Master in
Computer Vision
Barcelona

Module: Video Analysis

Lecture 10: Human pose & activity recognition

Lecturer: Javier Ruiz Hidalgo

Outline

- Introduction
- Gesture/action recognition system
  - Capture
  - Analysis
  - Classification
- Conclusions
- Demonstration

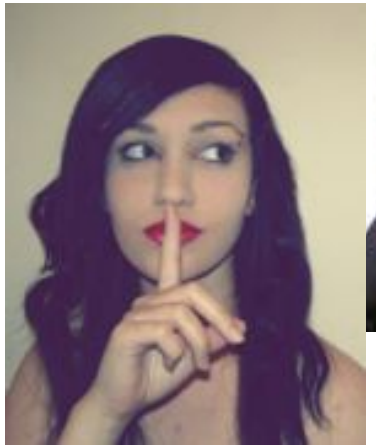
GOAL: Understanding of theoretical and practical aspects of the capturing and recognition of human activity, pose & gestures in video sequences

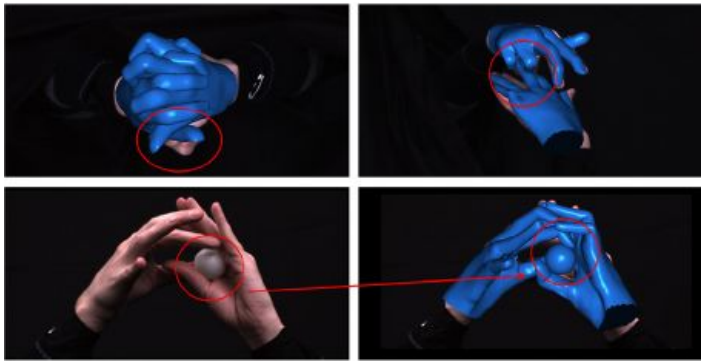
Introduction

- We humans are very good understanding human pose / motion
  - Gunnar Johansson: Gestalt of biological motion (1970)

Types of gestures (1)

- Traditional gestures
  - **STATIC (POSE)** vs dynamic
    - pose, skeleton





Types of gestures (2)

- Traditional gestures
  - static vs **DYNAMIC**

Types of gestures (3)

- Dynamic gestures
  - Simple activities: walking, running, seating, jumping

  - Complex activities, hierarchical roles (sports)

Humans can identify much more!

- Identity
- Social cues, dominance
- Emotion
- Sexual orientation
- Vulnerability to attack
- Intent to deceive

http://www.biomotionlab.ca/Demos/BMLwalker.html

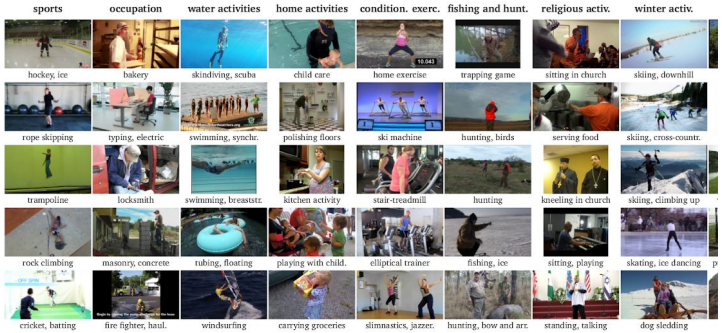
Applications

- Human-Computer Interaction (HCI)
- Deaf people assistance
- Synthesis and Animation (films, computer games)
- Surgery / Medical applications
- Virtual reality

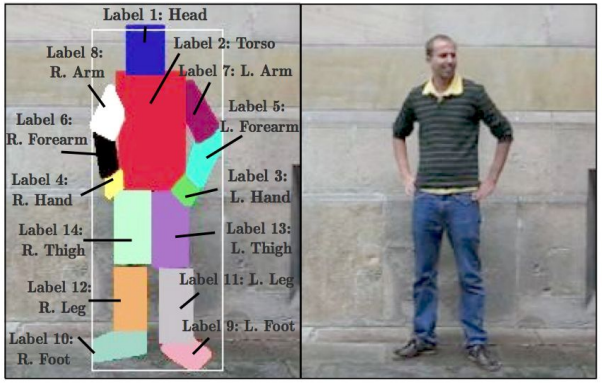
Oculus Rift

Minority Report

Datasets (1)



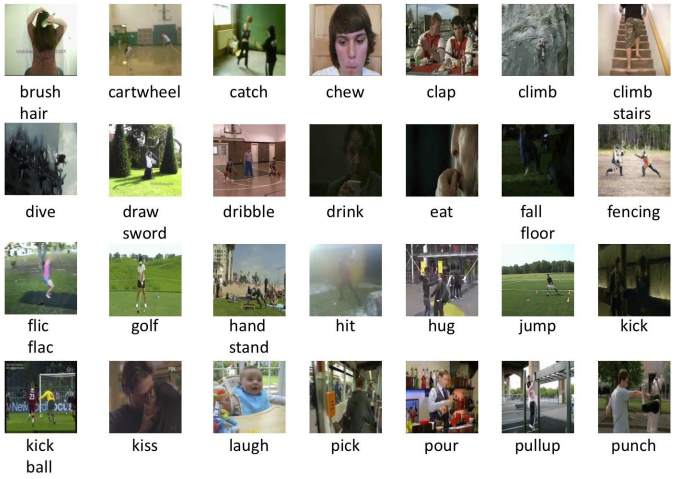
- Static / pose
  - COCO keypoint dataset
    - 250k people with annotated body joints.
  - MPII Human Pose
    - 40K people with annotated body joints.
  - Leeds Sports Pose
    - 2k pose annotated images of mostly sports people.
  - Chalearn Human Pose
    - 8k images labelled at pixel precisions with 14 limbs.
  - Hand gesture dataset
    - 2.5k images with ASL gestures.



Datasets (2)

- Dynamic
  - PoseTrack21
    - 177k annotated skeletons in videos with ID
  - Chalearn gesture recognition
    - 48k RGBD videos with 249 gestures labels performed by 21 different individuals.
  - Chalearn multimodal gesture recognition
    - 14k RGBD videos from a vocabulary of 20 Italian sign gesture categories.
  - Sheffield Kinect Gesture (SKIG)
    - 2k RGBD videos from 6 subjects and 10 categories of hand gestures.
  - National center sign language
    - 3k videos with signs of ASL

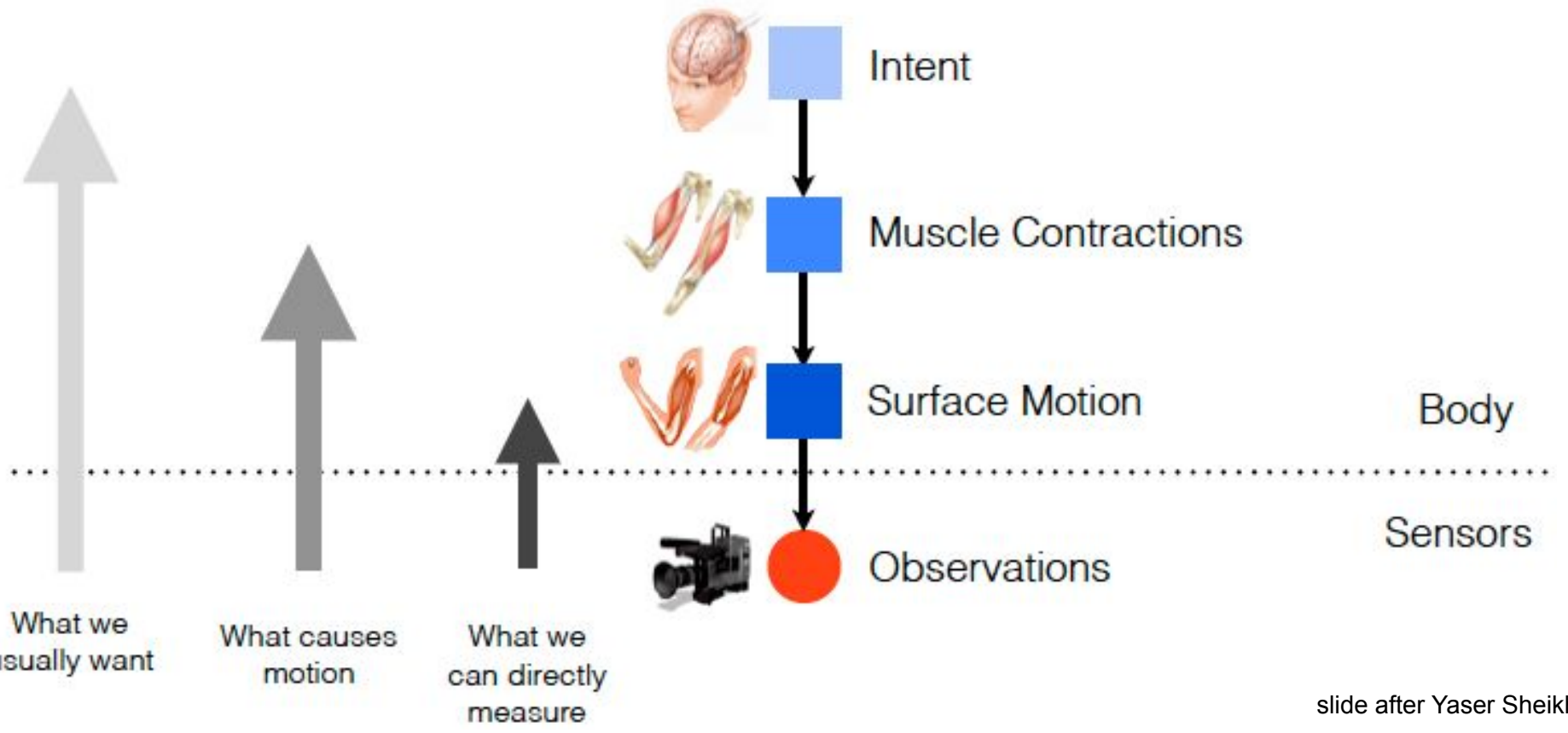
Datasets (3)



- Activities
  - UCF-101
    - 13k RGB videos with 101 actions.
  - HMDB-51
    - 7k RGB videos with 51 actions.
  - ActivityNet-200
    - 28k RGB videos with 200 actions.
  - Sports-1M
    - 1M YouTube videos with 487 sport classes. (ex.)
  - Kinetics
    - 300/400/600 action labels from YouTube videos

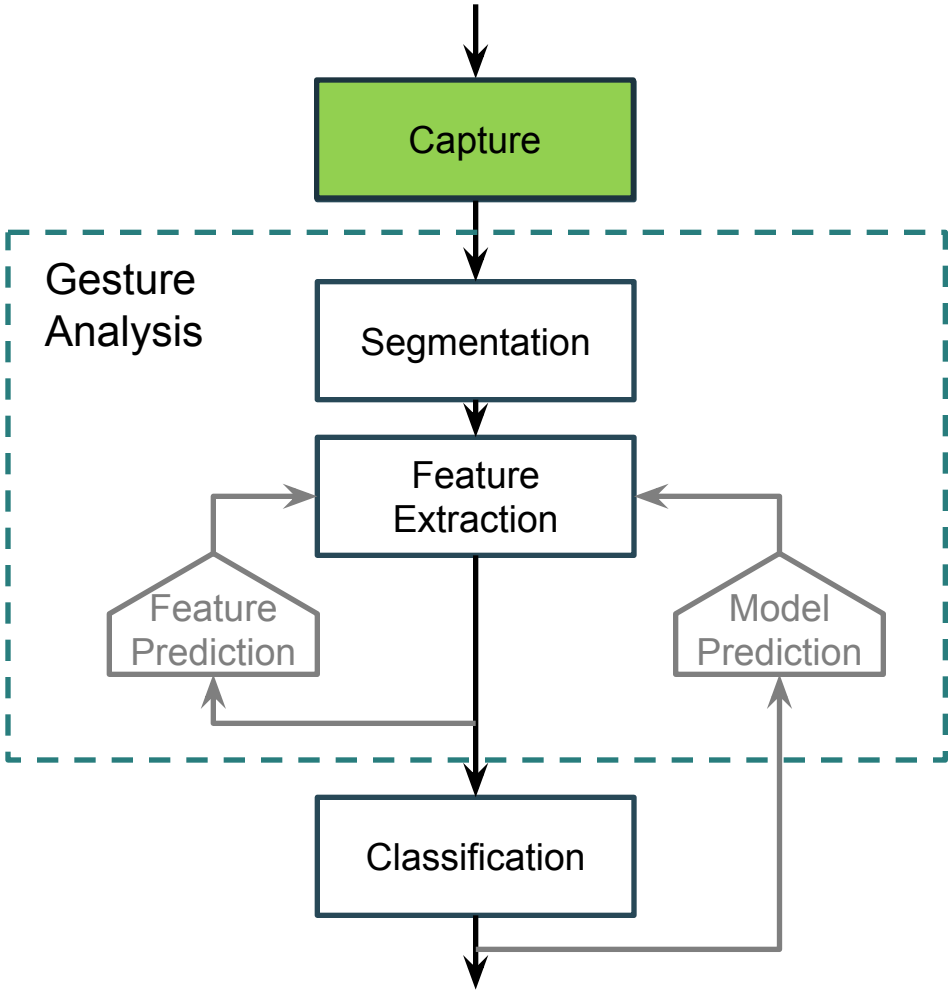
What makes Human Motion Hard to Analyse?

- Human motion is any muscular contraction of the human body
- Hard to measure because muscular contractions are **hidden**
- Intent is also **hidden**



slide after Yaser Sheikh

Pose/Gesture recognition general scheme



Capture (1)

- Marker/suit based
  - Electro-Mechanical/Magnetic
  - Inertial
  - Optical
    - Markers are positioned on the body

Rogue One

Avatar

Marker/suit motion capture (1)

Pro:

- Captures motion in high detail
- Easy to get data from complex motions and with subtle emotional content

Lord of the Rings

- Proved beneficial for:
  - Visual FX, Games, biomechanical analysis, recognition

slide after Yaser Sheikh

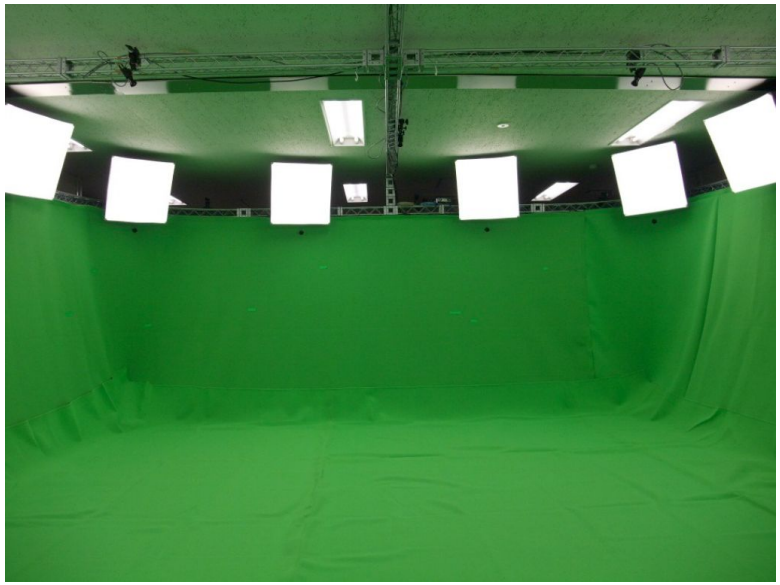
Marker/suit motion capture (2)

Con:

- Difficult to edit (animators)
- Unpopular for feature animated films (uncanny valley)
- Limited in use (position of cameras)
- Could be expensive (setup + hardware)
- Highly intrusive

Capture (2)

- Marker-less based
  - Video based motion capture
  - Sensors
    - Colour cameras: single, stereo, multi-view
    - Depth



Marker-less based motion capture (1)

- The marker systems could be very expensive, difficult to set up and, more importantly, very invasive.
- We would like to capture motion from video
- PERFECT SCENARIO: Use a single camera with no markers!

slide after Raquel Urtasun

Marker-less based motion capture (2)

- Why is it difficult?
  - Poor imaging: motion blurred, occlusions, bad correspondences
  - Mapping multimodal
    - One image observation can represent more than one pose!

Agarwal and Triggs, 2005

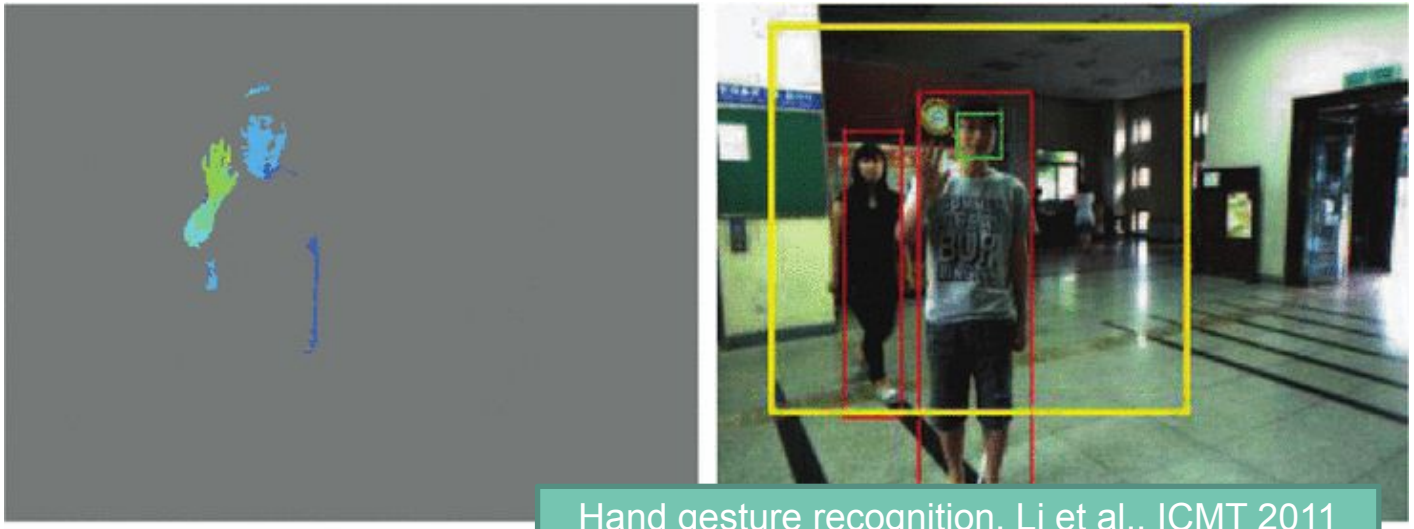
Sensors for marker-less motion capture (1)

- Colour cameras
  - Ideal, single sensor, easy setup
  - Occlusions, image quality, multimodal

- First solution → Multiple cameras
  - Stereo vision
  - Multi-view

Stereo cameras

- Stereo vision



Hand gesture recognition, Li et al., ICMT 2011

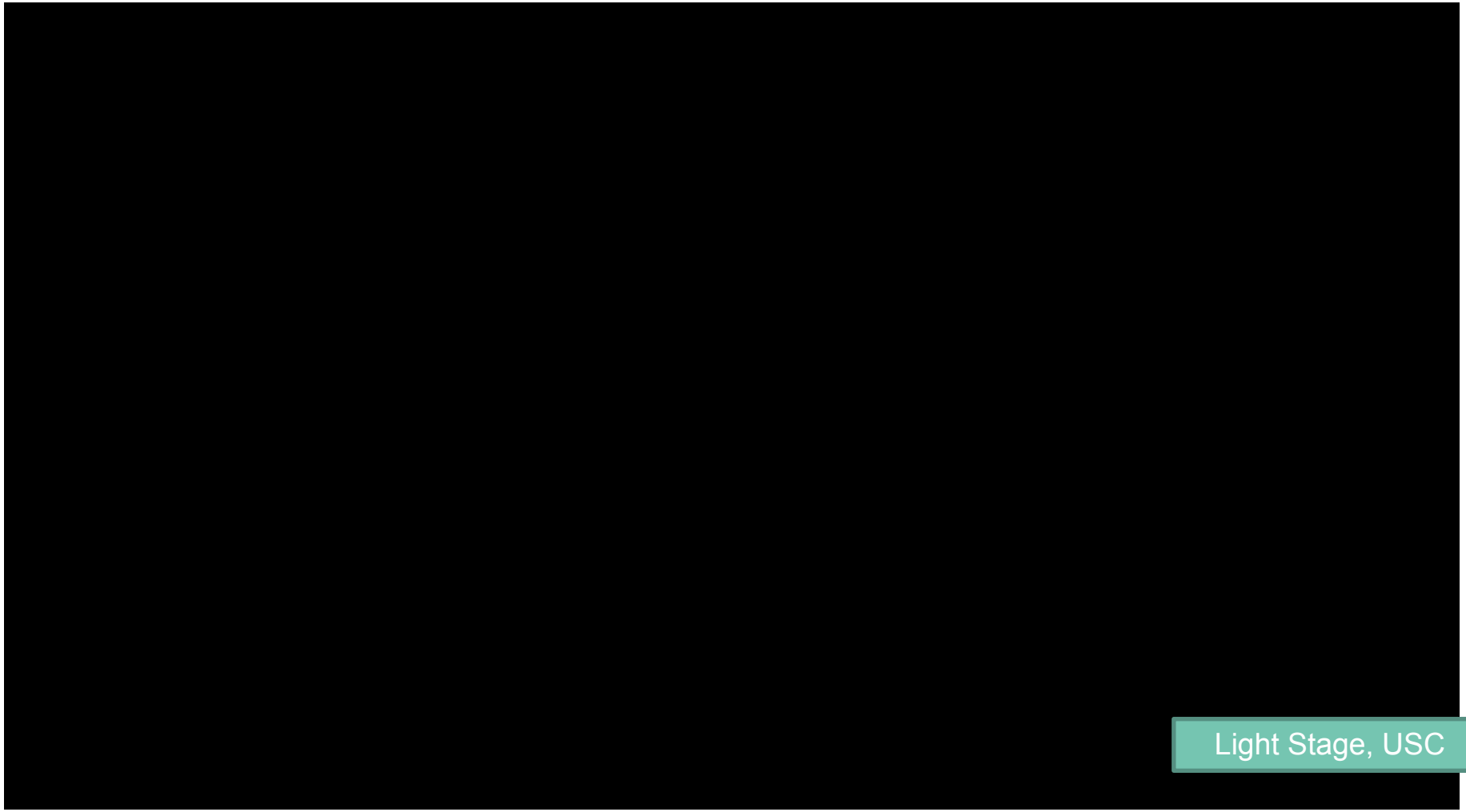
Multi-view setup (1)

- Array of cameras recording the subject
  - 6 cameras

SmartRoom, UPC

Multi-view setup (2)

- Array of cameras recording the subject



Light Stage, USC

Multi-view setup (3)

- Array of cameras recording the subject

On site 3D video capture, Nobuhara, 2009

- "Mobile"
- 16 cameras

Multi-view setup (4)

- Array of cameras recording the subject

Pro:

- High accuracy

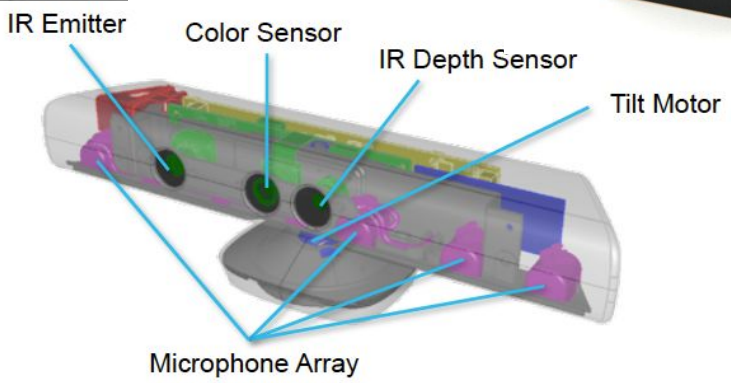
Con:

