SEEING PEOPLE
WITH DEEP LEARNING

GRAHAM TAYLOR

SCHOOL OF ENGINEERING
UNIVERSITY OF GUELPH

Deep Learning Summer School 2015
Montreal, Quebec

Source: Austin Granger @ Flickr

Source: ohadby @ Flickr

Seeing Humans

- Humans: dominant subject in nearly all video
- Better algorithms for interpreting their behaviour can
  - help understanding of people's use of public spaces
  - improve healthcare delivery and outcomes
  - augment people's interaction with the world
  - improve human-computer and human-robot interaction

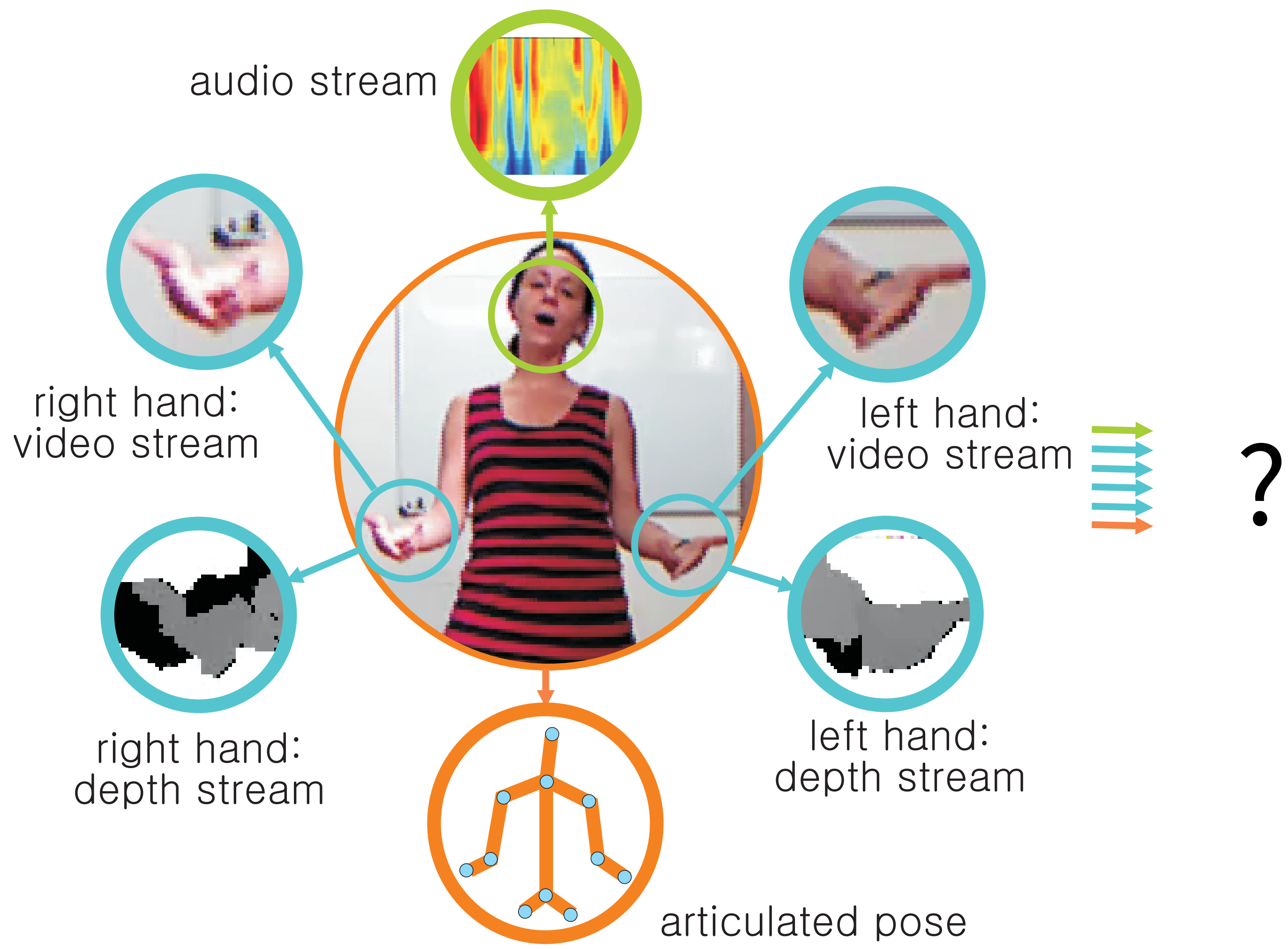
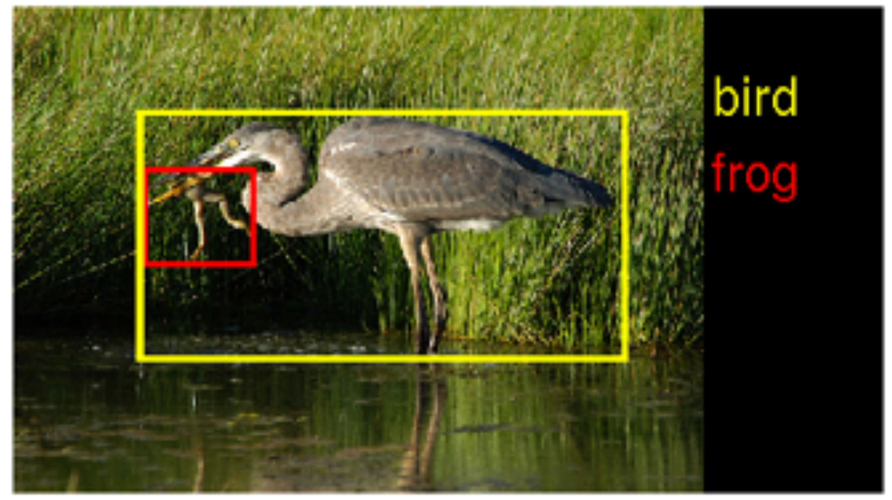


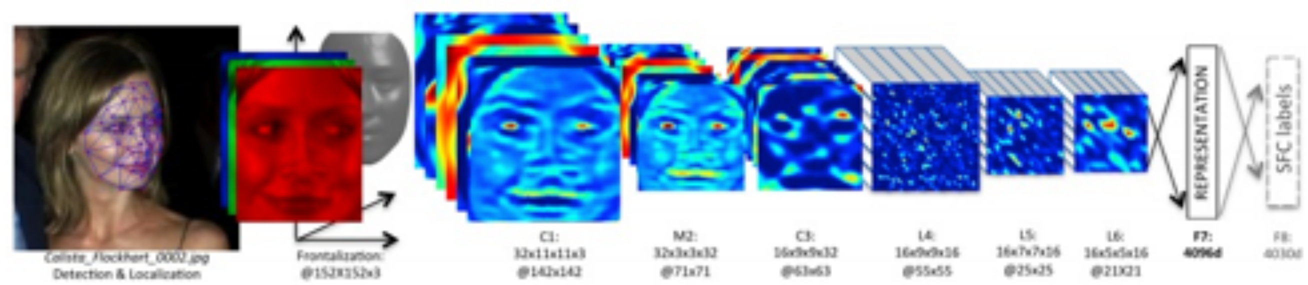
Image: Neverova et al. (2015)

Source: Daily dose of imagery

ML for Vision

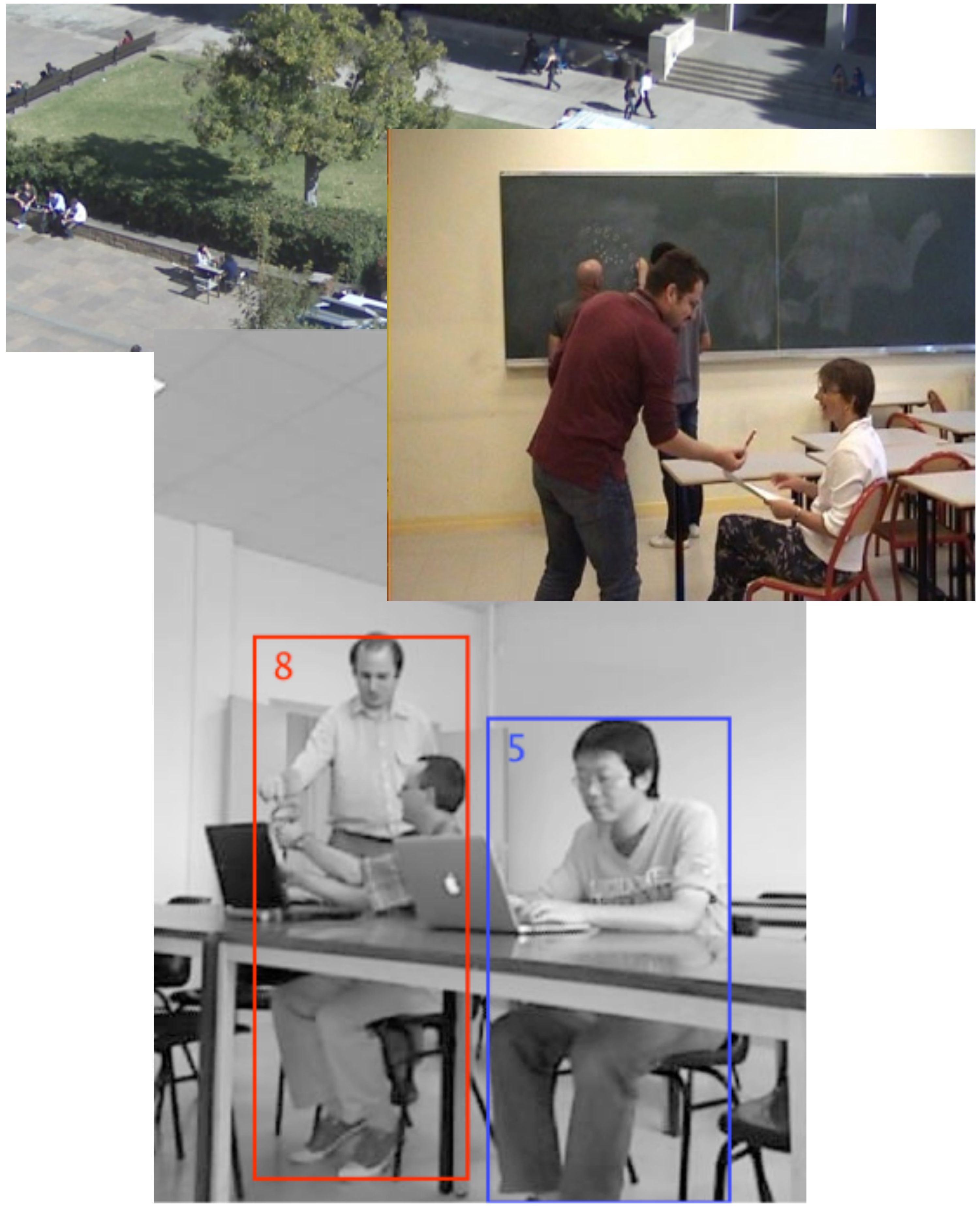


- Advances in vision have enabled "sci-fi" like applications: gesture recognition, face detection and recognition
- Machine learning is a major driving force behind this development
  - vast amounts of visual data, inherently large variations
  - emergence of new computational paradigms (GPUs)
- Deep learning has emerged as a major force in vision



Challenges Lie Ahead

- Many realistic situations are currently out of reach
  - person-person and person-object interactions
  - long-running dynamical behaviour in video
  - large-scale variation (e.g. deformable objects)



Images:Christian Wolf

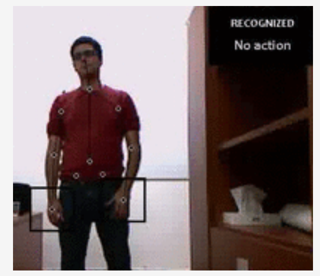
This Lecture

Focus on “seeing humans” in images and video using deep learning methods:

Pose Estimation

Tracking

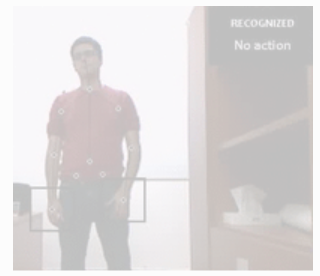
Activity /Gesture



Pose Estimation

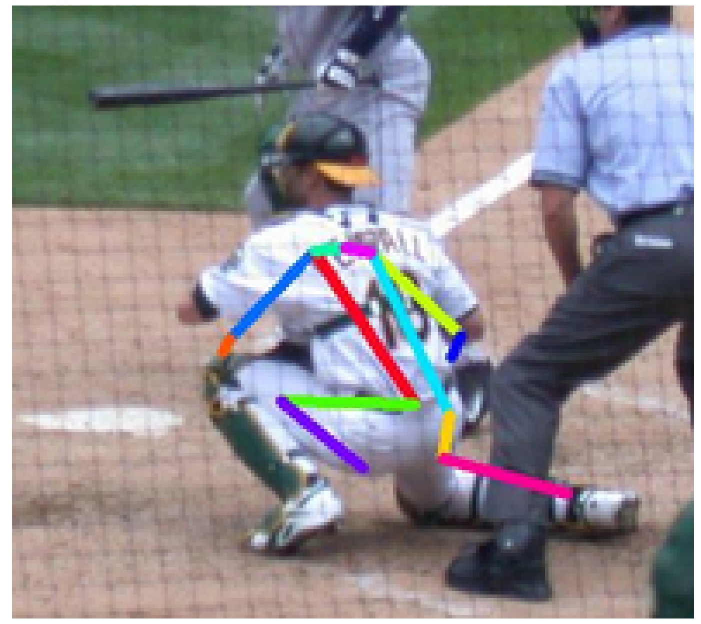
Tracking

Activity /Gesture



Pose Estimation

- Localization of joints
- Extreme variability in articulations
- Many joints barely visible
  - small # pixels
  - occlusions



Images: Toshev and Szegedy (2014)

DNNs for Precise Localization?

DNNs for Precise Localization?

- Most obvious approach: map input vector directly to a vector coding the articulated pose (e.g. unbounded 2-D or 3-D positions of joints or angles)

DNNs for Precise Localization?

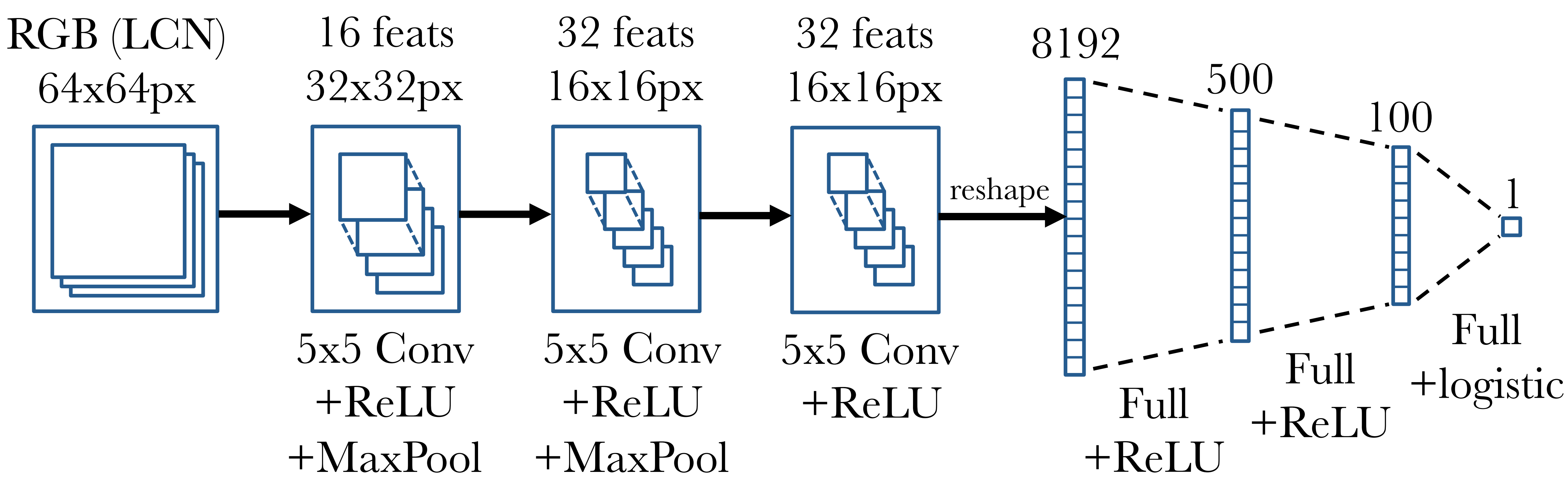
- Most obvious approach: map input vector directly to a vector coding the articulated pose (e.g. unbounded 2-D or 3-D positions of joints or angles)
  - Pooling, while useful for recognition, destroys precise spatial information
  - The mapping from input space to kinematic pose is highly nonlinear and not one-to-one
  - Valid poses represent a much lower-dimensional manifold in the high-dimensional space of configurations

CNNs for Pose Estimation

(Jain et al. 2014)

- Train multiple convnets to perform independent body-part classification

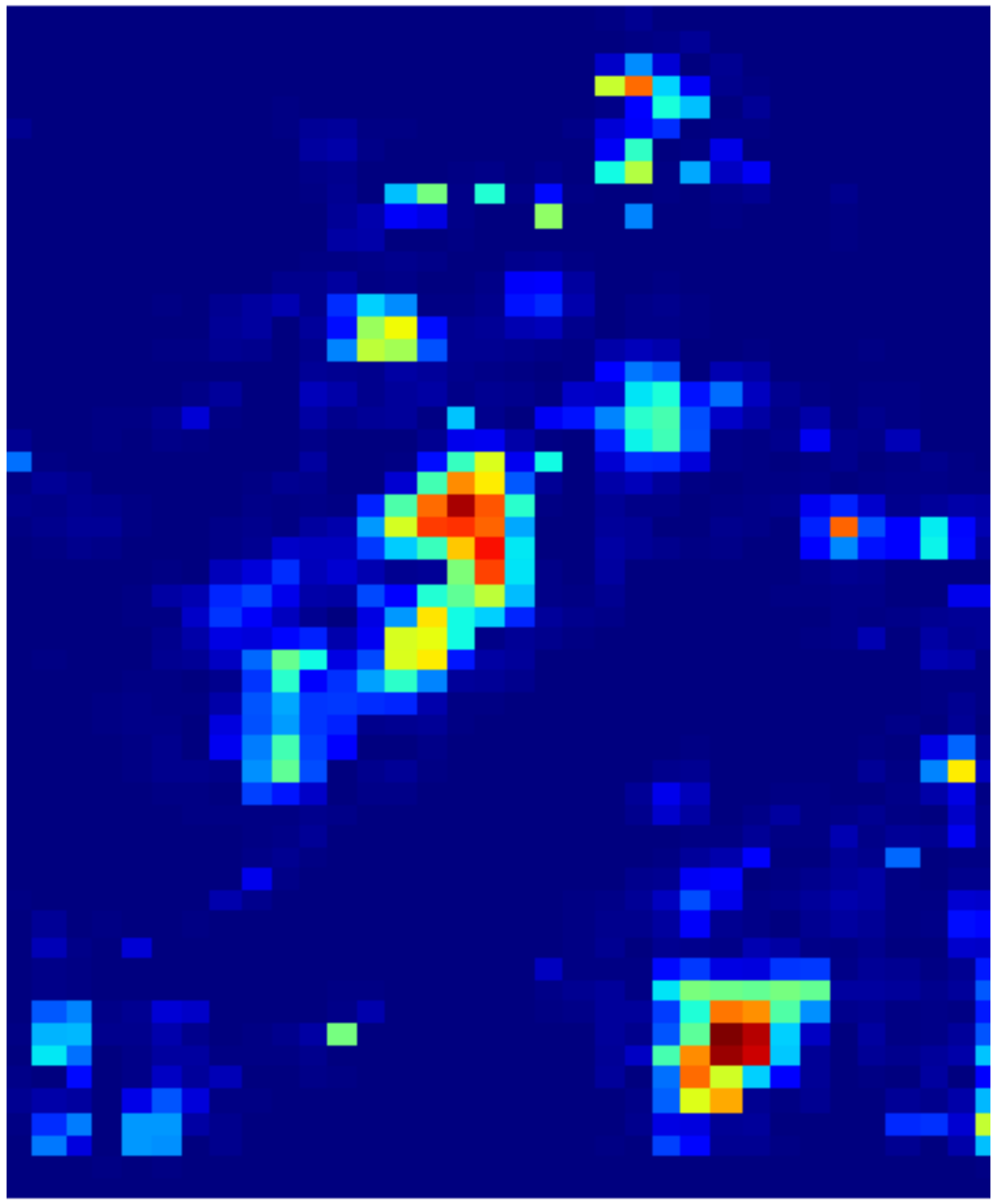
- Applied as sliding windows to input, map a window of pixels to a single binary output



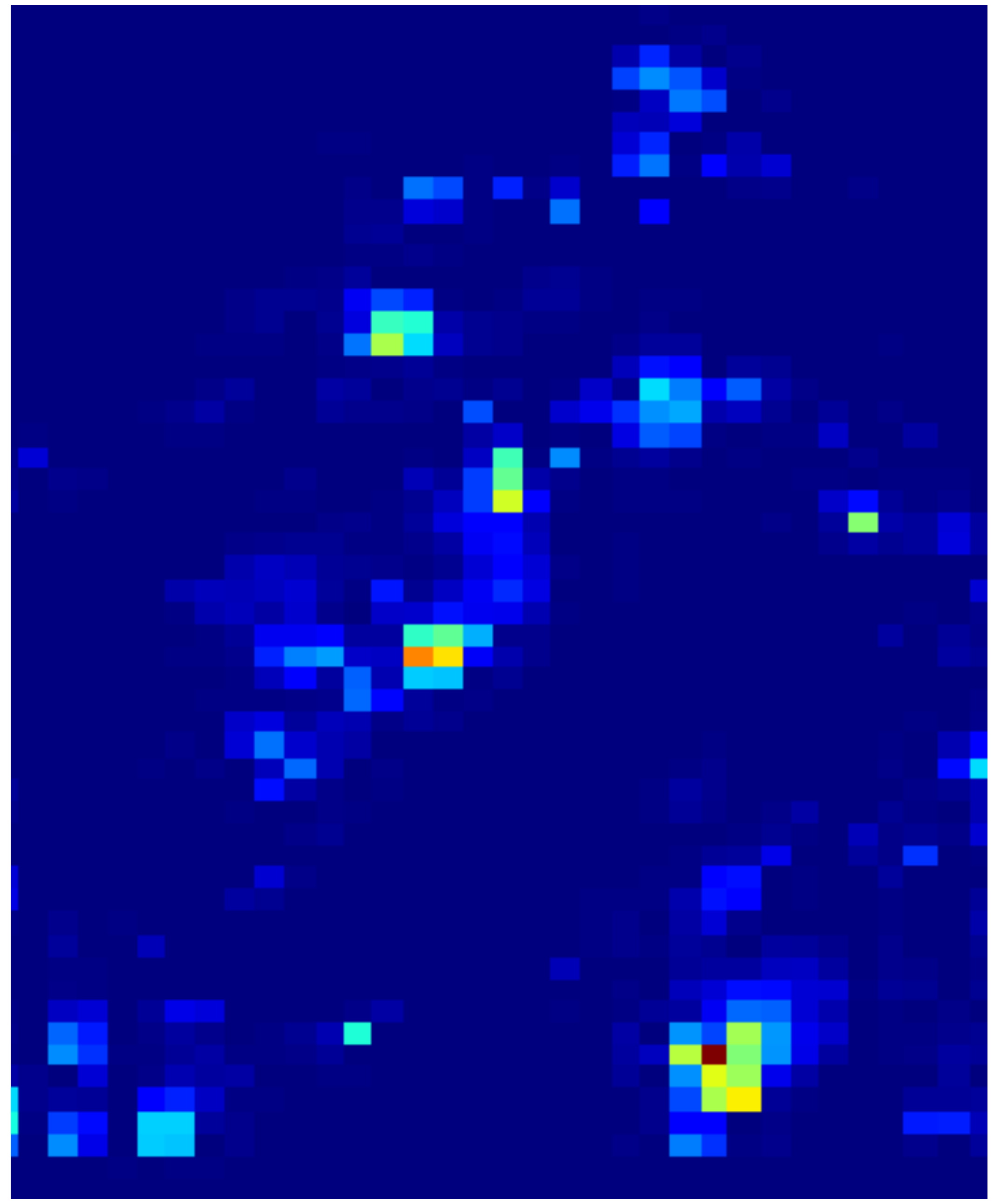
Output:Pose Confidence Maps

RGB and
joint predictions

Output before
Spatial Model

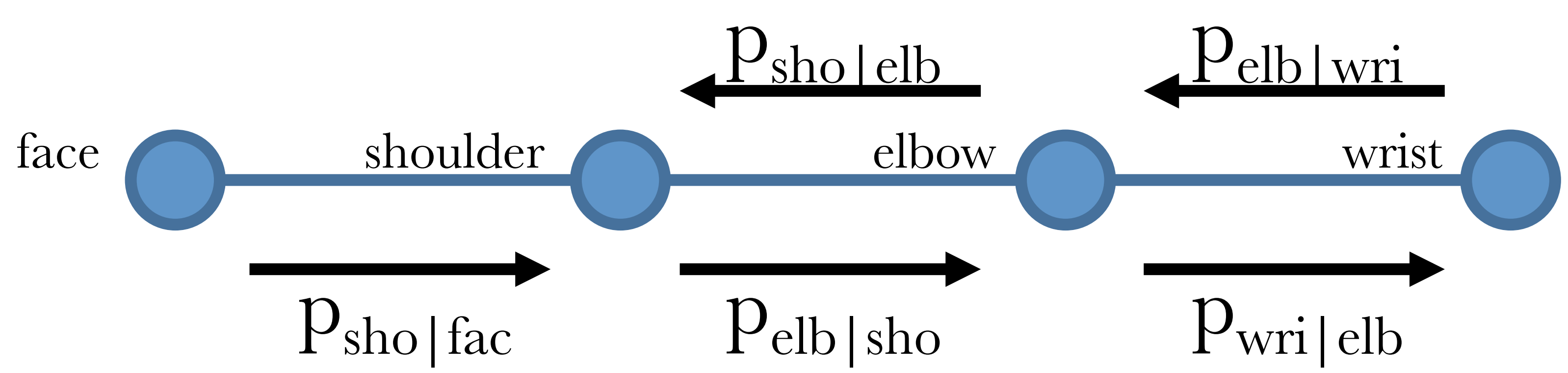


Output after
Spatial Model

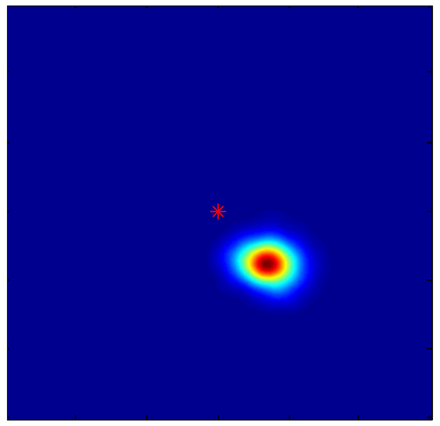


Spatial Model

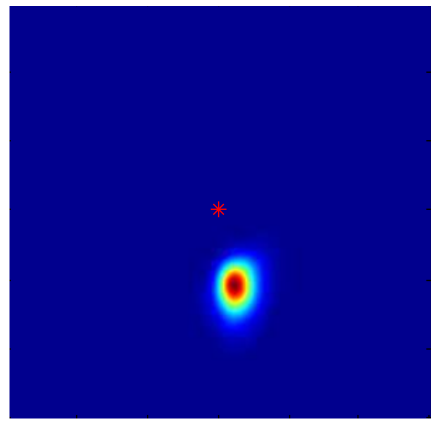
- Raw output of network produces many false positives
  - small image context
  - training set size limited
- Simple spatial model with body-pose priors can de-emphasize anatomically impossible poses
  - convnet provides unary distributions
  - body part priors fit to training data



