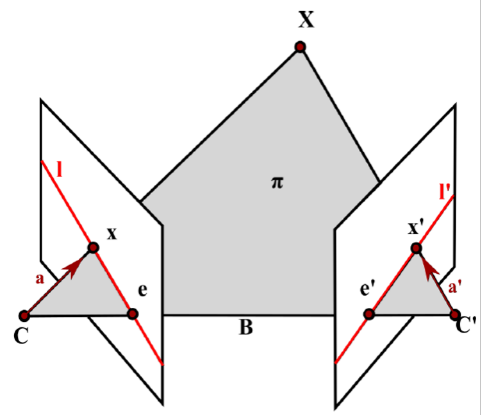
T-test shows significant differences between elevation angles computed for healthy and unhealthy legs

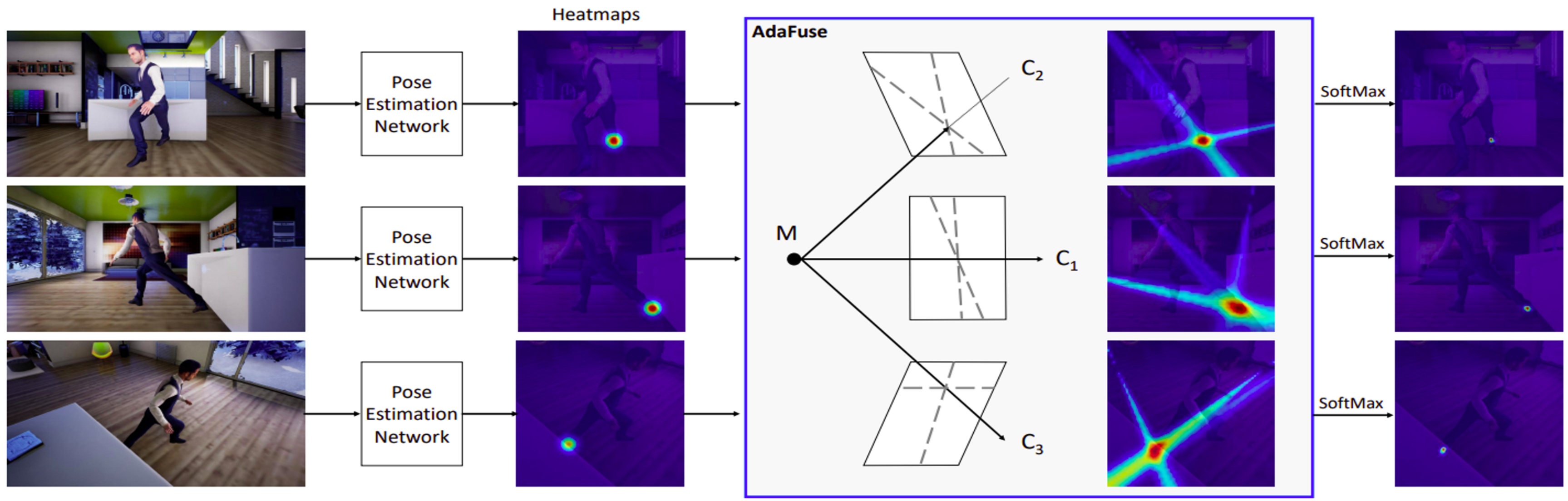
Moro Casadio Odone "Markerless gait analysis in stroke survivors based on computer vision and deep learning: a pilot study. ACM SAC 2020

UniGe | MaLGa

Detecting and studying motion primitives

3D Marker-less gait analysis

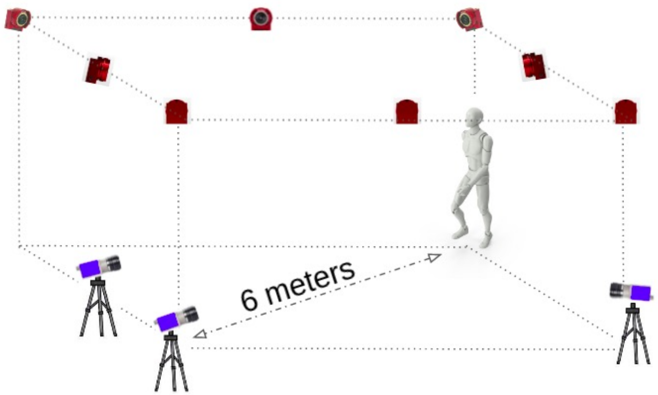


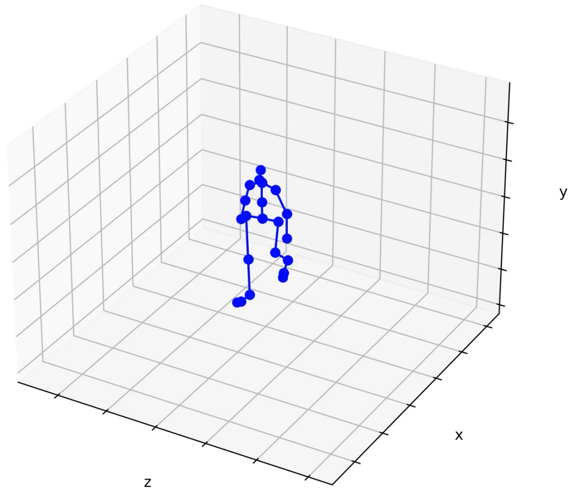


Markerless vs. marker-based gait analysis: A proof of concept study *M. Moro et al Sensors 2022*

Detecting and studying motion primitives

3D Marker-less gait analysis

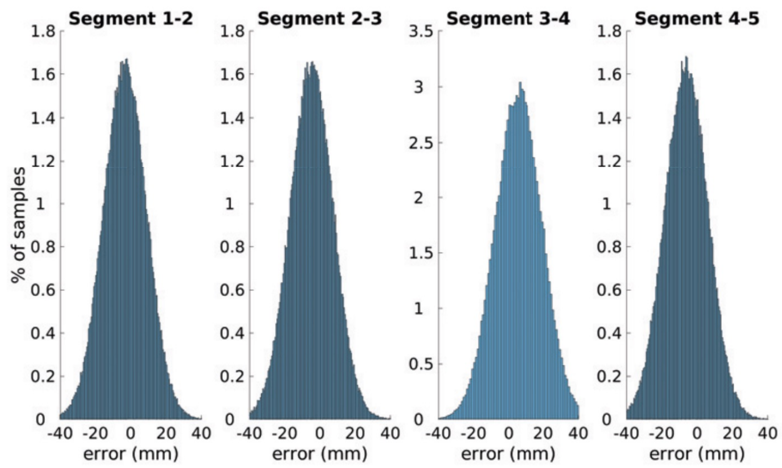




[Markerless vs. marker-based gait analysis: A proof of concept study] M. Moro et al Sensors 2022