- Need of calibration
- Complex setup
- High computational power

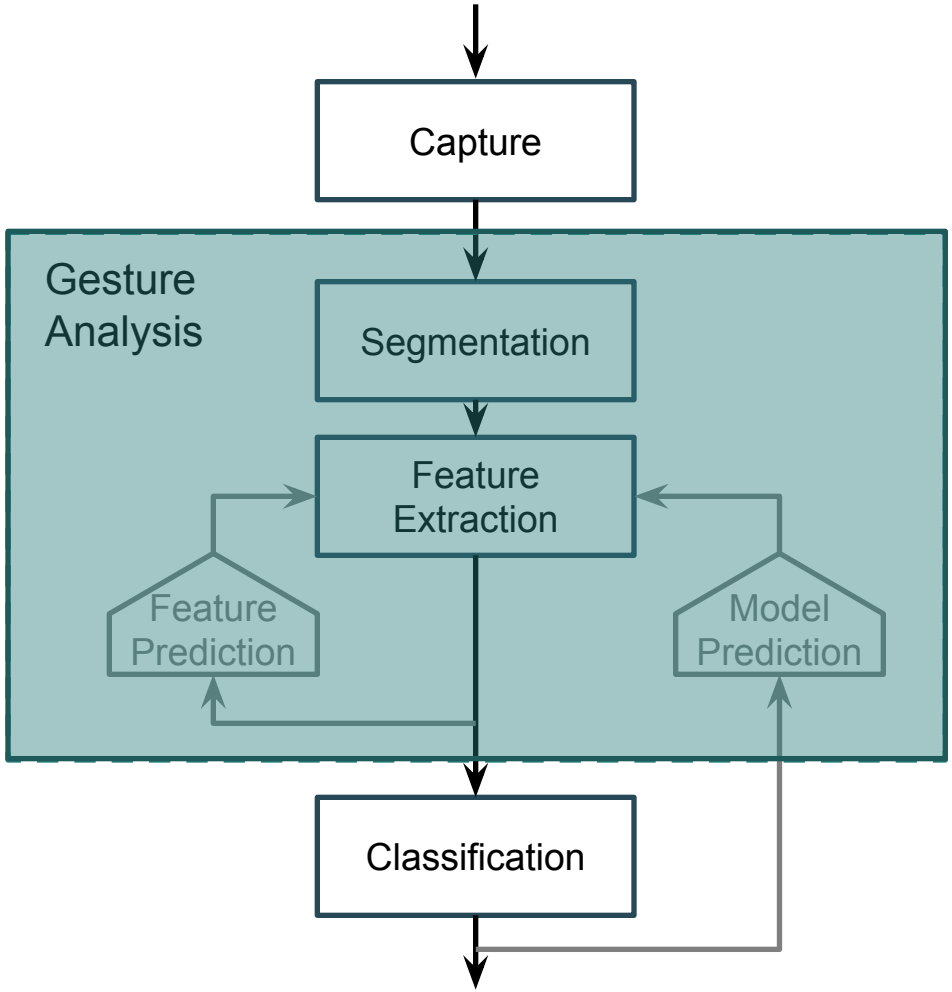
Hand capture, Ballan et al., ECCV 2012

Depth sensors

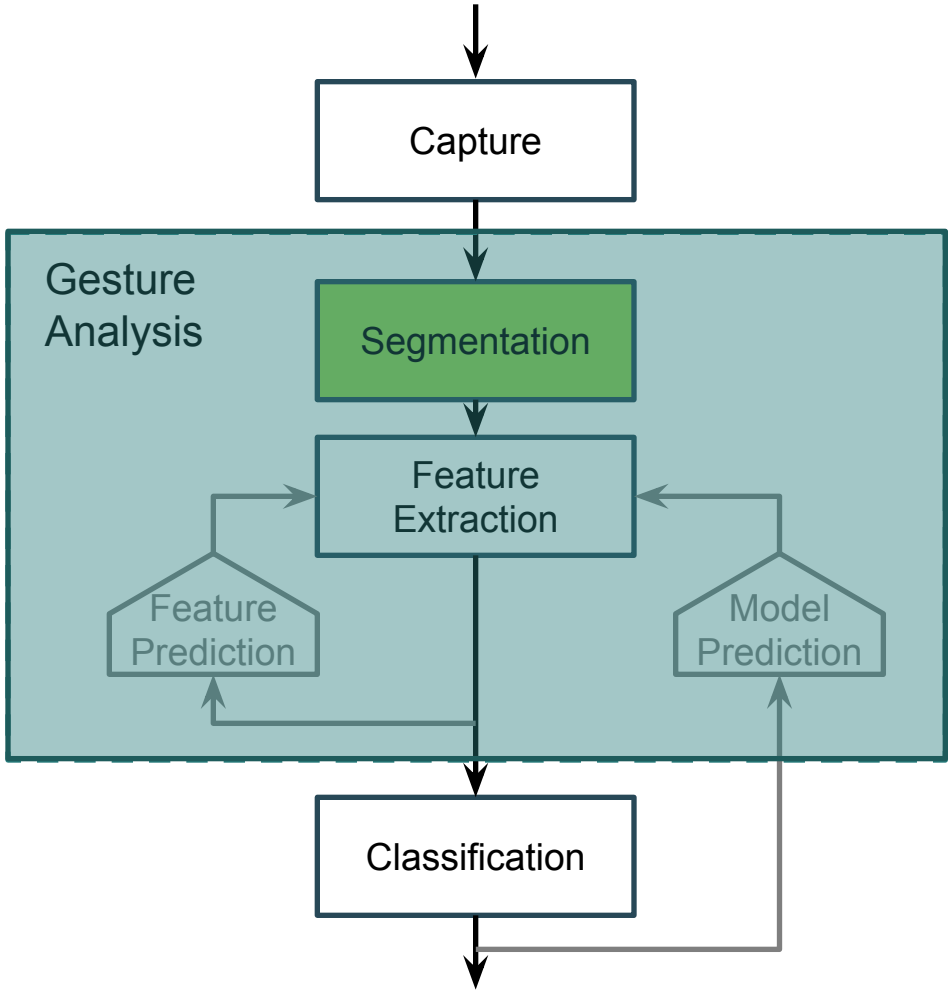
- Appearance of depth sensors have pushed the limits of gesture recognition technologies to move away from controlled scenarios
  - Time-of-flight, structured light, etc.



Gesture recognition general scheme



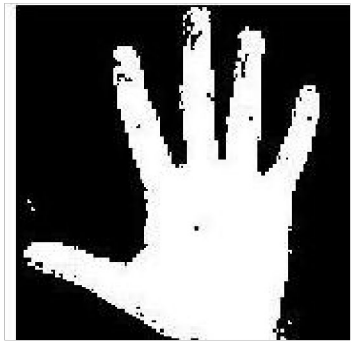
Gesture recognition general scheme



Gesture Analysis: Segmentation (1)

- Using colour cameras segmentation of humans/hands was very difficult
  - Background modelling (Stauffer & Grimson)
  - Use controlled scenarios (chroma-key)
  - Segmentation networks (pre-training, overfitting)

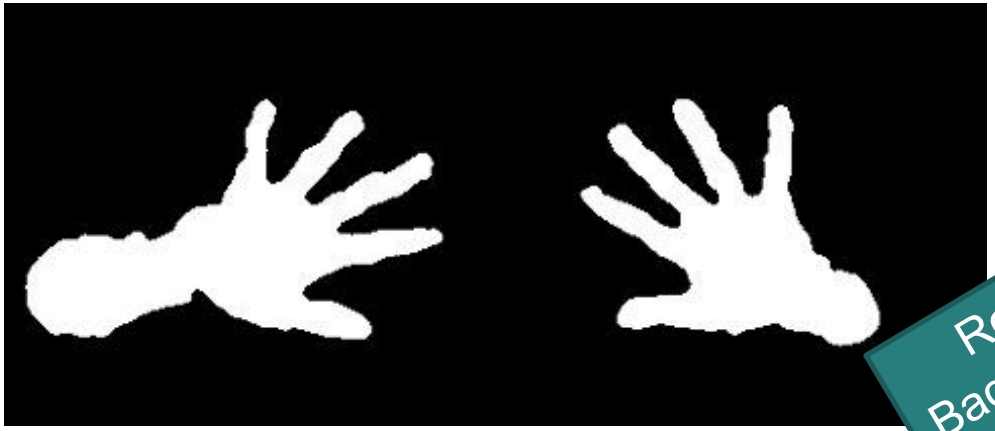
Original Image



Segmented Hand

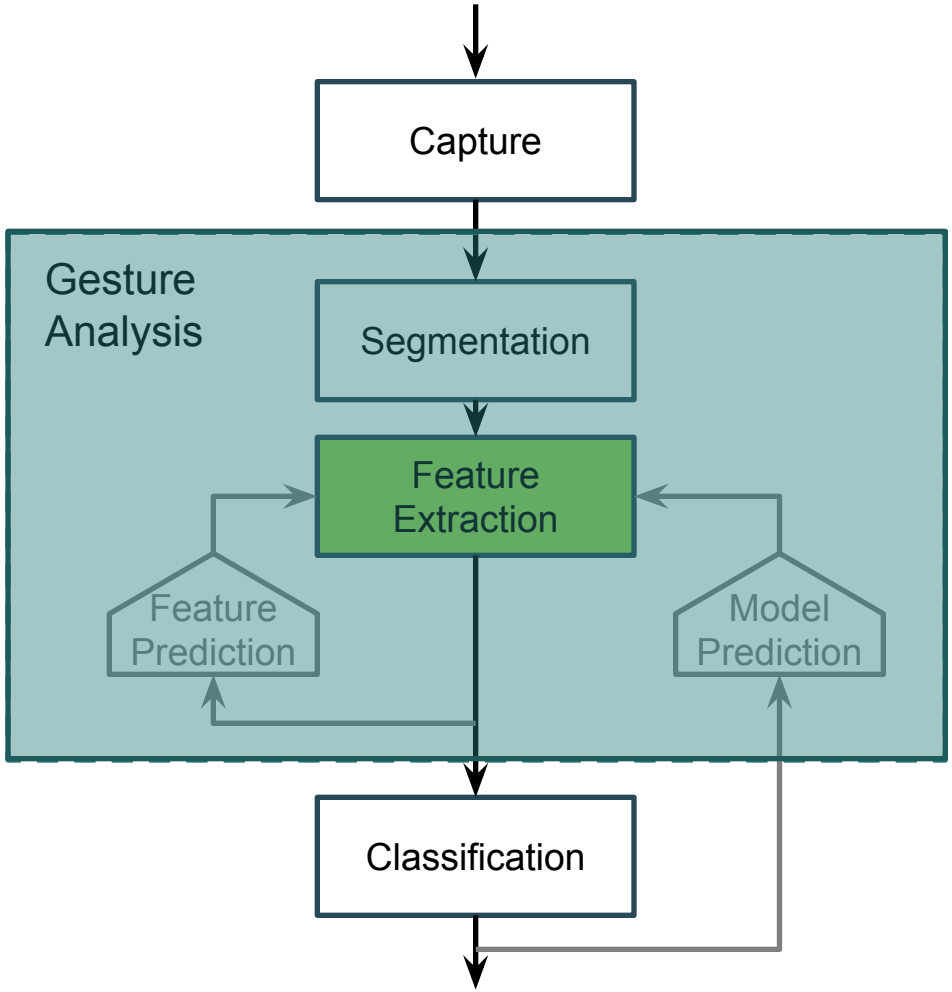
Gesture Analysis: Segmentation (2)

- Depth much better
  - Segmentation is almost straight-forward
  - Stereo & multiview → Depth
  - Use ToF or Kinect sensors



Robust against Background changes

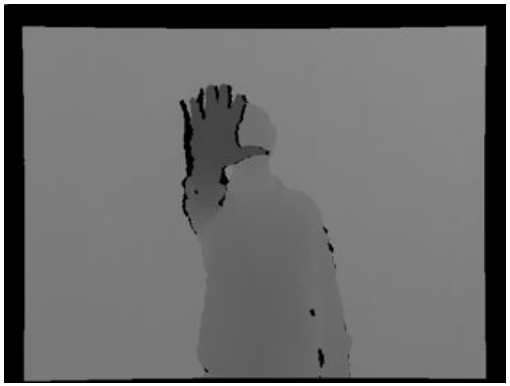
Gesture recognition general scheme



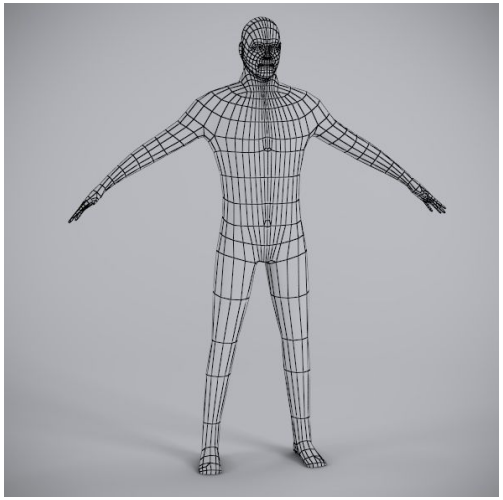
Gesture Analysis: Static features (1)

- 2D vs 3D
  - Extract features from:

colour



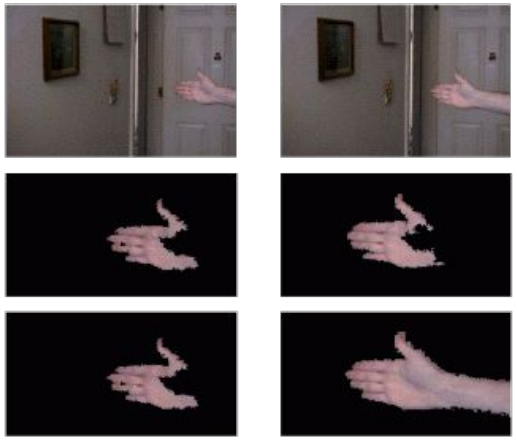
depth



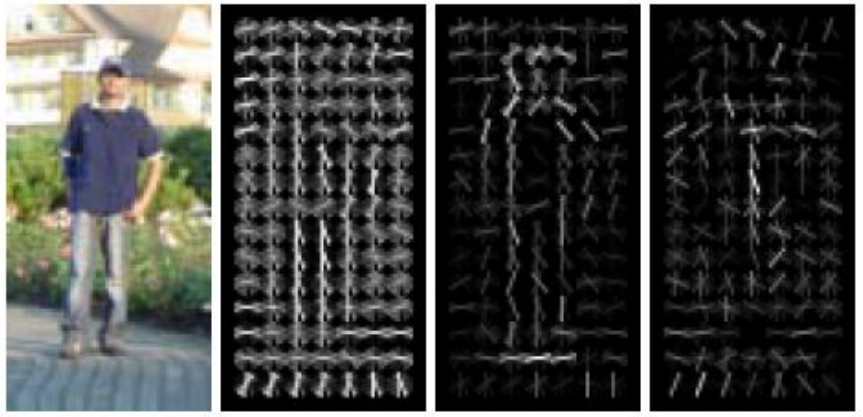
3D
(meshes, points, etc.)

Gesture Analysis: Static features (2)

- 2D features
  - Dominant colour (skin)

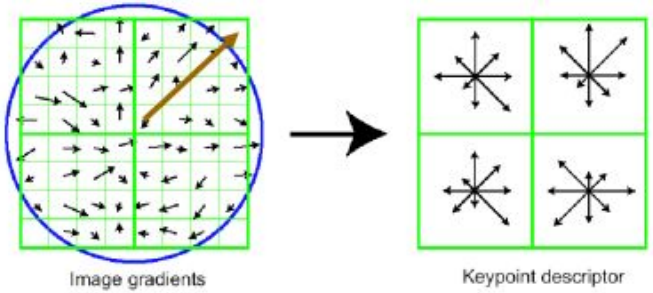


  - Histogram of Oriented Gradients (HOG) [Dalal and Triggs,2005]
    - Pyramid of HOG [Bosch et al., 2007]

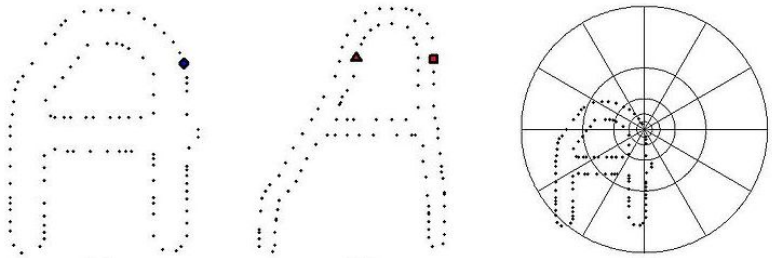


Gesture Analysis: Static features (3)

- 2D features
  - SIFT [Lowe, 2004]
    - Invariant to image translation, scaling and rotation
    - PCA to reduce dimensionality



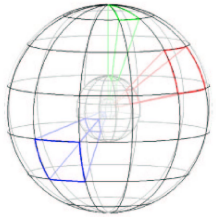
  - Shape context (Belongie and Malik, 2000)
    - Distribution of relative positions on the contours



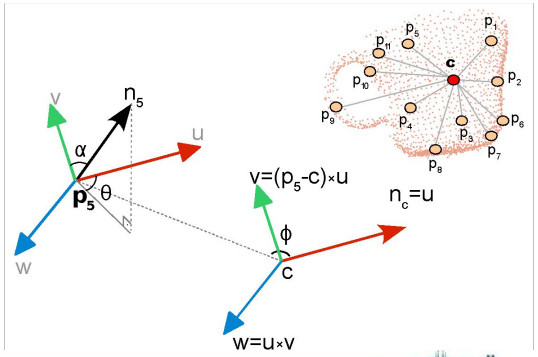
Gesture Analysis: Static features (4)

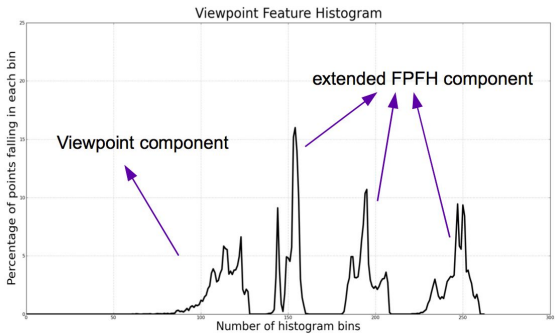
- 3D features
  - 2D descriptors (SIFT, etc.) are not suited for depth images
    - Low texture, sharp edges, low resolution

- 3D descriptors:
  - 3D Shape Context [Frome and Malik, 2004]
    - Extension to 3D



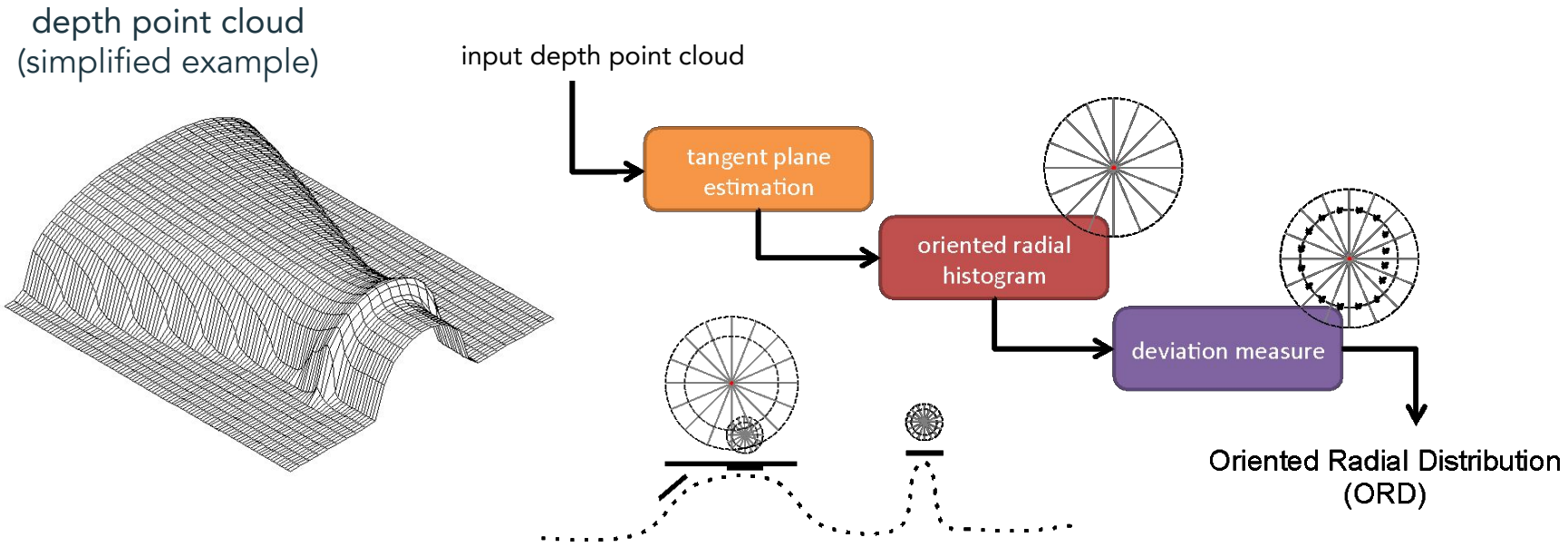
  - Viewpoint Feature Histograms (VFH) [Rusu et al. 2010]
    - Histogram of the relative angles between surface and centroid normals





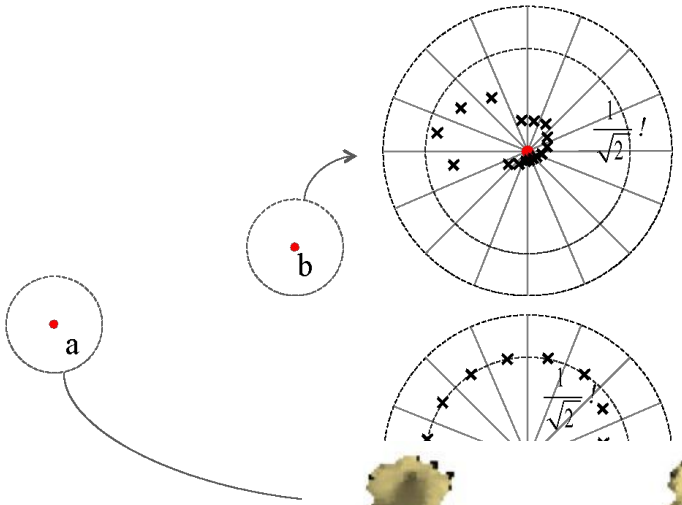
Gesture Analysis: Static features (5)

- 3D descriptors (cont.)
  - Oriented Radial Distribution (ORD) [Suau et al. 2012]
    - Measures the curvature (prominent or flat zones) of 3D points

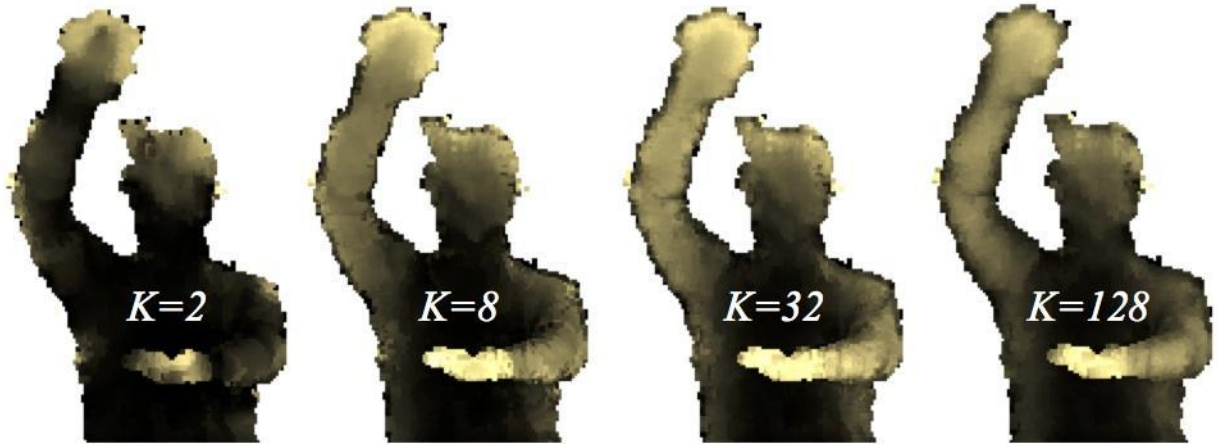


Gesture Analysis: Static features (6)

- Oriented Radial Distribution (ORD) cont.