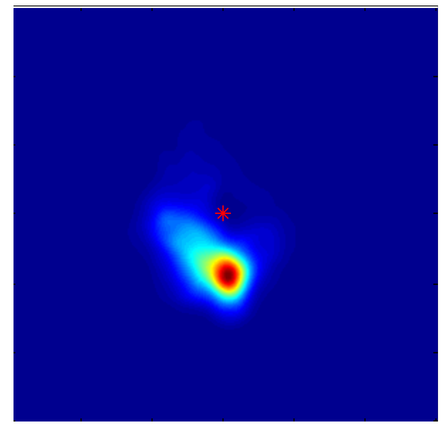
Spatial priors



$p_{\text{sho}|\text{fac}=\vec{0}}$



$p_{\text{elb}|\text{sho}=\vec{0}}$



$p_{\text{wri}|\text{elb}=\vec{0}}$

For a body part i with a set of neighbouring nodes U :

$$\hat{p}_i \propto {p_i}^{\lambda} \prod_{u \in U} \left(p_{i|u=\vec{0}} * p_u \right)$$

e.g. for the shoulder joint:

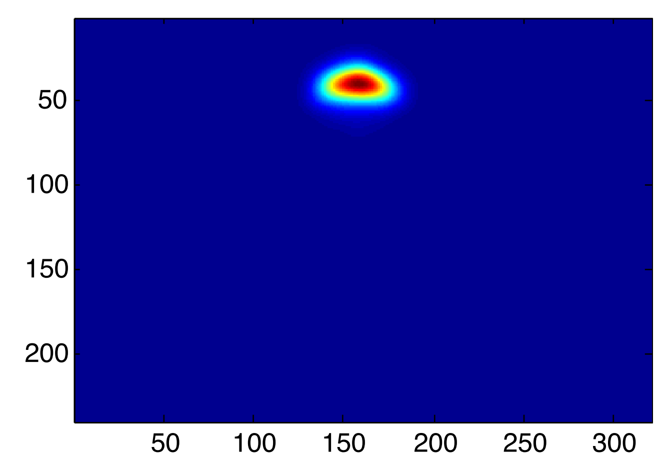
$$\log\left(\hat{p}_{\text{sho}}\right) \propto \lambda \ \log\left(p_{\text{sho}}\right) + \log\left(p_{\text{sho}|\text{fac}=\vec{0}} * p_{\text{fac}}\right) + \log\left(p_{\text{sho}|\text{elb}=\vec{0}} * p_{\text{elb}}\right)$$

$\lambda = 1$ in experiments

Face prior

- Incorporating image evidence from the shoulder joint to the filtered face distribution doesn't work
  - Due to the fact that the convnet already does a good job of localizing the face
  - Incorporating noisy evidence from the shoulder increases uncertainty
- Instead use a global position prior:

$$\log\left(\hat{p}_{\text{fac}}\right) \propto \lambda \ \log\left(p_{\text{fac}}\right) + \log\left(h_{\text{fac}}\right)$$

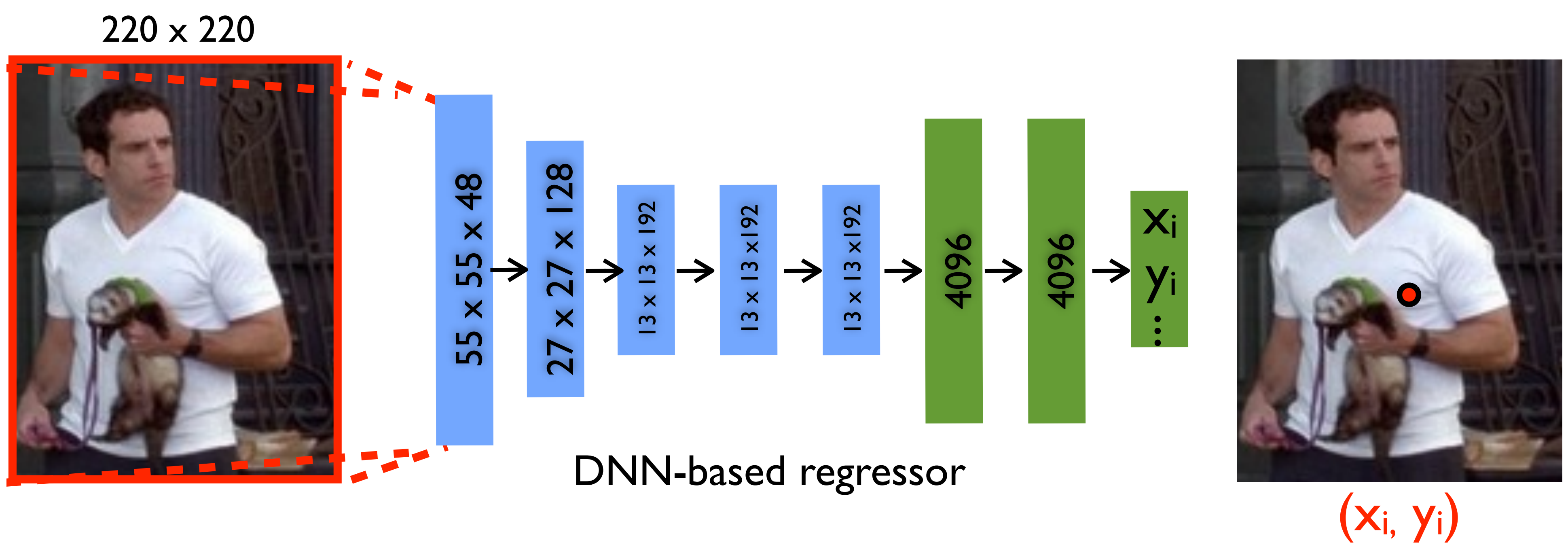


h_{fac}

DeepPose

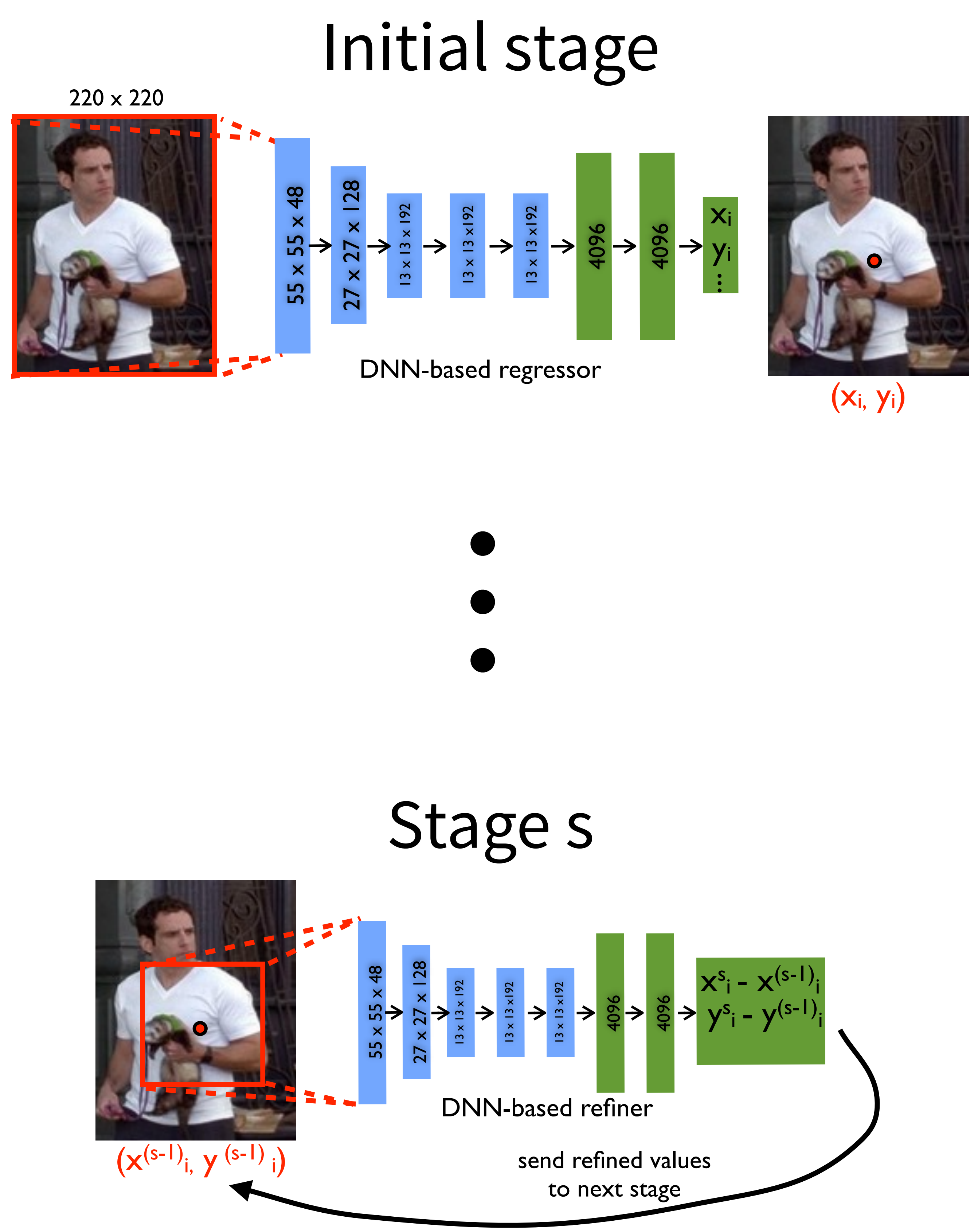
(Toshev and Szegedy 2014)

- Pose estimation as DNN-based regression
- Normalize joint co-ordinates w.r.t. human bounding box
- Normalize the image by the same box (crop human)
- “Alexnet” architecture



Cascade of pose regressors

- Joint estimation is based on the full image and therefore relies on context
- Fixed input size of 220 x 220, only captures pose at coarse scale
- Propose to train a cascade of regressors



Images: Toshev and Szegedy (2014)

Pose Estimation Datasets

- Frames Labeled In Cinema (FLIC, Sapp and Taskar 2013)
  - 6,543 training images, 1,016 test images
  - 10 upper-body joints
- Leeds Sports Dataset (Johnson and Everingham, 2010, 2011)
  - 11,000 training and 1,000 test images
  - 14 full-body joints

MPII Human Pose

(Andriluka et al. 2014)

- Addresses appearance variability and complexity
- YouTube as a data source
- Many activities, indoor and outdoor scenes, variety of imaging conditions

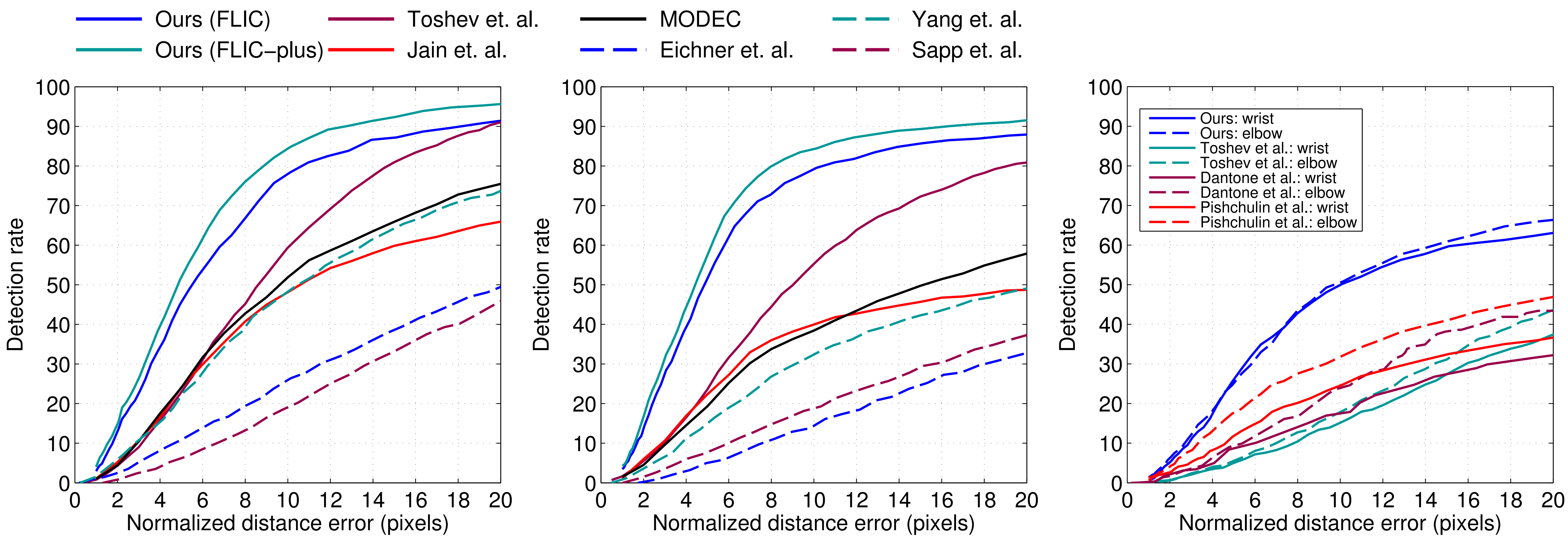
Dataset	#training	#test	img. type
Full body pose datasets			
Parse [16]	100	205	diverse
LSP [12]	1,000	1,000	sports (8 types)
PASCAL Person Layout [6]	850	849	everyday
Sport [21]	649	650	sports
UIUC people [21]	346	247	sports (2 types)
LSP extended [13]	10,000	-	sports (3 types)
FashionPose [2]	6,530	775	fashion blogs
J-HMDB [11]	31,838	-	diverse (21 act.)
Upper body pose datasets			
Buffy Stickmen [8]	472	276	TV show (Buffy)
ETHZ PASCAL Stickmen [3]	-	549	PASCAL VOC
Human Obj. Int. (HOI) [23]	180	120	sports (6 types)
We Are Family [5]	350 imgs.	175 imgs.	group photos
Video Pose 2 [18]	766	519	TV show (Friends)
FLIC [17]	6,543	1,016	feature movies
Sync. Activities [4]	-	357 imgs.	dance / aerobics
Armlets [9]	9,593	2,996	PASCAL VOC/Flickr
MPII Human Pose (this paper)	**28,821**	**11,701**	diverse (491 act.)

Metrics

- Percentage of Correct Parts (PCP)
  - measures detection rate of limbs
  - penalizes shorter limbs
- Percent of Detected Joints (PDJ)
  - distance b/w detected and true joint within certain (varying) fraction of the torso diameter

State-of-the-art

(Jain et al. 2014)



(a) FLIC: Elbow (b) FLIC: Wrist (c) LSP: Wrist and Elbow

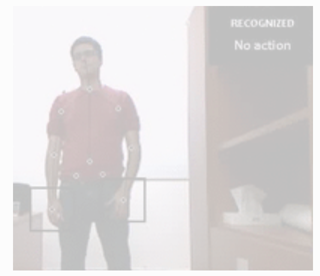
Enhanced version of the model described earlier:

- more efficient sliding-window convnet
- learn spatial prior model structure

Pose Estimation

Tracking

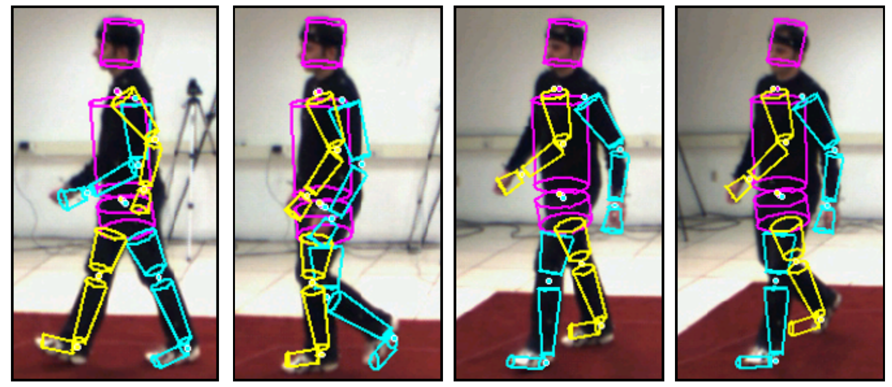
Activity /Gesture



3-D Human Pose Tracking

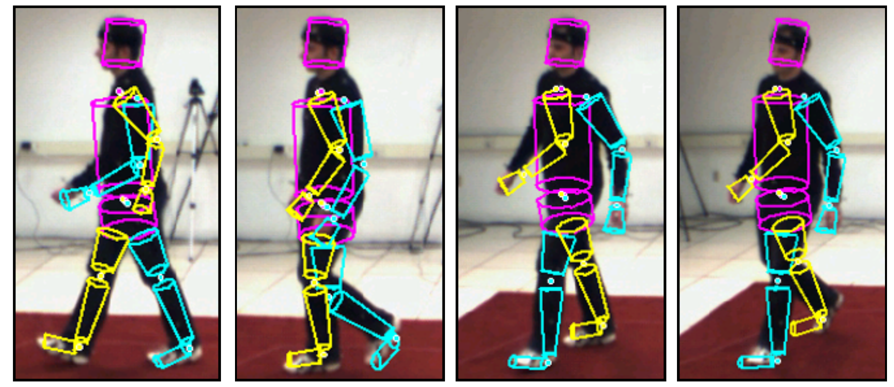
3-D Human Pose Tracking

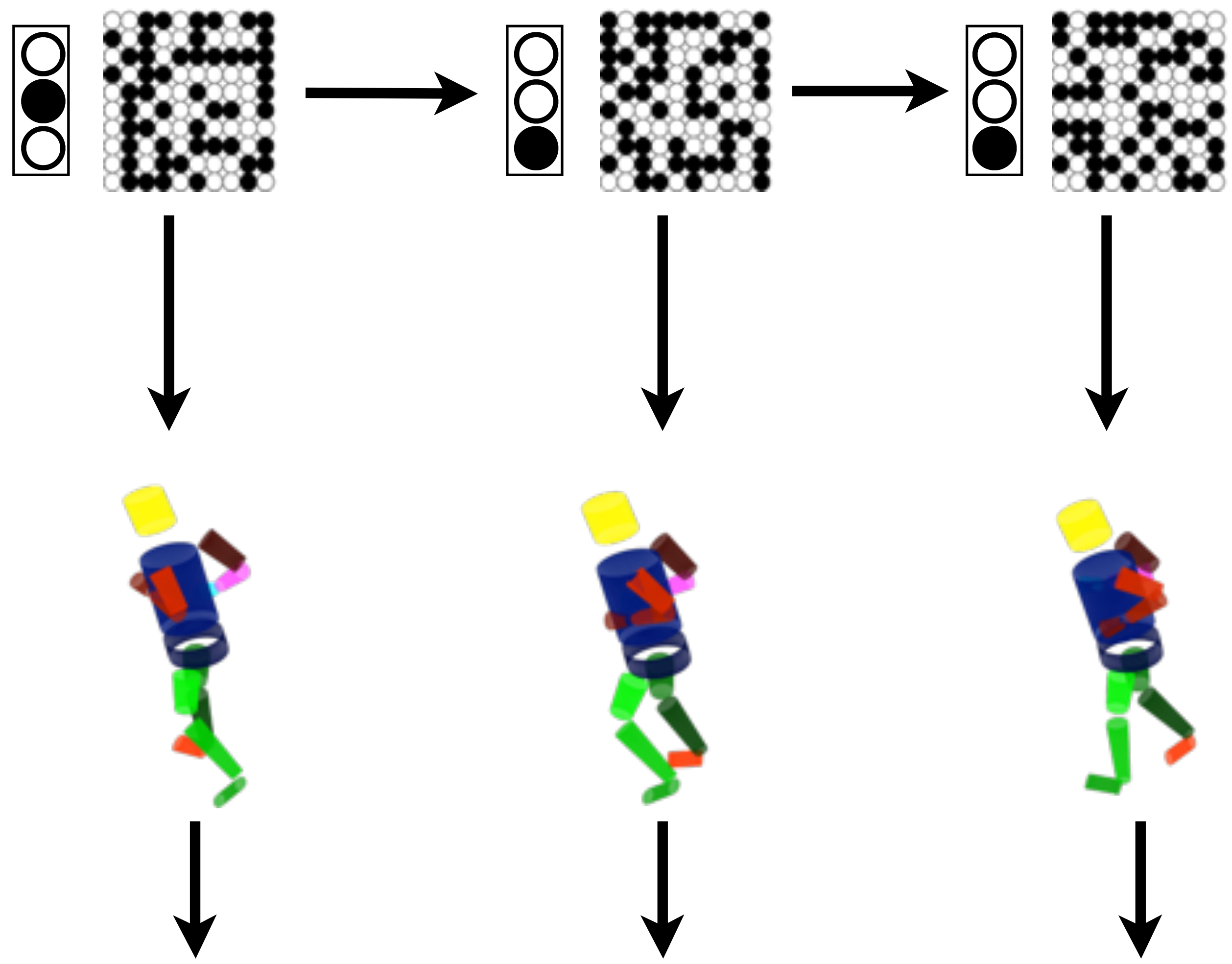
- Pose estimation + time element



3-D Human Pose Tracking

- Pose estimation + time element
- We will investigate methods which learn a dynamical prior using motion capture data
    - intuition: if you understand the way people move, you can make a good prediction of where they will be at the next frame





Prior Models of Human Pose and Motion

Prior work	Limitations
Linear models (*Sidenbladh et al. ‘00, Balan et al. ‘05, Deutscher & Reid ‘05*)	• Nonlinear dynamics not captured
Switching LDS (*Pavlovic et al. ‘99*)	• Inference is complicated • Difficulty modeling transitions
Nonlinear dimension reduction *(Sminchisescu & Jepson ‘04, Lee & Elgammal ‘07, Lu & Carreira-Perpinan ‘07, Li et al. ‘07)*	• Poor generalization
GPLVM / GPDM (*Urtasun et al. ‘05,’06*)	• Only small training corpora