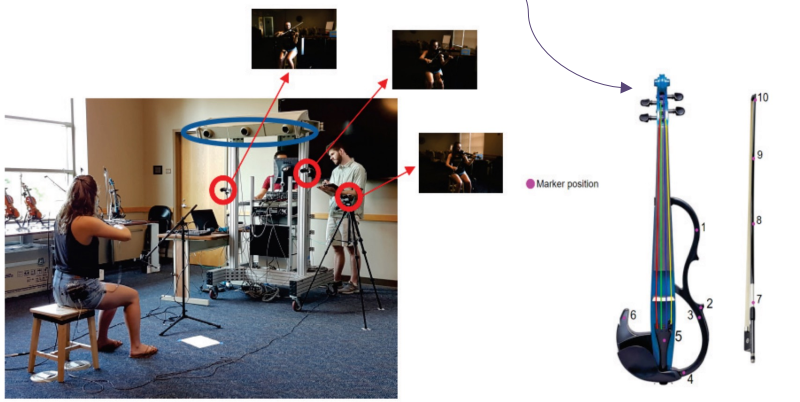
On the precision of the 3D measurements

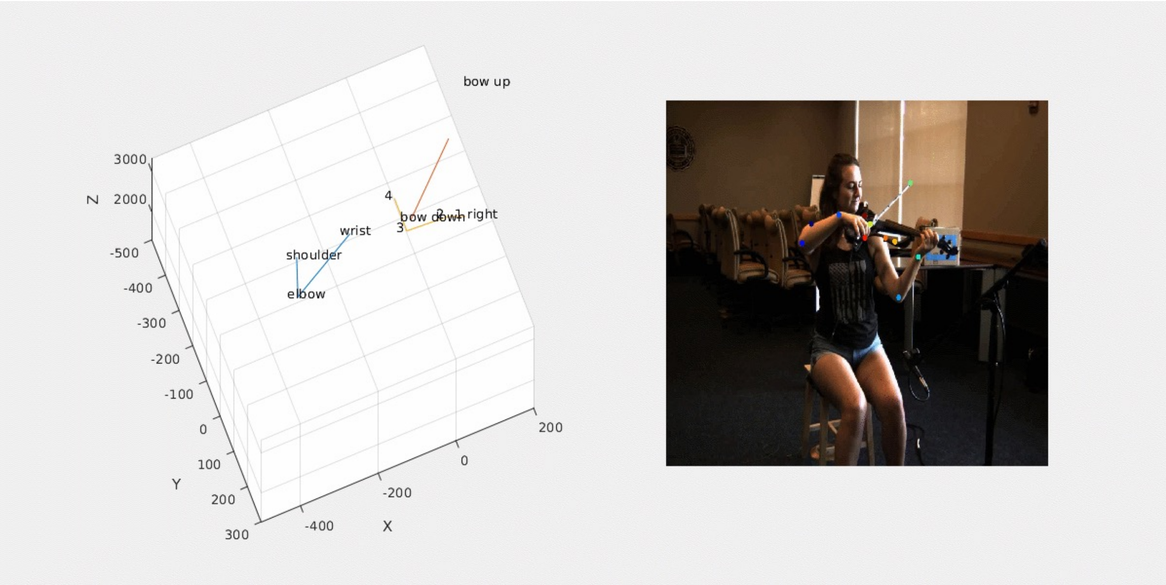
A study case on violin playing



Quality assessment is carried out in an indirect manner, by comparing the metric distance between pairs of marker based and marker-less keypoints

Markers are placed on the violin and the bow





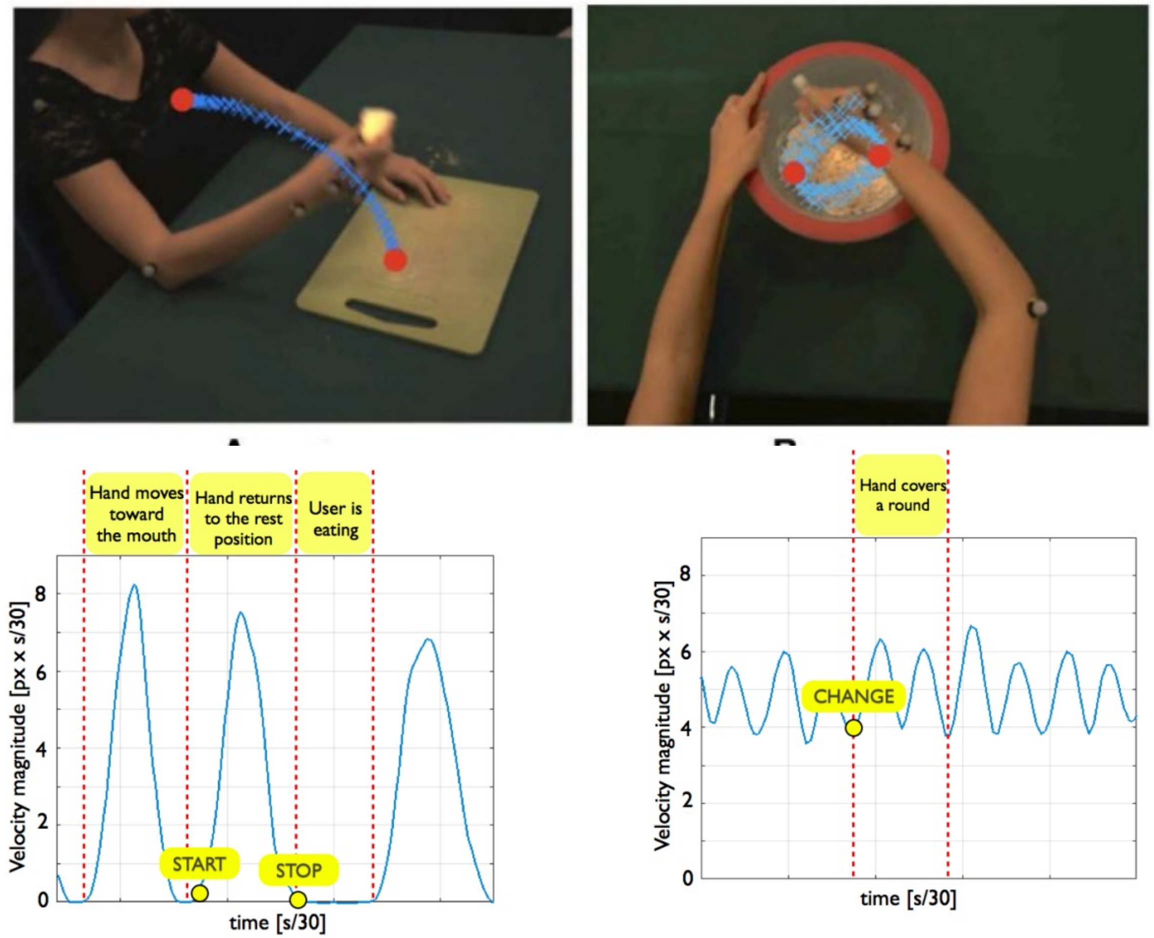
On The Precision Of Markerless 3d semantic features: An experimental study on violin playing M Moro ICIP 2021