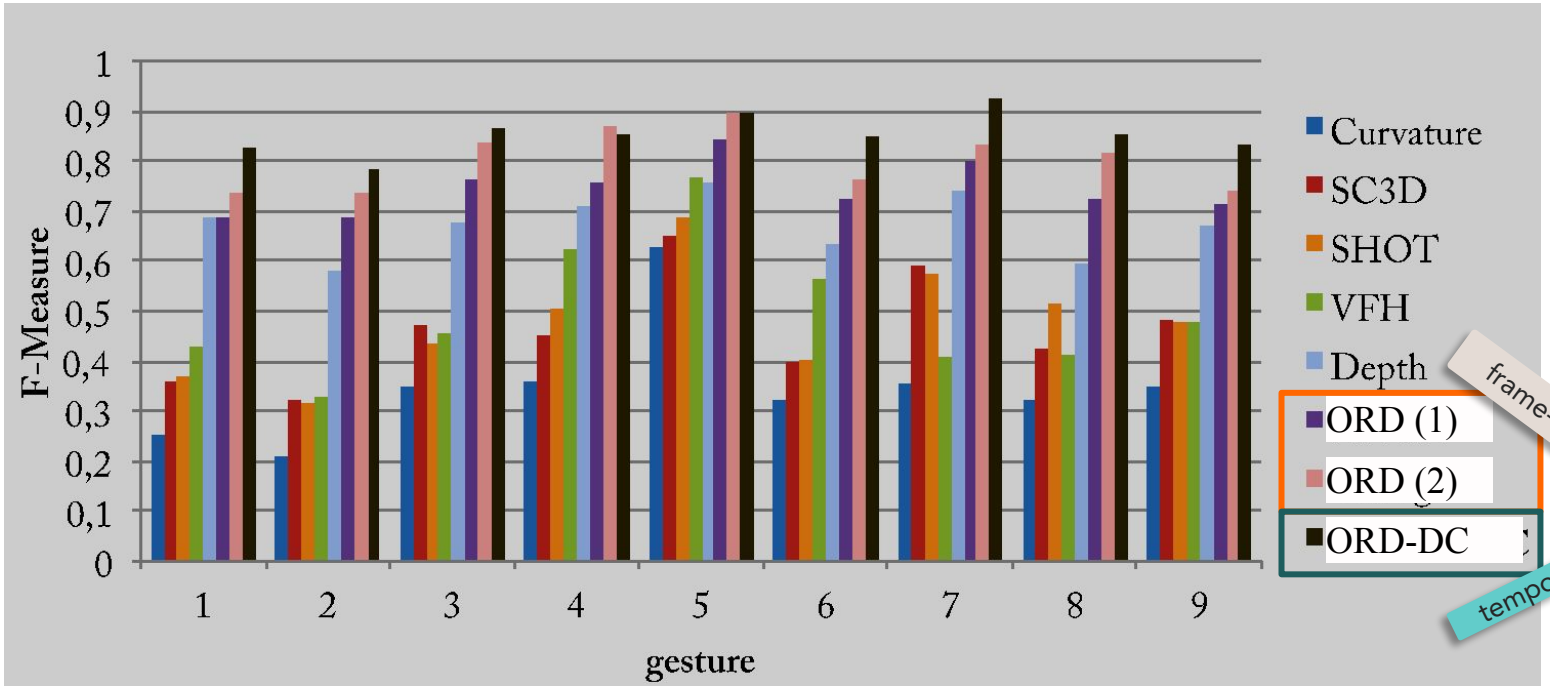


a : low deviation → FLAT
b : high deviation → PROMINENT



Gesture Analysis: Static features (7)

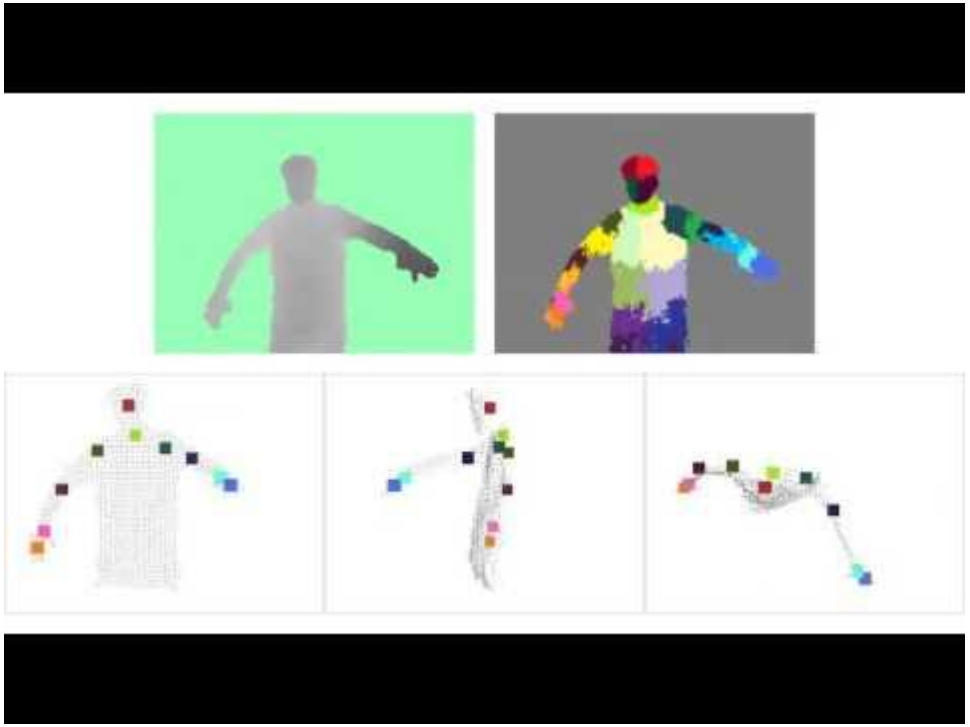
Gesture Recognition
Comparison between using a 3D feature benchmark and ORD



Suau et al., 2012

Gesture Analysis: Learning

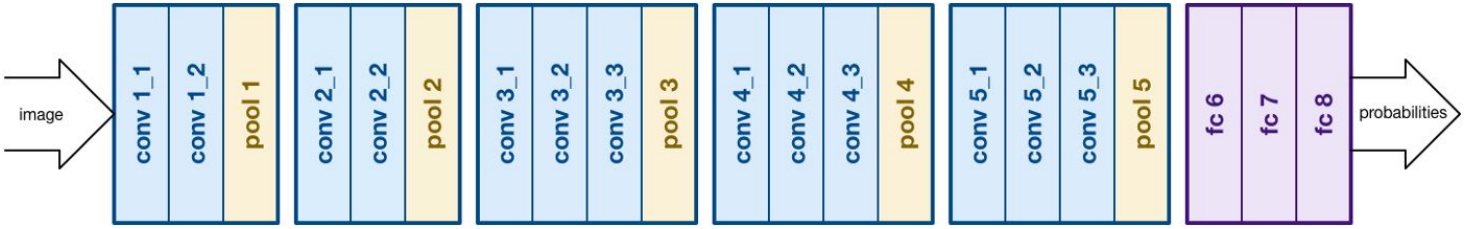
- Machine learning approaches (training database)
- Work of Shotton et al. 2011 fundamental:
  - Use kinect (depth data)
  - Gigantic DB (900k images of synthetic data) as ground truth
  - Train random forest to label body parts



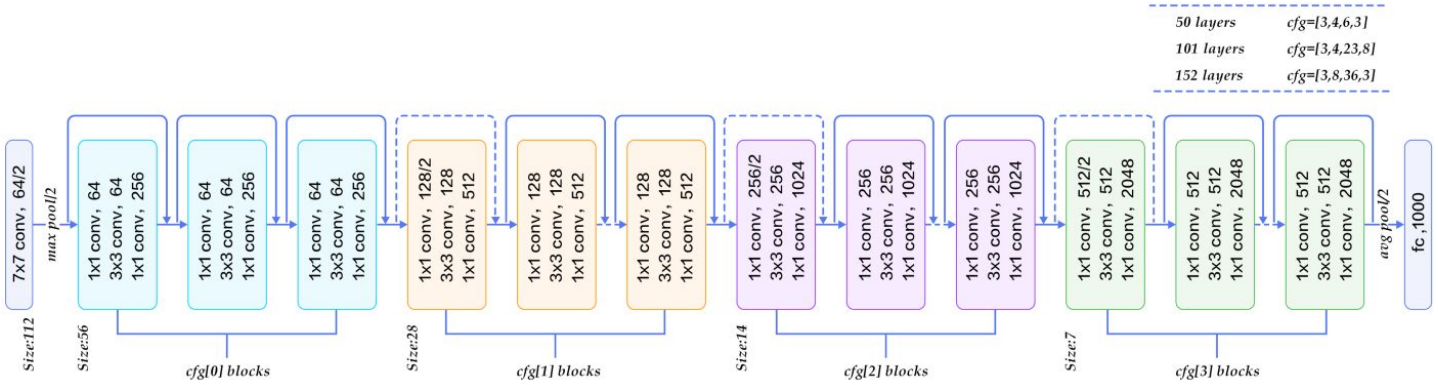
https://www.youtube.com/watch?v=IntbRsi8lU8

Gesture Analysis: Static features (7)

- Learned features (2015+)
  - VGG (pretrained on ImageNet)
    - K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, CoRR, 2014.



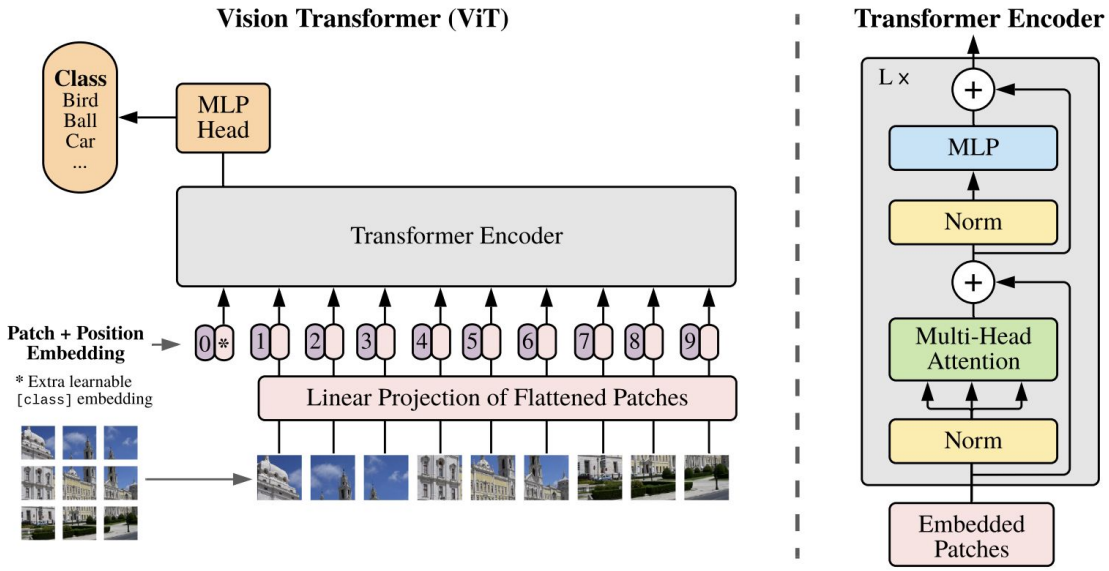
  - ResNet (pretrained on ImageNet)
    - H. Kaiming et al., Deep Residual Learning for Image Recognition, 2015.



Credit images

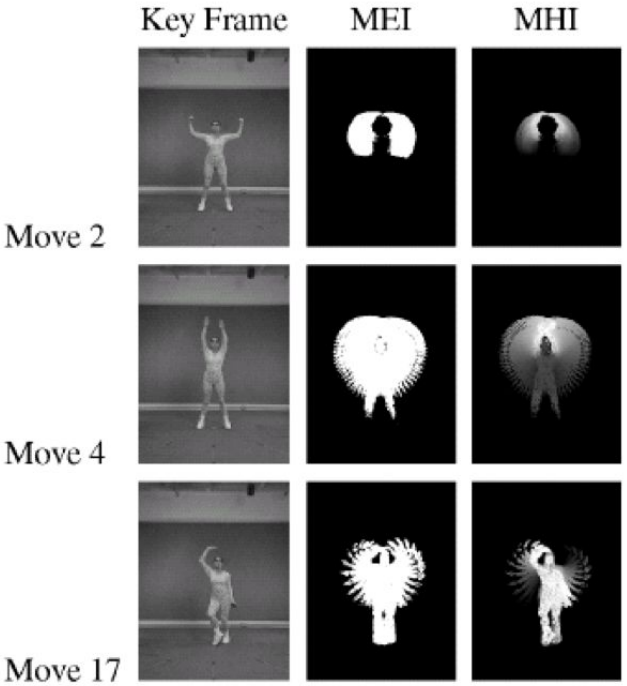
Gesture Analysis: Static features (8)

- Learned features (2021+)
  - ViT (pretrained on ImageNet)
    - A. Dosovitskiy, et al., An Image is worth 16x16 words: Transformers for image recognition at scale, ICLR, 2021.



Gesture Analysis: Dynamic features (1)

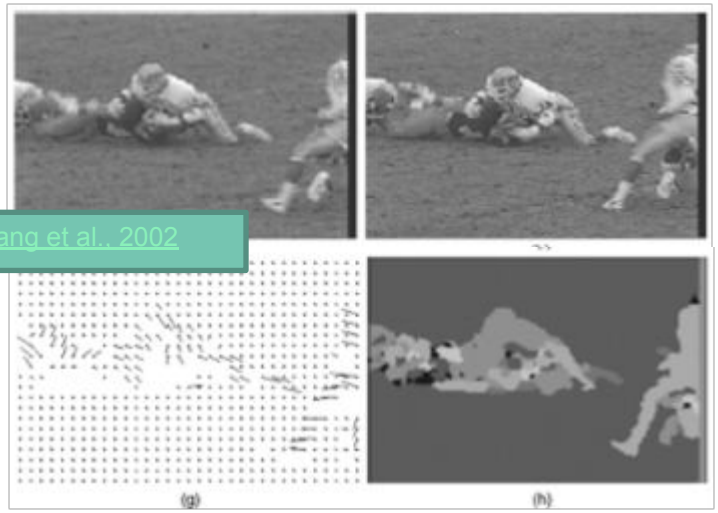
- Temporal templates
  - Representation of movement (static background)
    - Motion Energy Image (MEI)
      - stores info **where** motion has occurred in the image
    - Motion History Image (MHI)
      - Stores info **how** the motion is performed



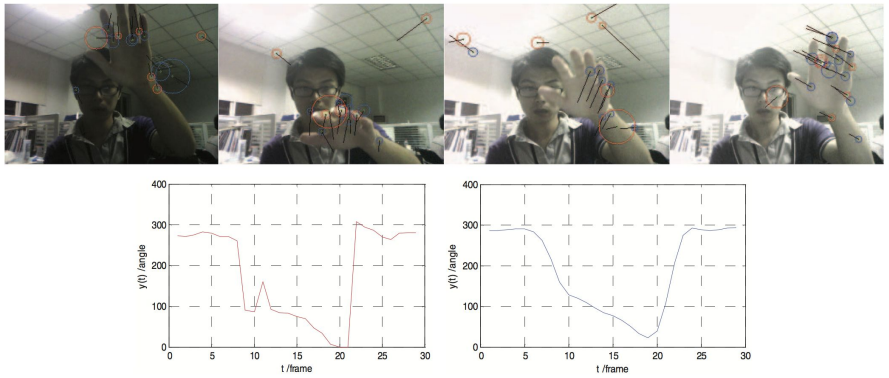
Bobick, et al. 1997

Gesture Analysis: Dynamic features (2)

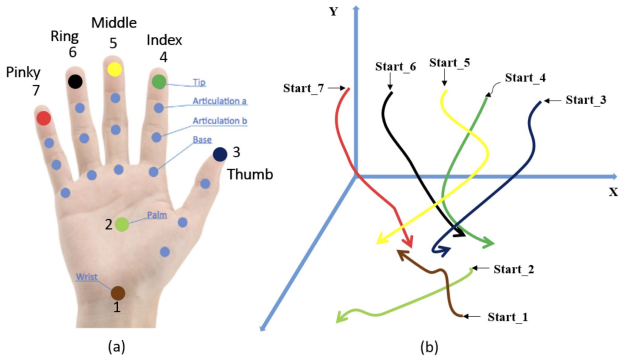
- Optical flow, trajectories,

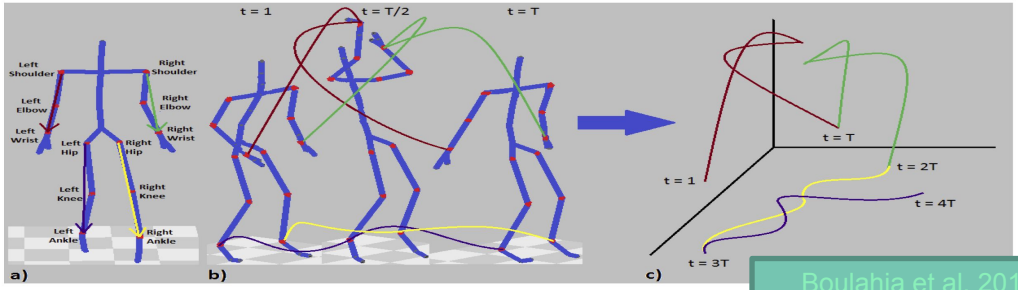


Yang et al., 2002



Bao, et al. 2011





Boulahia et al. 2017

Gesture Analysis: Dynamic features (3)

- Learned features (2015+)
  - C3D (pretrained on Sports-1M)
    - D. Tran, et al. Learning Spatiotemporal Features with 3D Convolutional Networks, CVPR, 2015.

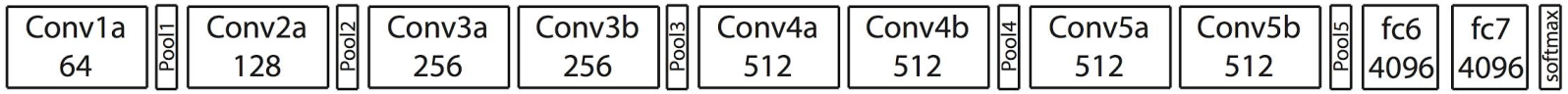
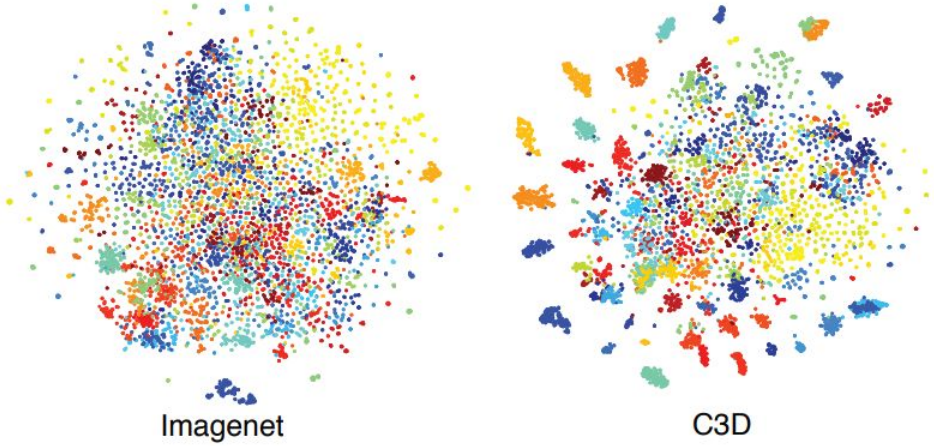


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.



Gesture Analysis: Dynamic features (4)

- Learned features (2021+)
  - Video Transformers (pre-trained in Kinetics and UCF101)
    - Z. Tong, et al. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, CVPR, 2022.