Implicit Mixtures of CRBMs

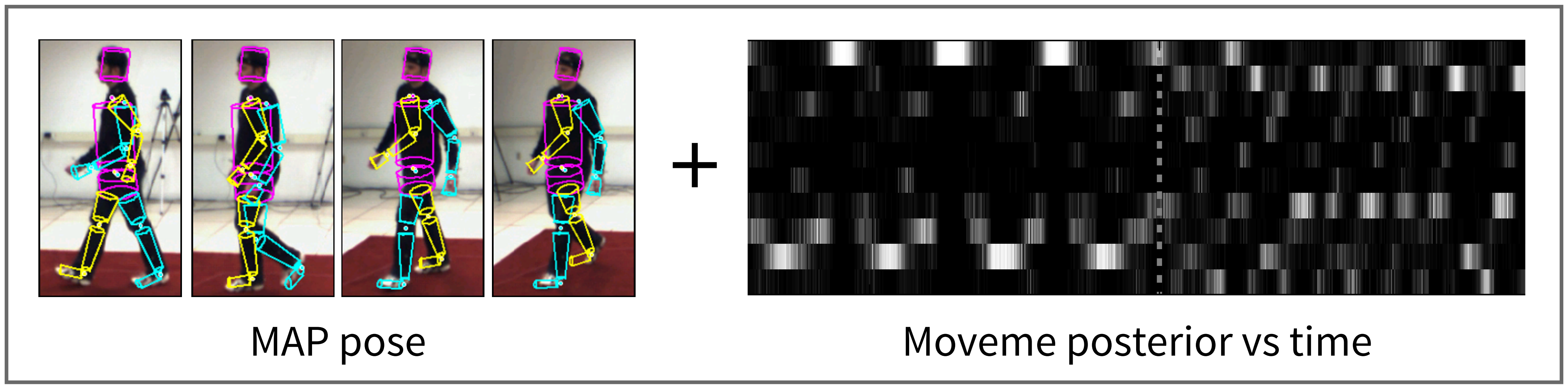
(Taylor et al. 2010)

- Very large datasets, stylistic diversity and multiple activities
- Supervised with activity labels, or unsupervised with automatic discovery of atomic motions ("movemes")
- Simultaneous inference of pose and activity

Implicit Mixtures of CRBMs

(Taylor et al. 2010)

- Very large datasets, stylistic diversity and multiple activities
- Supervised with activity labels, or unsupervised with automatic discovery of atomic motions ("movemes")
- Simultaneous inference of pose and activity



Bayesian Filtering w/ imCRBM

Latent variables:

$\mathbf{q}$: discrete activity

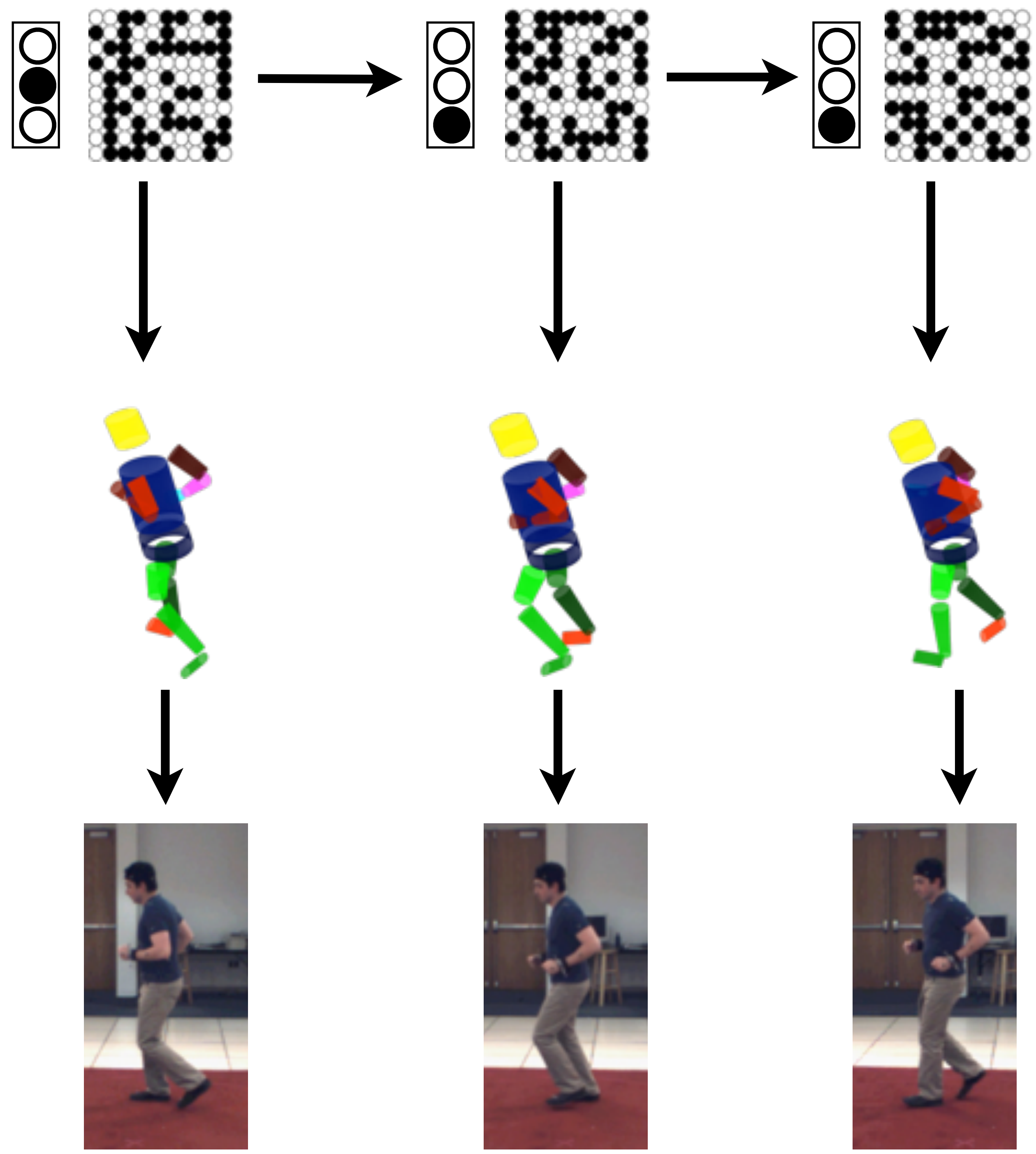
$\mathbf{z}$: multivariate binary (shared among activities)

3D pose: $\mathbf{x}$

- observed for learning
- latent during tracking

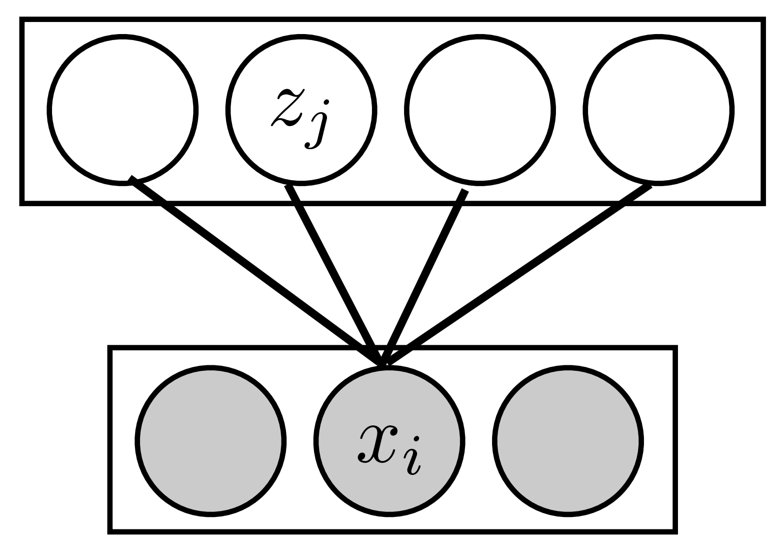
Image features: $\mathbf{y}$

- always observed



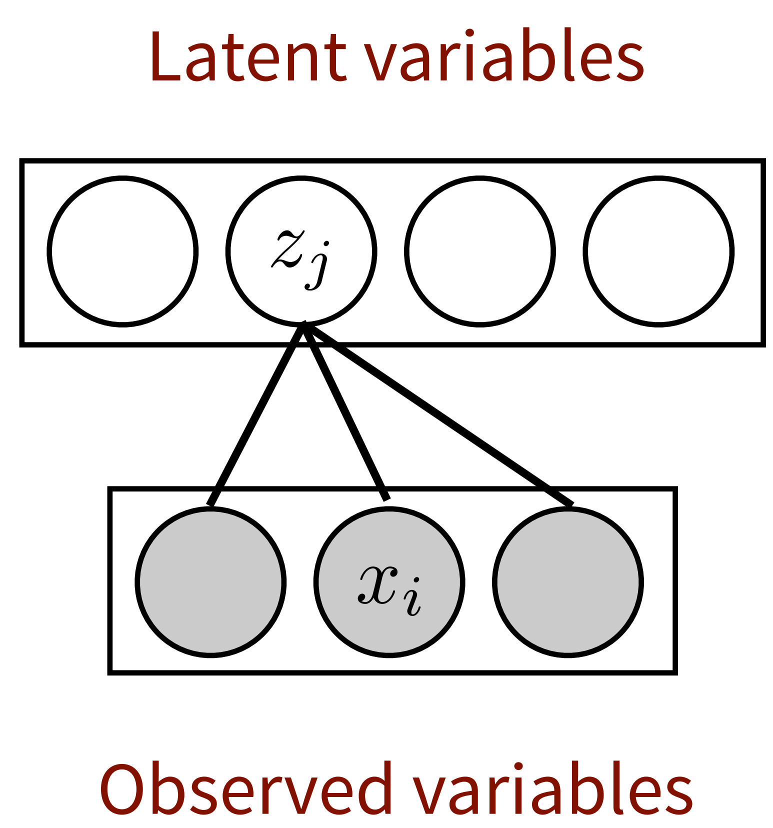
Restricted Boltzmann Machines (RBM) - Review

- Continuous observed variables (pose)
- Binary latent variables (capture pose/dynamics)
- Efficient, exact inference (bipartite connectivity)
- Can be stacked



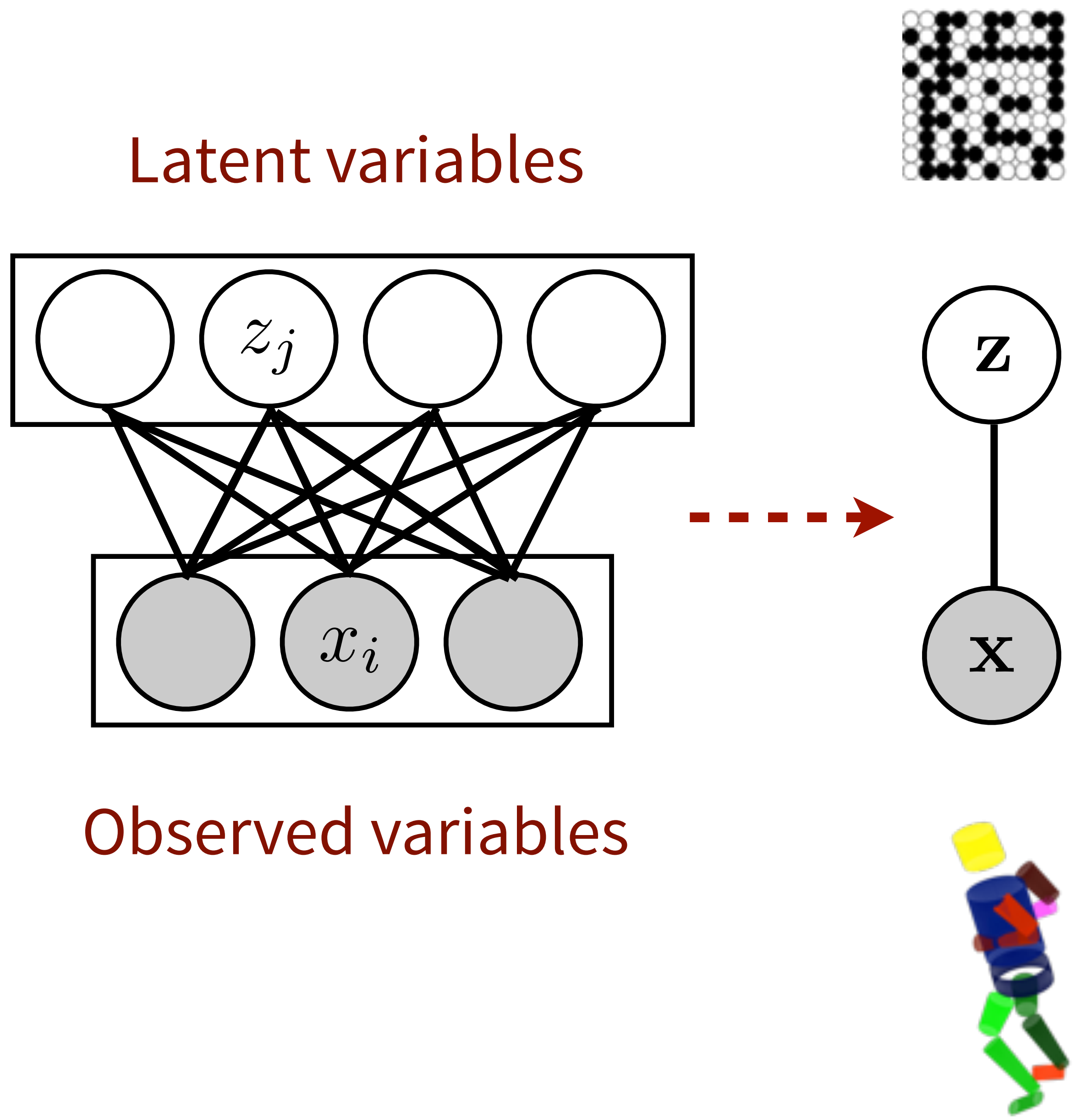
Restricted Boltzmann Machines (RBM) - Review

- Continuous observed variables (pose)
- Binary latent variables (capture pose/dynamics)
- Efficient, exact inference (bipartite connectivity)
- Can be stacked

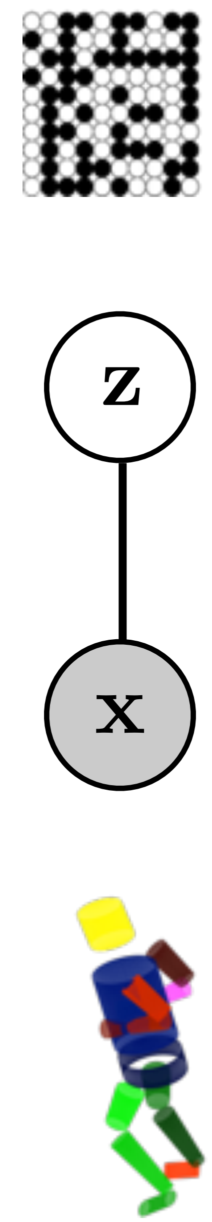


Restricted Boltzmann Machines (RBM) - Review

- Continuous observed variables (pose)
- Binary latent variables (capture pose/dynamics)
- Efficient, exact inference (bipartite connectivity)
- Can be stacked

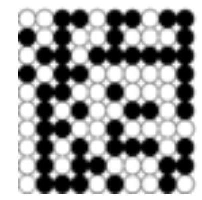


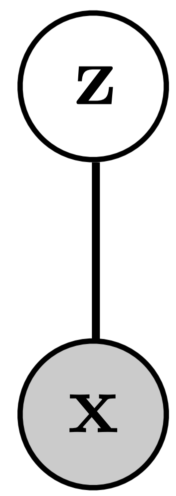
Conditional Restricted Boltzmann Machines (CRBM)



Conditional Restricted Boltzmann Machines (CRBM)

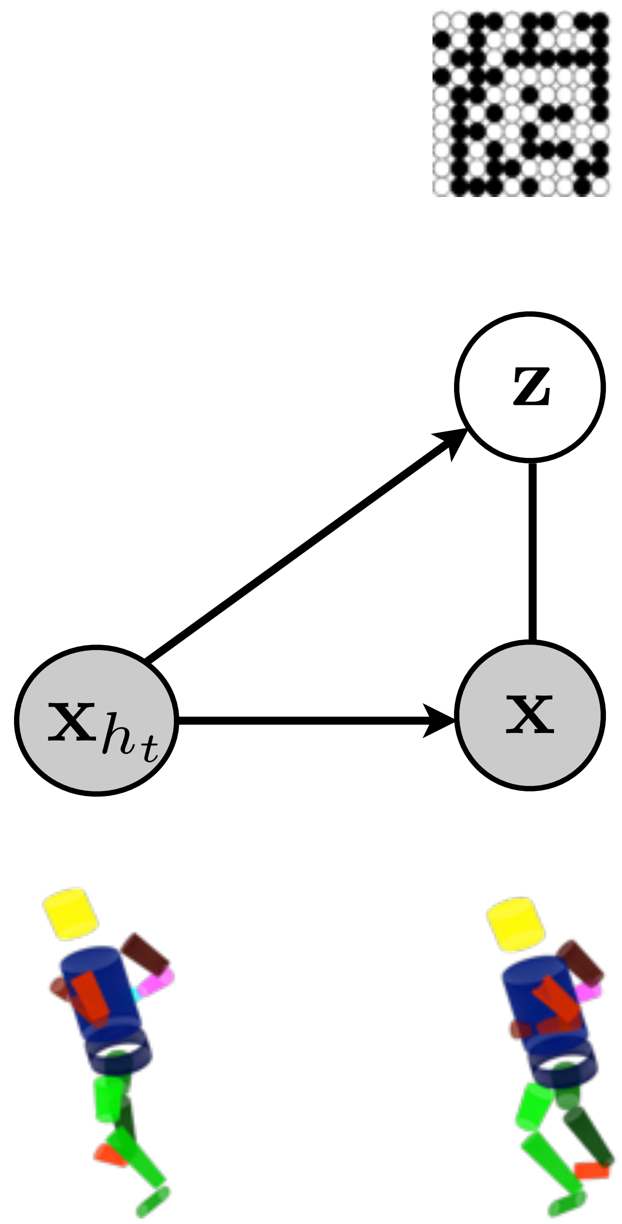
- Extend RBM to capture temporal dependencies





Conditional Restricted Boltzmann Machines (CRBM)

- Extend RBM to capture temporal dependencies
- Observed and latent variables conditioned on the observation history

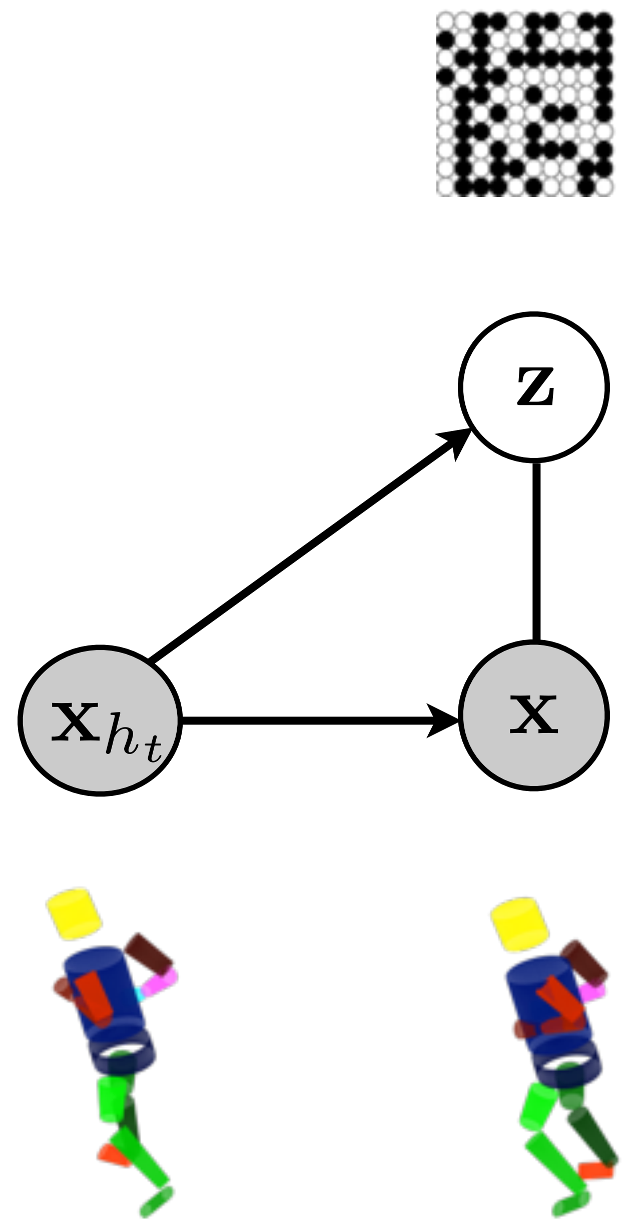


Conditional Restricted Boltzmann Machines (CRBM)

- Extend RBM to capture temporal dependencies

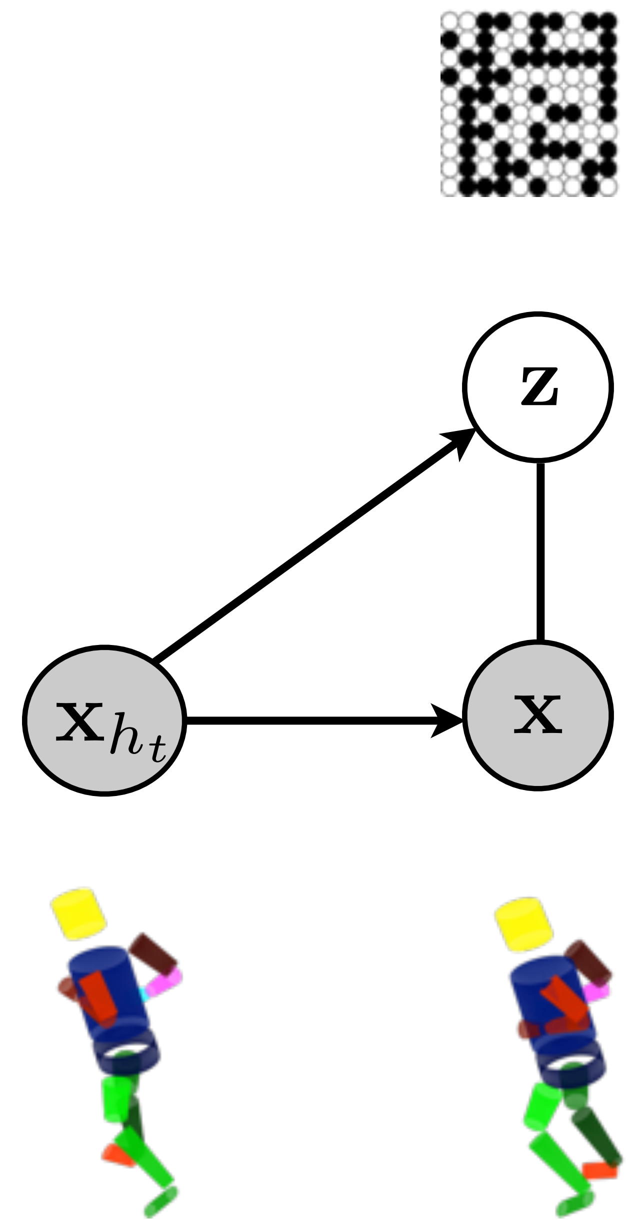
- Observed and latent variables conditioned on the observation history