Detecting action primitives

Besides gait, other (several) human actions present a repetitive pattern

We represent a dynamic event as a sequence of velocities and we segment the sequence detecting dynamic instants

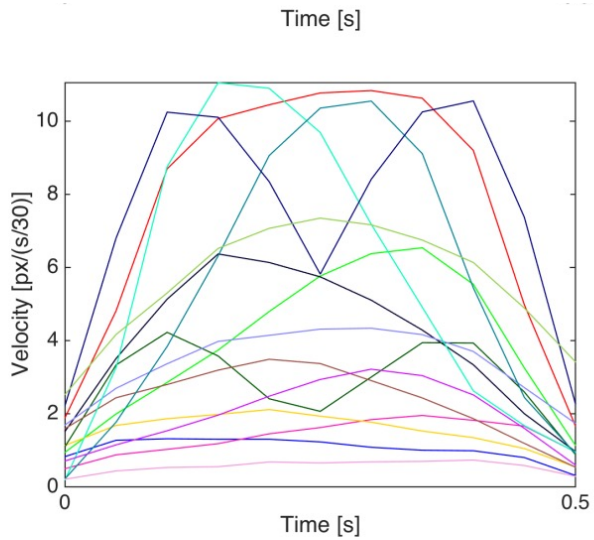
Dynamic instants are defined as local minima of the velocity profiles



UniGe | MaLGa

F.Rea et al *Human motion understanding for selecting action timing in collaborative human-robot interaction. Frontiers in Robotics and AI, 2019*

Representing Action primitives



- We approach dictionary learning as an unsupervised problem using K-Means

$$\min_{\mathbf{D},\mathbf{U}} \|\mathbf{X} - \mathbf{D}\mathbf{U}\|_F^2 \ . \ \text{s.t.} \quad \text{Card}(\mathbf{u}_i) = 1, |\mathbf{u}_i| = 1,$$
$$\mathbf{u}_i \geq 0, \forall i = 1, \dots, T$$

where $\mathbf{X}$ is the training set, $\mathbf{U}$ are the clusters membership codes, and $\mathbf{D}$ is the dictionary with K atoms

- We use Sparse Coding to derive a sparse representation using the dictionary

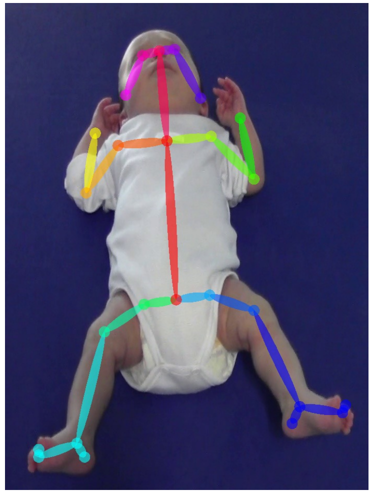
$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \|\mathbf{x} - \mathbf{D}\mathbf{u}\|^2 + \lambda\|\mathbf{u}\|_1$$

Learning dictionaries of kinematic primitives for action classification A Vignolo et al ICPR 2020

Integrating local information with graphs

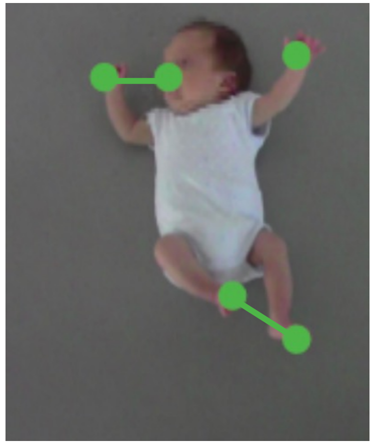
Local keypoints can describe motion only partially

For this reason a common approach is to rely on full-body pose estimators on a pre-defined skeleton model (OpenPose, MediaPipe, ...)



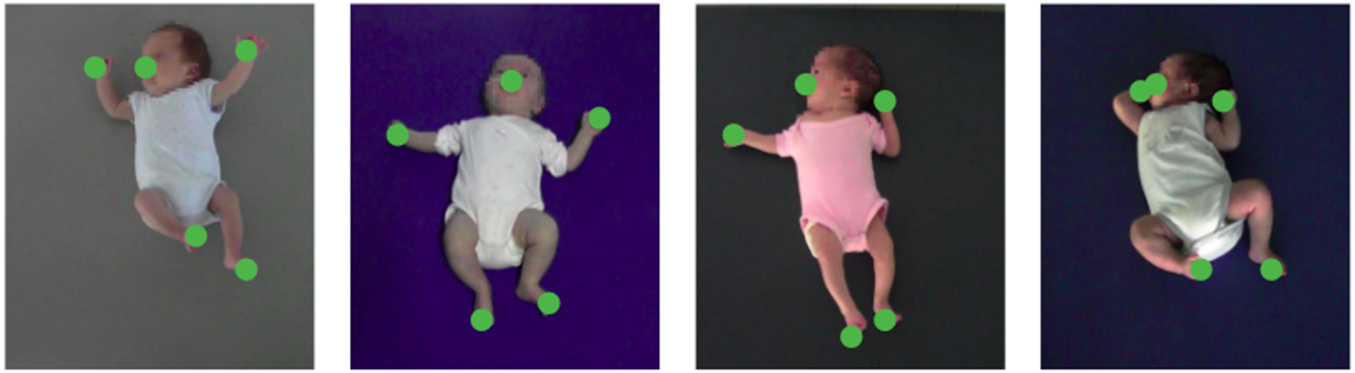
In our research we also explored the possibility of adopting more adaptive graph representations

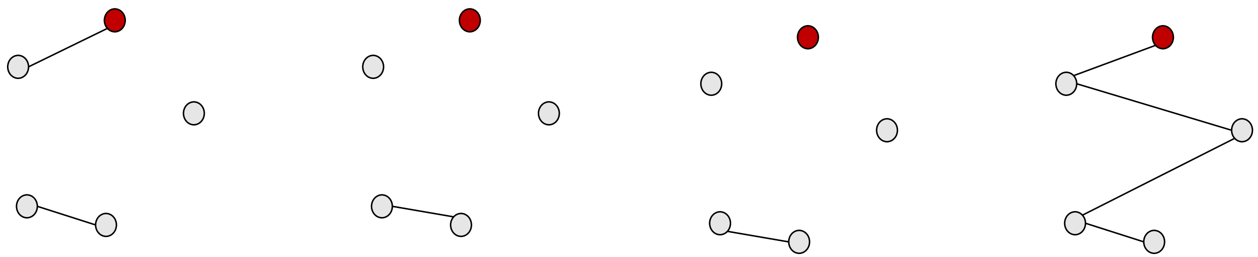
- we describe a body configuration for each frame considering landmark points as nodes of a network and connecting them depending on their proximity.
- Each configuration can be described by means of *attributed graphettes*.