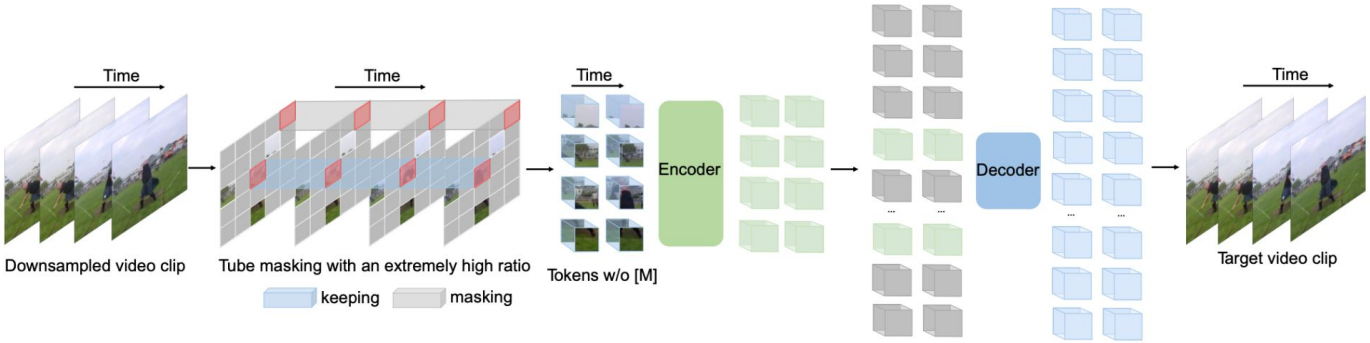
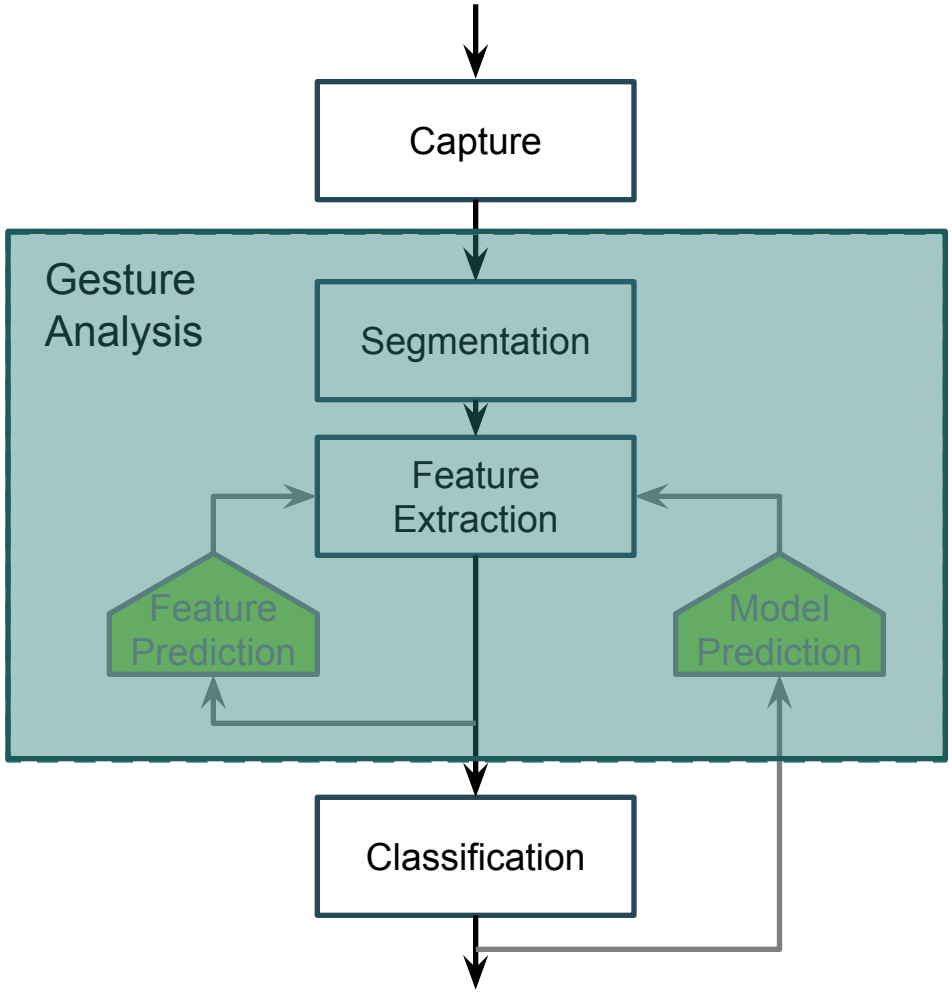


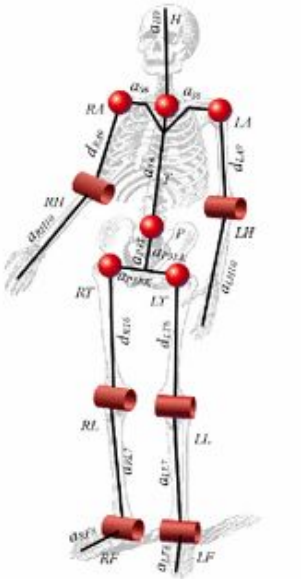
Figure 1: **VideoMAE** performs the task of masking random cubes and reconstructing the missing ones with an asymmetric encoder-decoder architecture. Due to high redundancy and temporal correlation in videos, we present the customized design of tube masking with an extremely high ratio (90% to 95%). This simple design enables us to create a more challenging and meaningful self-supervised task to make the learned representations capture more useful spatiotemporal structures.

Gesture recognition general scheme

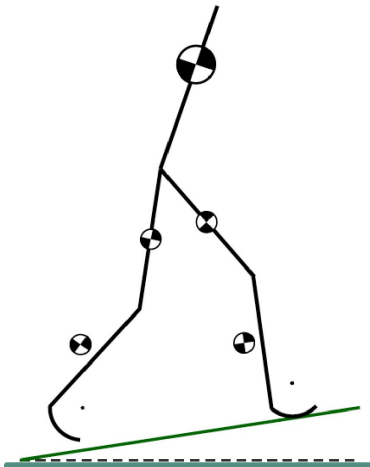


Gesture Analysis: Models (1)

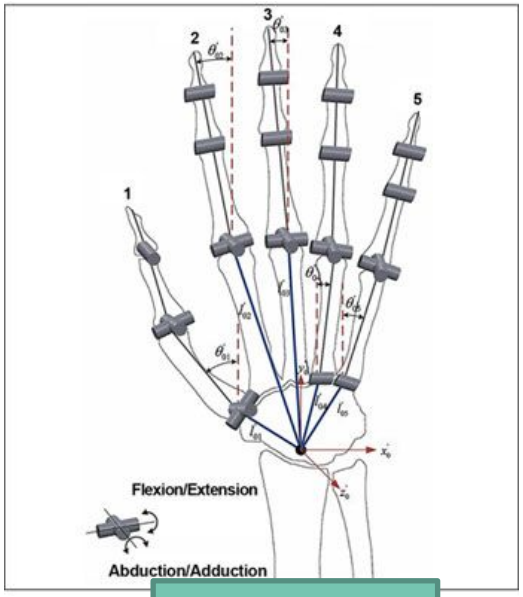
- Model vs. Appearance
  - Appearance based techniques use directly the extracted features (2D or 3D)
  - Model based techniques use the features to fit a model
    - 3D human models (hand & body)
    - Approximated by Kinematic chain (various DOF)



Koritnik et al., 2010



Brubaker and Fleet, 2008



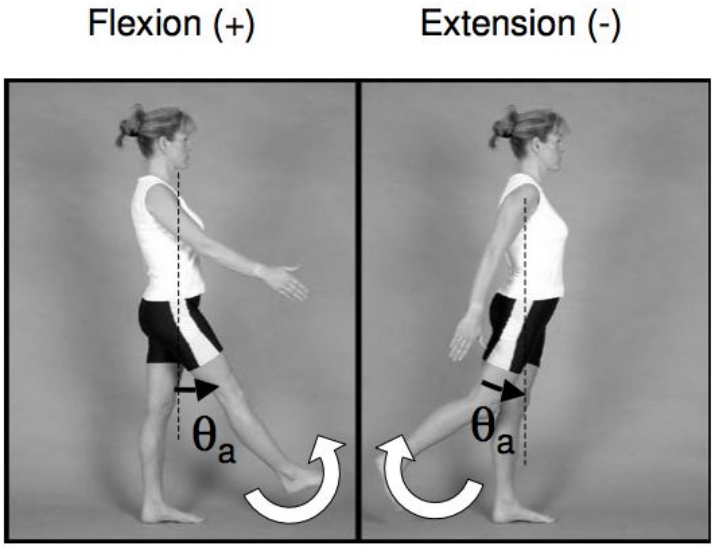
Li et al., 2011

Gesture Analysis: Models (2)

- 3D human models (hand & body)

Pro:

- Easier to generalize
- Able to put priors (physical constraints)
- Independent of marker placement / feature selection



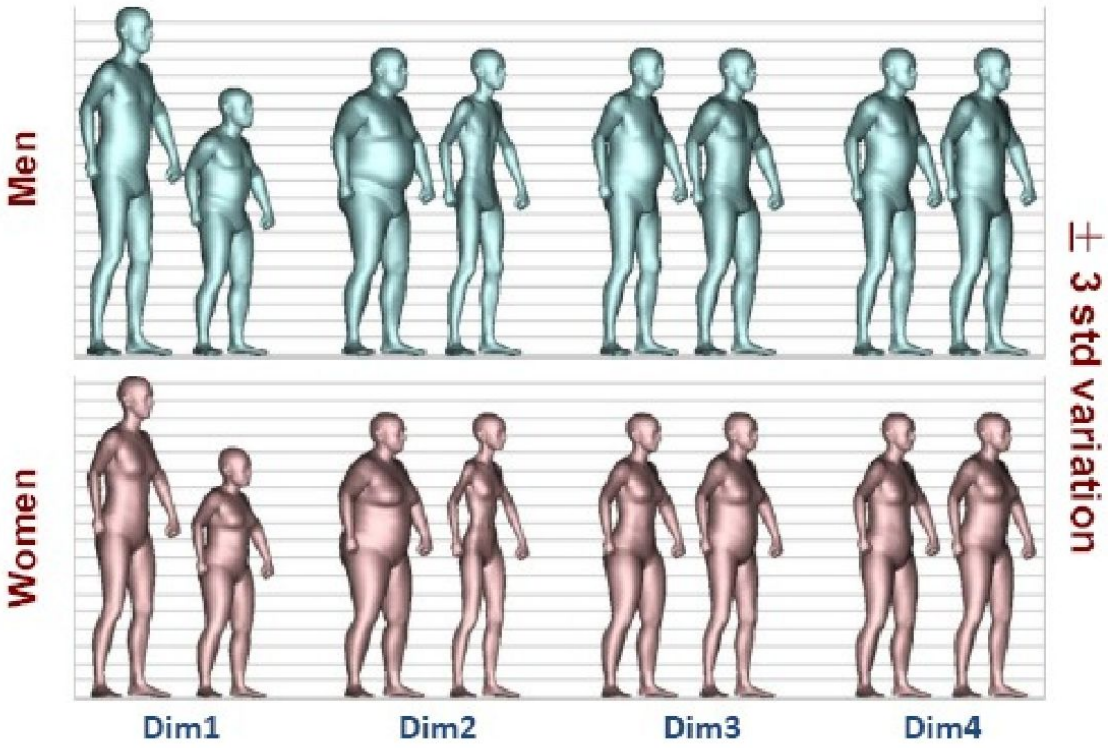
Capello, 2006

Con:

- Initialization is usually needed
- Need of tracking (previous M6 lectures)
- Prone to lose track

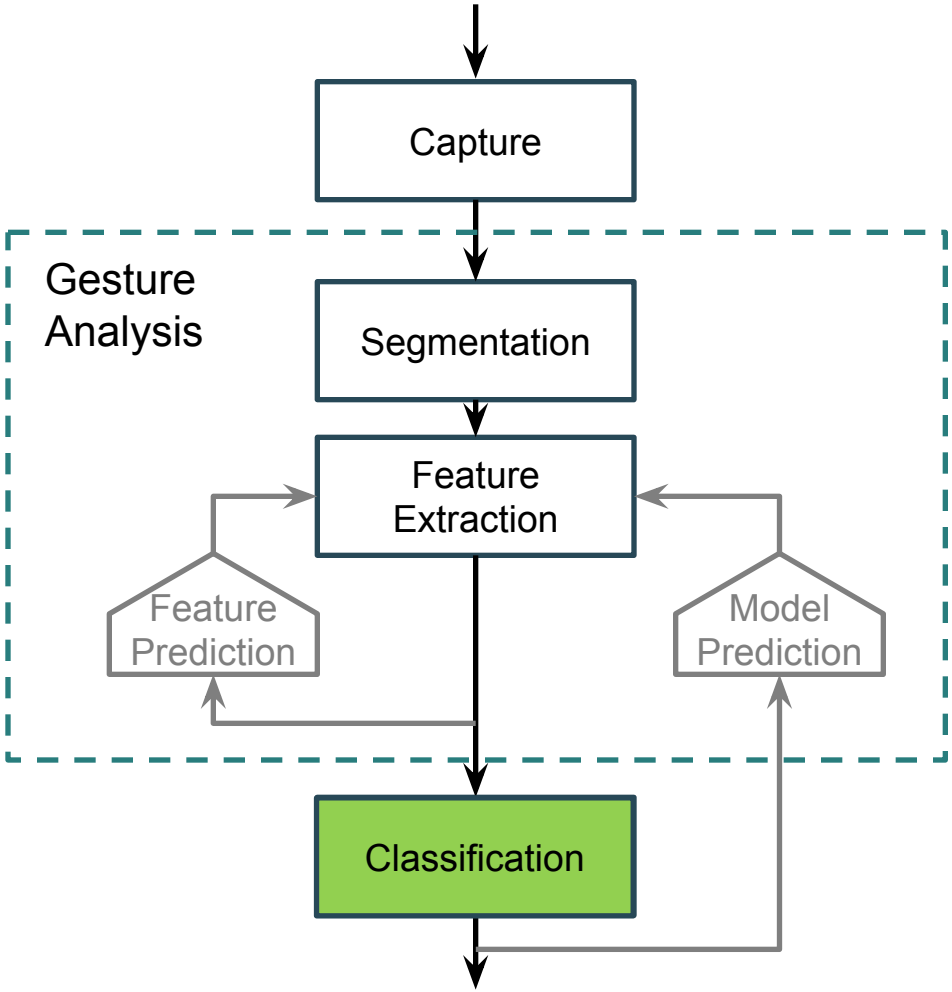
Gesture Analysis: Models (3)

- 3D human models (hand & body)
  - Flesh priors



Scape model
Anguelov, 2005

Gesture recognition general scheme



Classification: Introduction

- Once the different features are extracted from the images
  - 2D or 3D features (reduced dimensionality)
  - Joint angles and/or displacements in the case of 3D models
  - Fusion (features+images)

- Apply machine learning techniques → **M3 module**
  - Usually the classes (gestures/activities) are known and supervised methods are used
  - Classification of specific gestures / activities
  - Regression for motions

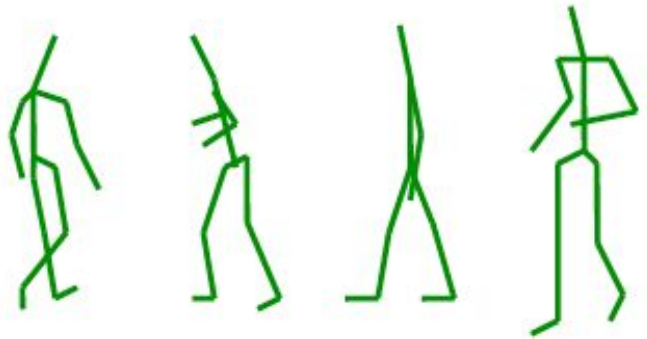
- Techniques
  - Nearest neighbour (NN), Mixture of experts, Hidden Markov Models (HMM), Support Vector Machines (SVM), Gaussian Processes, Ensemble methods (boosting, bagging, random forests) ...
  - Deep Learning

Classification: Nearest Neighbour (1)

- One simply searches in a database the example that it's close to the query under some metric.

Shakhnarovich, Viola, Darrell, 2003

Given a **large** database of image-pose pairs



Test image

Recognized action

slide after Leonid Sigal

Classification: Nearest Neighbour (2)

- One simply searches in a database the example that it's close to the query under some metric.

Pro:

- Simple to implement
- One can do metric learning to learn similarities

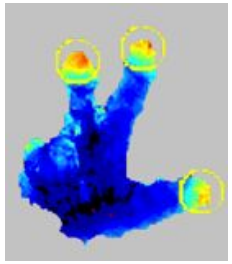
Con:

- Generalization:
  - Amount of training data required is very large
- Computing NN might be very slow.

slide after Raquel Urtasun

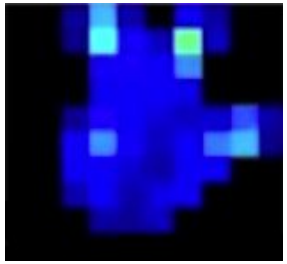
Classification: NN example (1)

Suau et al. 2014

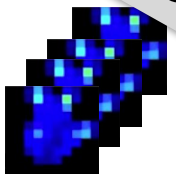


Segmented hand patch (π)
Oriented Radial Distribution
finger scale

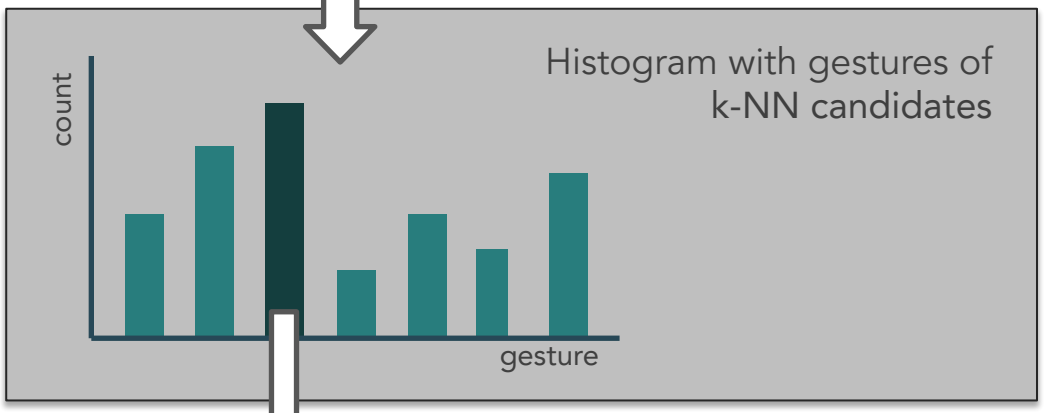
re-sampling
m x m



compare
L2 distance
using kd-tree



DataBase



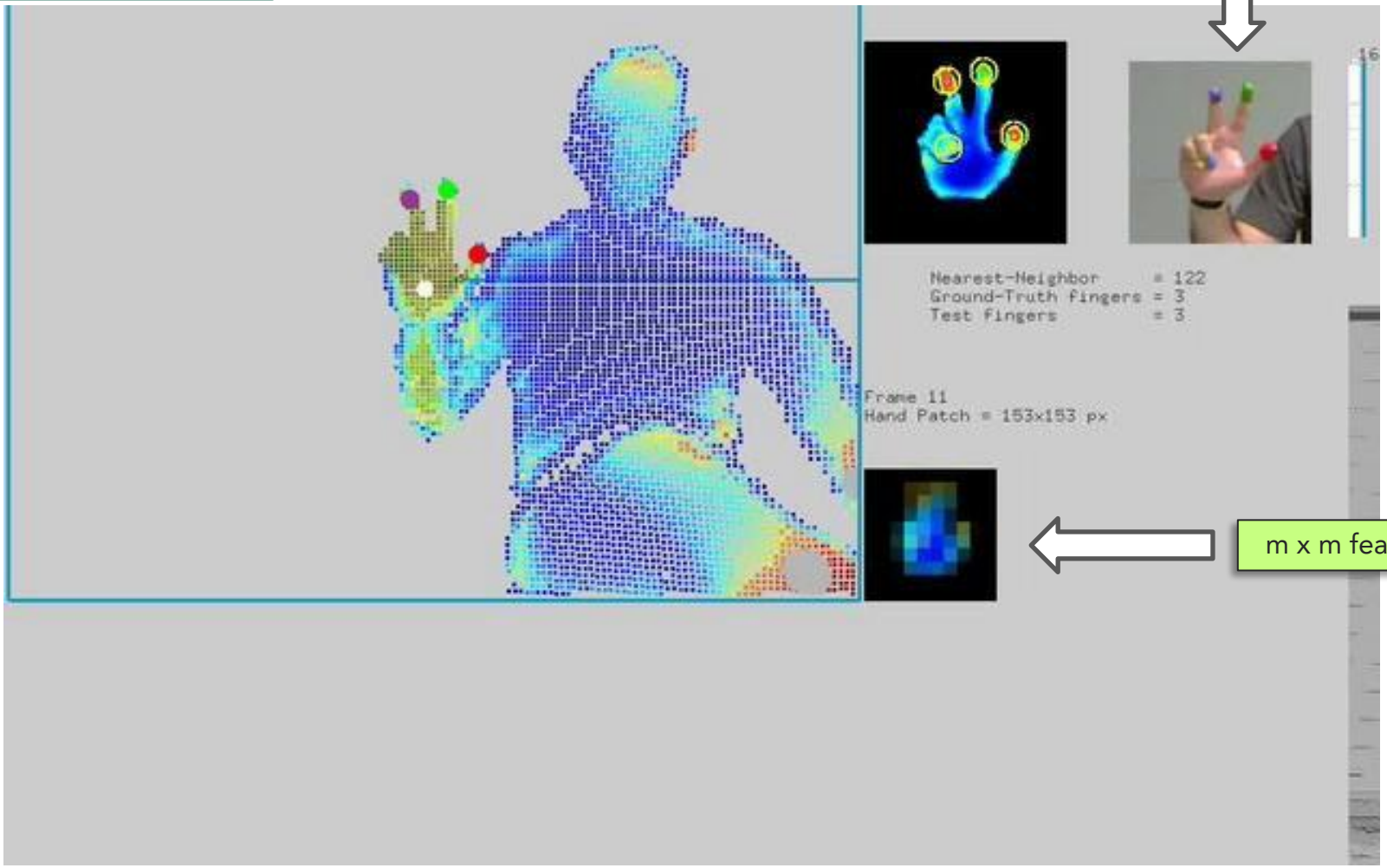
Best candidate in the dataset

Dynamically Constrained
K-NN DC

(keep trace of the last Q gestures)

Classification: NN example (2)

Suau et al., UPC, 2014

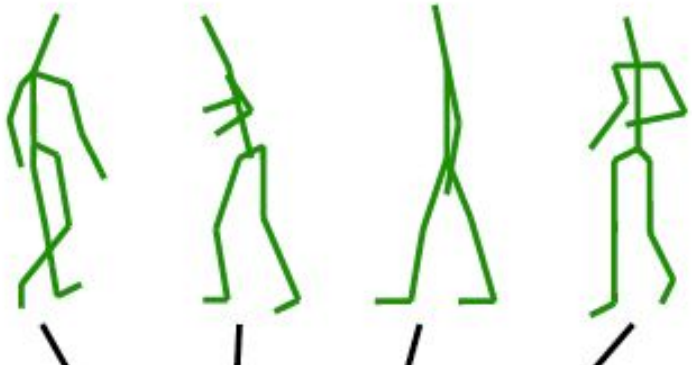


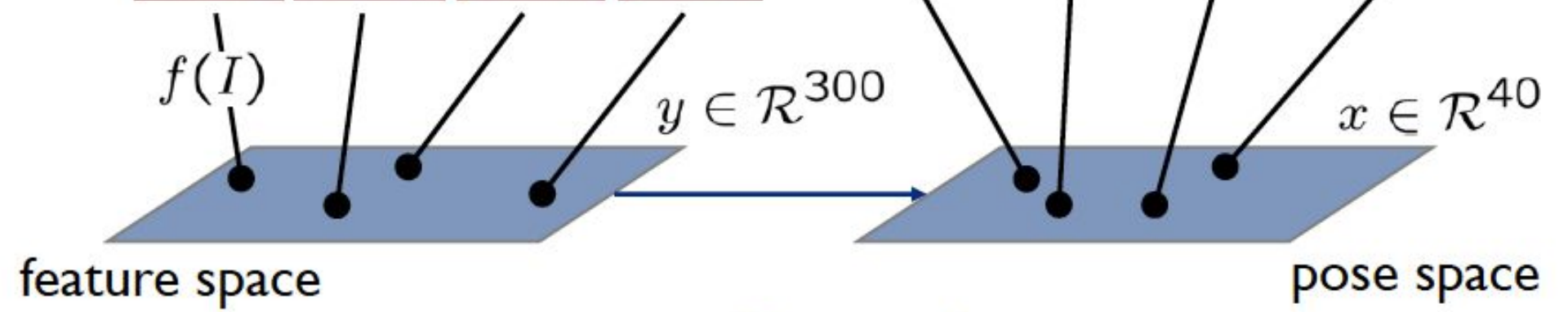
Classification: Linear regression

Agarwal and Triggs, 2004

Learn a functional mapping from features to pose

(e.g. Linear Regression: $x = g(y) = \mathbf{A}y + b$)