- Inference and learning unchanged



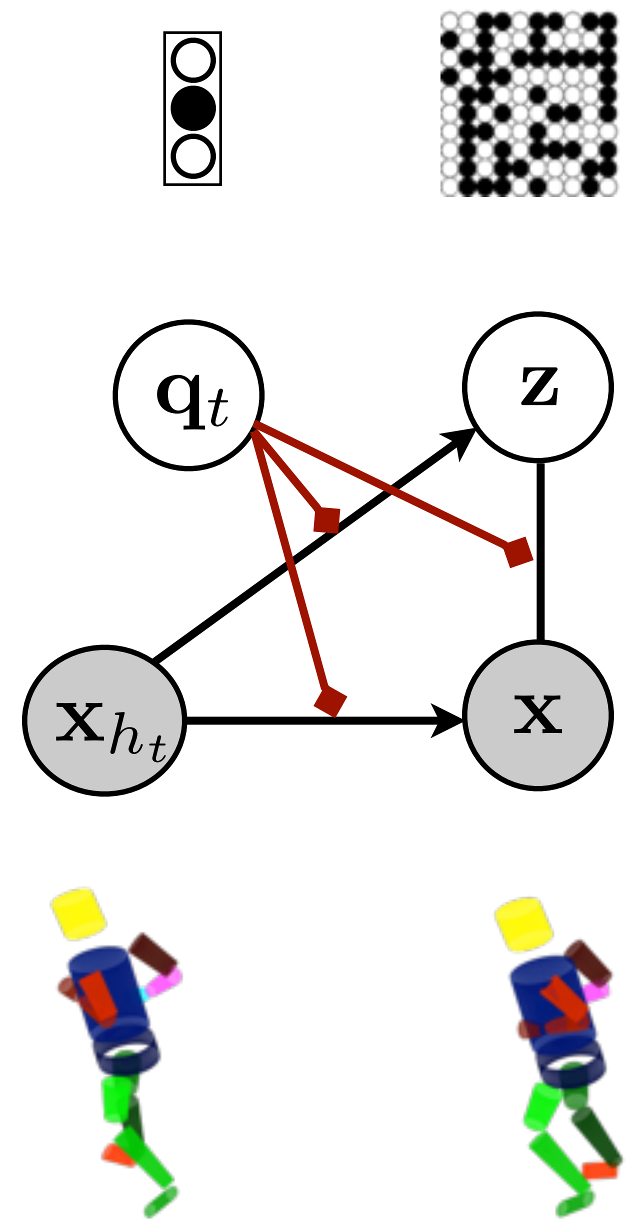
Conditional Restricted Boltzmann Machines (CRBM)

- Extend RBM to capture temporal dependencies
- Observed and latent variables conditioned on the observation history
- Inference and learning unchanged
- Proposed for motion synthesis (Taylor et al. 2006)



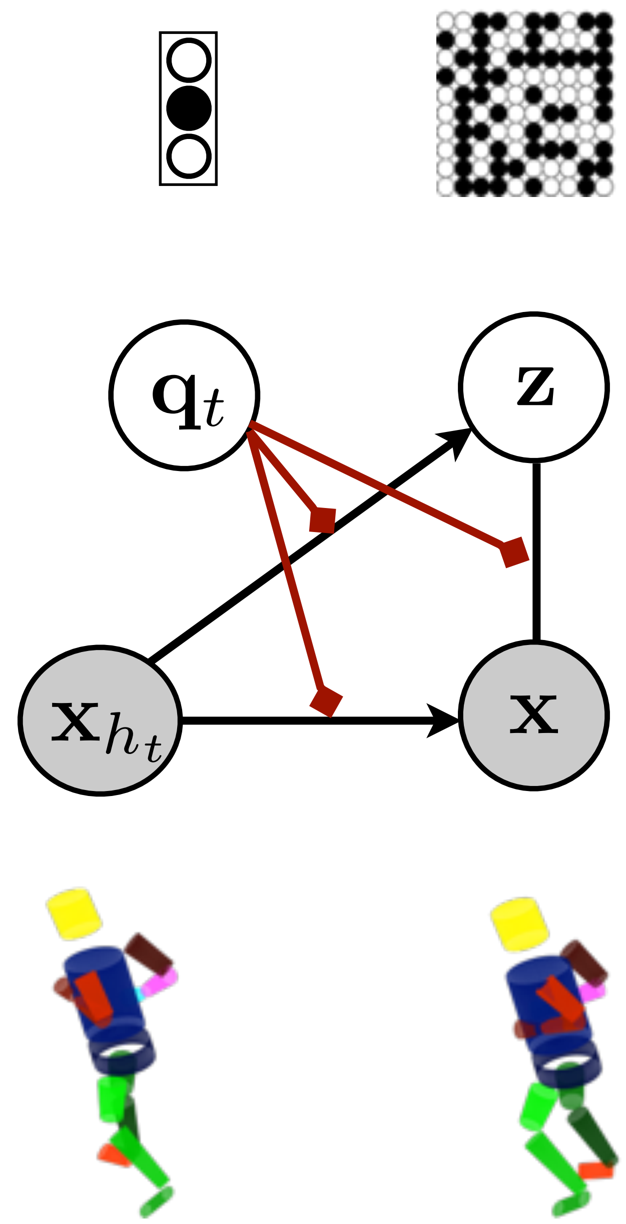
Implicit mixture of CRBMs (imCRBM)

Discrete component variable sets the "effective" CRBM



Implicit mixture of CRBMs (imCRBM)

Discrete component variable sets the "effective" CRBM



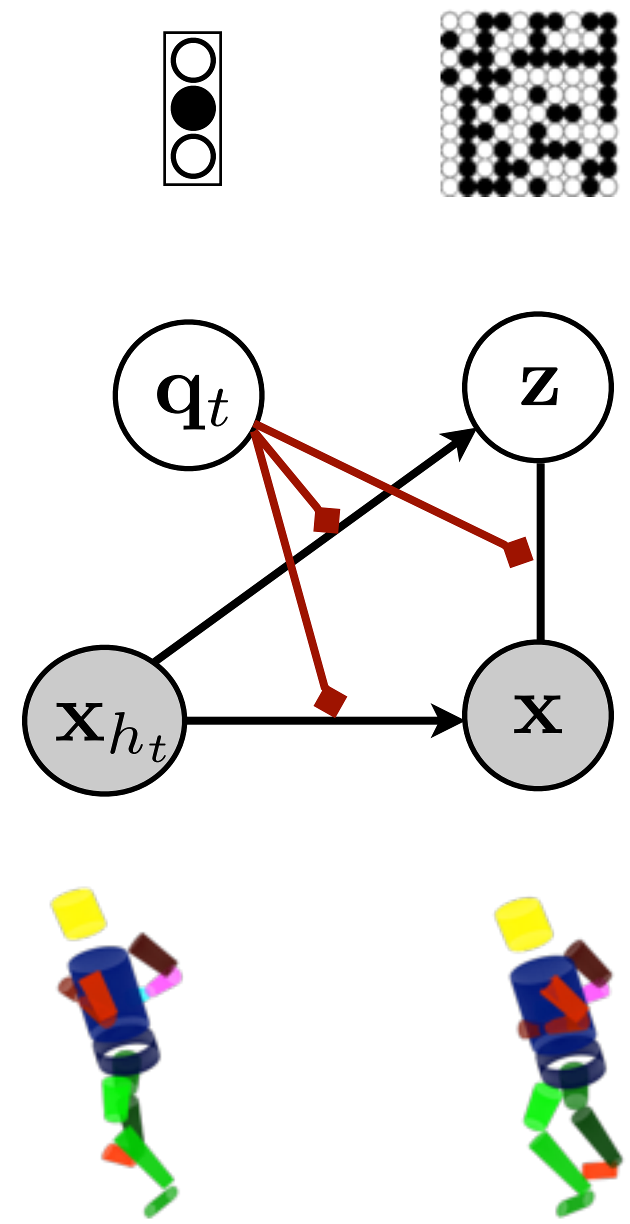
Marginalize over latent variables to obtain dynamical mixture model

$$p(\mathbf{x}_t|\mathbf{x}_{h_t}) = \sum_{\mathbf{z}_t,\mathbf{q}_t} p(\mathbf{x}_t, \mathbf{z}_t, \mathbf{q}_t|\mathbf{x}_{h_t})$$

$$= \sum_{k=1}^{K} p(\mathbf{q}_t = k) \sum_{\mathbf{z}_t} p(\mathbf{x}_t, \mathbf{z}_t|\mathbf{q}_t = k, \mathbf{x}_{h_t})$$

Advantages of the imCRBM

- Approximate learning by contrastive divergence (or PCD, or Minimum Probability Flow, or…)
- Can be trained on 10^6 frames in a few hours (minutes on GPUs)
- Gibbs sampling is simple and fast for synthesis (at 60Hz)
- Training can be done with and without activity labels



Tracking via Bayesian Filtering

Filtering distribution:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t)\, p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$$

Tracking via Bayesian Filtering

Filtering distribution:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t)\,p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$$

posterior likelihood prediction

Tracking via Bayesian Filtering

Filtering distribution:

$$p(\mathbf{x}_t | \mathbf{y}_{1:t}) \propto p(\mathbf{y}_t | \mathbf{x}_t) \, p(\mathbf{x}_t | \mathbf{y}_{1:t-1})$$

Predictive distribution:

$$p(\mathbf{x}_t | \mathbf{y}_{1:t-1}) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t | \mathbf{x}_{t-1}) \, p(\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1}) \, d\mathbf{x}_{t-1}$$

dynamical model

posterior

Tracking via Bayesian Filtering

Filtering distribution:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t)\, p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$$

Predictive distribution:

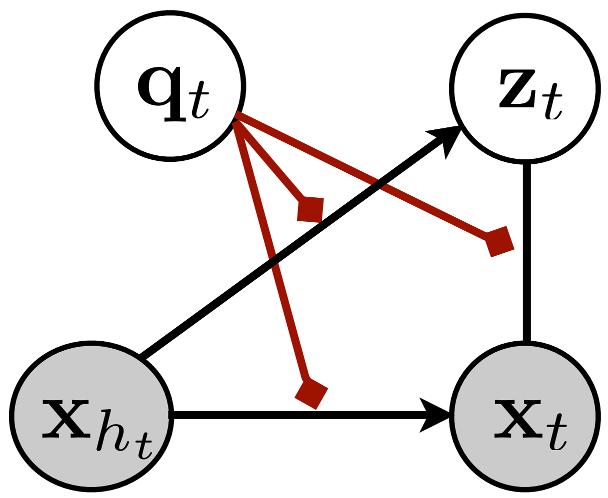
$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})\, p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\, d\mathbf{x}_{t-1}$$

dynamical model

posterior

Inference: Particle filter

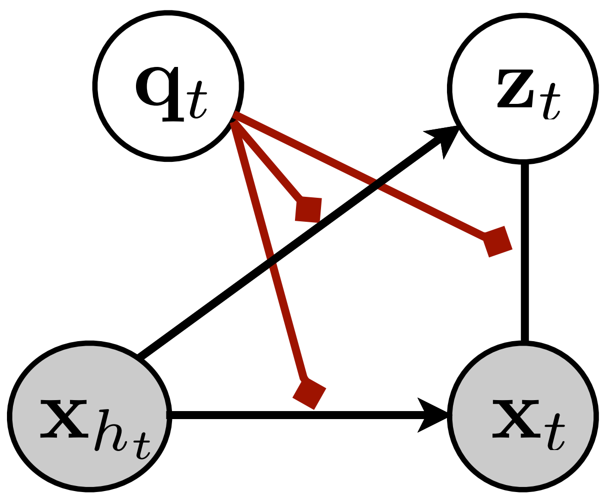
Bayesian Filtering



Dynamical Model:

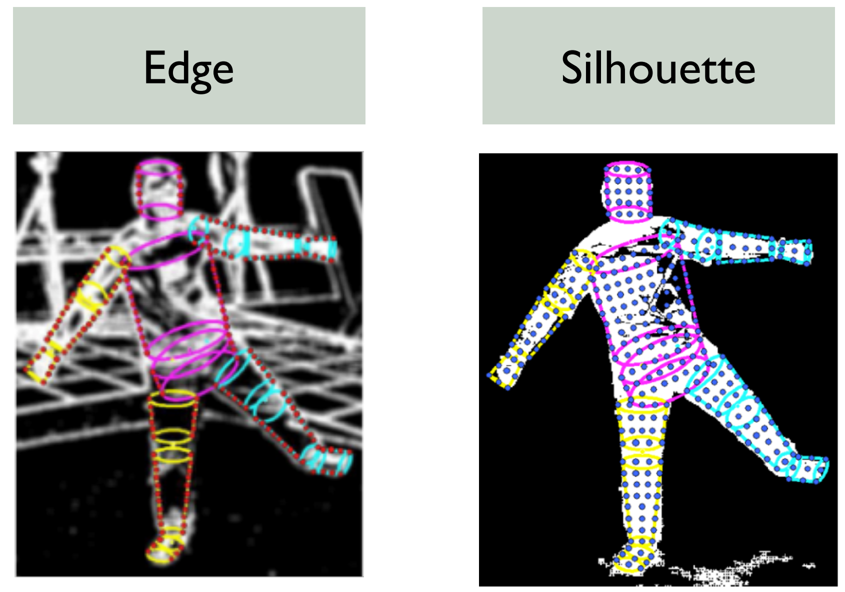
$$p(\mathbf{x}_t \mid \mathbf{x}_{h_t})$$

Bayesian Filtering



Dynamical Model:

$$p(\mathbf{x}_t \mid \mathbf{x}_{h_t})$$



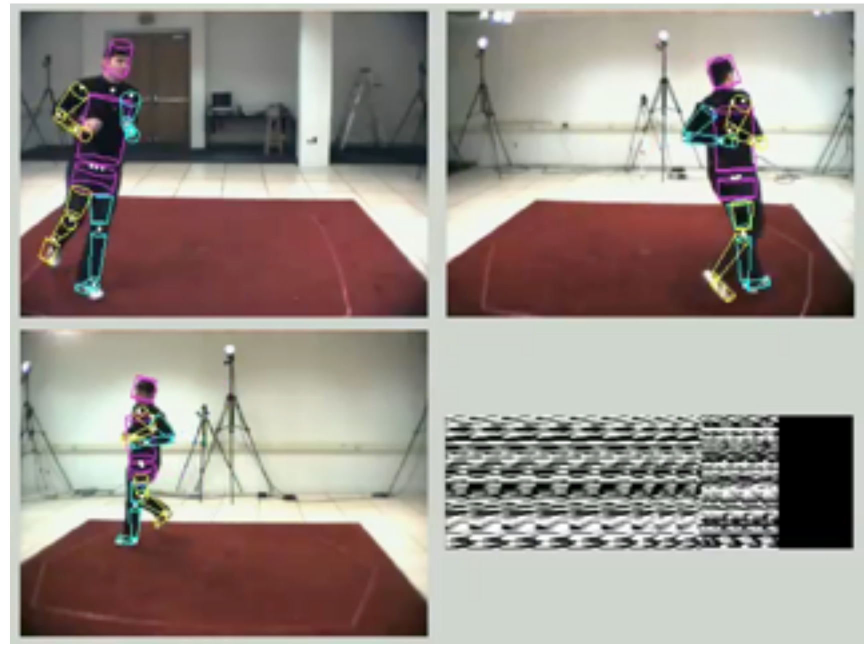
Likelihood:

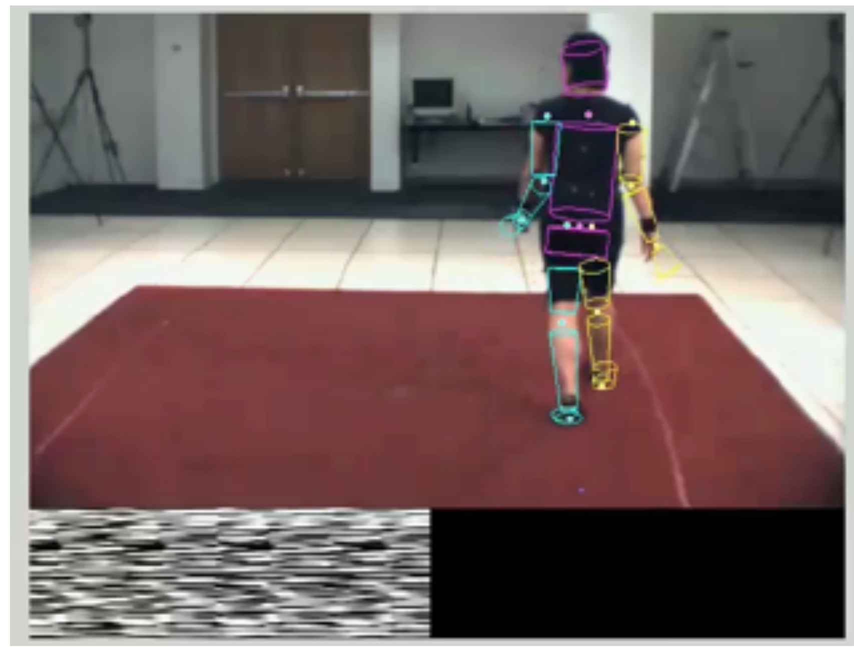
$$p(\mathbf{y}_t \mid \mathbf{x}_t)$$

(*Deutscher & Reid '05, Balan et al. '05*)

Experiments

- Multi-view and monocular 3D tracking
- HumanEva: multi-view sequences with synchronized mocap data for training and quantitative evaluation
- Comparisons: annealed particle filter with smooth zero-order dynamics (baseline) and other state-of-the-art methods
- Performance measure: Average joint location error (mm)





Multi-view: Walking + Jogging with Transitions

Model	Error (mm)
Baseline	164.2±25.0
CRBM	81.9±12.4
imCRBM-2L	**60.2±1.2**
imCRBM-2L*	75.5±1.8
imCRBM-10U	**75.8±1.7**
imCRBM-10U*	84.7±1.1

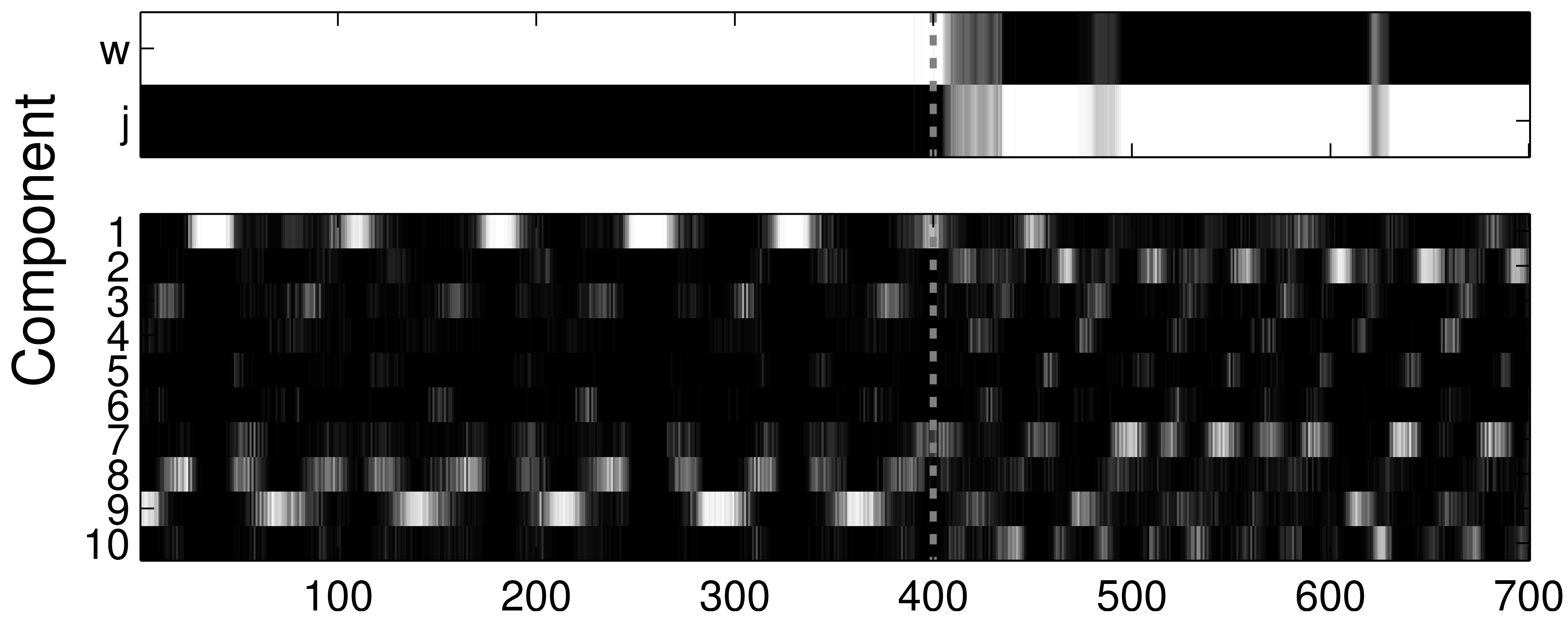
Pose estimation and segmentation:

imCRBM-2L (supervised)

Multi-view: Walking + Jogging with Transitions

Model	Error (mm)
Baseline	164.2±25.0
CRBM	81.9±12.4
imCRBM-2L	**60.2±1.2**
imCRBM-2L*	75.5±1.8
imCRBM-10U	**75.8±1.7**
imCRBM-10U*	84.7±1.1

Pose estimation and segmentation:

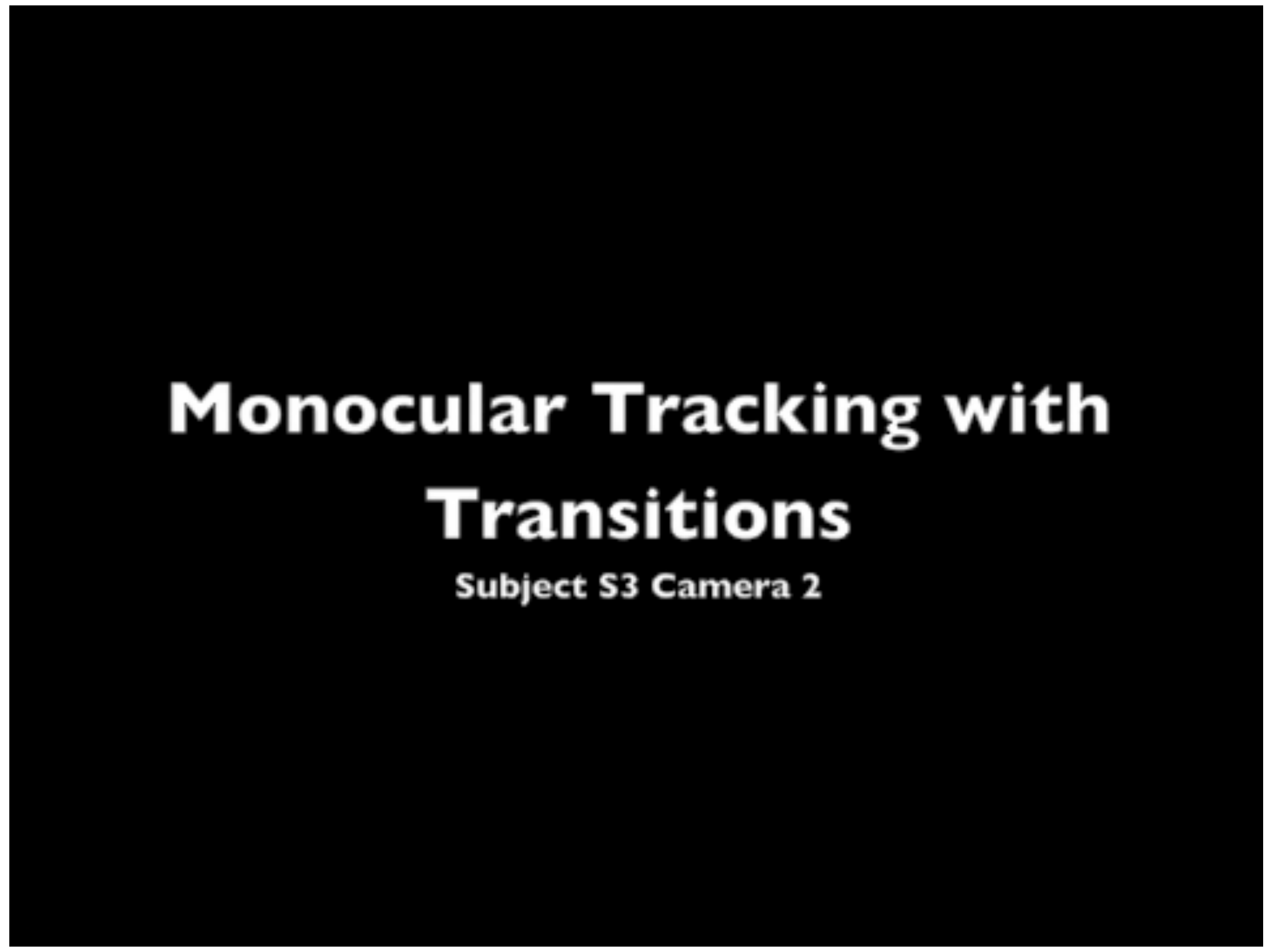


imCRBM-2L (supervised)

imCRBM-10U (unsupervised)

Monocular tracking with transitions (imCRBM-2L)

- This is a very challenging scenario at which both the baseline and CRBM fail
- We track with imCRBM-2L on each of the 3 views independently and report performance averaged over 5 runs