Attributed Graphettes-Based Preterm Infants Motion Analysis D. Garbarino Complex Networks 2022

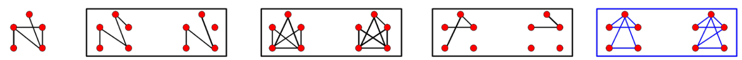
Graphettes-based analysis





- We describe a body configuration for each frame considering landmark points as nodes of a network and connecting them depending on their proximity.
- Proximity is computed by the Euclidean distance, normalized across the whole video
- Most common configurations allow us to provide an "interpretable" description of common and abnormal patterns

5 most important configurations in 40 weeks infants

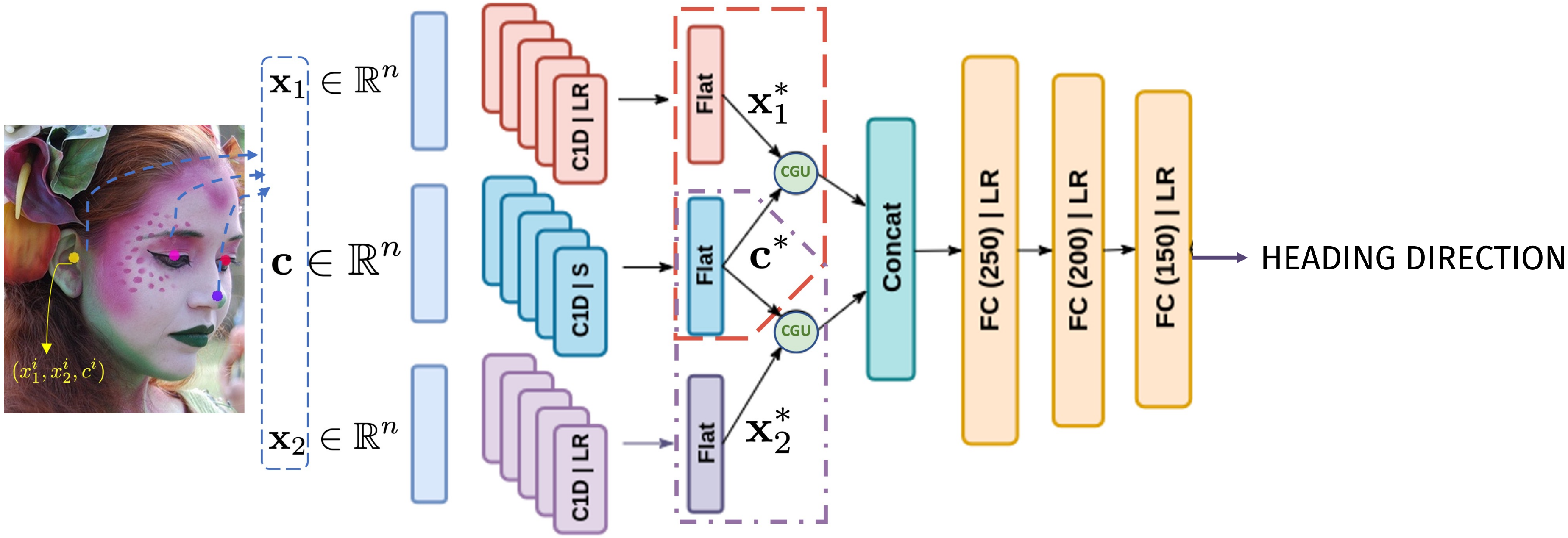


UniGe | MaLGa

High level image and video analysis

Towards high level human motion analysis

Gaze / heading estimation

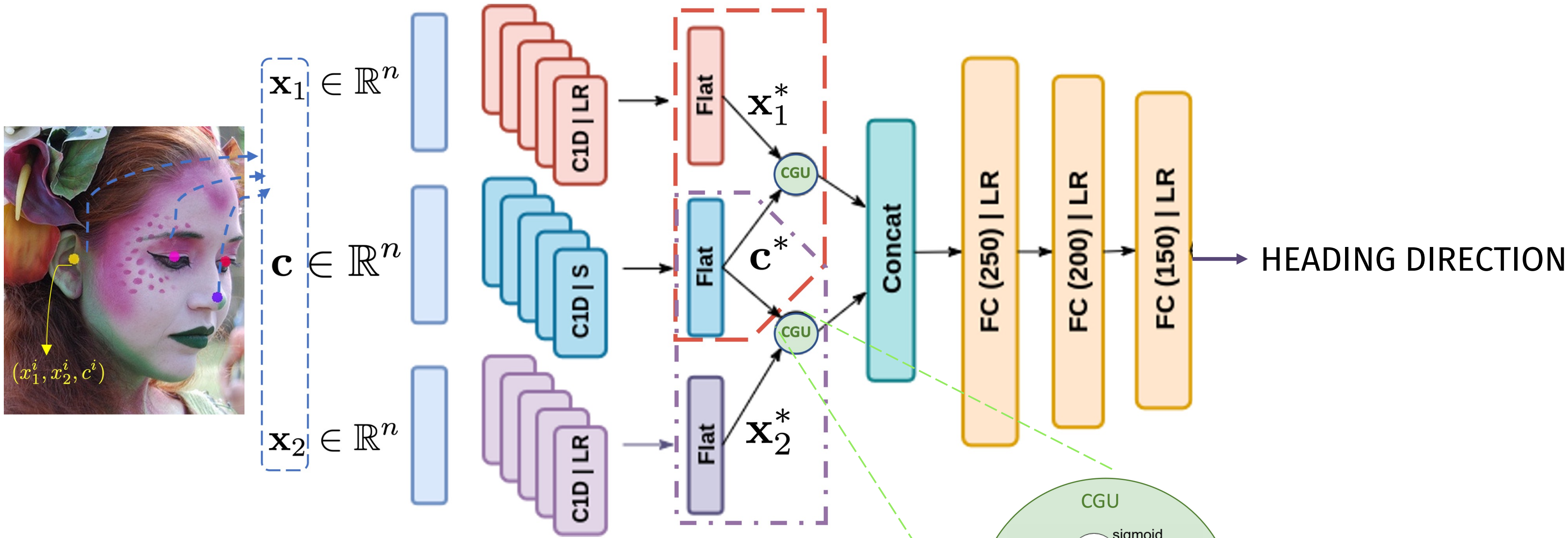


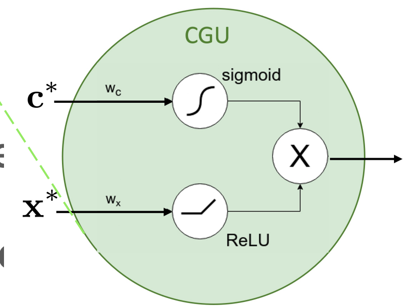
- We have explored the possibility of estimating heading from semantic features only
- We have designed a very small architecture that can be used as a plug-in to common pose estimation / semantic segmentation algorithms

HHP-Net: A light Heteroscedastic neural network for Head Pose estimation with uncertainty FF Tomenotti, et al WACV 2022