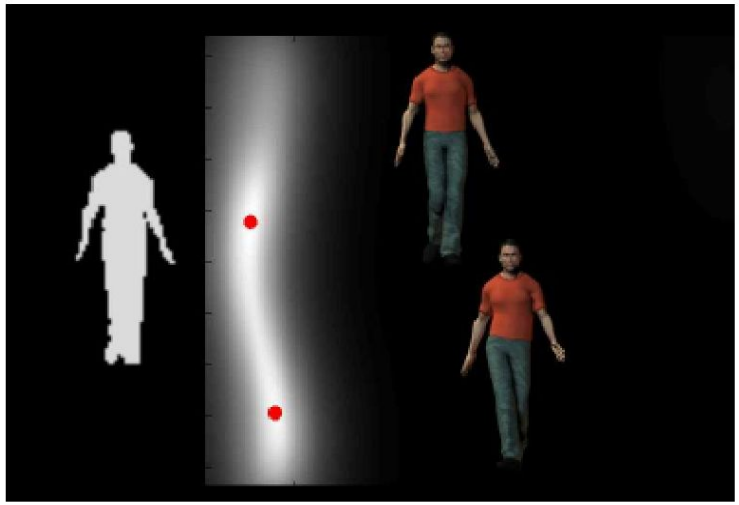


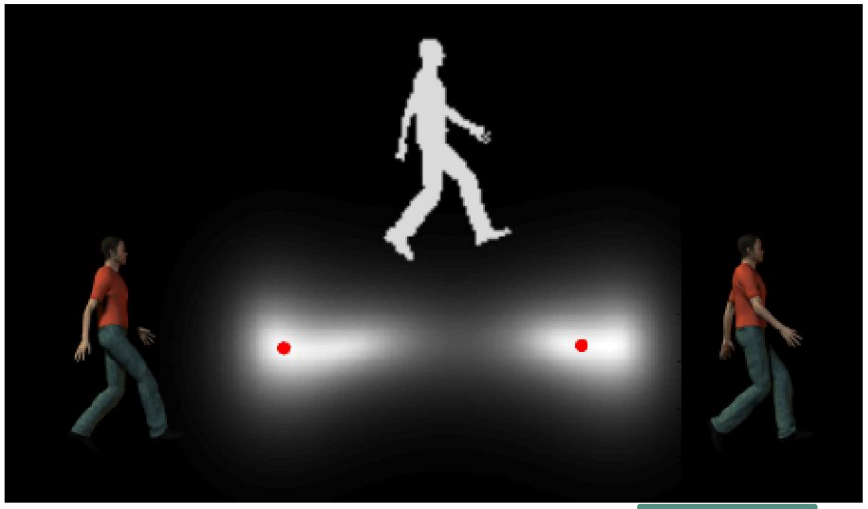
pose = g (features)

slide after Leonid Sigal

Classification: Mixture of experts (1)

- Regression cannot model multiple mappings



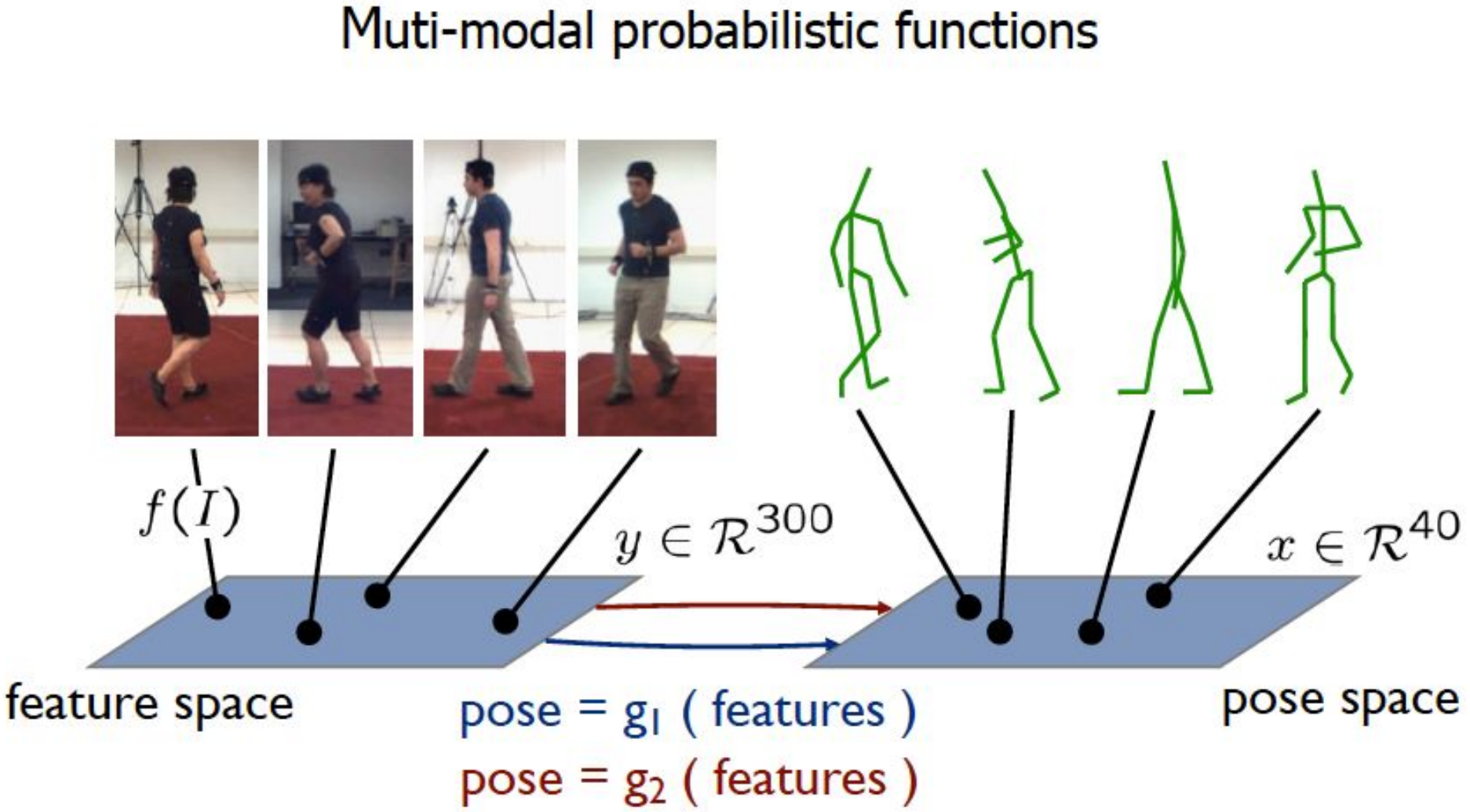


Ek, 2009

- Solution is to use mixture of experts

Classification: Mixture of experts (2)

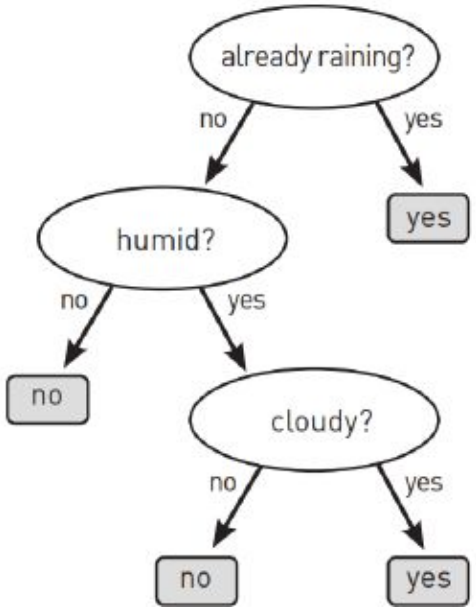
Sminchisescu et al, 2007
Bo et al, 2008



Classification: Ensemble methods

- Bagging
- Boosting
- Random forests
  - Randomized Decision trees
    - Questions: Use features from data
    - Forest: train many trees

Should you take an umbrella?



Classification: RF example (1)

López-Méndez et al. 2014

- Recognize human gestures for TV control
  - FascinatE EU project
  - 5 gestures
  - Depth data (kinect) as input
    - 15 trees, max depth 20
    - ~10.000 patches per tree
    - 85x85 pixel patches, sampled at every 4th pixel


Take control


Mute


Volume Up

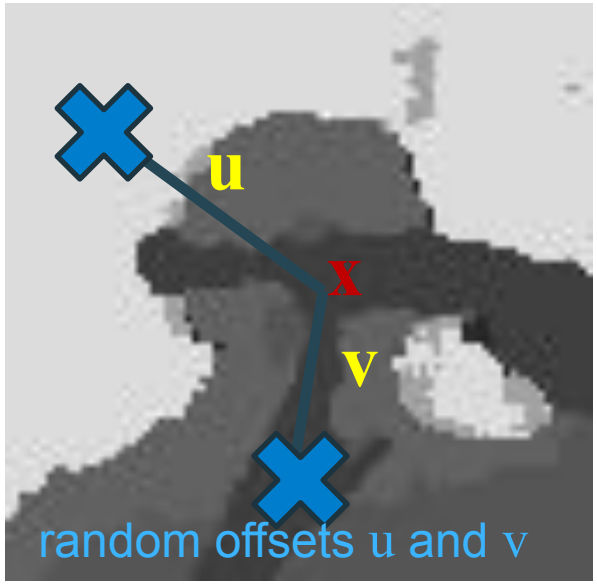

Volume Down


Pause

Classification: RF example (2)

López-Méndez et al., UPC, 2014

- RF questions
  - Function over depth data, such that a binary decision is taken for each input patch



Two random offsets (u, v)
Depth distance values (d_I)

Binary response (random threshold θ)

Random forests

Test patch
(depth data)

Ensemble of M randomized
decision trees

decision test
(binary)

. . . .

Gesture class (*c*)

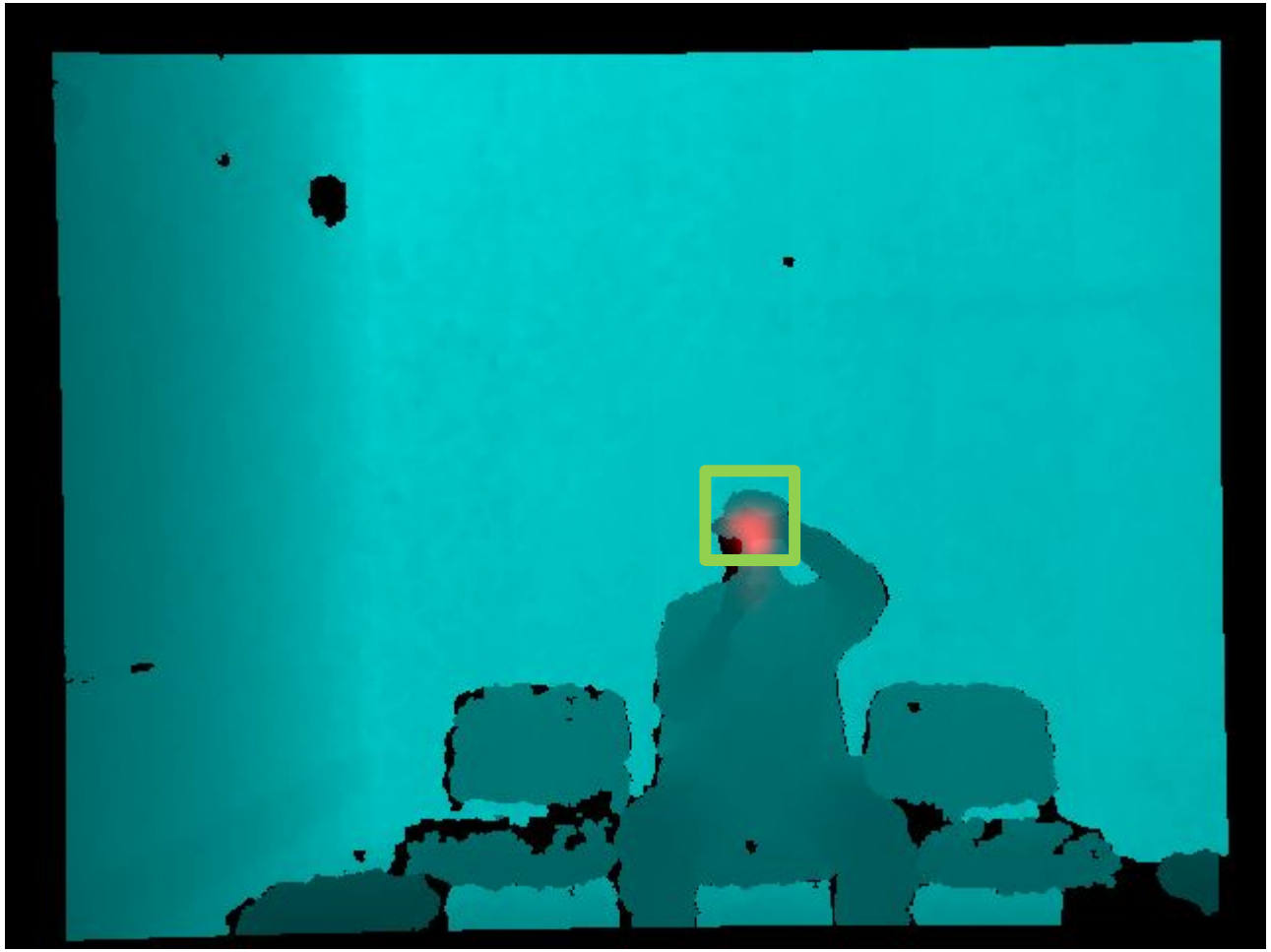
Predicted class
distributions

Averaged for classification:

$$p(c|\mathbf{I}_t, \mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} p_{l_m}(c|\mathbf{I}_t, \mathbf{x})$$

Classification: RF example (3)

López-Méndez et al., UPC, 2014



Frame I_t (with depth map)

↓

Patches are classified by RF

↓

aggregated with Parzen estimator

Classification: RF example (4)

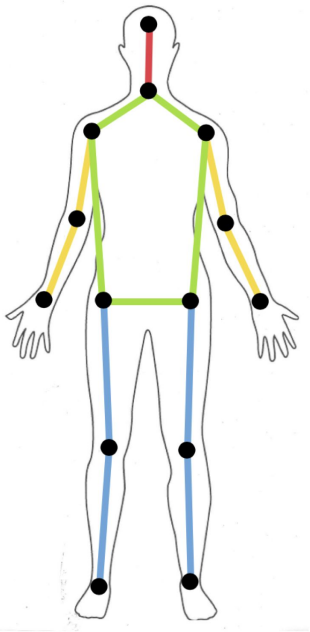
López-Méndez et al., UPC, 2014

Deep Learning

- Machine learning approaches (learning based) are the state-of-the-art techniques for gesture/activity recognition

Deep Learning: Static gestures (1)

- CNN approaches
  - Bearman, et.al., Human Pose Estimation and Activity Classification Using Convolutional Neural Networks, 2015



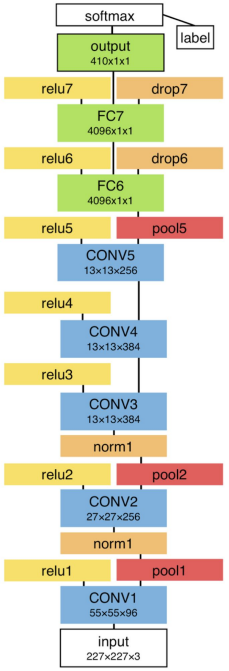


Figure 5. Classification CNN architecture for activity classification. The final layer outputs the probabilities of each of the 410 activity types.

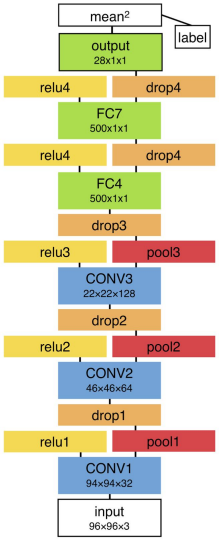
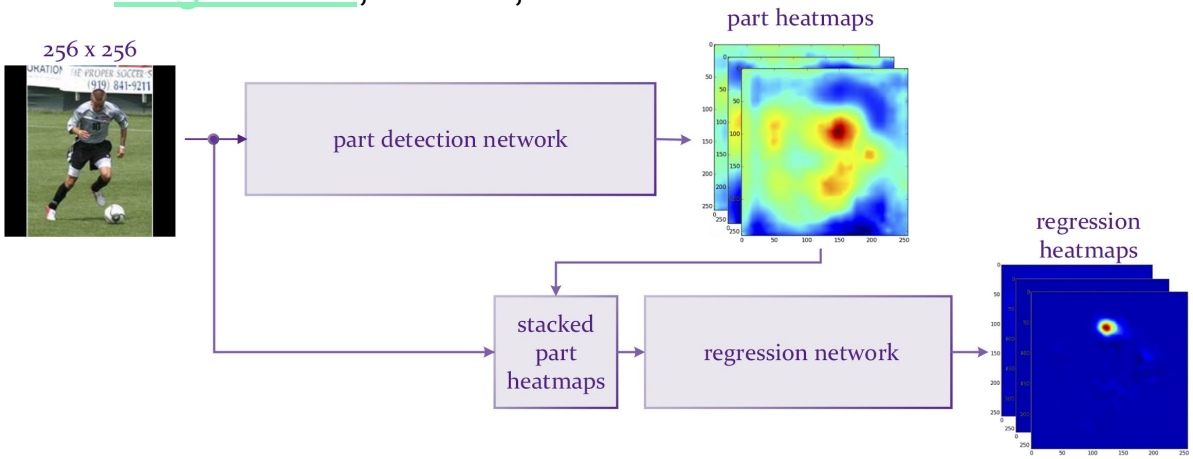


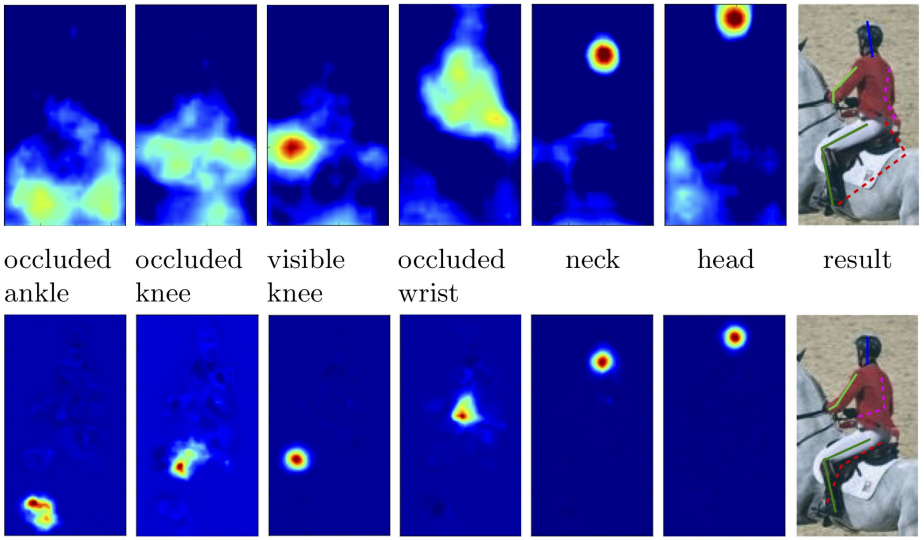
Figure 6. Regression CNN architecture for keypoint location estimation. The final layer outputs a 28-dimensional vector representing the x and y coordinates of each of the 14 keypoints.

Deep Learning: Static gestures (2)

- A. Bulat, et al., Human pose estimation via Convolutional Part Heatmap Regression, ECCV, 2016

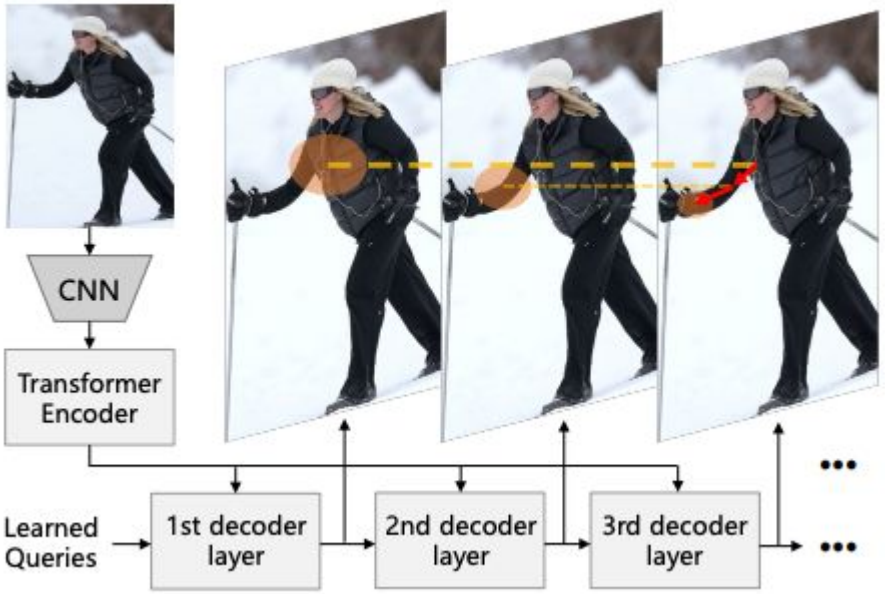


- Both sub-networks are based on VGG-16



Deep Learning: Static gestures (5)

- Transformers
  - K. Li, et al., Pose Recognition with Cascade Transformers, CVPR, 2021



- Regression based
- Gradual refinement for the keypoints across different Transformer decoder layers

Deep Learning: Static gestures (3)

- A. Toshev, et al., DeepPose: Human Pose Estimation via Deep Neural Networks, CVPR, 2014.
- W. Shen, et al. Object Skeleton Extraction in Natural Images by Fusing Scale-associated Deep Side Outputs, CVPR, 2016.
- A. Newell, et al., Stacked Hourglass Networks for Human Pose Estimation, ECCV, 2016.
- L. Pishchulin, et al., DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation, 2016.
- X. Chu, et al., Multi-Content Attention for Human Pose Estimation, CVPR, 2017.
- C. Wan, et al., Crossing Nets: Combining GANs and VAEs with a Shared Latent Space for Hand Pose Estimation, CVPR, 2017.
- R.A. Güler, et al. DensePose: Dense Human Pose Estimation In The Wild, 2017.
- Nie, et al. Single-Stage Multi-Person Pose Machines, ICCV, 2019.
- Dandan Shan, et al. Understanding Human Hands in Contact at Internet Scale, CVPR 2020.
- Z. Geng, et al. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression, CVPR 2021.

Deep Learning: Static gestures (4)

- OpenPose, Z. Cao, et al., Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR, 2017.

Deep Learning: 3D Static gestures (1)

- 3D poses
  - Tekin, et al. Structured prediction of 3D human pose with Deep Neural Networks, BMVC, 2016.