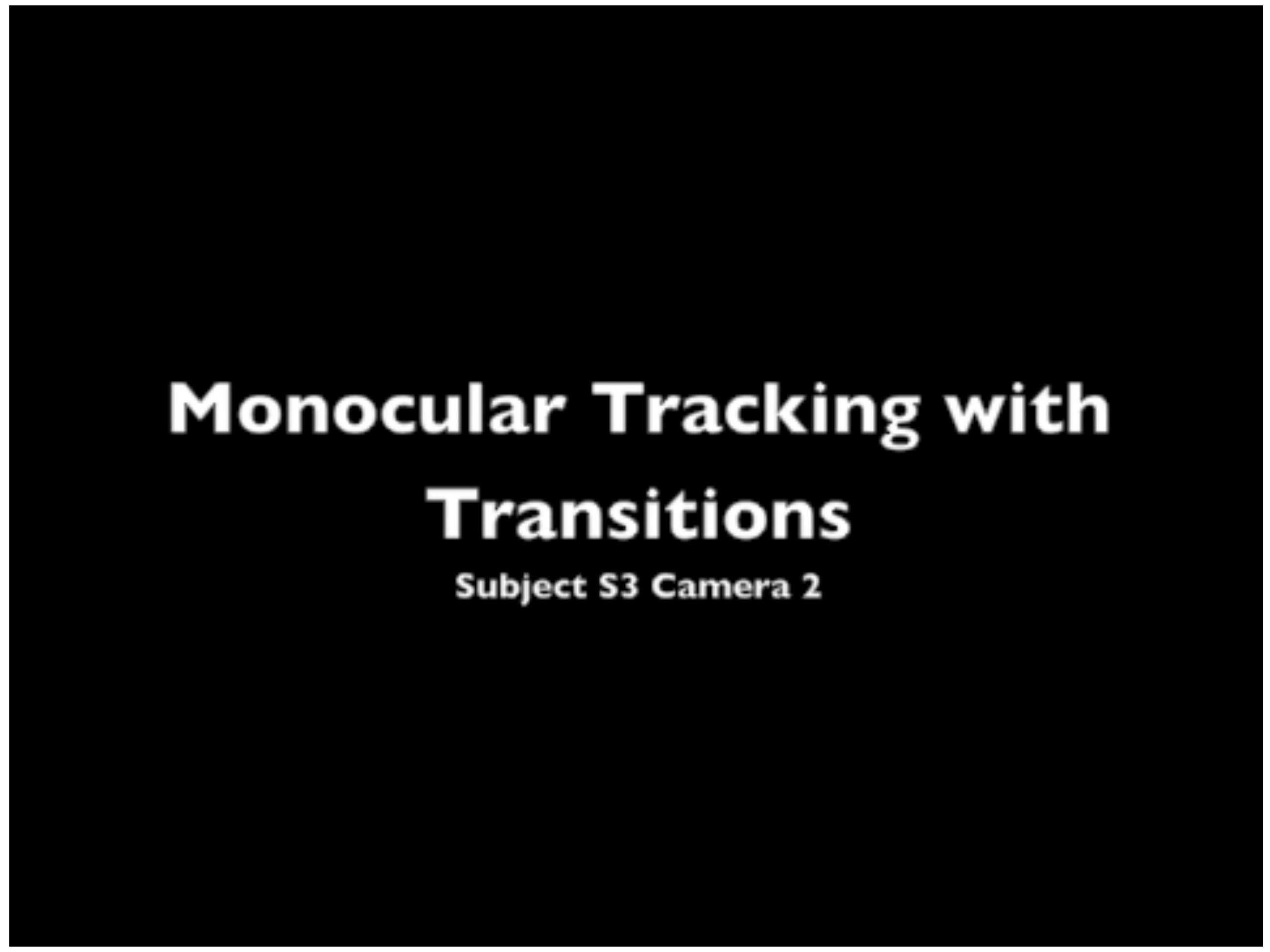


	Relative Error (mm)
Camera 1	118.9±33.1
Camera 2	84.26±6.9
Camera 3	90.4±7.6

Monocular tracking with transitions (imCRBM-2L)

- This is a very challenging scenario at which both the baseline and CRBM fail

- We track with imCRBM-2L on each of the 3 views independently and report performance averaged over 5 runs

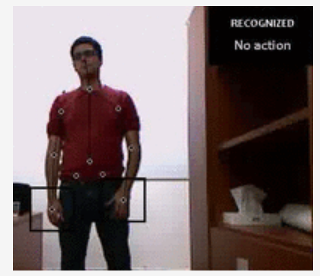


	Relative Error (mm)
Camera 1	118.9±33.1
Camera 2	84.26±6.9
Camera 3	90.4±7.6

Pose Estimation

Tracking

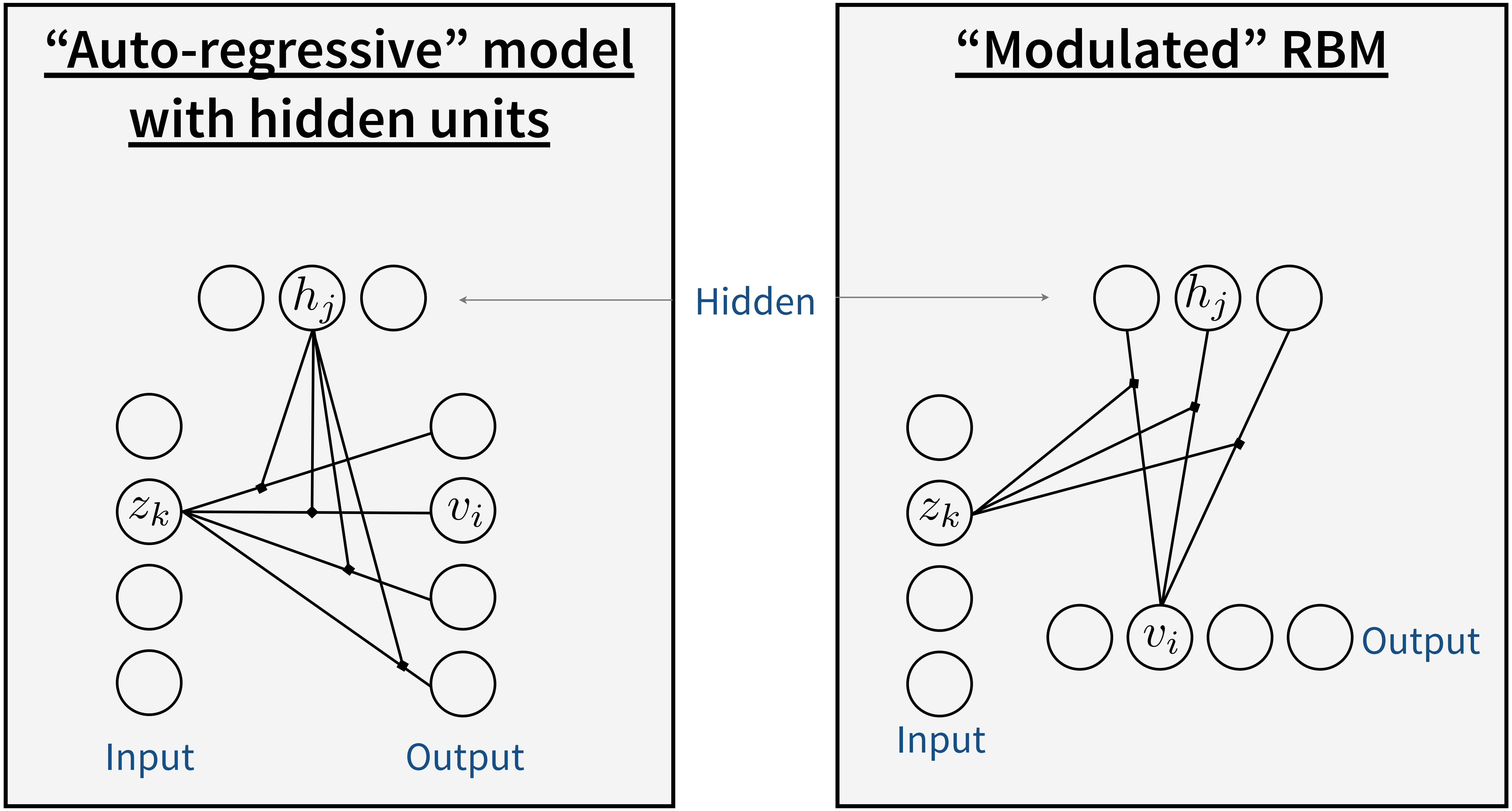
Activity /Gesture



Hybrid Unsupervised/Supervised

Gated RBM (Two views)

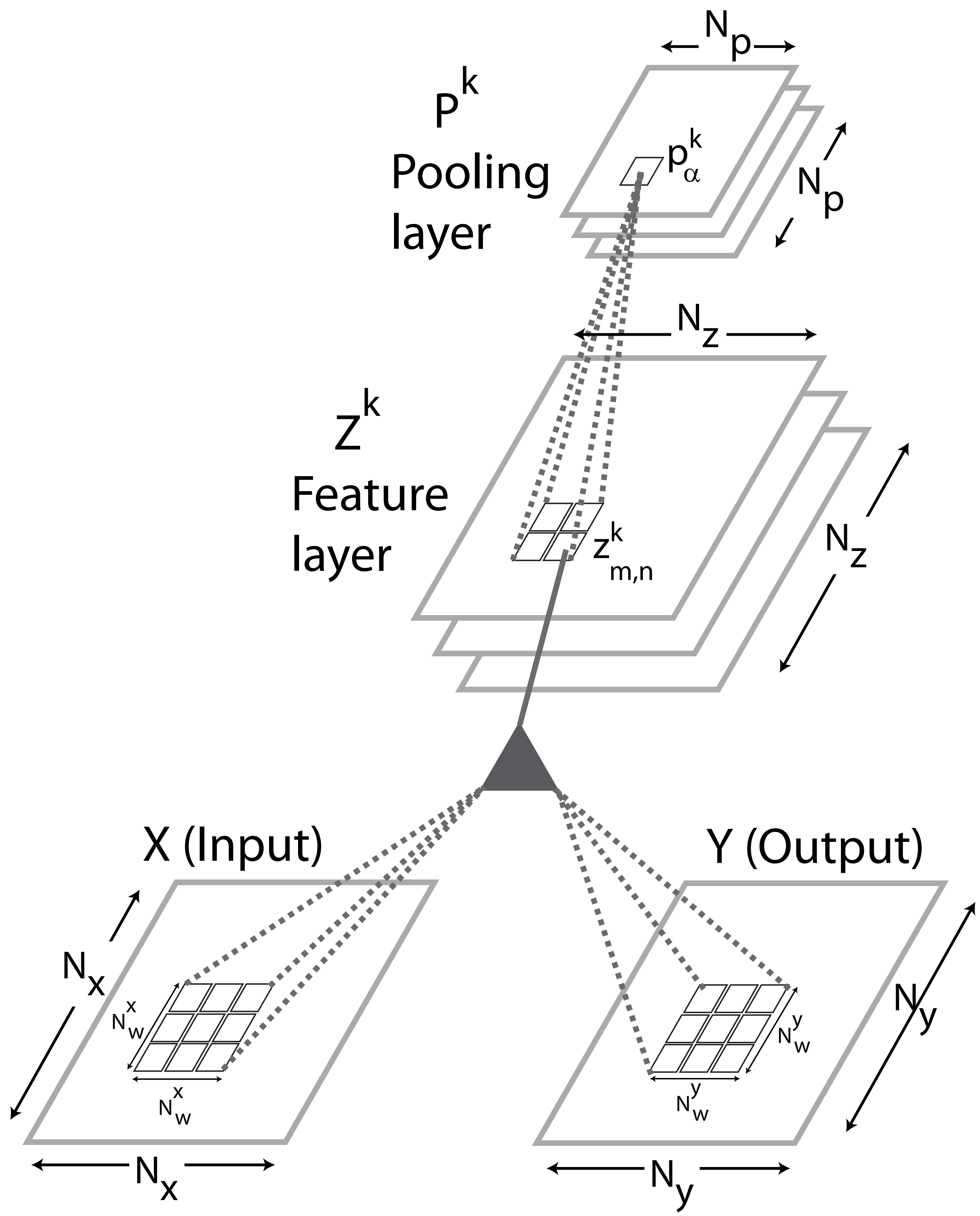
(Memisevic and Hinton, 2007)



Convolutional Gated RBM

(Taylor et al. 2010)

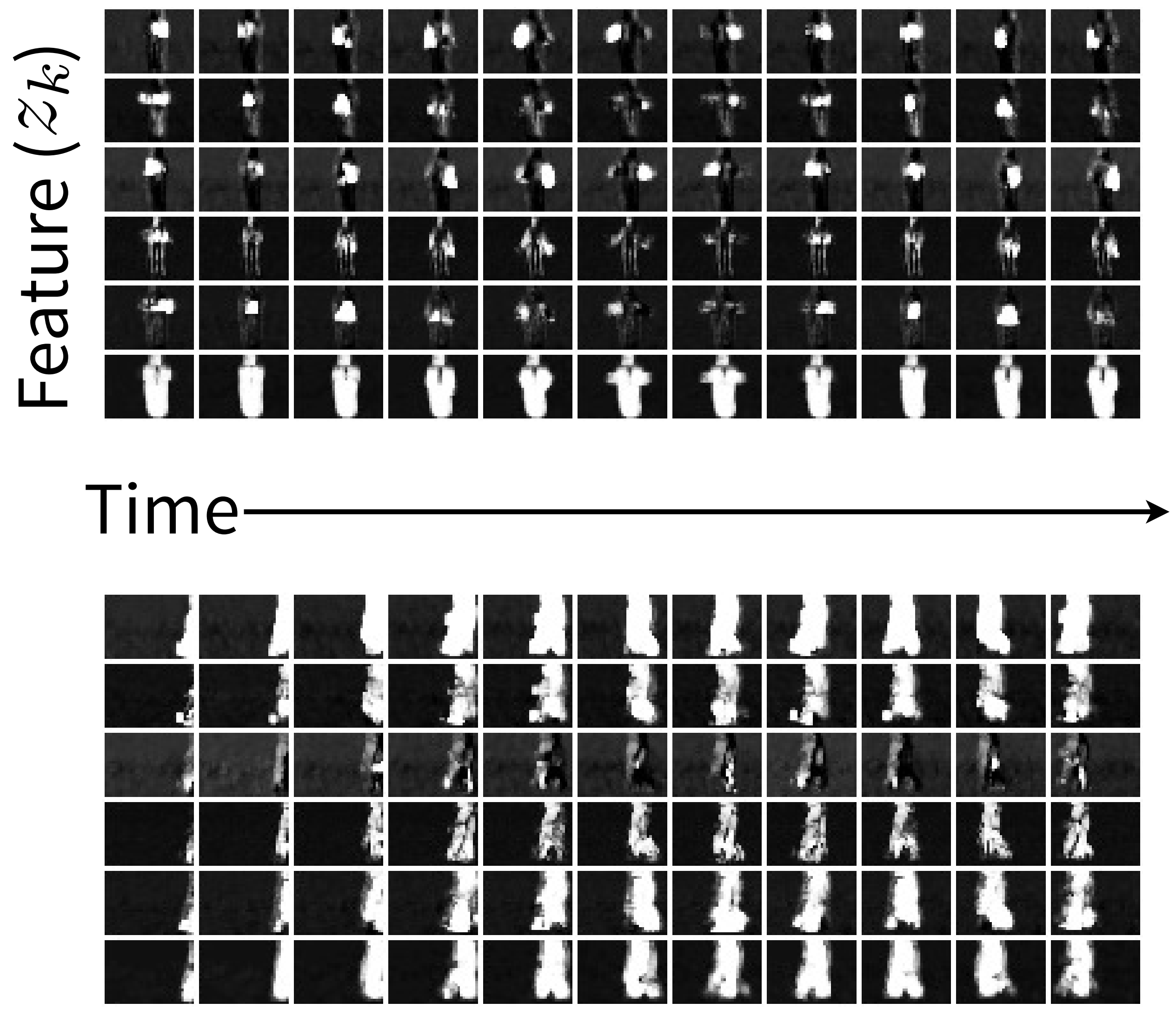
- Like the GRBM, captures third-order interactions
- Shares weights at all locations in an image
- As in a standard RBM, exact inference is efficient
- Inference and reconstruction are performed through convolution operations



Feature extraction examples

- We learn 32 feature maps
- 6 are shown here
- KTH contains 25 subjects performing 6 actions under 4 conditions
- Only preprocessing is local contrast normalization

Hand clapping

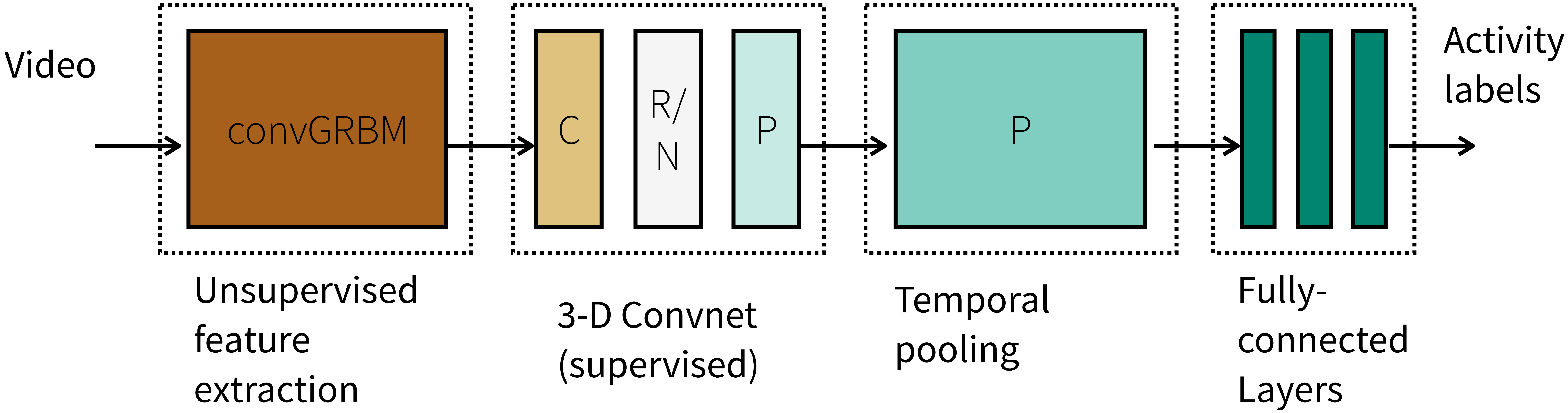


Walking

- Motion sensitive features (1,3)
- Edge features (4)
- Segmentation operator (6)

Recognition Architecture

Pipeline



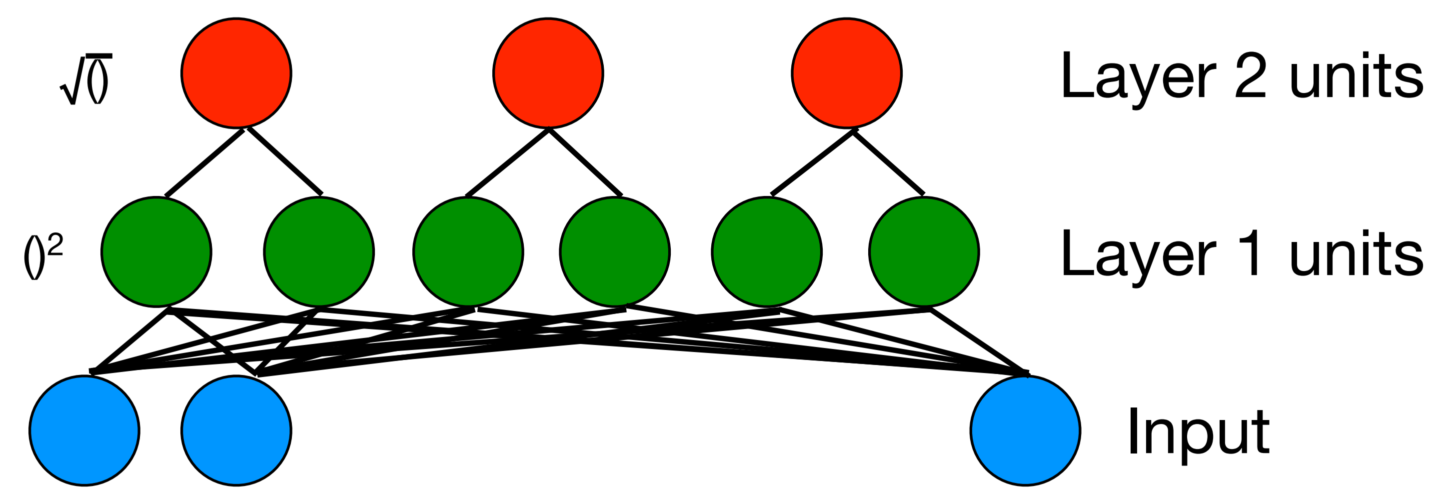
KTH Results

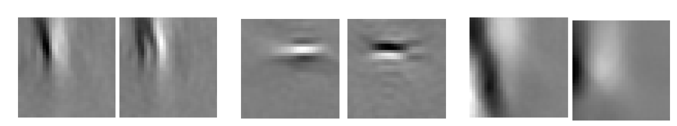
Prior Art	Acc (%)	Convolutional architectures	Acc. (%)
HOG3D+KM+SVM	85.3	convGRBM+3D-convnet+logistic reg.	88.9
HOG/HOF+KM+SVM	86.1	convGRBM+3D convnet+MLP	**90**
HOG+KM+SVM	79	3D convnet+3D convnet+logistic reg.	79.4
HOF+KM+SVM	88	3D convnet+3D convnet+MLP	79.5

Stacked Convolutional Independent Subspace Analysis (ISA)

(Le et al. 2011)

- Use of ISA (right) as a basic module
- Learns features robust to local translation; selective to frequency, rotation and velocity
- Key idea: scale up ISA by applying convolution and stacking





Typical filters learned by ISA when trained on static images (organized in pools - red units above)

Images: Le et al. (2011)

Convolution and Stacking

- The network is built by “copying” the learned network and “pasting” it to different parts of the input data (analagous to convnet)
- Outputs are then treated as the inputs to a new ISA network
- PCA is used to reduce dimensionality

Simple example: 1D data

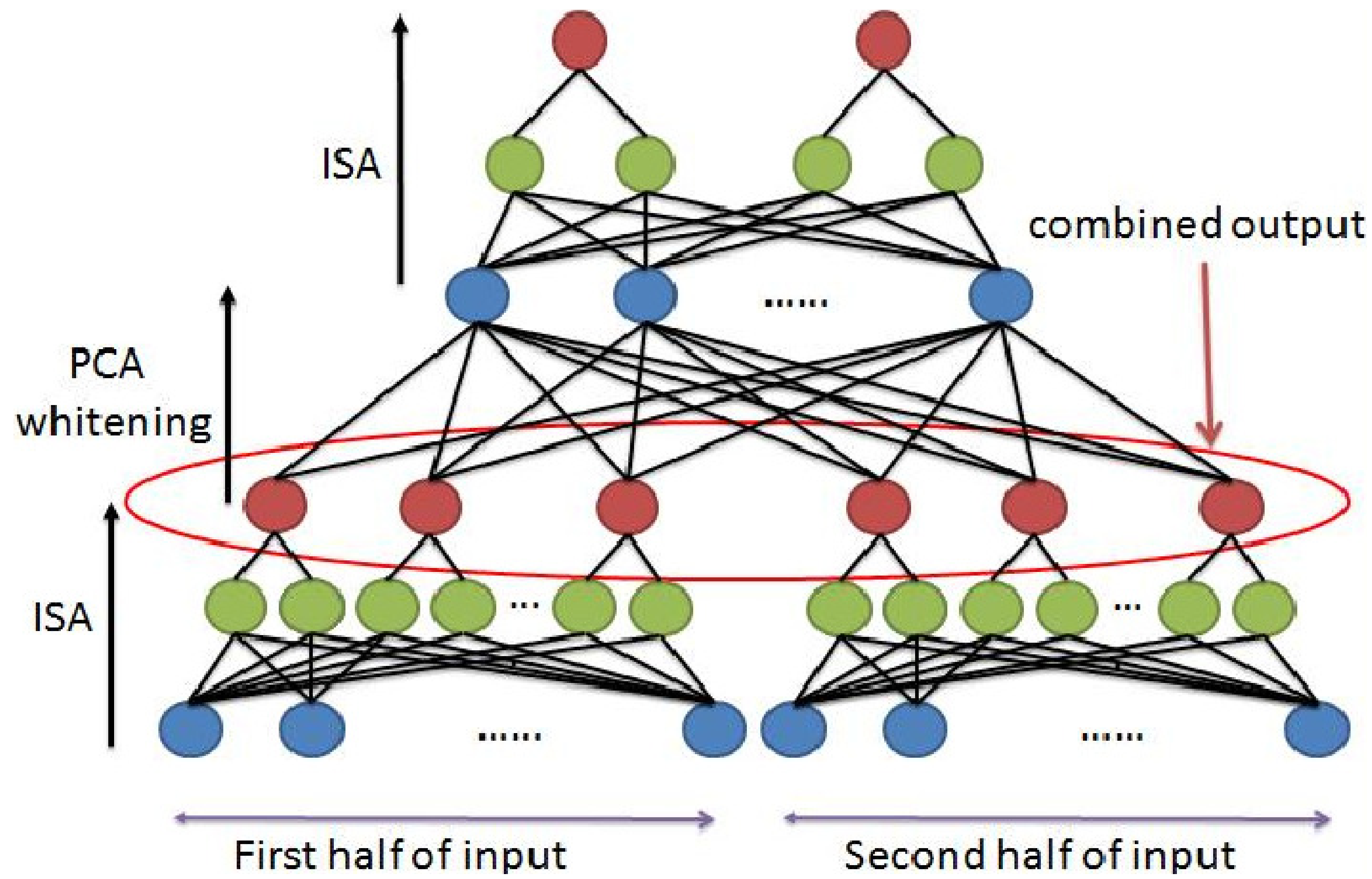


Image: Le et al. (2011)