Towards high level human motion analysis

Gaze / heading estimation

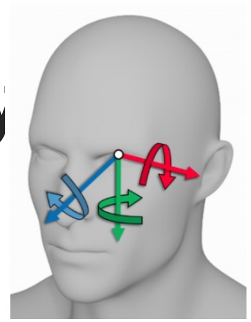


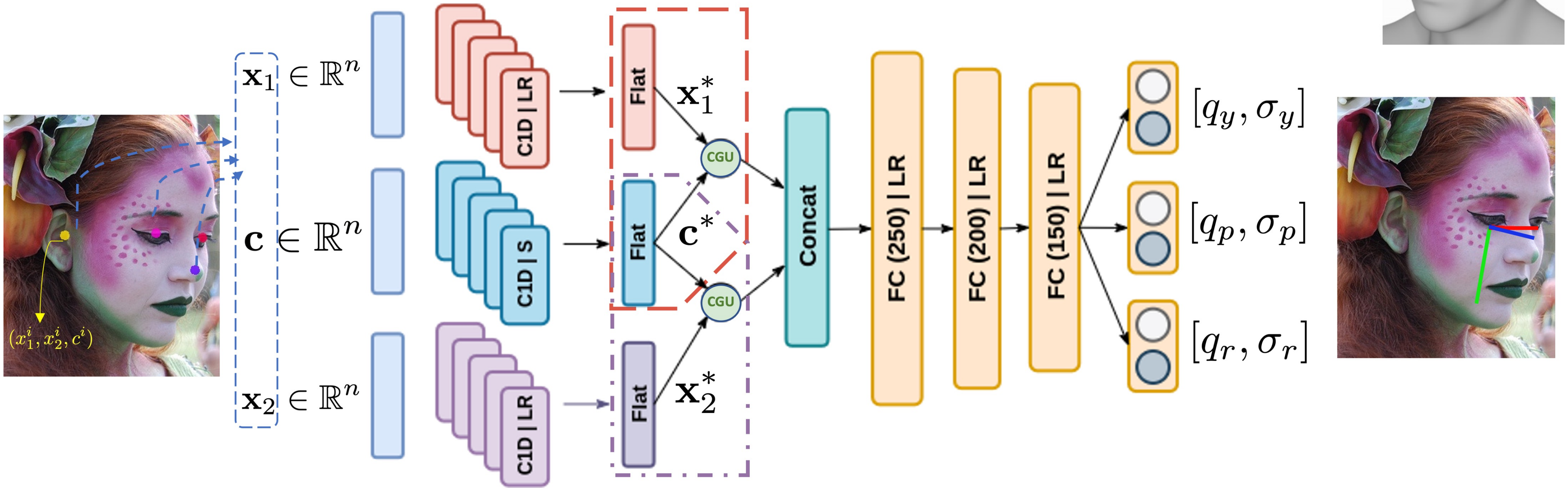


- We have explored the possibility of estimating he features only
- We have designed a very small architecture that in to common pose estimation / semantic segmentation algorithms

HHP-Net: A light Heteroscedastic neural network for Head Pose estimation with uncertainty FF Tomenotti, et al WACV 2022

Estimating yaw,pitch, roll, with uncertainty



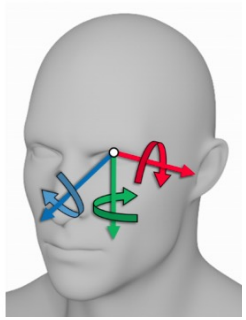


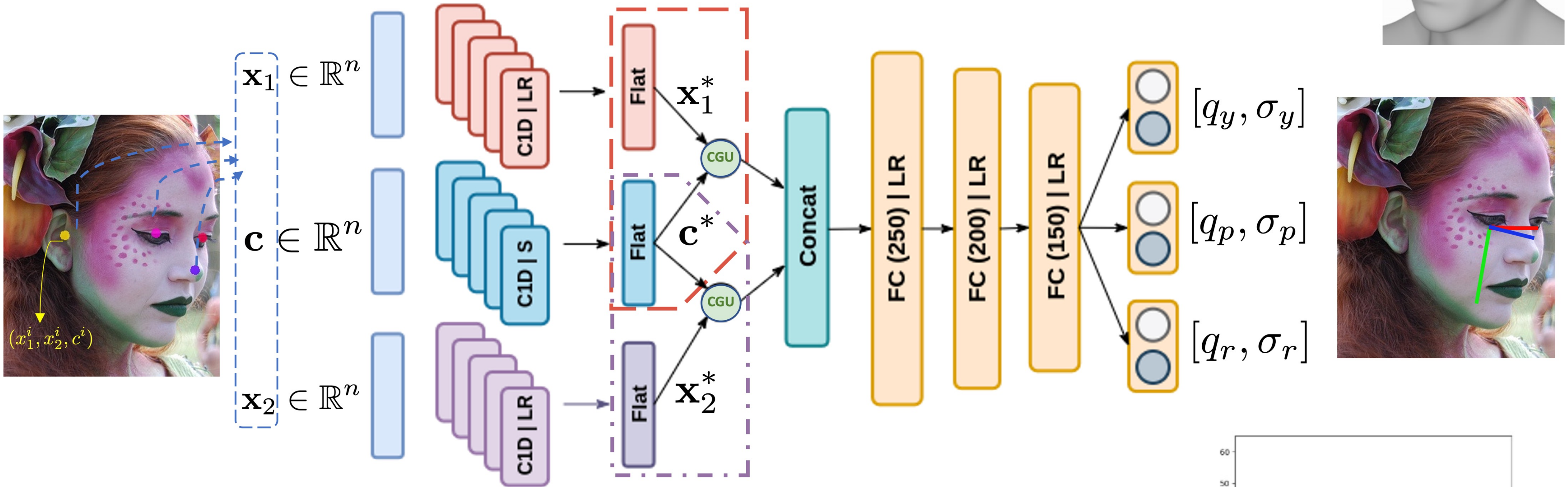
For training the network we adopt a multi-task loss function incorporating heteroscedastic aleatoric uncertainty to provide an estimate of the uncertainty of each prediction.