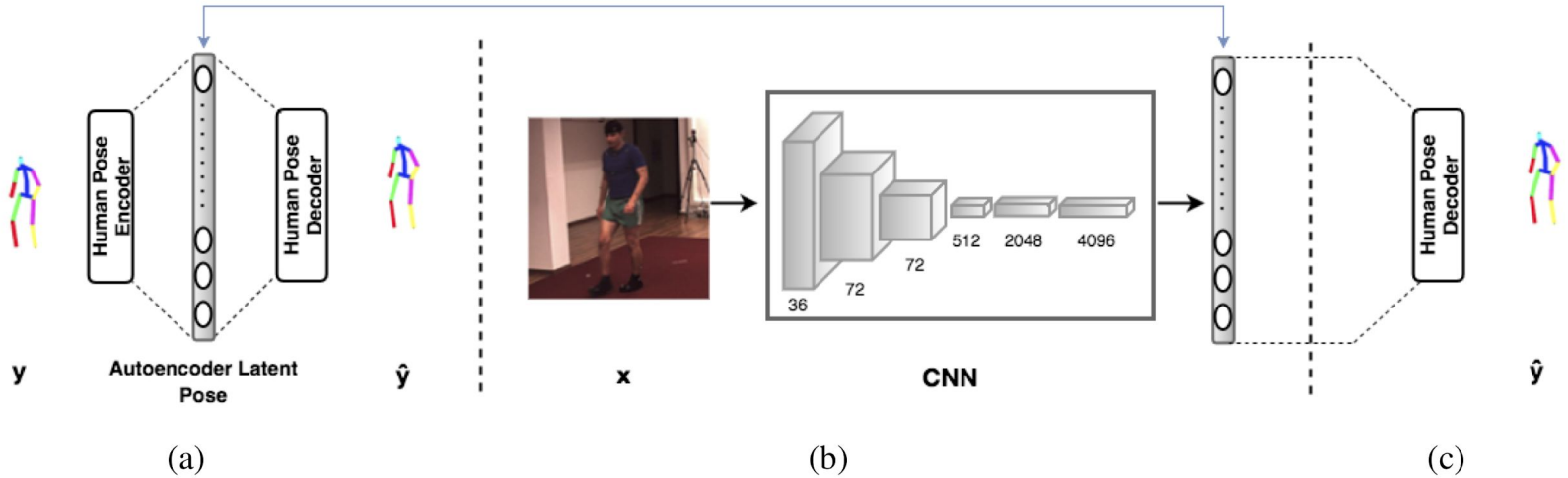


Figure 1: Our architecture for the structured prediction of the 3D human pose. **(a)** An auto-encoder whose hidden layers have a larger dimension than both its input and output layers is pretrained. In practice we use either this one or more sophisticated versions that are described in more detail in Section 3.1 **(b)** A CNN is mapped into the latent representation learned by the auto-encoder. **(c)** the latent representation is mapped back to the original pose space using the decoder.

Deep Learning: 3D Static gestures (2)

- 3D poses (multiview)
  - Kobacas, et al. Self-Supervised Learning of 3D Human Pose using Multi-view Geometry, CVPR, 2019.

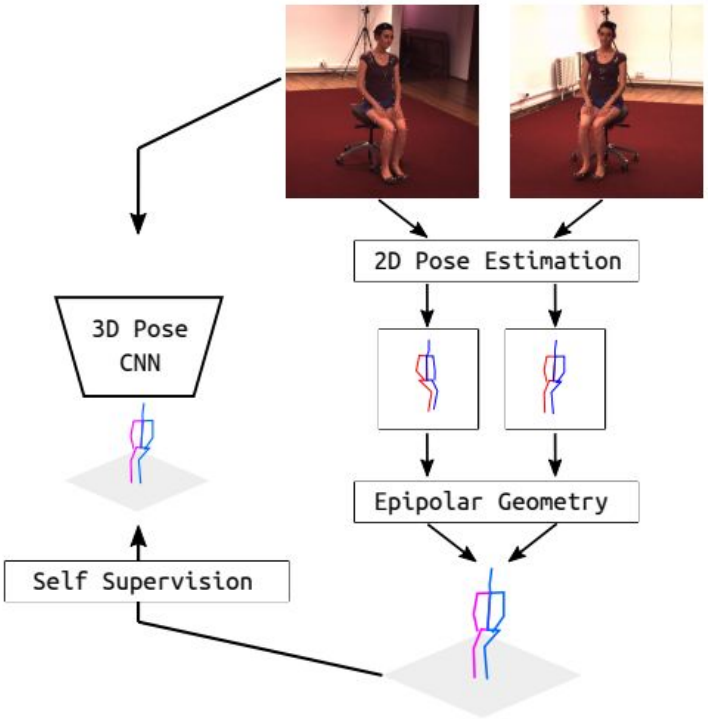
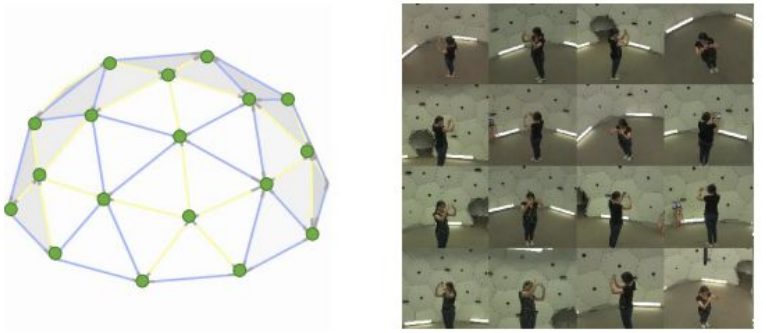


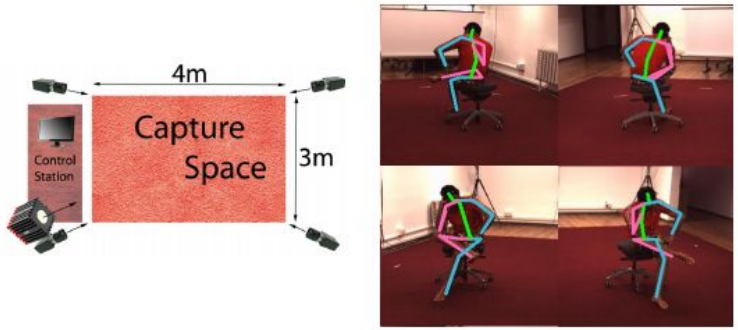
Figure 1. **EpipolarPose** uses 2D pose estimation and epipolar geometry to obtain 3D poses which are subsequently used to train a 3D pose estimator.

Deep Learning: 3D Static gestures (3)

- 3D poses (multiview)
  - Rongchang Xie, et al. MetaFuse: A Pre-trained Fusion Model for Human Pose Estimation, CVPR, 2020.



(a) Large-Scale Pretraining of *MetaFuse* from many camera views.



(b) Efficient Adaptation of *MetaFuse* to Unseen Camera placement.

Deep Learning: 3D Static gestures (4)

- 3D poses (depth):
  - L. Ge, et al.,3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images, CVPR, 2017.

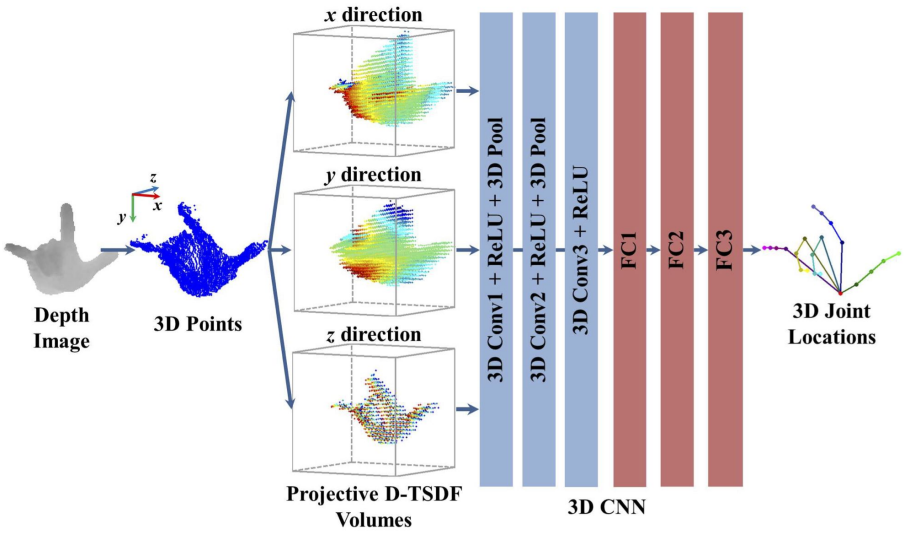


Figure 1: Overview of our proposed 3D CNN based hand pose estimation method. We generate the 3D volumetric representation of hand with projective D-TSDF from the 3D point cloud. 3D CNN is trained in an end-to-end manner to map the 3D volumetric representation to 3D hand joint relative locations in the 3D volume.

Deep Learning: 3D Static gestures (5)

- 3D poses:
  - Xiong, et al., A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation from a Single Depth Image, ICCV, 2019.
  - Chen, et al., SO-HandNet: Self-Organizing Network for 3D Hand Pose Estimation with Semi-supervised Learning, ICCV, 2019
  - Yang, et al., Aligning Latent Spaces for 3D Hand Pose Estimation, CVPR 2019.
  - Alldieck, et al., Tex2Shape: Detailed Full Human Body Geometry From a Single Image, ICCV, 2019.
  - Bowen Cheng, et al., HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation, CVPR, 2020.
  - B. Wandt, et al., CanonPose: Self-Supervised Monocular 3D Human Pose Estimation in the Wild, CVPR, 2021.
  - J. Lin, et al., Multi-View Multi-Person 3D Pose Estimation with Plane Sweep Stereo, CVPR, 2021.

Deep Learning: Mesh Static gestures

- Mesh:
  - Xu, et al.,DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare, ICCV, 2019.

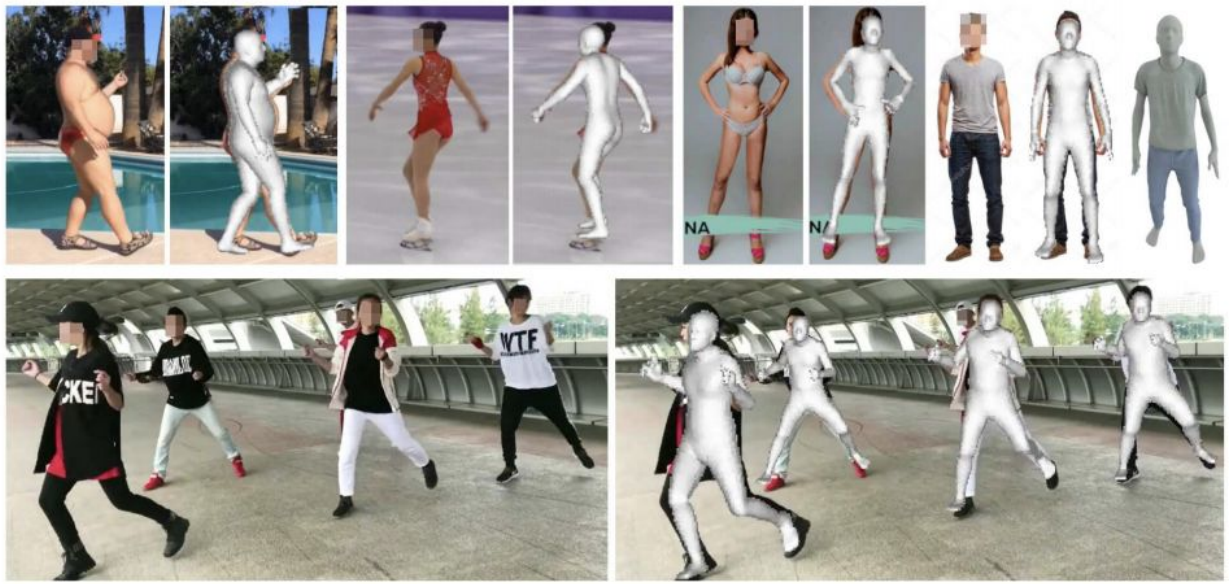


Figure 1. DenseRaC estimates 3D human poses and body shapes given people-in-the-wild images. The proposed framework handles scenarios with multiple people, all genders, and various clothing in real time. Here, we show results on Internet images [1].

Deep Learning: Dynamic gestures

- Architectures

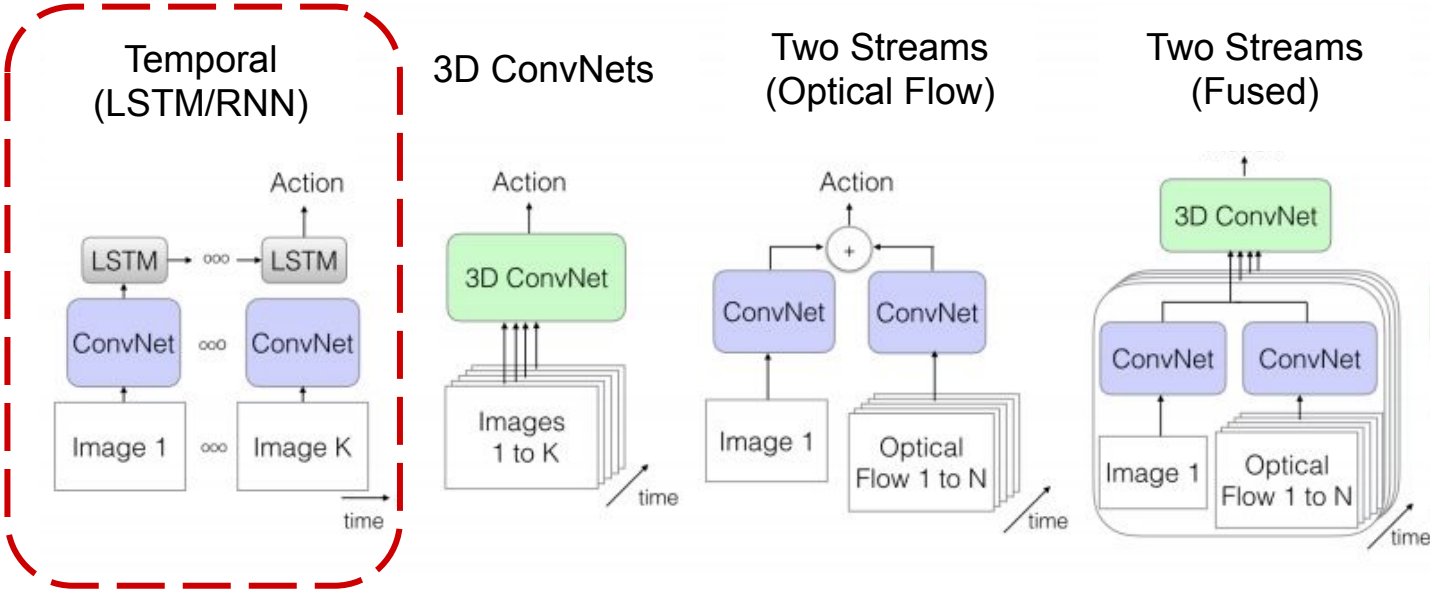


Figure from J. Carreia and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, 2018.

Deep Learning: Temporal (1)

- Temporal models: RNN and LSTM
  - E. Tsironi, et al. Gesture recognition with a convolutional long short term memory recurrent neural network, ESANN, 2015.

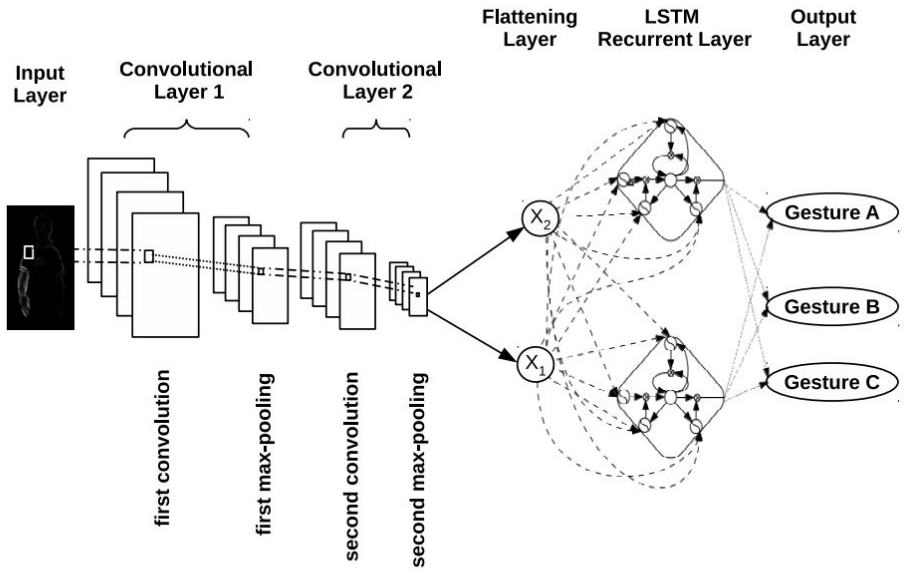
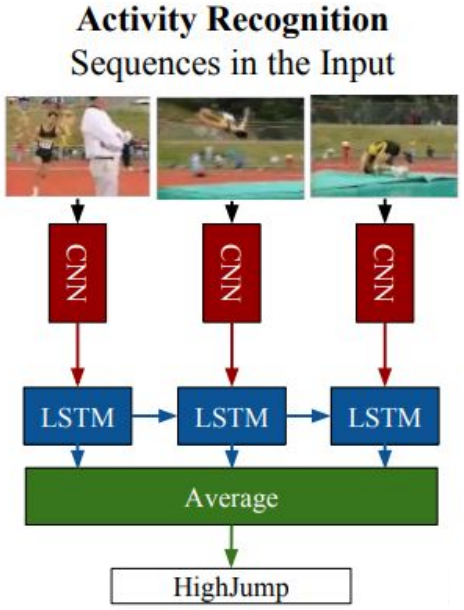
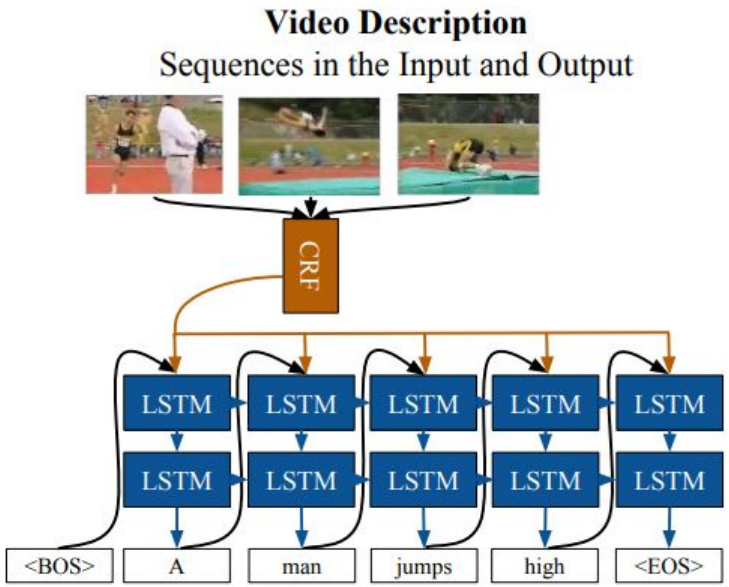


Fig. 1: The architecture of the proposed CNNLSTM.

Deep Learning: Temporal (2)

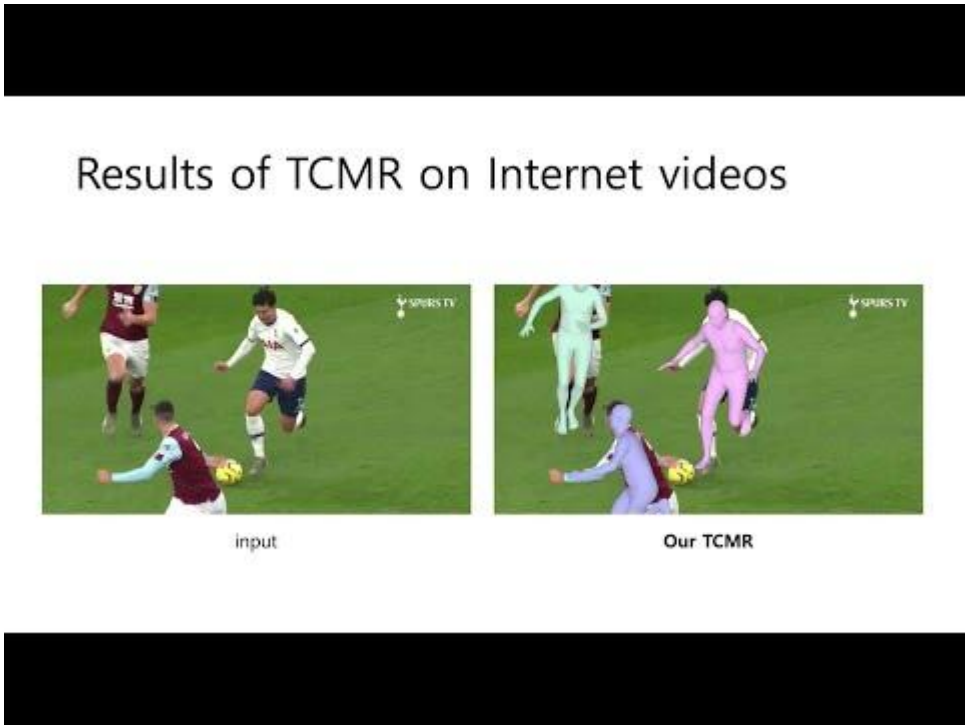
- Temporal models: RNN and LSTM
  - J. Donahue, et al. Long-term recurrent convolutional networks for visual recognition and description, CVPR, 2015.





Deep Learning: Temporal (3)

- Temporal models: GRU
  - H. Choi, et al.,Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video, CVPR, 2021.



Deep Learning: Dynamic gestures

- Architectures

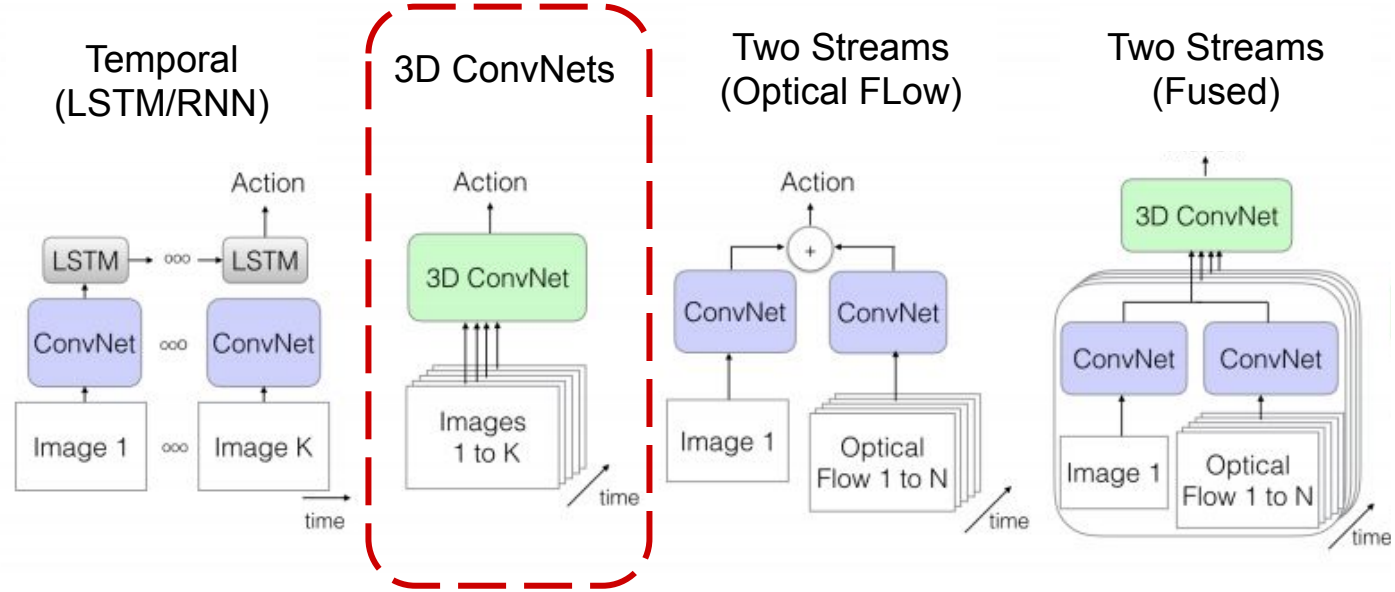


Figure from J. Carreia and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, 2018.