Spatio-Temporal Feature Extraction

- Inputs to the network are blocks of video
- Each block is vectorized and processed by ISA
- Features from Layer 1 and Layer 2 are combined prior to classification

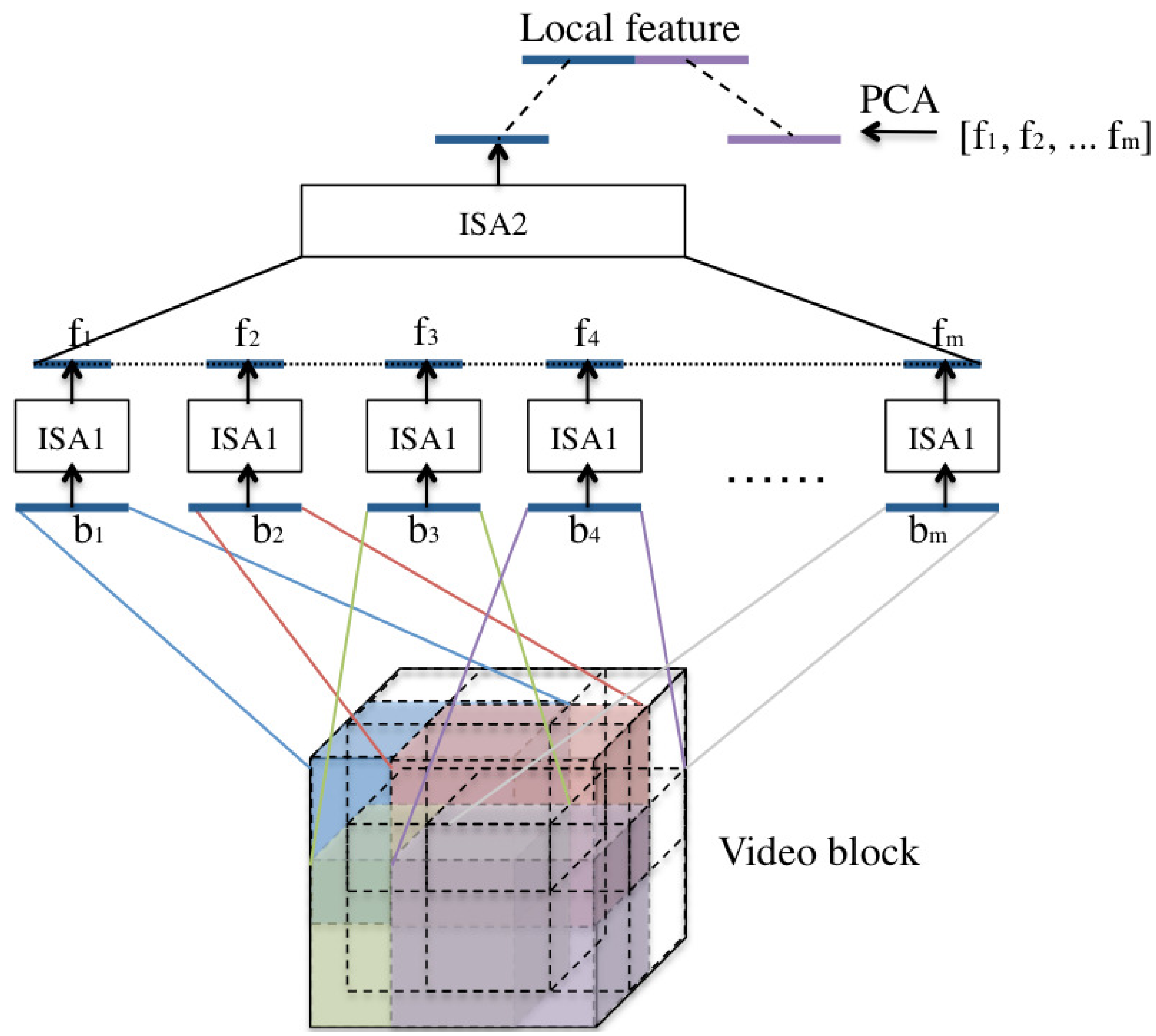
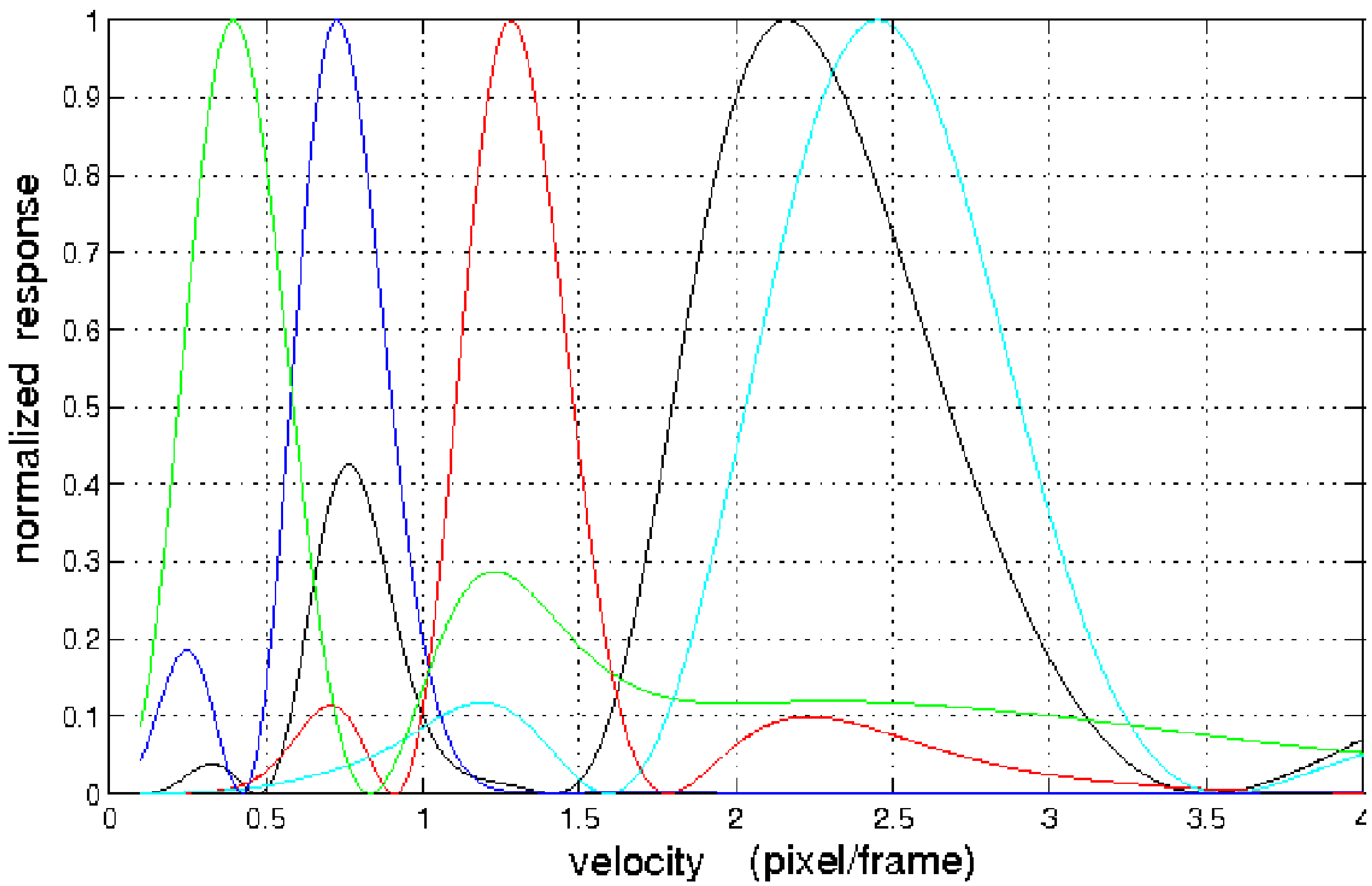
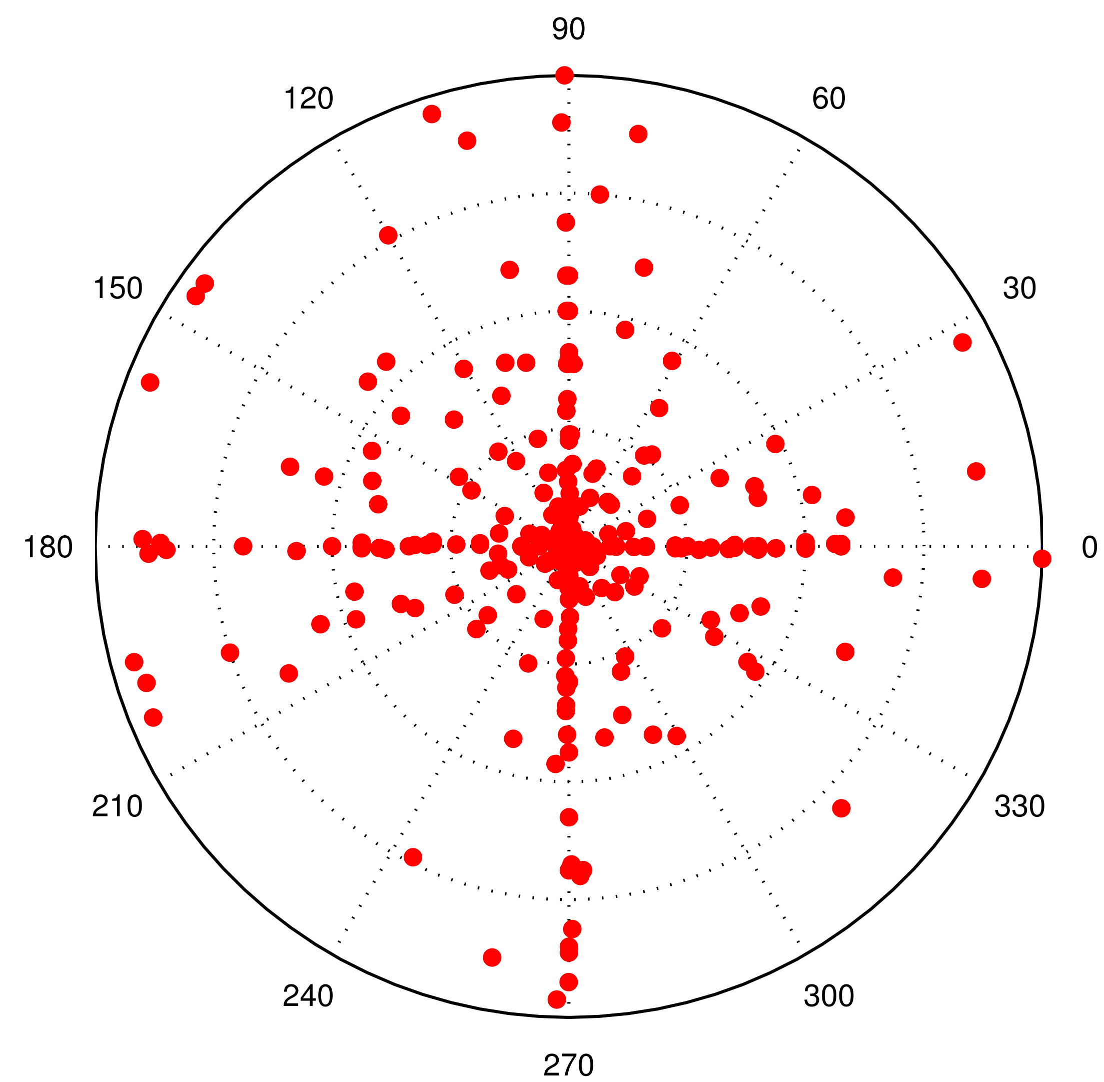


Image: Le et al. (2011)

Velocity and Orientation Selectivity





Velocity tuning curves for five neurons in an ISA network trained on Hollywood2 data

Edge velocities (radius) and orientations (angle) to which filters give maximum response
Outermost velocity: 4 pixels per frame

Coupling of motion and invariance

- Traditional motion energy models (Adelson & Bergen, 1985) and cross-correlation models (Arndt et al, 1995, Fleet et al., 1996) are closely related and <u>they confound representing transformations and encoding invariance</u>

- (Konda et al. 2014): decouple by computing motion by "synchrony detection" and achieving content-invariance by pooling

Motion synchrony

(Konda et al. 2014)

- Say, two images are related by an orthogonal image warp

$$\mathbf{x}_2 = P\mathbf{x}_1$$

- To detect the transformation:
  - Choose a filter pair, such that it is an example of that transformation

$$\mathbf{w}_2 = P\mathbf{w}_1$$

  - Determine whether the two filters yield equal responses when applied in sequence to two frames

$$\mathbf{w}_2^T\mathbf{x}_2 = \mathbf{w}_1^T\mathbf{x}_1$$

Practically: how to check for synchrony?

- Necessary to detect equality of transformed filter responses across time
- Can't use standard sum of filter responses + thresholding
- Can use multiplicative (gating) interactions between filter responses

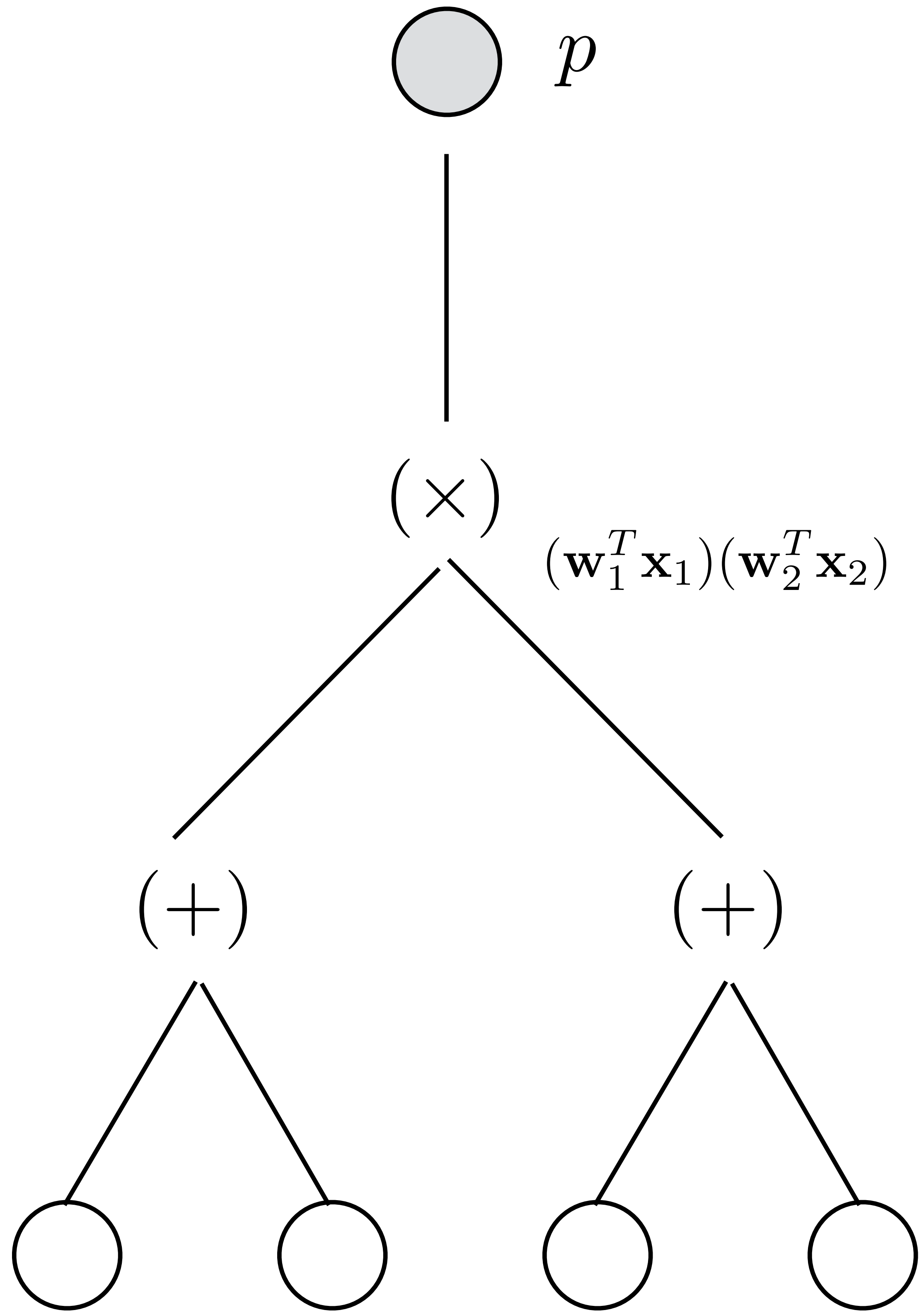


Image: Konda et al. (2014)

Learning to detect synchrony

Synchrony autoencoder

- Learn a gated autoencoder with tied weights, trained to reconstruct $\mathbf{x}_2$ from $\mathbf{x}_1$ and vice-versa
- Use a contractive regularization term

Synchrony K-means

- Filters are learned by a temporal variant of online K-means (Coates et al. 2011, Rumelhart & Zipser, 1986)
- Gradient descent-based optimization

Note: neither method is trained with pooling.
A pooling layer may be learned separately.

Results

KTH Dataset

Method	Accuracy (%)
SAE (Konda et al. 2014)	93.5
SK-means (Konda et al. 2015)	93.6
Conv-ISA (Le et al. 2011)	**93.9**
Conv-GRBM (Taylor et al. 2010)	90.0

UCF Sports

Method	Accuracy (%)
SAE (Konda et al. 2014)	86.0
SK-means (Konda et al. 2015)	84.7
Conv-ISA (Le et al. 2011)	**86.5**

Hollywood 2

Method	Mean A.P.
SAE (Konda et al. 2014)	51.8
SK-means (Konda et al. 2015)	50.5
Conv-ISA (Le et al. 2011)	**53.3**
Conv-GRBM (Taylor et al. 2010)	43.3

Training Time

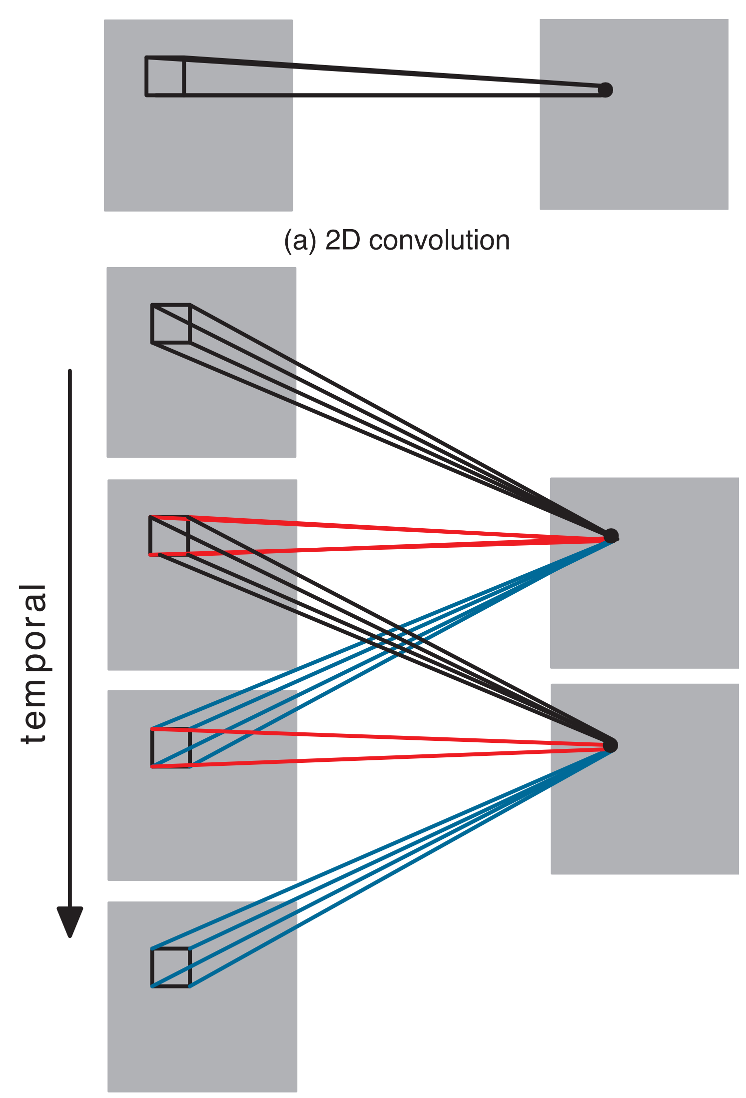
Method	Mean A.P.
SK-means (Konda et al. 2015) (GPU)	2 min
SK-means (Konda et al. 2015) (CPU)	3 min
SAE (Konda et al. 2014) (GPU)	1 - 2 hr
Conv-ISA (Le et al. 2011)	1-2 hr
Conv-GRBM (Taylor et al. 2010)	2 - 3 days

Results from (Konda et al. 2014), (Le et al. 2011), (Taylor et al. 2010)

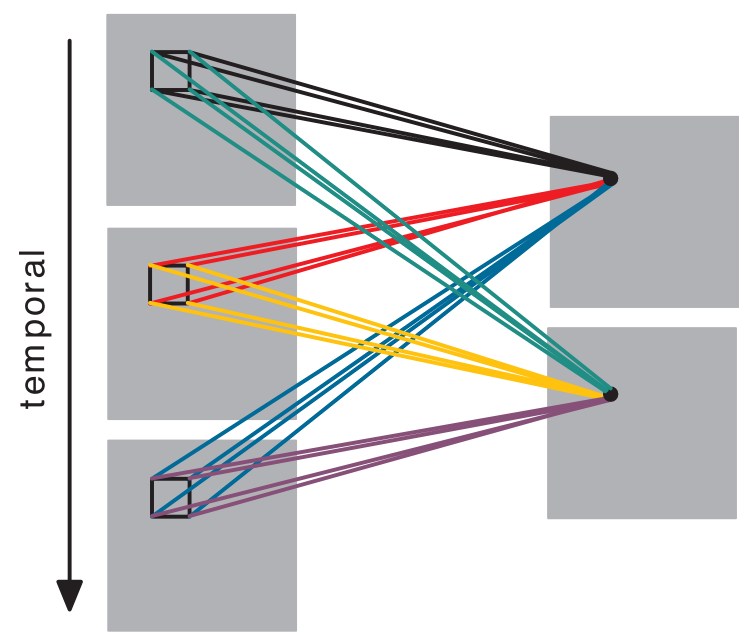
End-to-end Supervised

3D Convnets for Activity Recognition

- One approach: treat video frames as still images (LeCun et al. 2005)
- Alternatively, perform 3D convolution capturing discriminative features across space and time



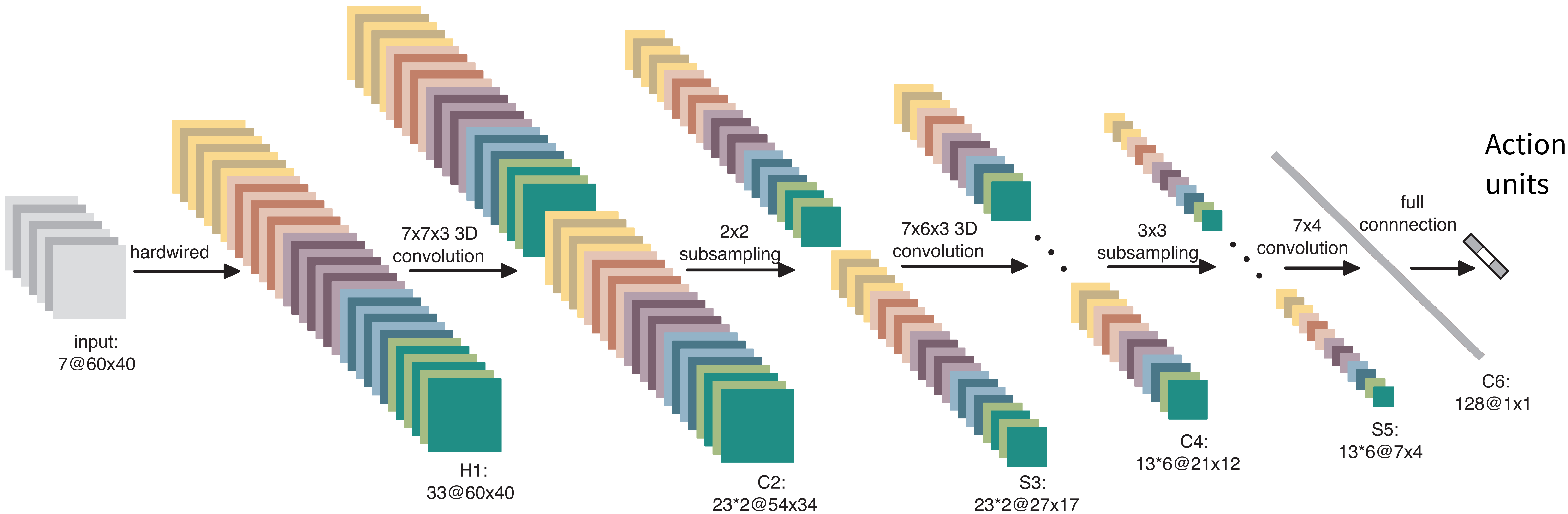
(b) 3D convolution



Multiple convolutions applied to contiguous frames to extract multiple features

Figure: Ji et al. (2010)