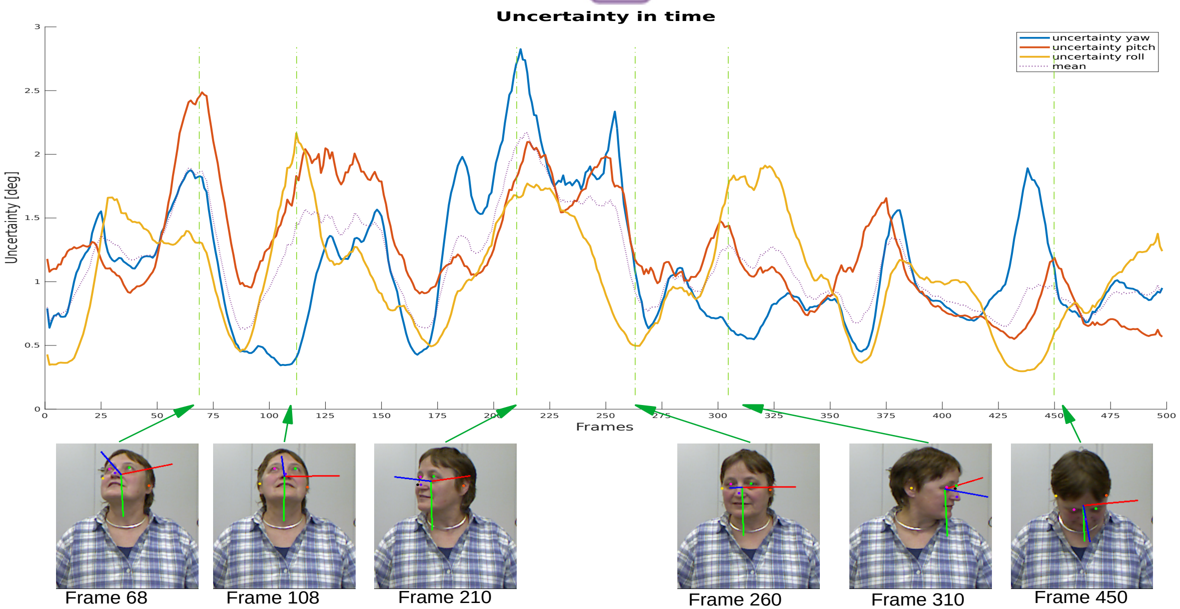
$$\sum_{i\in\{y,p,r\}} \left(\frac{1}{2}\exp\left(-s_i\right)\|q_i - f_i\left(\mathbf{x}_1,\mathbf{x}_2,\mathbf{c}\right)\|^2 + \frac{1}{2}s_i\right) \qquad s_i = \log\sigma_i(\mathbf{x}_1,\mathbf{x}_2,\mathbf{c})^2$$

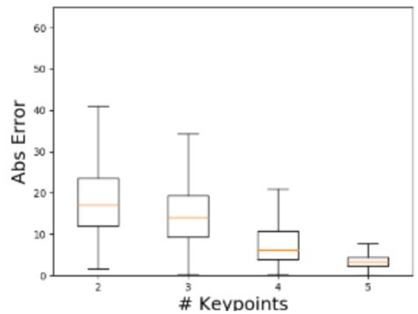
This is useful to capture noise within input observations: in our case it is related with inherent keypoints detection which may be affected by difficult viewpoints or occlusions.

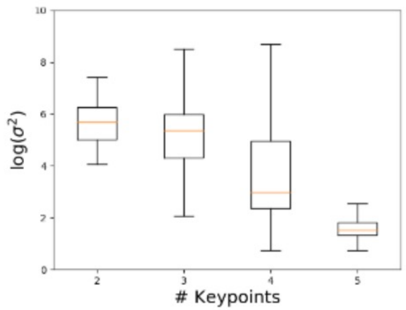
Estimating yaw,pitch, roll, with uncertainty









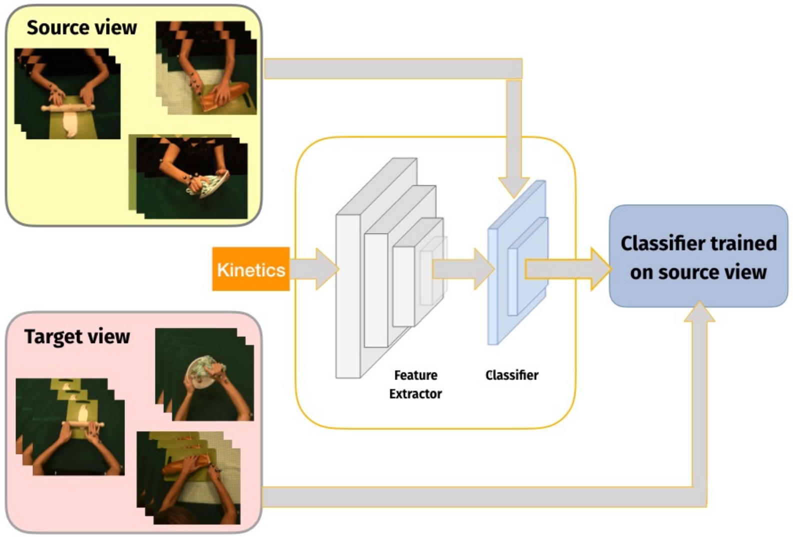


Experiments with end-to-end architectures

View-invariant action recognition

We do not possess massive datasets making view-point information explicit, to this purpose

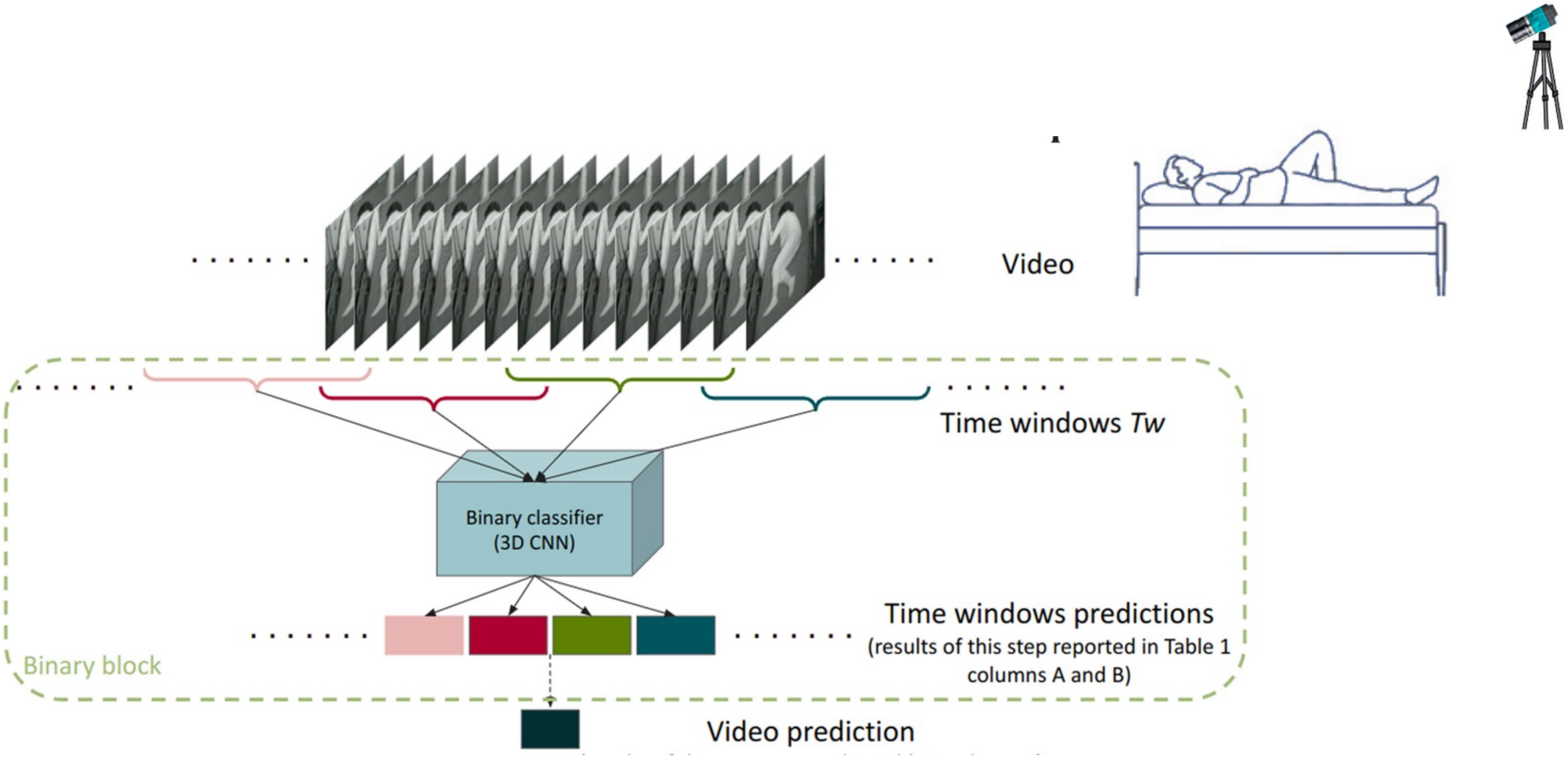
- We have explored the power of transfering high level deep features from large dimensional multi-view datasets (e.g., Kinetics)