Deep Learning: 3D ConvNets (1)

S. Ji, et al. 3D convolutional neural networks for human action recognition, PAMI, 2013.

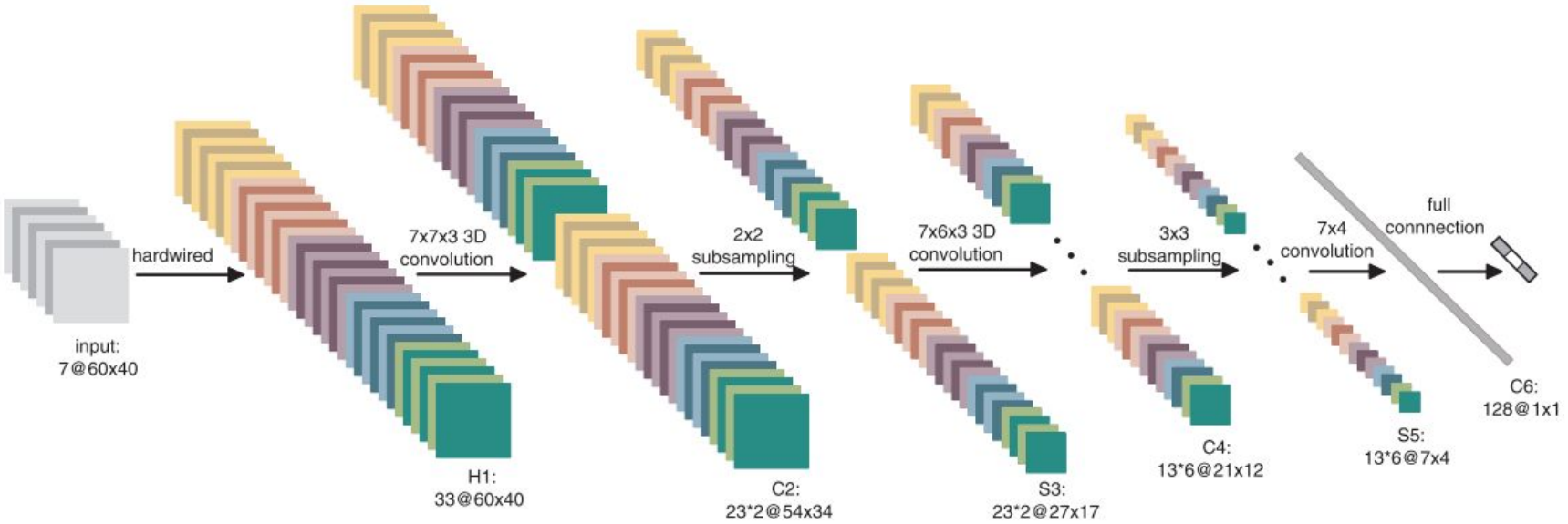


Fig. 3. A 3D CNN architecture for human action recognition. This architecture consists of one hardwired layer, three convolution layers, two subsampling layers, and one full connection layer. Detailed descriptions are given in the text.

Deep Learning: 3D ConvNets (2)

A. Karpathy, et al. Large-scale video classification with convolutional neural networks, CVPR, 2014.

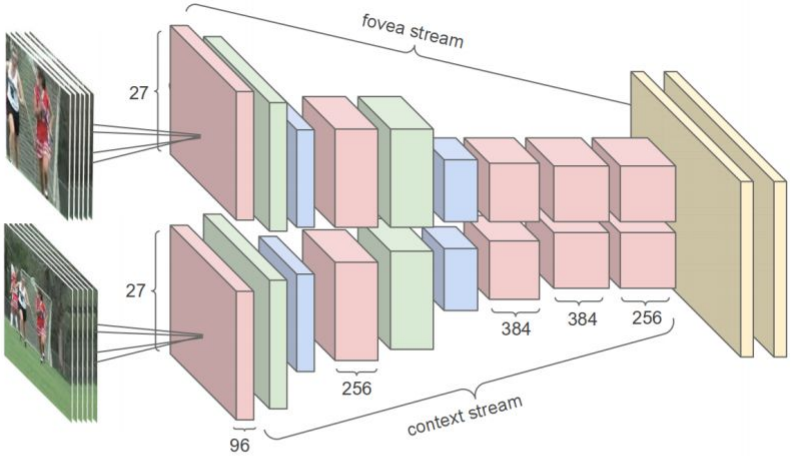


Figure 2: Multiresolution CNN architecture. Input frames are fed into two separate streams of processing: a *context stream* that models low-resolution image and a *fovea stream* that processes high-resolution center crop. Both streams consist of alternating convolution (red), normalization (green) and pooling (blue) layers. Both streams converge to two fully connected layers (yellow).

Deep Learning: 3D ConvNets (3)

P. Molchanov, et al. Hand gesture recognition with 3d convolutional neural networks, CVPR, 2015.

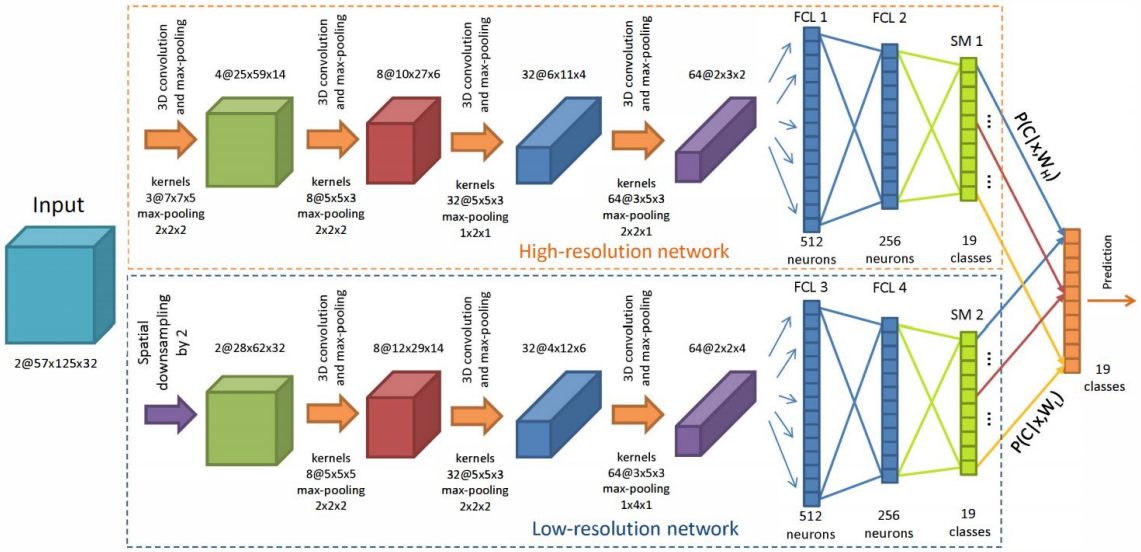


Figure 1: **Overview of our CNN classifier** We used a CNN-based classifier for hand gesture recognition. The inputs to the classifier were $57 \times 125 \times 32$ sized volumes of image gradient and depth values. The classifier consisted of two sub-networks: a high-resolution network (HRN) and a low-resolution network (LRN). The outputs of the sub-networks were class-membership probabilities $P(C|x, \mathcal{W}_H)$ and $P(C|x, \mathcal{W}_L)$, respectively. The two networks were fused by multiplying their respective class-membership probabilities element-wise.

Deep Learning: 3D ConvNets (4)

J. Huang, et al. Sign Language Recognition using 3D convolutional neural networks, ICME, 2015.

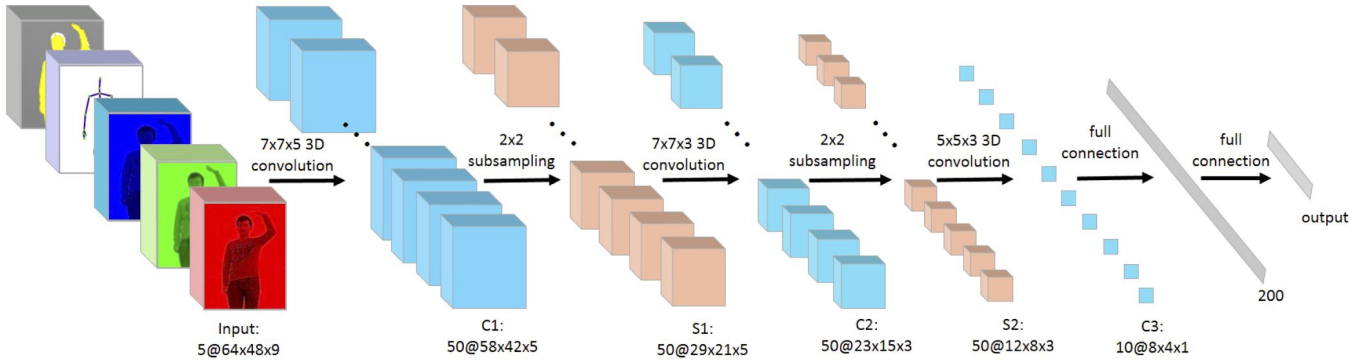


Fig. 4. Our 3D CNN architecture for sign language recognition. This architecture consists of five types of data as input, three convolution layers, two subsampling layers, and two full connection layer. Descriptions in detail are given in the text.

Deep Learning: 3D ConvNets (5)

G. Varol, et al. Long-term temporal convolutions for action recognition, PAMI, 2017.

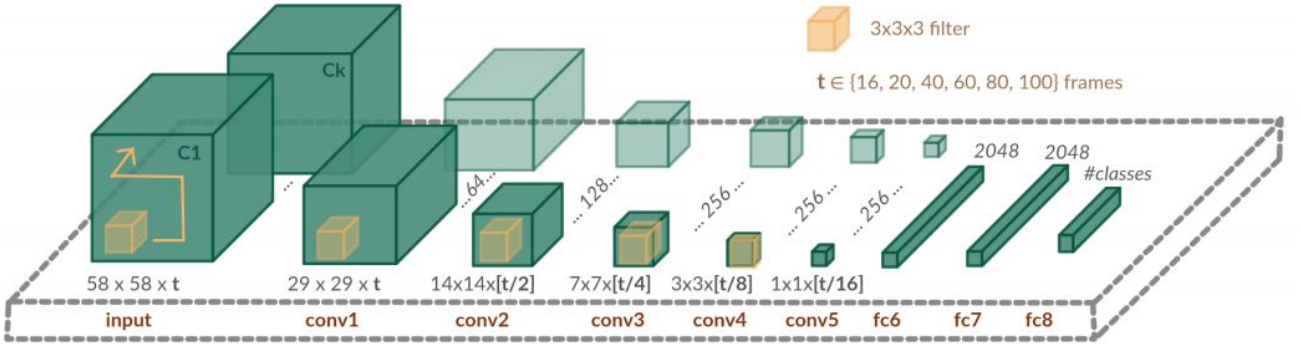


Fig. 2. Network architecture. Spatio-temporal convolutions with 3x3x3 filters are applied in the first 5 layers of the network. Max pooling and ReLU are applied in between all convolutional layers. Network input channels $C1...Ck$ are defined for different temporal resolutions $t \in \{20, 40, 60, 80, 100\}$ and either two-channel motion (*flow-x*, *flow-y*) or three-channel appearance (*R*,*G*,*B*). The spatio-temporal resolution of convolution layers decreases with the pooling operations.

Deep Learning: Dynamic gestures

- Architectures

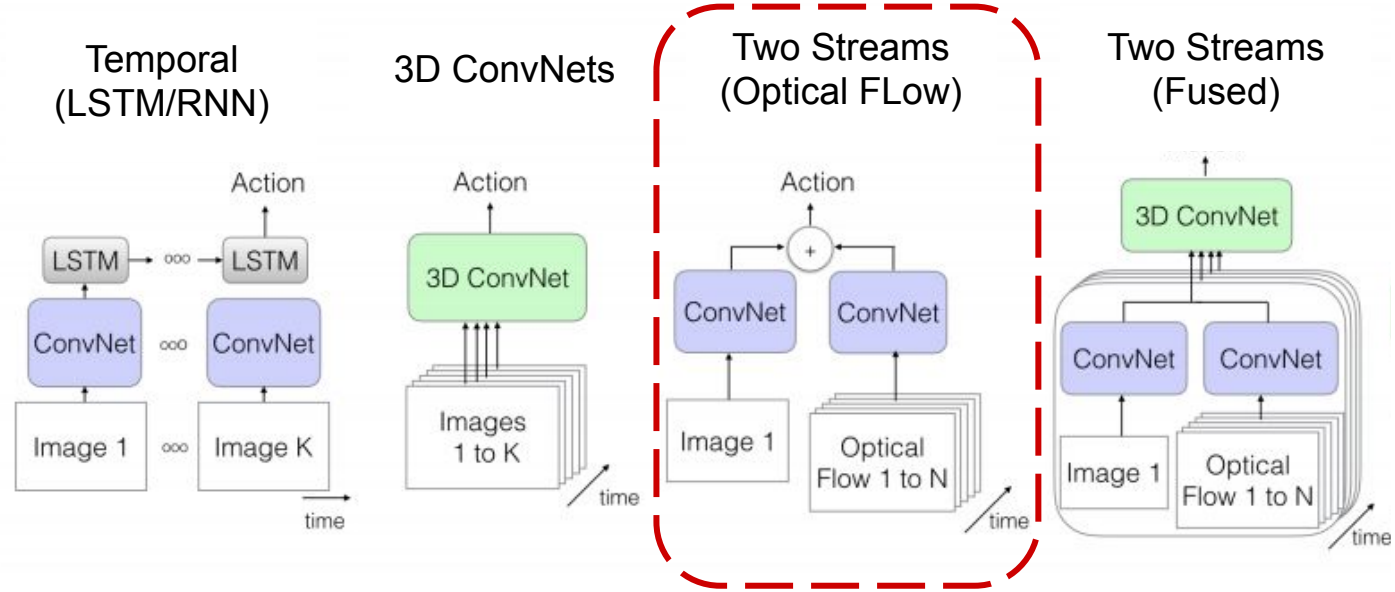


Figure from J. Carreia and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, 2018.

Deep Learning: Two streams (1)

- Motion based features
  - K. Simonyane and A. Zisserman. Two-stream convolutional networks for action recognition in videos, In Advances in Neural Information Processing Systems, 2014.

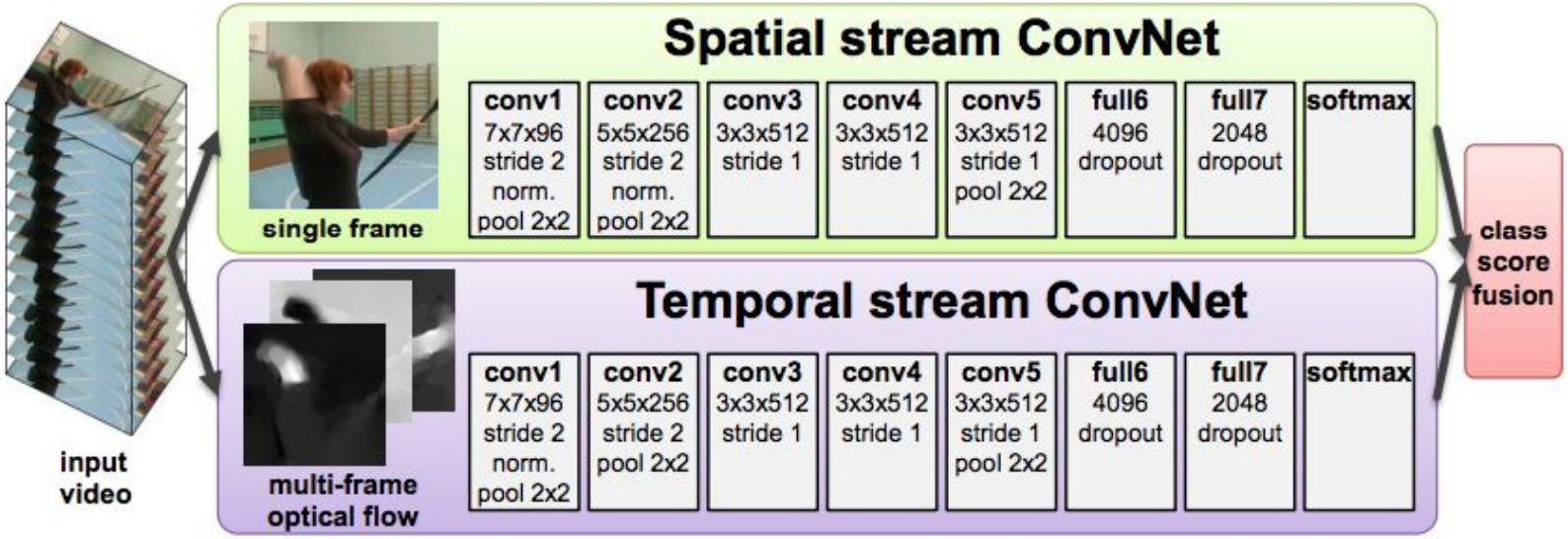


Figure 1: **Two-stream architecture for video classification.**

Deep Learning: Two streams (2)

- Motion based features
  - A. Jain, et al. MoDeep: A deep learning framework using motion features for human pose estimation, ACCV, 2015.

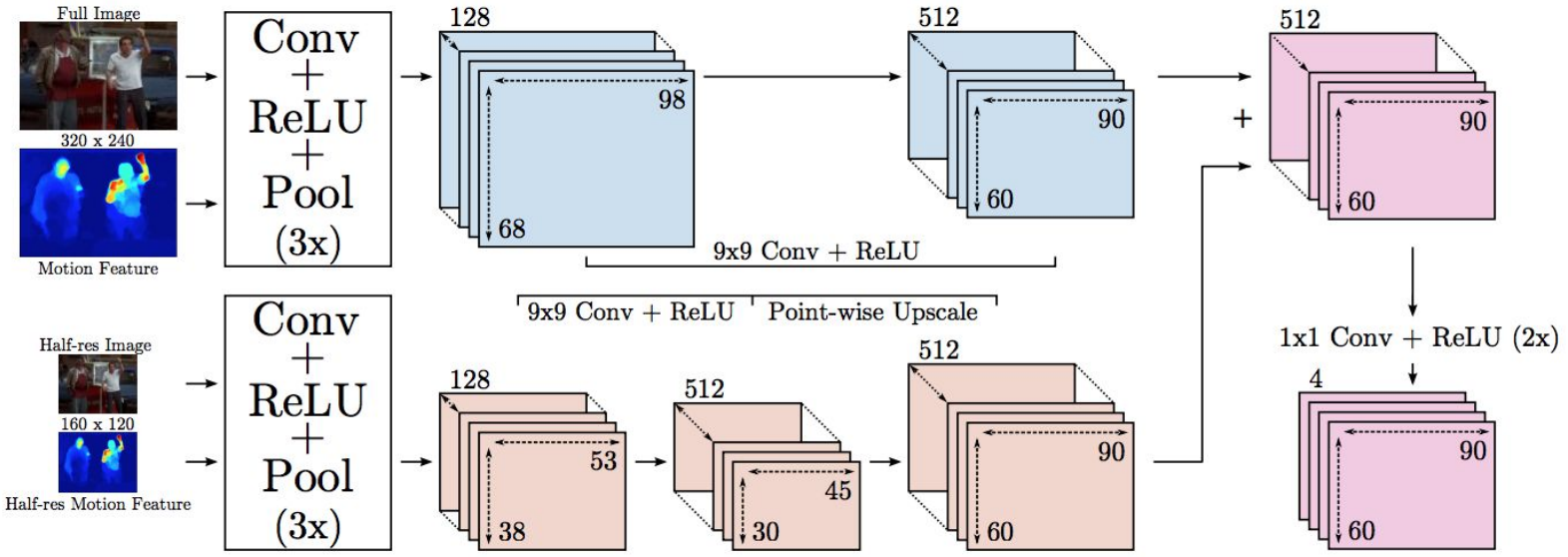


Fig. 4: Multi-resolution efficient sliding window model

Deep Learning: Dynamic gestures

- Architectures

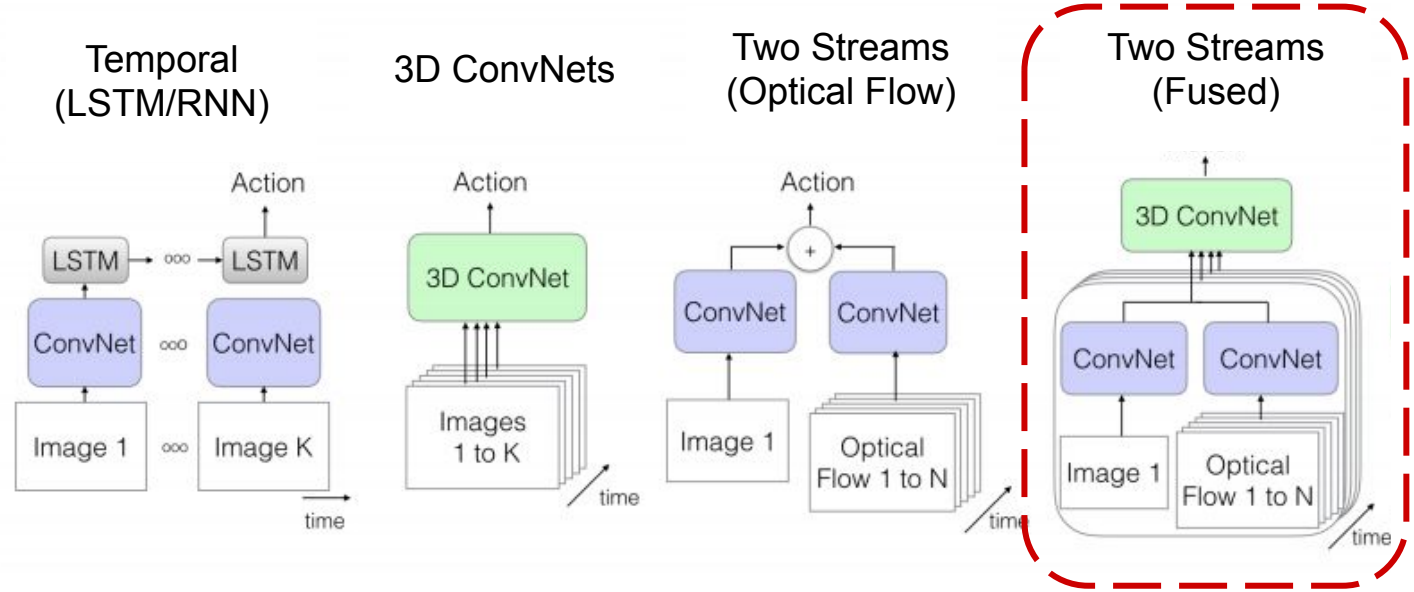


Figure from J. Carreia and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, 2018.

Deep Learning: Two streams fused (1)

- Motion based features
  - C. Feichtenhofer, et al. Convolutional two-stream network fusion for video action recognition, CVPR, 2016.