Early CNN Architecture



Hardwired to extract:
1)grayscale
2)grad-x
3)grad-y
4)flow-x
5)flow-y

2 different 3D filters applied to each of 5 blocks independently

Subsample spatially

3 different 3D filters applied to each of 5 channels in 2 blocks

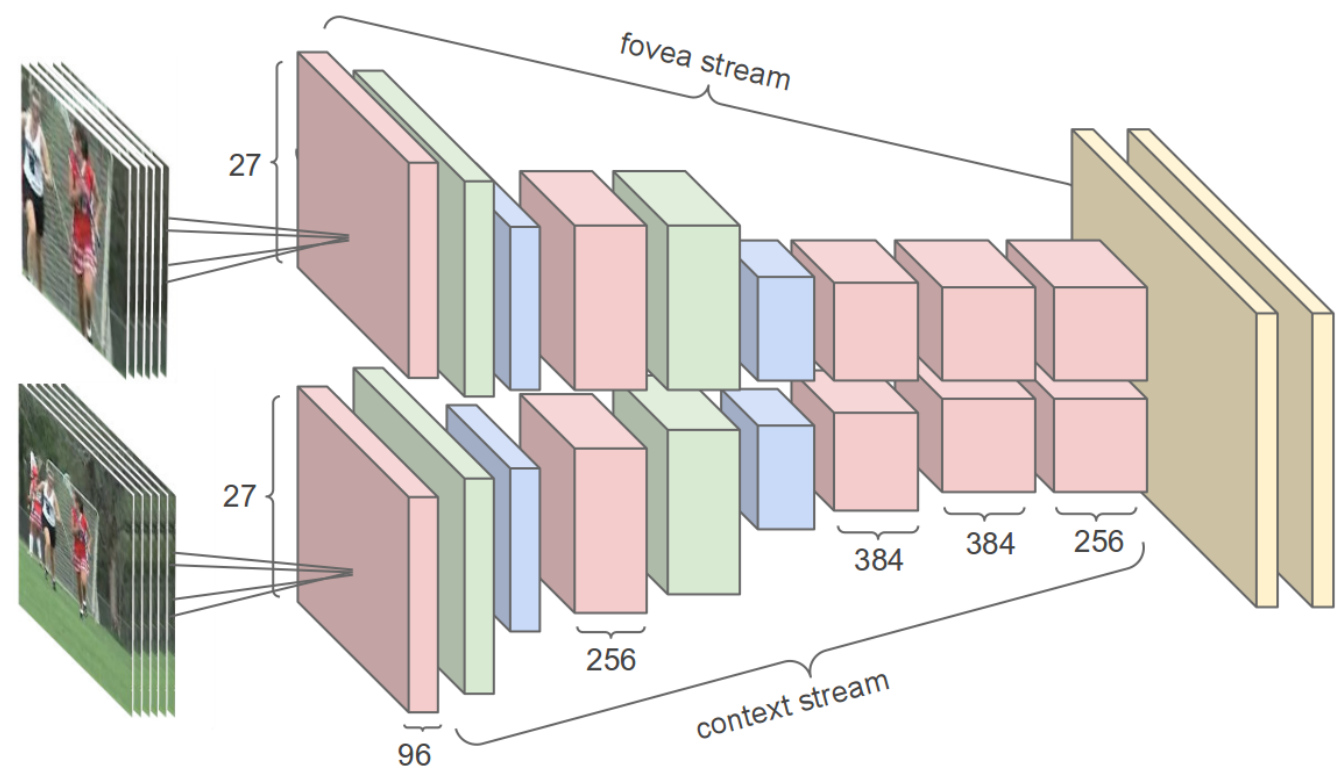
Two fully-connected layers

Figure: Ji et al. (2010)

State-of-the-art CNN Architecture

(Karpathy et al. 2014)

- Multi-resolution, foveated architecture
- Released Google Sports-1M dataset, 487 classes
- Significant performance compared to feature-based baselines
- Modest improvement compared to single-frame architectures

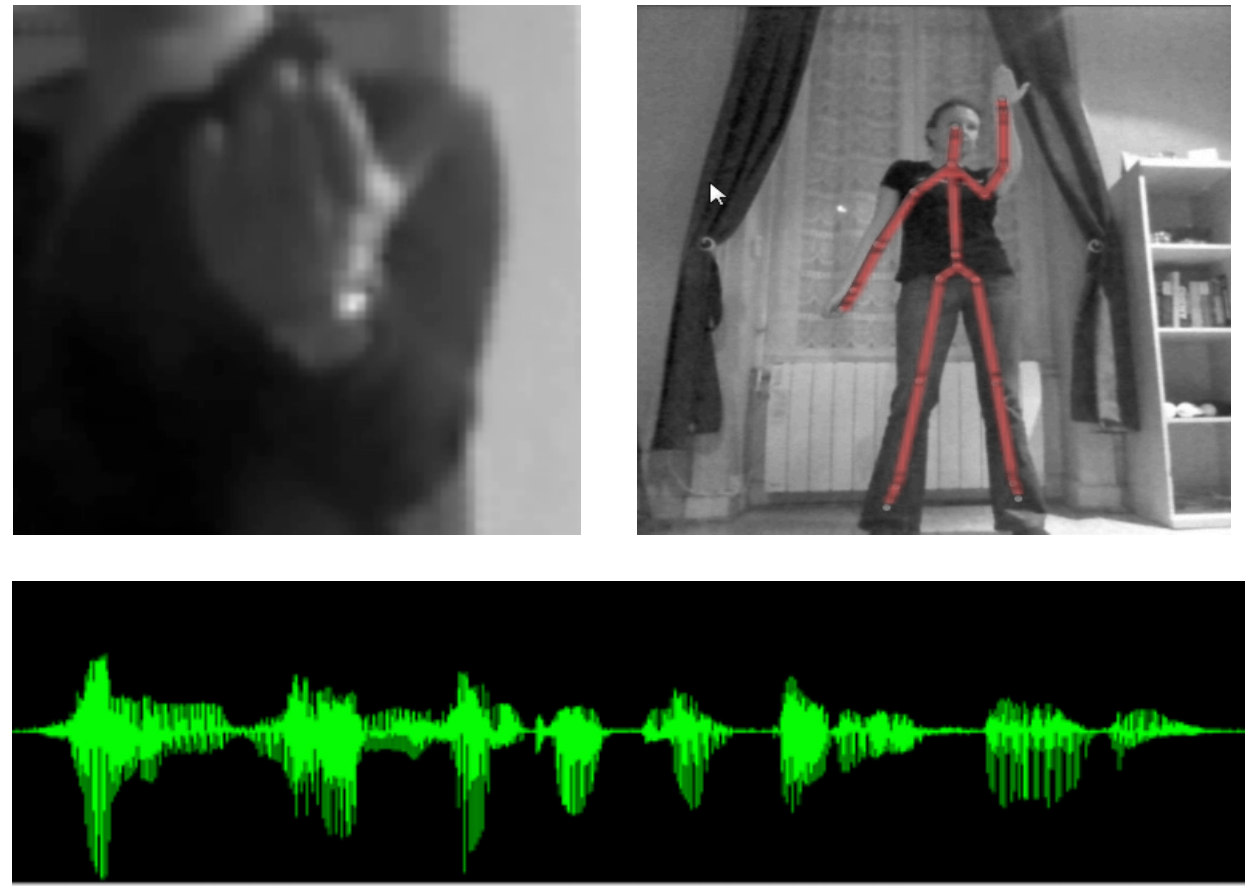


Also see: Simonyan and Zisserman, 2014

Recognizing intentional gestures

(Neverova et al. 2015)

- Communicative gestures
- Multiple modalities:
  - colour and depth video
  - skeleton (articulated pose)
  - audio
- Multiple scales:
  - full upper-body motion
  - fine hand articulation
  - short and long-term dependencies

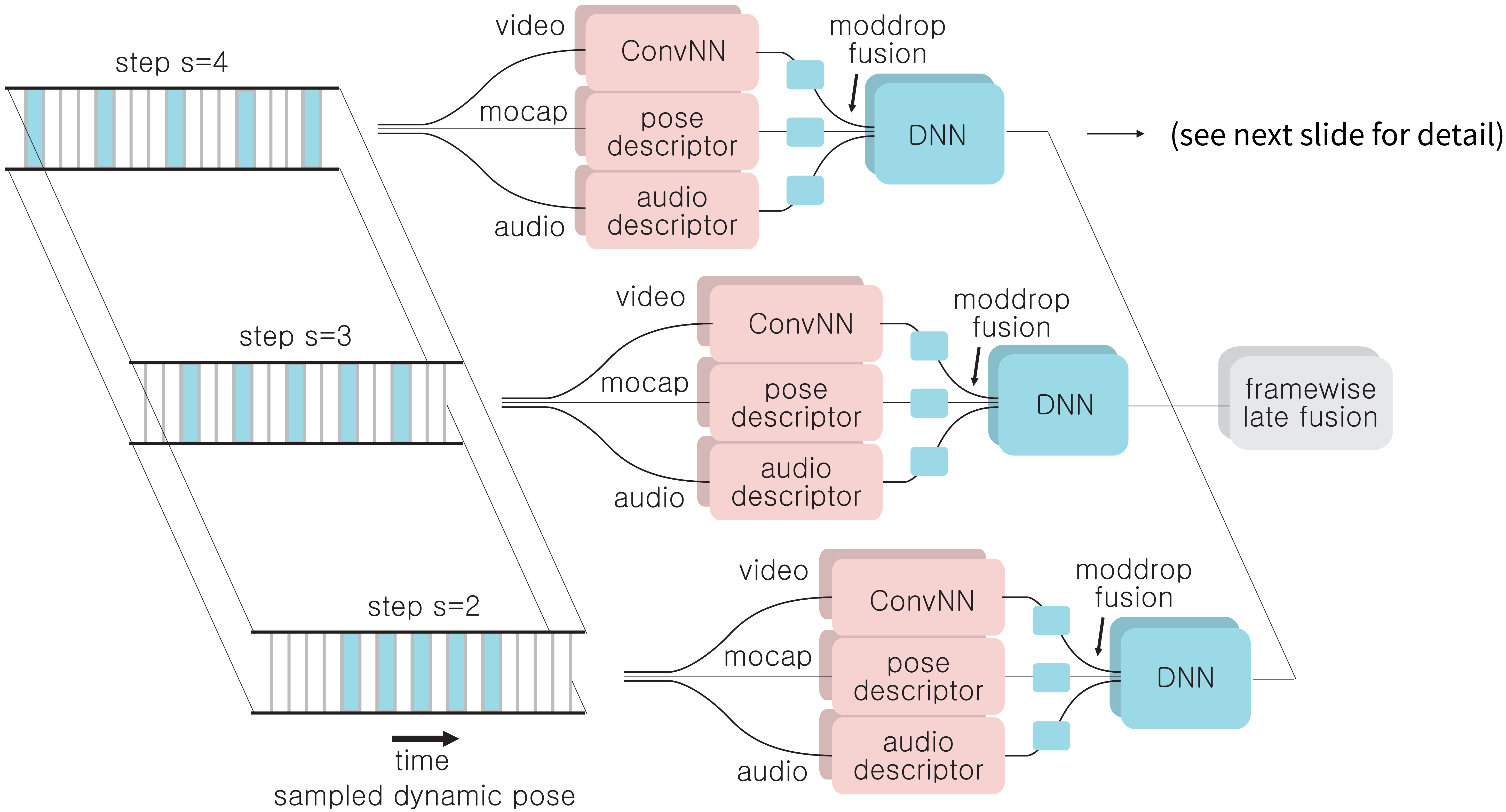


This gesture can be fully characterized by upper-body motion

Here, subtle finger movements play the primary role

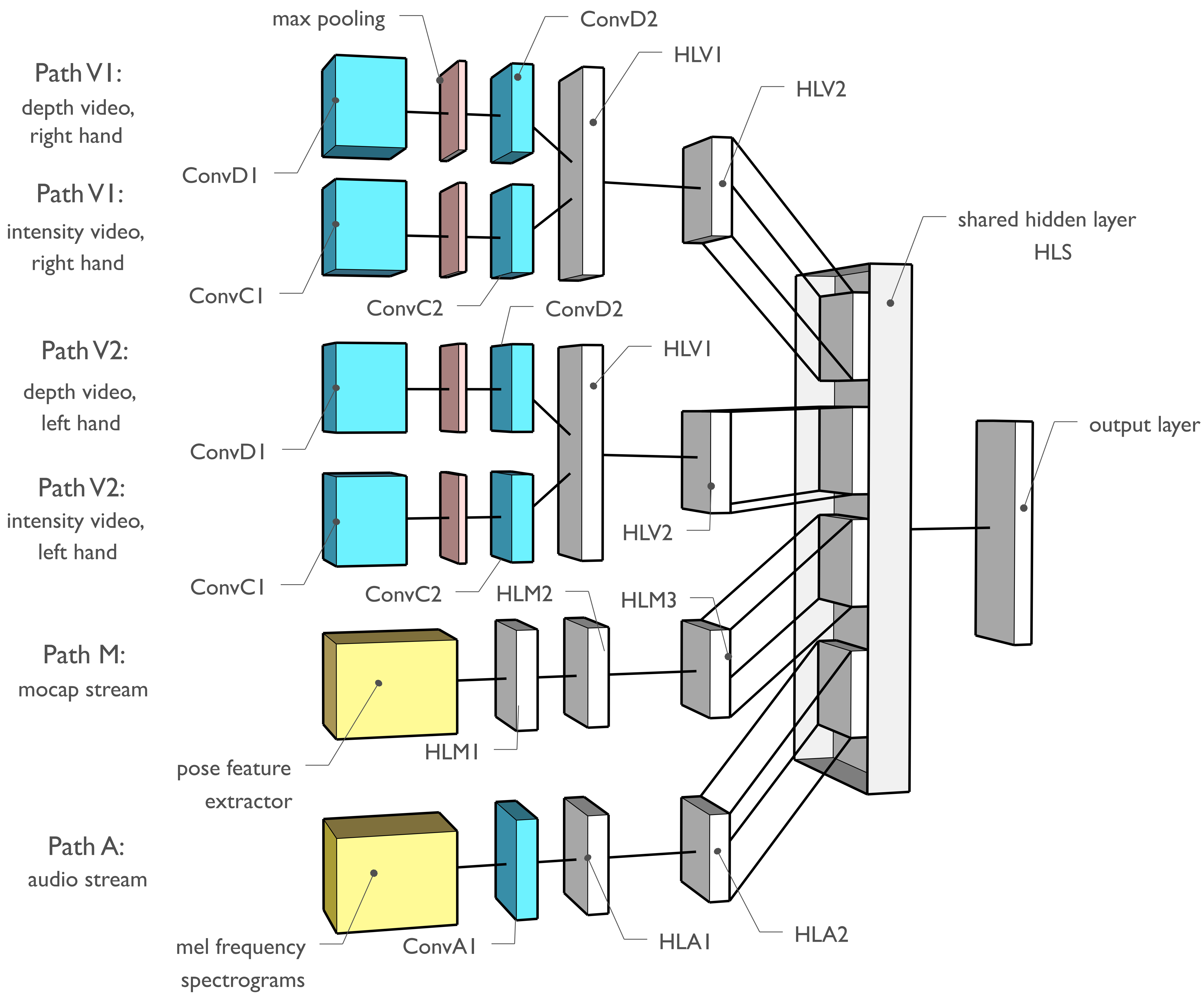
PhD work of Natalia Neverova (here!) and co-advisor Christian Wolf (INSA-Lyon)

A multi-scale architecture



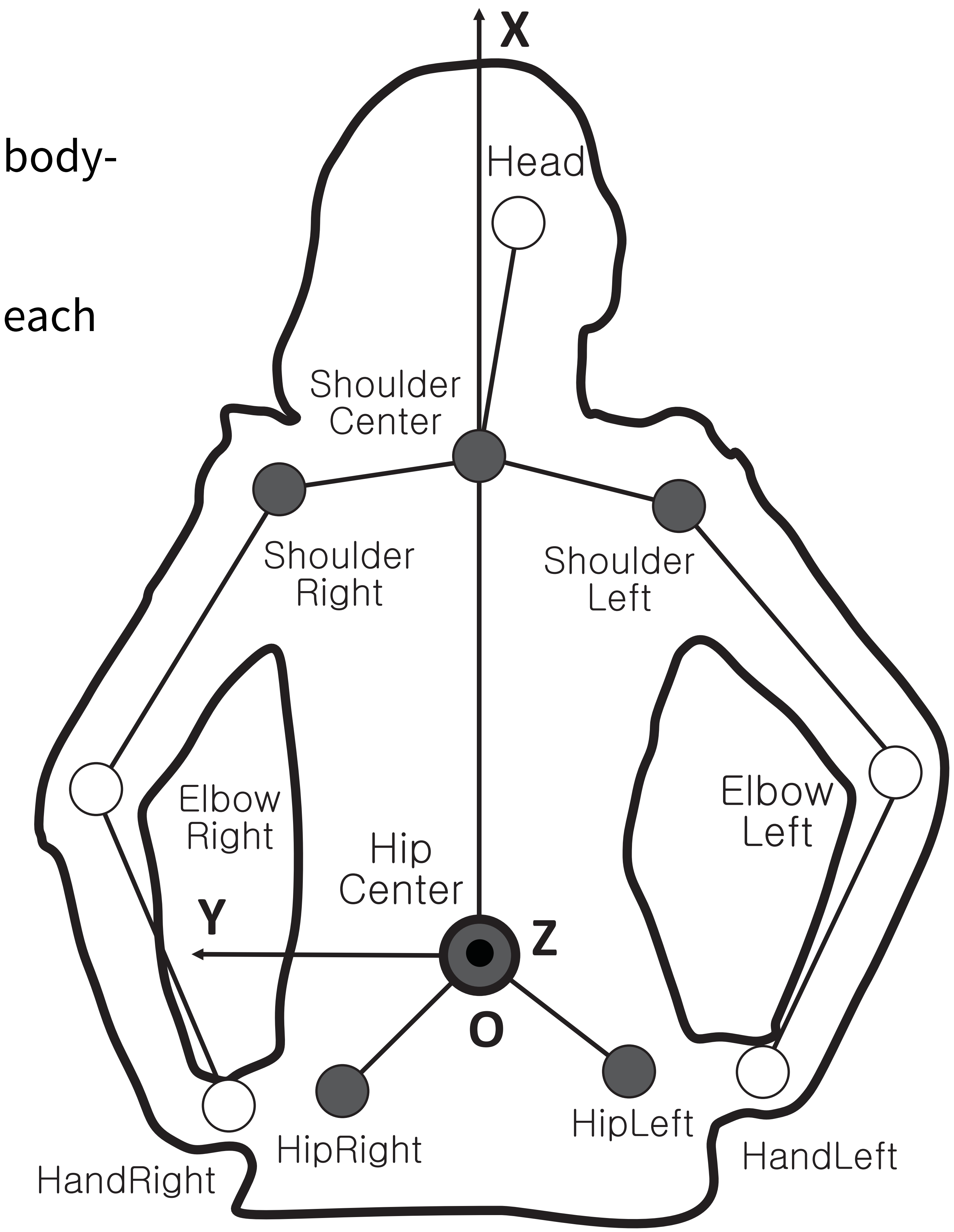
Operates at 3 temporal scales
corresponding to dynamic poses of 3 different durations

Single-scale deep architecture



Articulated Pose: Input

- Extract 11 joints from full-body skeleton (Kinect)
- Position normalization: HipCentre is an origin of a body-centred co-ordinate system
- Size normalization by the mean distance between each pair of joints (compensate for different body sizes, proportions, and shapes)
- Final representation (183-D descriptor)[1]
  - Joint positions, velocities, and accelerations
  - Inclination angles
  - Azimuth angles
  - Bending angles
  - Pairwise distances



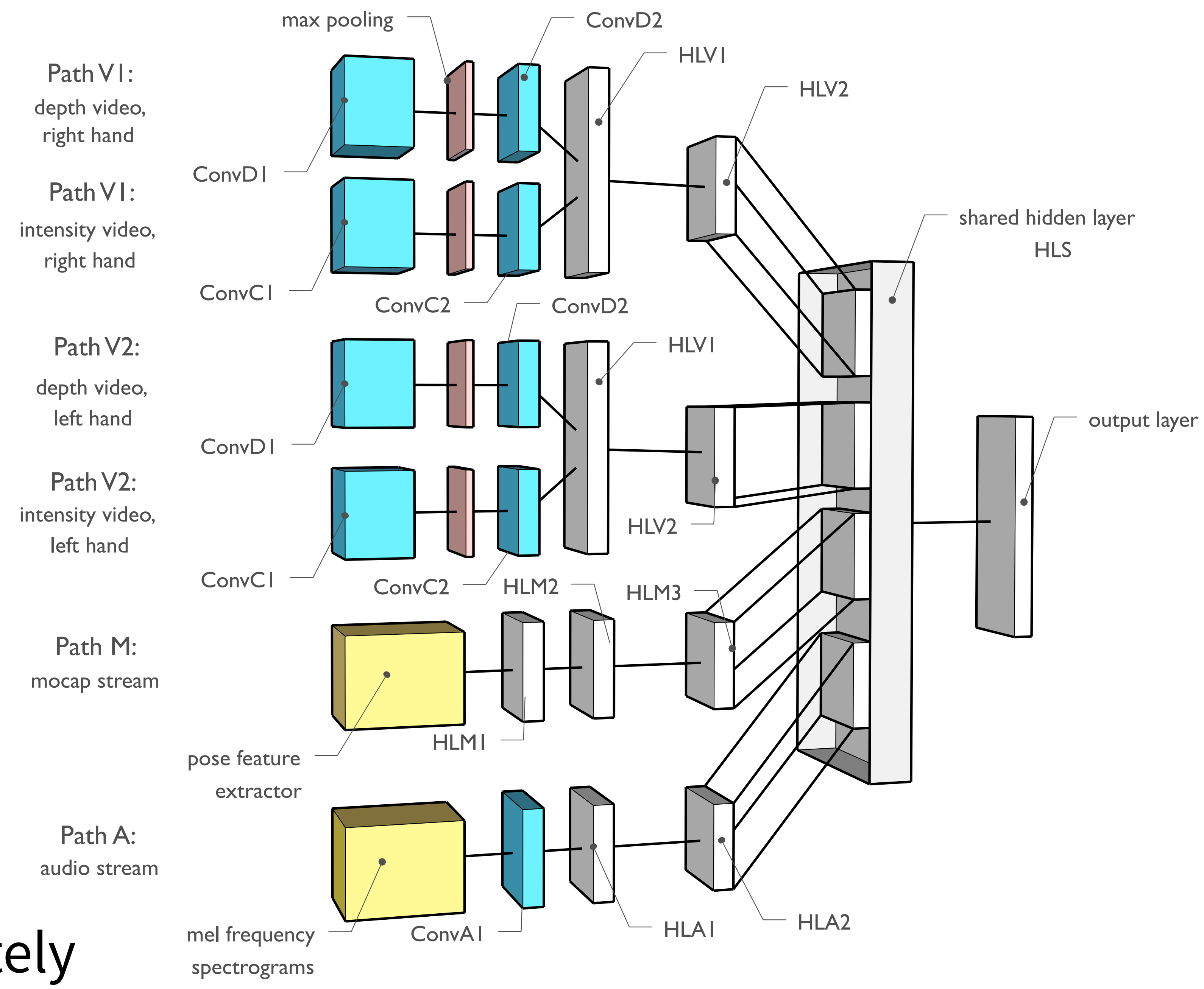
[1]Zanfir M., Leordeanu, M., Sminchisescu, C., "The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection", ICCV 2013

Depth Video Stream

- Interested in capturing fine movements of palms and fingers
- Extract a bounding box around RHand, LHand centred at hand positions provided by skeleton
- Subtract background by thresholding along depth axis
- Apply local contrast normalization