•Cross-view action recognition with small-scale datasets G. Goyal et al ImaVis 2022

Experiments with end-to-end architectures

Classification of epileptic seizures



[Automatic Video Analysis and Classification of Sleep-related Hypermotor seizures and Disorders of Arousal](#)

[M. Moro et al, Epilepsia, 2023](#)

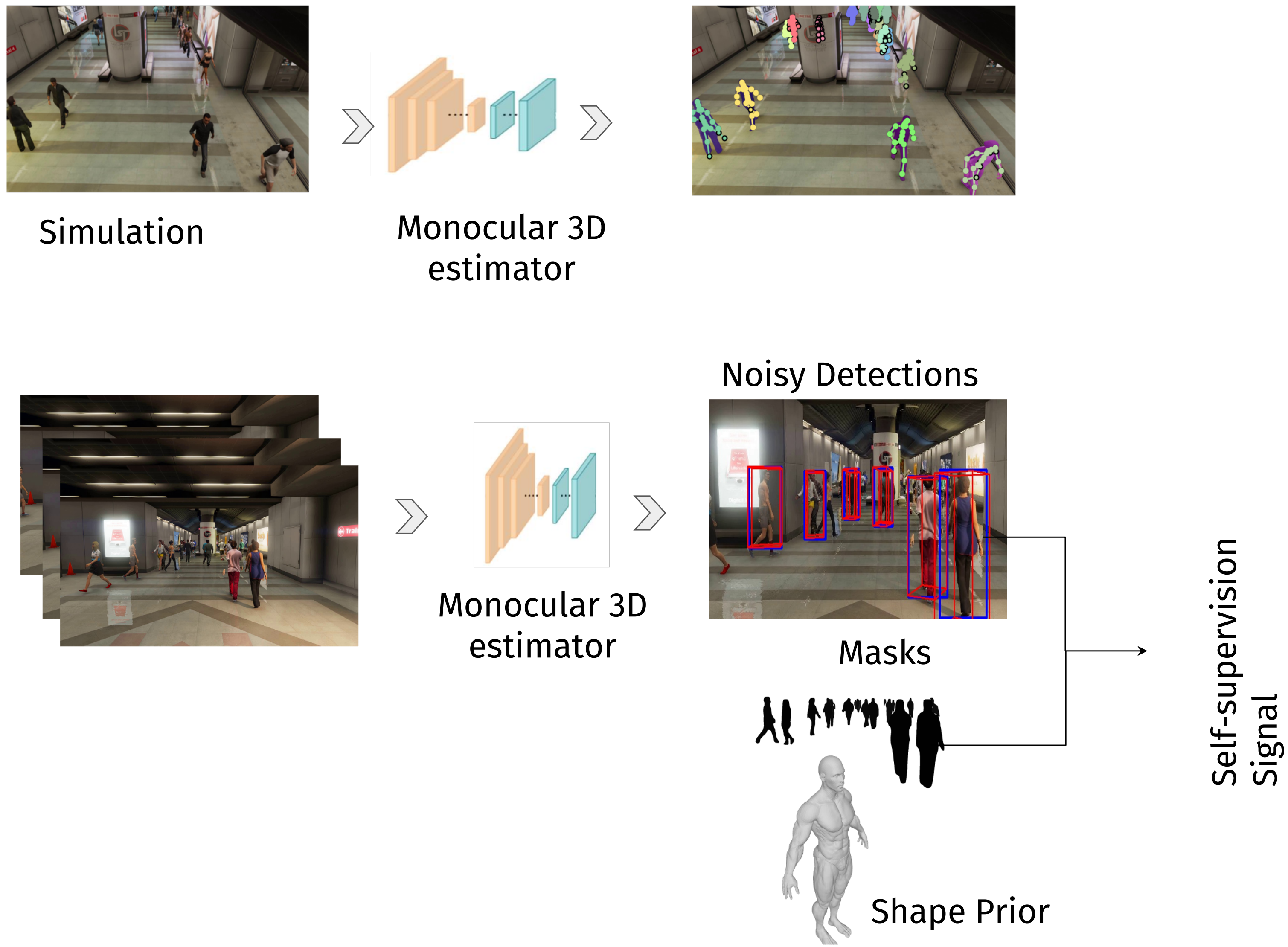
UniGe | MaLGa

Current directions

Lack of labels (self supervision)

Supervised pre-training on Synthetic dataset

Self-Supervised fine-tuning on Real dataset

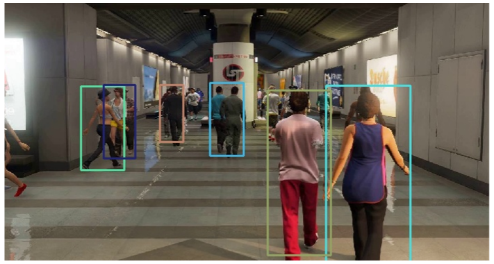


Current directions

Self supervision with humans: challenges

Rendering deformable humans

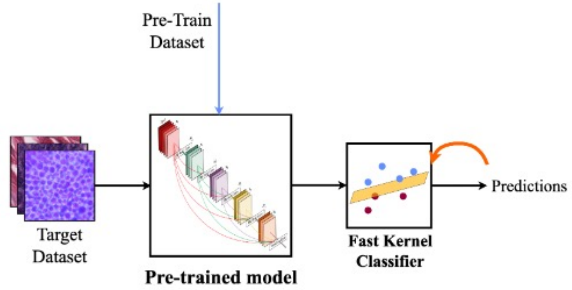
Viewpoint shift



Tracking in crowded scenes

Current directions

Lack of data and efficient training



- Address limited availability of data by **transferring pre-trained features to new task**
- Control training time by adopting efficient kernel-based algorithms
- We obtained comparable results to complex fine-tuning modalities in image classification
- We are currently addressing a similar procedure in action classification