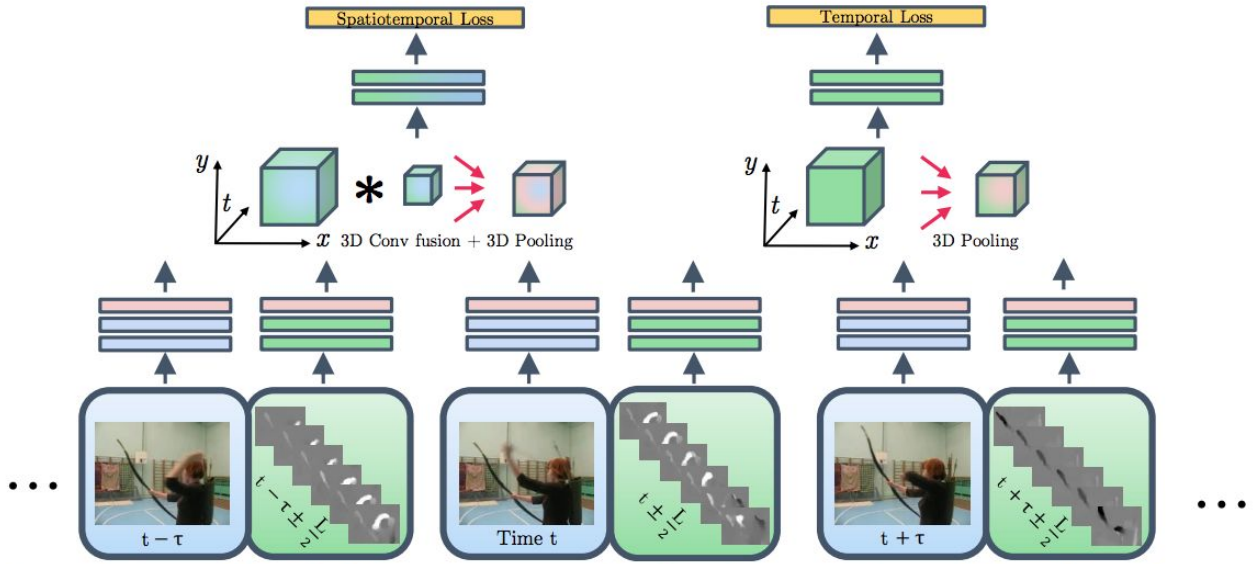


Figure 4. Our spatiotemporal fusion ConvNet applies two-stream ConvNets, that capture short-term information at a fine temporal scale ($t \pm \frac{L}{2}$), to temporally adjacent inputs at a coarse temporal scale ($t + T\tau$). The two streams are fused by a 3D filter that is able to learn correspondences between highly abstract features of the spatial stream (blue) and temporal stream (green), as well as local weighted combinations in x, y, t. The resulting features from the fusion stream and the temporal stream are 3D-pooled in space and time to learn spatiotemporal (top left) and purely temporal (top right) features for recognising the input video.

Deep Learning: Two streams fused (2)

- Motion based features
  - J. Xiao, et al. Learning from Temporal Gradient for Semi-supervised Action Recognition, CVPR, 2022.

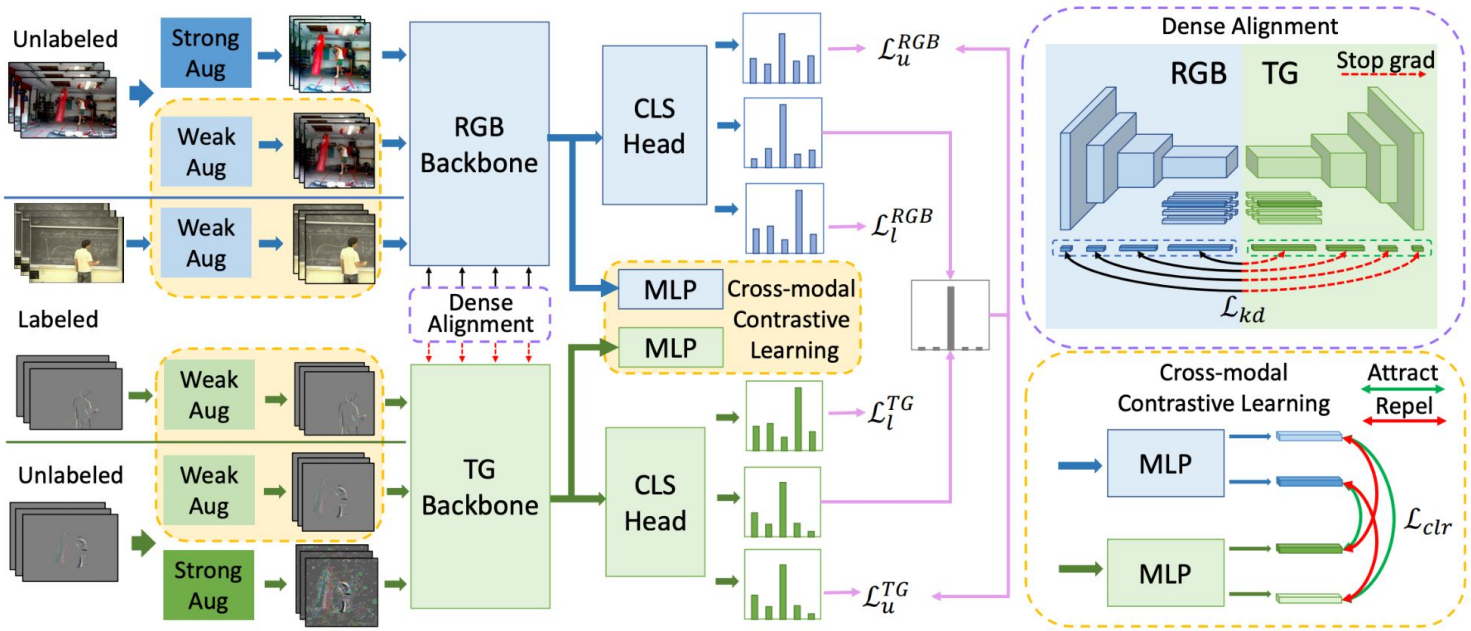
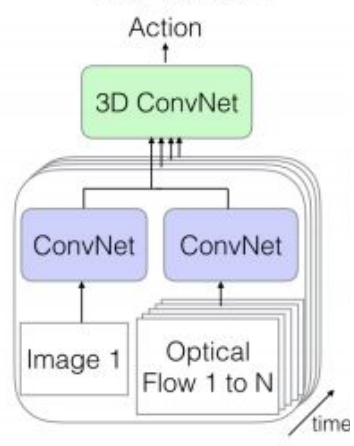


Figure 2. **An overview of our proposed framework.** Our method consists of two parallel models with different input modalities (*i.e.*, RGB and TG) of video clips. The entire framework is jointly optimized with (1) two parallel FixMatch frameworks with pseudo-labeling, (2) cross-modal dense feature alignment, and (3) cross-modal contrastive learning.

Deep Learning: Two streams fused (2)

J. Carreira and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (I3D), CVPR, 2017.



Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	63.3	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	56.1	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	62.2	52.4	65.6
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	–	–	67.2
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**71.1**	**63.4**	**74.2**

Table 2. Architecture comparison: (left) training and testing on split 1 of UCF-101; (middle) training and testing on split 1 of HMDB-51; (right) training and testing on Kinetics. All models are based on ImageNet pre-trained Inception-v1, except 3D-ConvNet, a C3D-like [31] model which has a custom architecture and was trained here from scratch. Note that the Two-Stream architecture numbers on individual RGB and Flow streams can be interpreted as a simple baseline which applies a ConvNet independently on 25 uniformly sampled frames then averages the predictions.

Deep Learning: Dynamic gestures

- Architectures

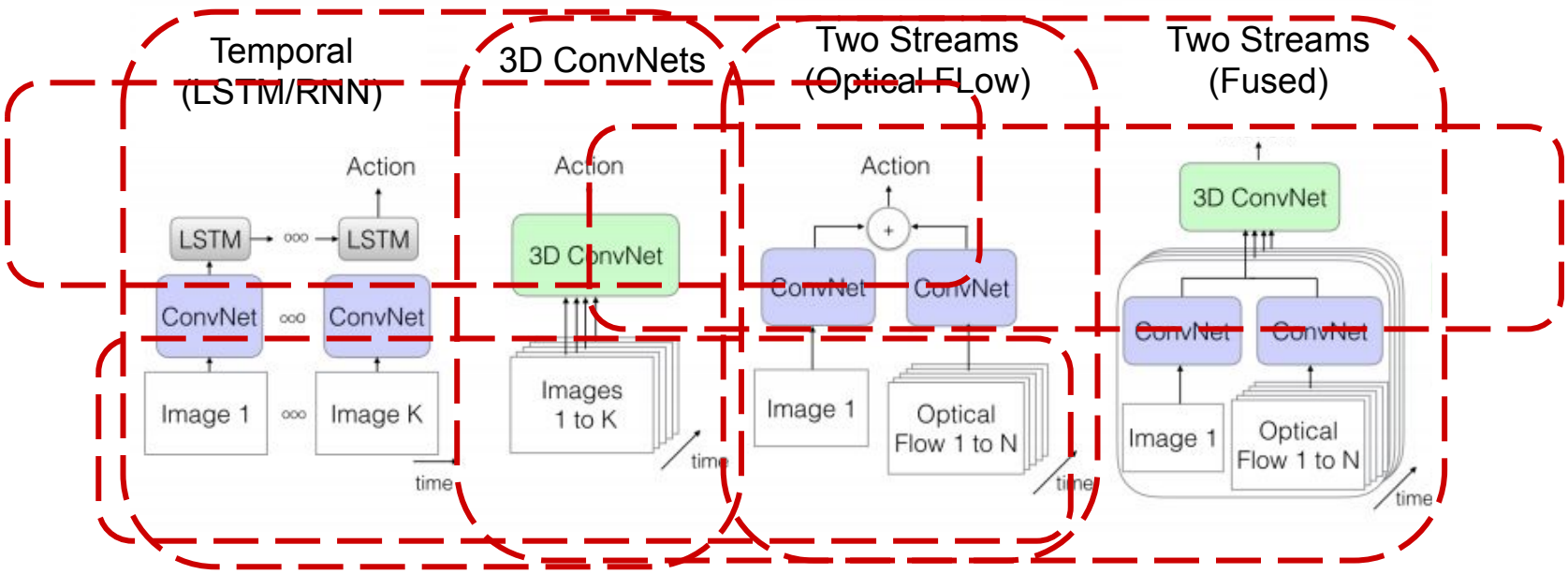


Figure from J. Carreia and A. Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, 2018.

Deep Learning: Mix (1)

- CNN + Flow + LSTM
  - J. Yue-Hei Ng, et al. Beyond short snippets: Deep networks for video classification, CVPR, 2015.

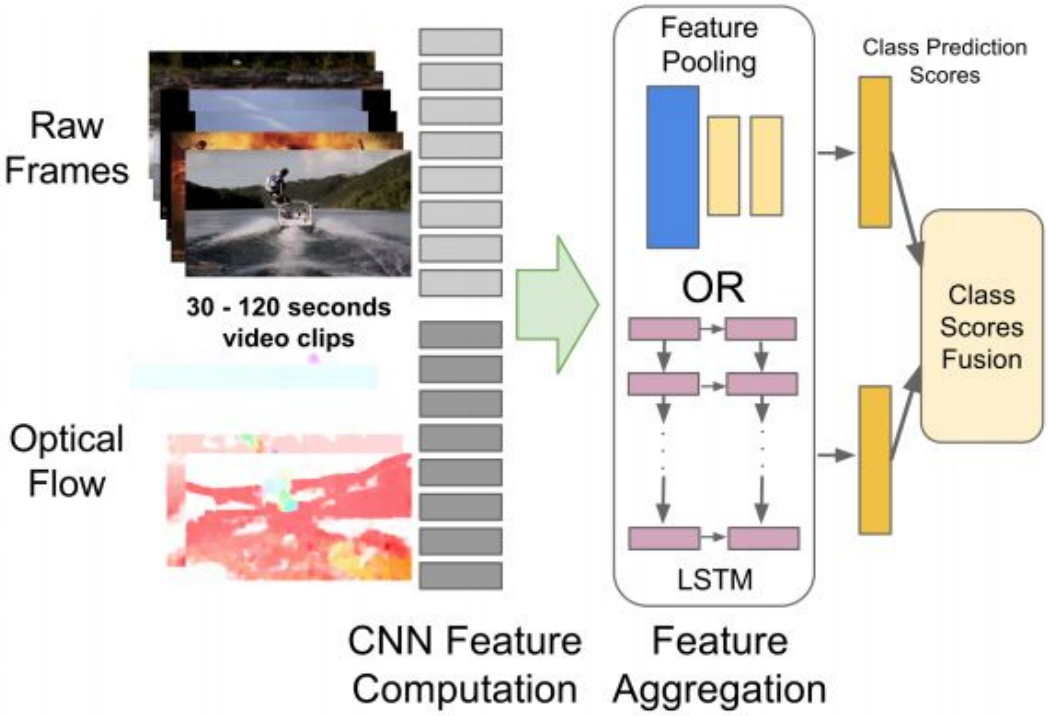


Figure 1: Overview of our approach.

Deep Learning: Mix (2)

- 3D-CNN + RNN
  - P. Molchanov, et al. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks, CVPR, 2016.

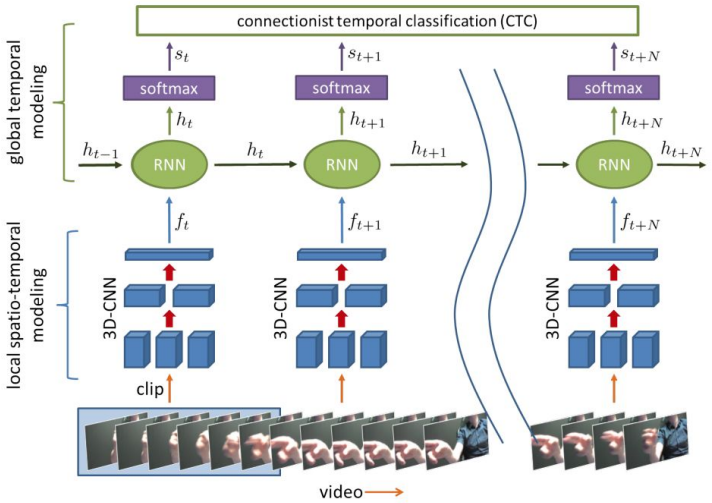


Figure 1: Classification of dynamic gestures with R3DCNN. A gesture video is presented in the form of short clips C_t to a 3D-CNN for extracting local spatial-temporal features, $\mathbf{f}_t$. These features are input to a recurrent network, which aggregates transitions across several clips. The recurrent network has a hidden state $\mathbf{h}_{t-1}$, which is computed from the previous clips. The updated hidden state for the current clip, $\mathbf{h}_t$, is input into a softmax layer to estimate class-conditional probabilities, $\mathbf{s}_t$ of the various gestures. During training, CTC is used as the cost function.

Deep Learning: Mix (3)

- Multi-Stream
  - H. Bilen, et al. Action Recognition with Dynamic Image Networks, PAMI, 2018.

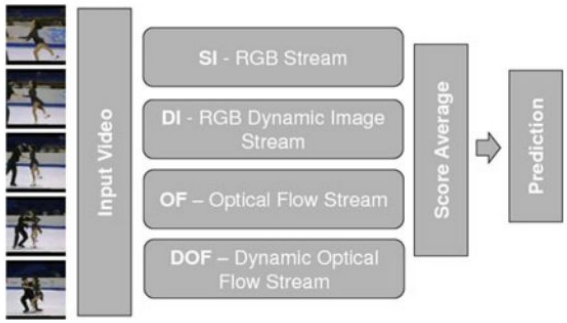


Fig. 5. The illustration of four stream dynamic image architecture that combines RGB data, optical flow with dynamic images and dynamic optical flow.

  - Dynamic images: Summary of a video into a single image (using ranking functions)

Deep Learning: Mix (4)

- Multi-stream + Combination of early & late fusion
  - K. Gavrilyuk, et al., Actor-Transformers for Group Activity Recognition, CVPR, 2020.

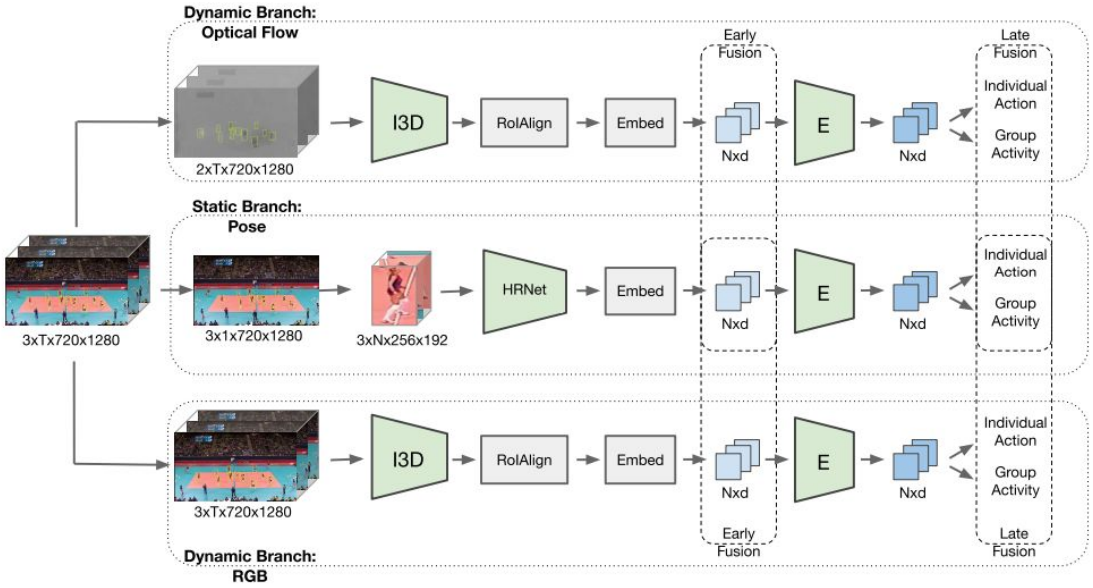


Figure 2: **Overview of the proposed model.** An input video with T frames and N actor bounding boxes is processed by two branches: static and dynamic. The static branch outputs an HRNet [51] pose representation for each actor bounding box. The dynamic branch relies on I3D [7], which receives as input either stacked RGB or optical flow frames. To extract actor-level features after I3D we apply a RoIAlign [24] layer. A transformer encoder (E) refines and aggregates actor-level features followed by individual action and group activity classifiers. Two fusion strategies are supported. For early fusion we combine actor-level features of the two branches before E, in the late fusion we combine the classifier prediction scores.

Deep Learning: Mix (5)

- Temporal + projections
  - Cheng, et al., Occlusion-Aware Networks for 3D Human Pose Estimation in Video, ICCV, 2019.

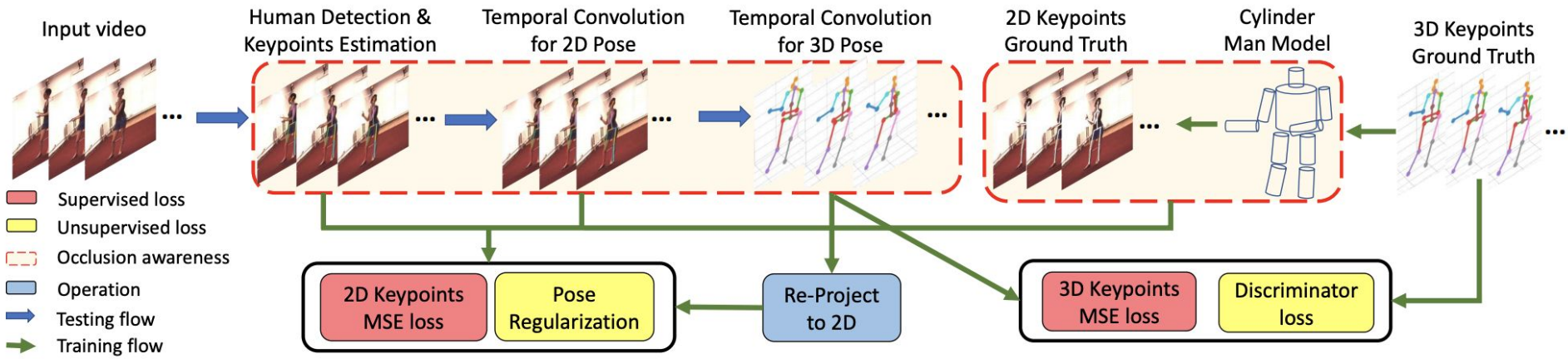


Figure 2. The framework of our approach, best viewed in color.

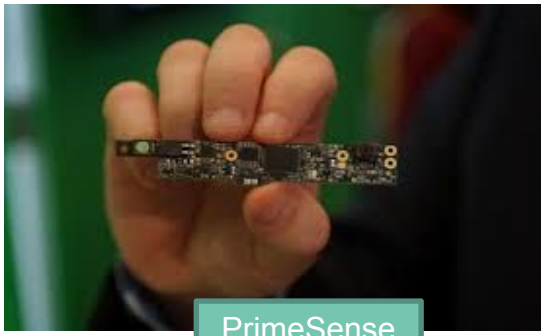
Conclusion (1)

- Many applications
- Mocap is still needed and widely used in synthesis (movie animations)
- Marker-less based methods with very high accuracy for controlled multiview scenarios
  - Single view still to improve

- **Deep learning approaches → state-of-the-art**
  - **Video-Transformers (2020+)**

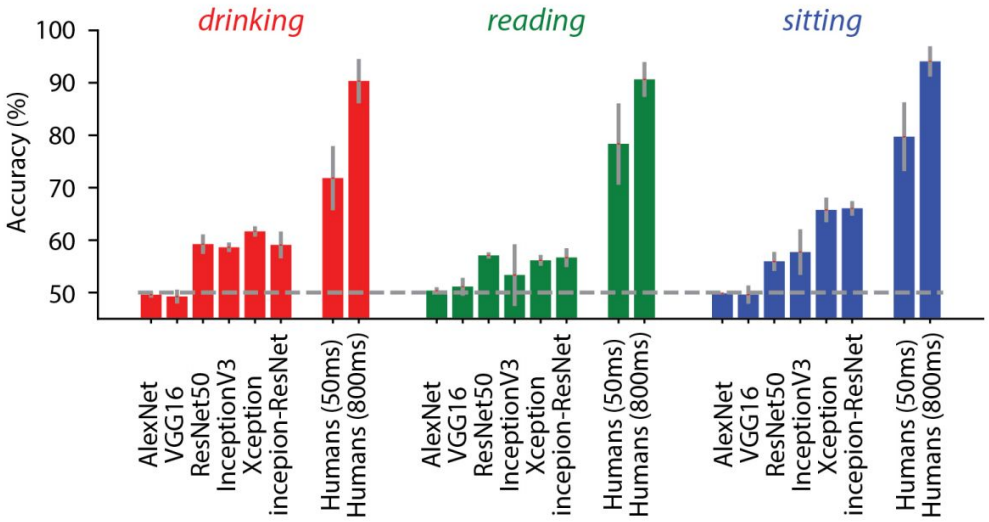
Conclusion (2)

- Challenges remaining
  - Limited gesture/action classes recognized
  - Temporal segmentation
  - Non controlled scenarios (occlusions)
  - 3D pose from 2D images
- New sensors (smaller, higher accuracy/resolution)
- Deep Learning Architectures



Challenges (1)

- Limited gesture/action classes recognized
  - V. Jacquot, et al. Can Deep Learning Recognize Subtle Human Activities? CVPR, 2020.
  - L. Zhang, et al. ZSTAD: Zero-Shot Temporal Activity Detection, CVPR, 2020.



- Temporal segmentation:
  - D. Zhang, et al. METAL: Minimum Effort Temporal Activity Localization in Untrimmed Videos, CVPR, 2020.

Challenges (2)

- Occlusions
  - Non-controlled scenarios
  - On the wild recognition
  - Chalearn challenge

- Unsupervised / Semi-supervised
  - L. Schmidtke, et al., Unsupervised Human Pose Estimation through Transforming Shape Templates, CVPR, 2021.


LIVE DEMO
STOP