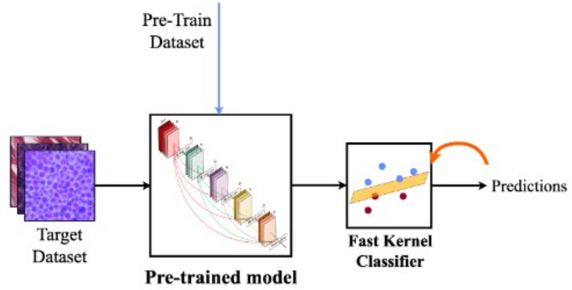
$$\underbrace{\Phi_{C+L} \circ \ldots \Phi_{C+1}}_{\text{Fully connected layers}} \circ \underbrace{\Phi_C \circ \ldots \circ \Phi_1(x)}_{\text{Convolutional layers}}.$$

$$\Phi_{TT} = \underbrace{\Psi}_{\text{Kernel feature map}} \circ \overbrace{\underbrace{\Phi_C \circ \ldots \circ \Phi_1(x)}_{\text{Convolutional layers}}}^{\textit{Feature vector Z}}.$$

$$\hat{f}(z) = \min_W \sum_{i=1}^{n} \|W z_i - \mathbf{y}_i\|^2 + \lambda \|W\|_F^2$$

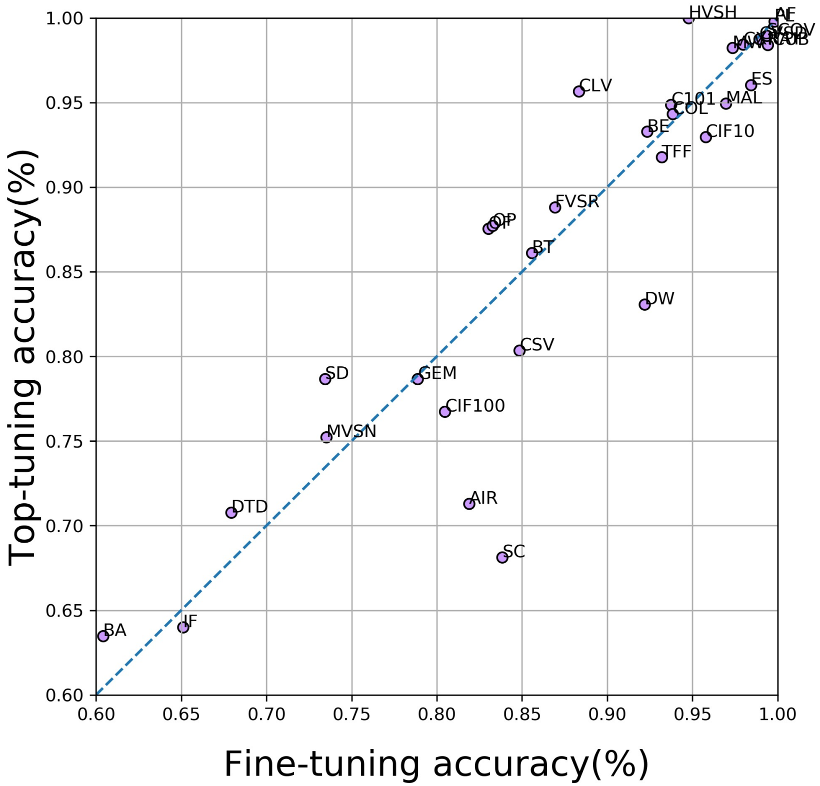
Fine-tuning or top-tuning? Transfer learning with pretrained features and fast kernel methods
PD Alfano, et al - arXiv preprint arXiv:2209.07932, 2022

Current directions



Lack of data and efficient training

Dataset name	#images (Tr/Te)	Img. size mean	#classes
AFHQ (AF)[58]	13.167/1.463	512 × 512	3
Beans (BE)[59]	1.167/128	500 × 500	3
Best artworks (BA)[60]	7.896/878	980 × 921	50
Boat types (BT)[61]	1.315/147	905 × 1234	9
Caltech-101 (C101)[62]	3.060/6.084	251 × 282	102
Cassava (CSV)[63]	7.545/1.885	573 × 611	5
Cats vs Dogs (CVSD) [64]	20.935/2.327	365 × 410	2
Chest xray (CXRAY) [65]	4.708/524	968 × 1321	2
CIFAR10 (CIF10) [66]	50.000/10.000	32 × 32	10
CIFAR100 (CIF100) [66]	50.000/10.000	32 × 32	100
Citrus leaves (CLV) [67]	534/60	256 × 256	4
Colorectal hist (COL) [68]	4.500/500	150 × 150	8
Deep weeds (DW) [69]	15.758/1.751	256 × 256	9
DTD (DTD)[70]	3.760/1.880	453 × 500	47
EuroSAT (ES) [71]	24.300/2.700	64 × 64	10
FGVC Aircraft (AIR)[72]	6.667/3.333	353 × 1056	100
Footb vs Rugby (FVSR) [73]	2.203/245	618 × 788	2
Gemstones (GEM) [74]	2.571/286	330 × 335	87
Hors or Hum (HVSH) [75]	1.027/256	300 × 300	2
iCubWorld subset (ICUB)[38]	86.400/9.600	256 × 256	10
Indian Food (IF) [76]	3.600/400	550 × 610	80
Make No Make(MVSN)[77]	1.355/151	211 × 246	2
Malaria (MAL) [78]	24.802/2.756	133 × 132	2
Meat quality (MQA) [79]	1.706/190	720 × 1280	2
Oxford Flowers (OF) [80]	2.040/6.149	538 × 624	102
Oxford-IIIT Pets (OP) [81]	3.680/3.669	383 × 431	37
Plankton (PL) [82]	4.500/500	106 × 120	10
Sars Covid (SCOV) [83]	2.232/249	260 × 350	2
Stanford Cars (SC) [84]	8.144/8.041	308 × 573	196
Stanford Dogs (SD) [85]	12.000/8.580	386 × 443	120
Tensorflow Flowers(TFF) [86]	3.303/367	272 × 365	5
Weather (MW) [87]	1.012/113	335 × 506	4



Fine-tuning or top-tuning? Transfer learning with pretrained features and fast kernel methods
PD Alfano, et al - arXiv preprint arXiv:2209.07932, 2022

Current directions

Anticipation

Task: understand motion cues anticipating the goal of an action

Here we show preliminary results obtained by reasoning on the direction of sight (approximated by heading direction) and the structure of the scene (the presence of table and objects)

Wrap up and directions

Understanding human motion involves analysis at multiple levels

Massive amount of data are allowing us to address tasks in an end-to-end manner, often relying on transfer learning or on generative techniques

In some tasks the ability of accessing intermediate outcomes is crucial. So far we have done it in a composite manner, the road is open to extract heterogeneous intermediate information from large networks

UniGe

MaLGa

http://malga.unige.it