Training algorithm

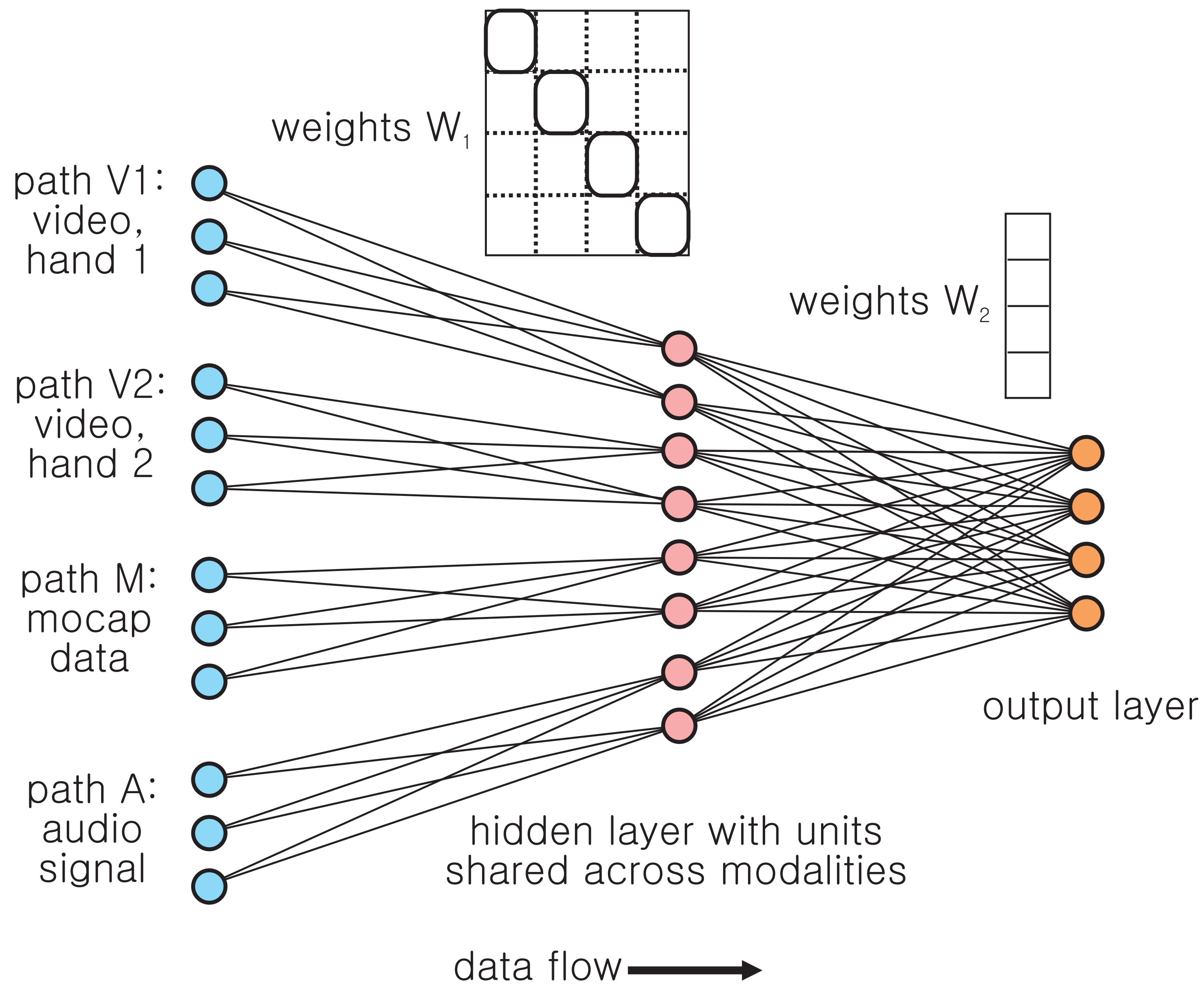
- Difficulties:
  - Number of parameters:
    - ~12.4M per scale
    - ~37.2M total
  - Number of training gestures: ~10,000
- Proposed solution:
  - Structured weight matrices
  - Pretraining of individual channels separately
  - Careful initialization of shared layers
  - Iterative training algorithm which gradually increases # of parameters



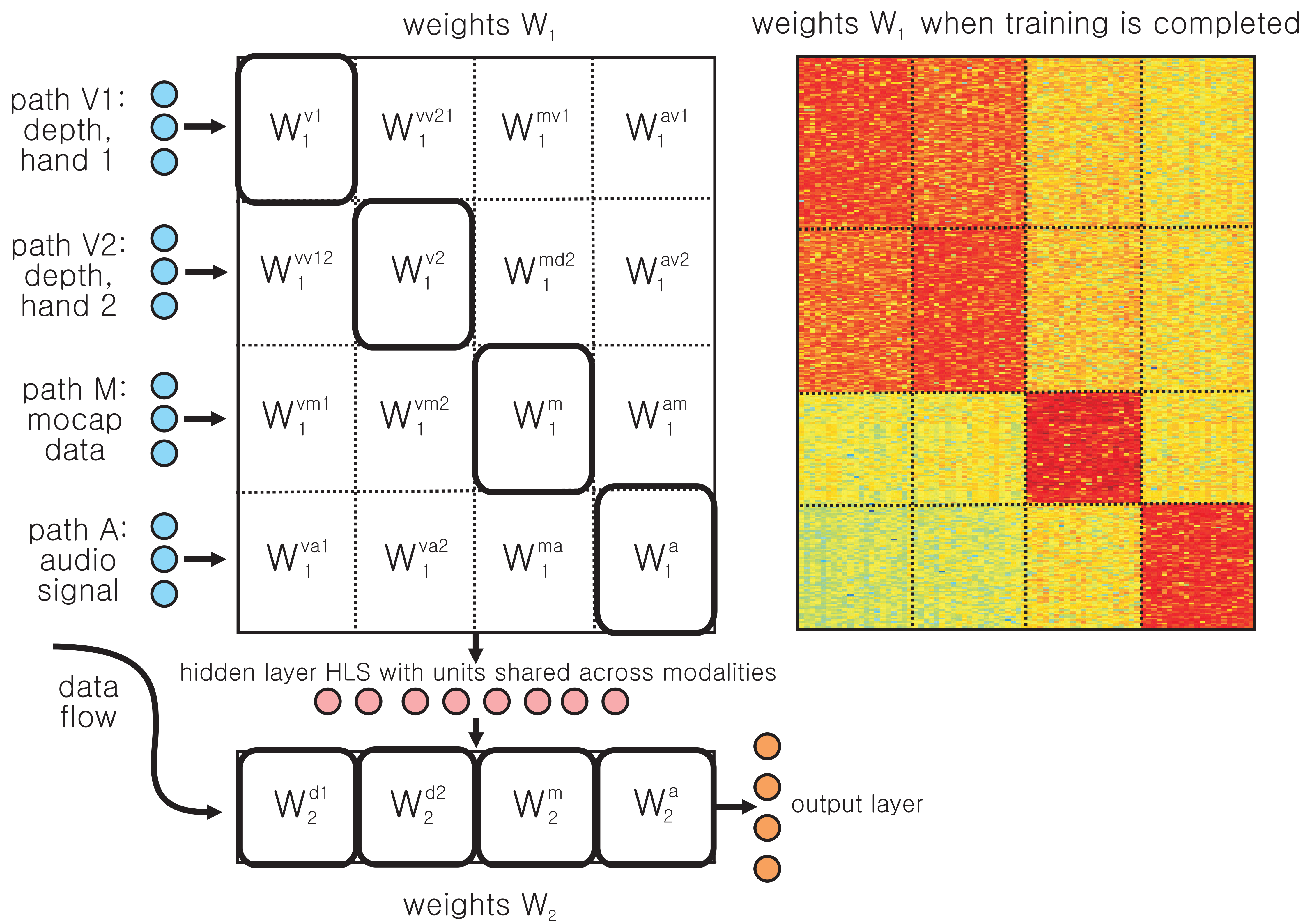
Initialization: structured weights

- Top hidden layer from each path is initially wired to a subset of neurons in the shared layer

- During fusion, additional connections between paths and the shared hidden layer are added

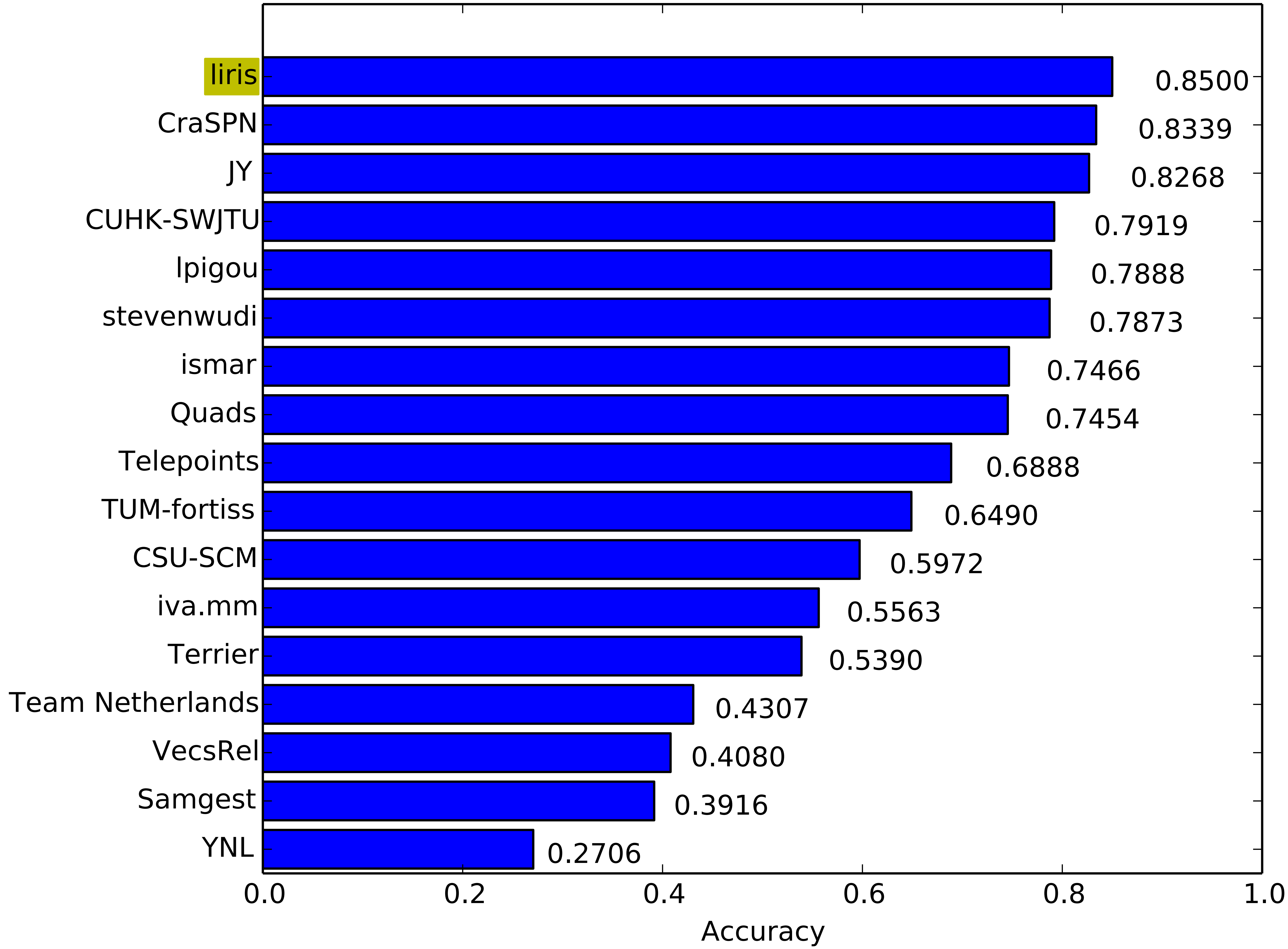


Slightly different view



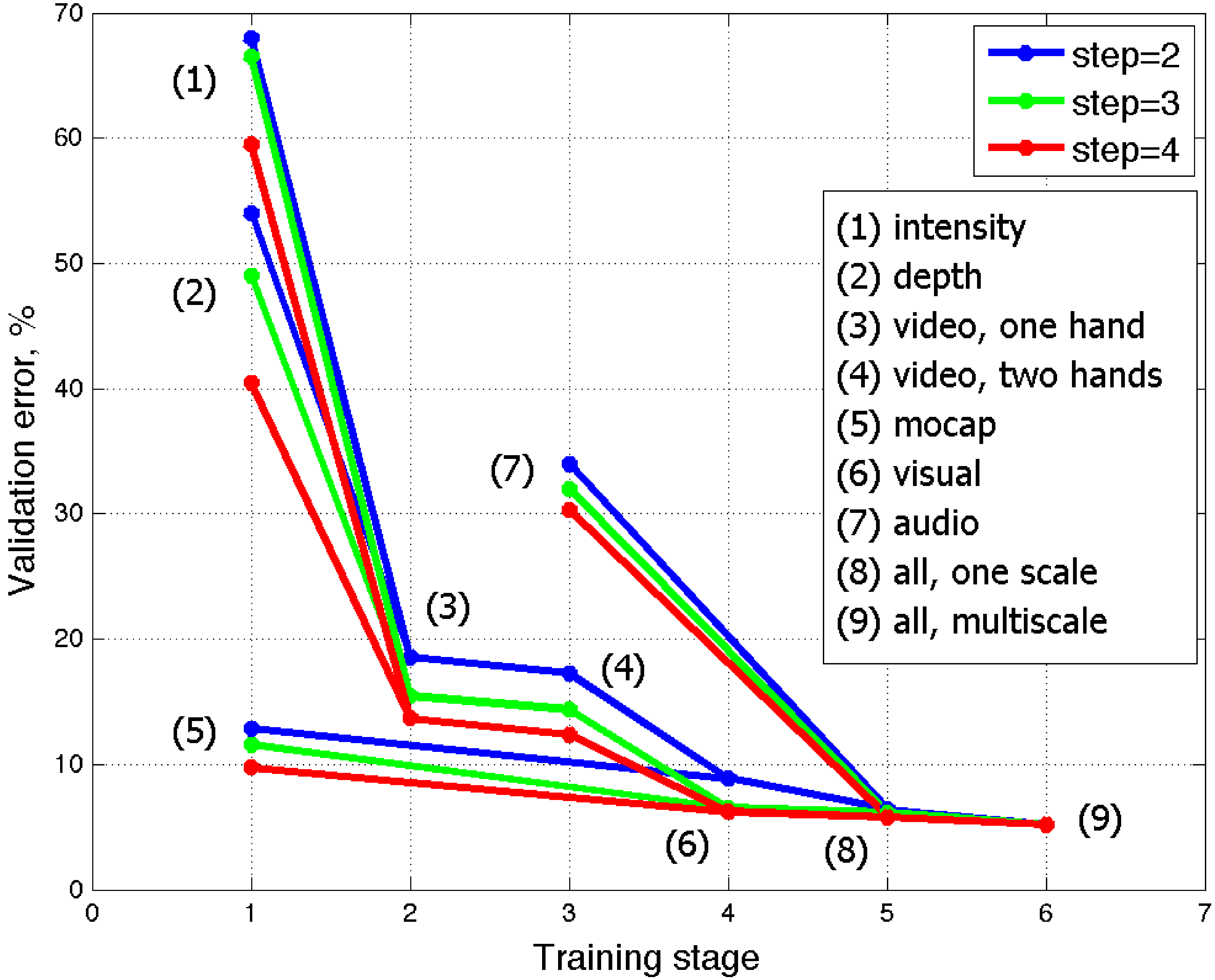
Blocks of the weight matrices are learned iteratively
after proper initialization of the diagonal elements

2014 ChaLearn Looking at People Challenge (ECCV)



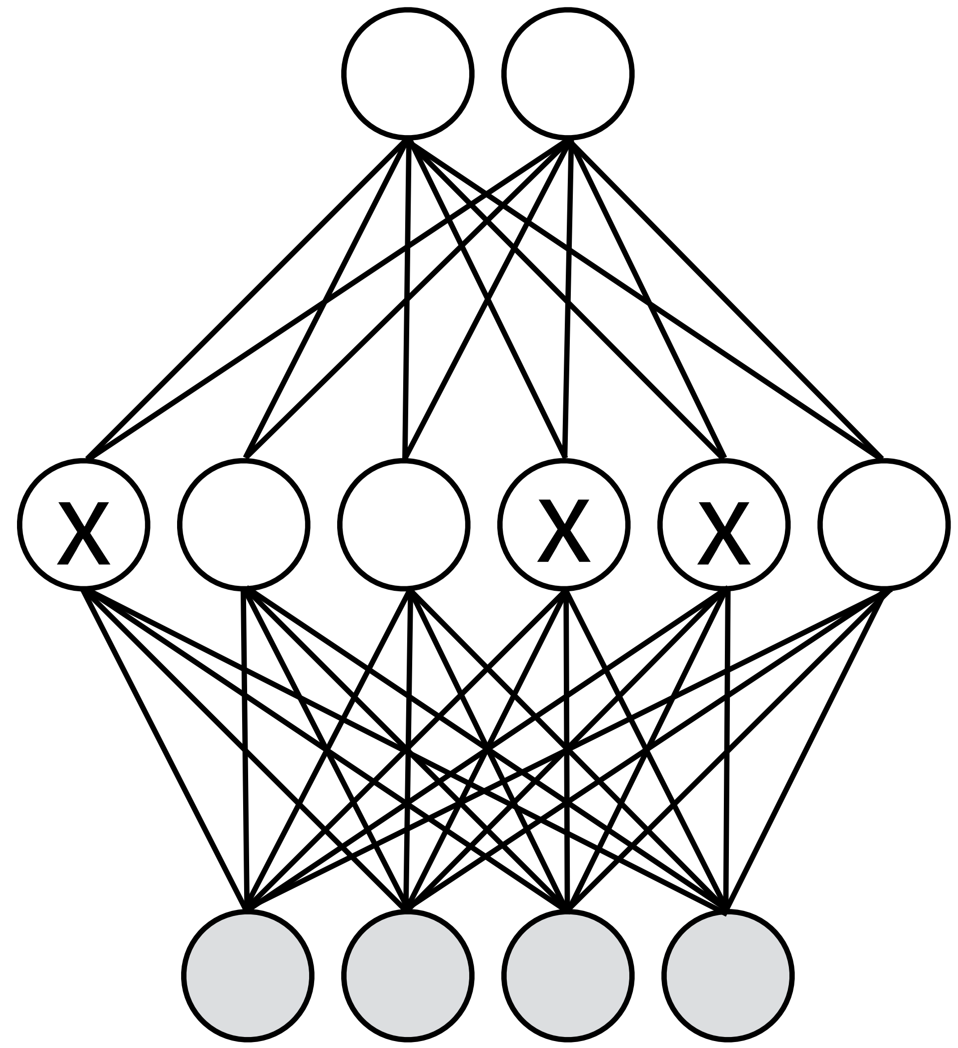
Metric is mean Jaccard Index (intersection over union)

Error evolution during iterative training

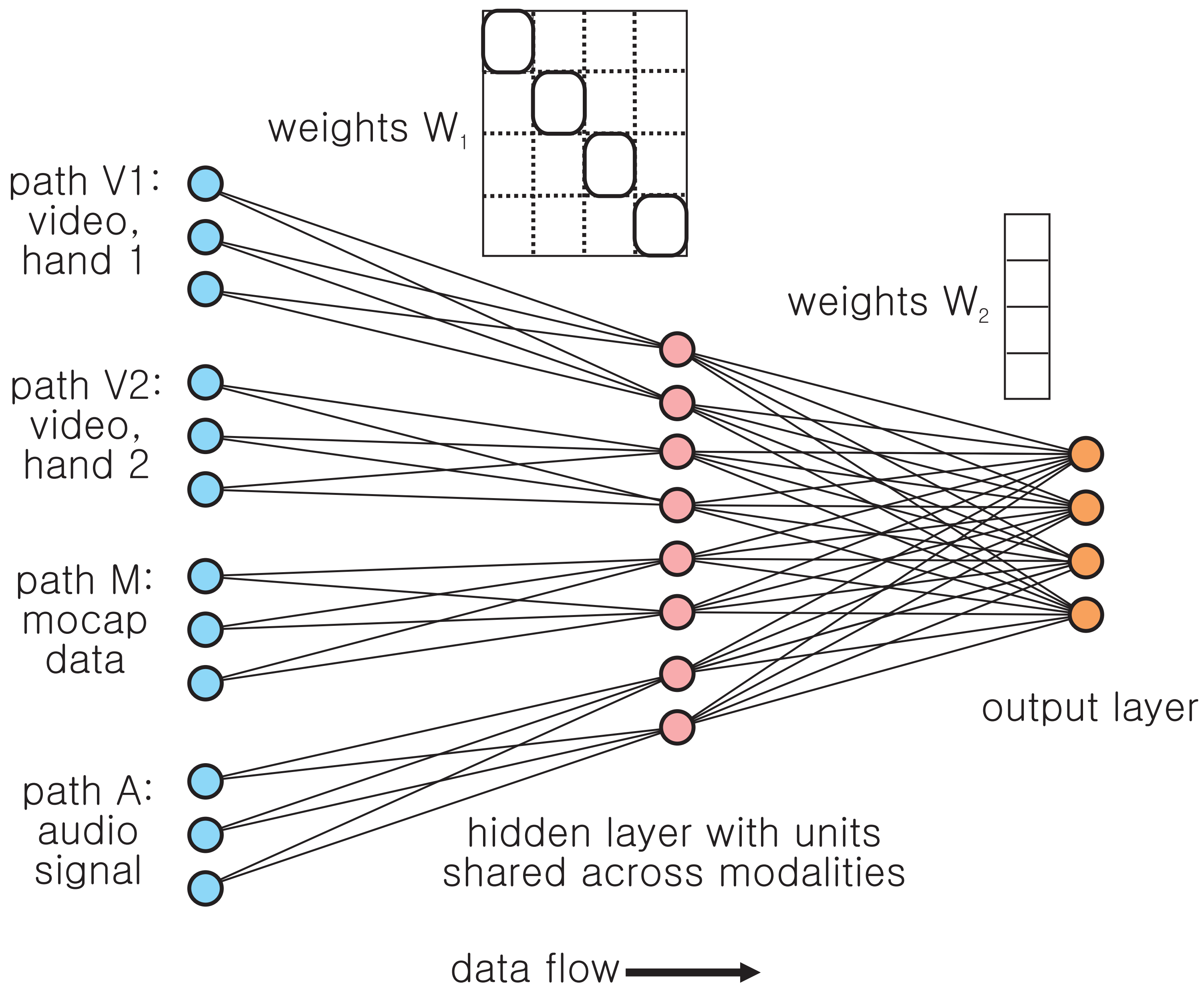


Dropout (review)

- Introduced in 2012, made famous by ImageNet
- During training, for each training sample, “drop out” 50% of hidden unit activities
- Punishes co-adaptation of units
- Can be viewed as very efficient model averaging



Moddrop - dropout on shared layer



$$h_j^{(k)} = \sigma\left[\sum_{i=1}^{F_k} w_{i,j}^{(k,k)} x_i^{(k)} + \gamma \sum_{\substack{n=1 \\ n \neq k}}^{K} \sum_{i=1}^{F_n} w_{i,j}^{(n,k)} x_i^{(n)} + b_j^{(k)}\right]$$

Moddrop: modality-wise dropout

- Punish co-adaptation of individual units (like dropout)
- Train a network which is robust/resistent to dropping of individual modalities (e.g. fail of audio)

$$\bar{h}_j^{(k)} = \sigma\left[\sum_{i=1}^{F_k} w_{i,j}^{(k,k)} x_i^{(k)} + \sum_{\substack{n=1 \\ n\neq k}}^{K} \delta^{(k)} \sum_{i=1}^{F_n} w_{i,j}^{(n,k)} x_i^{(n)} + b_j^{(k)}\right]$$

Bernoulli selector

$$P(\delta^{(k)} = 1) = p^{(k)}$$

Moddrop results

Classification accuracy on the validation set
(dynamic poses)

Modalities	Dropout (%)	Dropout + Moddrop (%)
All	96.77	96.81
Mocap missing	38.41	92.82
Audio missing	84.10	92.59
Hands missing	53.13	73.28

Jacquard index on test set (full gestures)

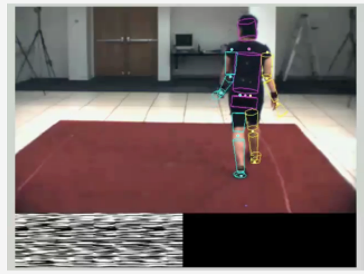
Modalities	Dropout (%)	Dropout + Moddrop (%)
All	87.6	88.0
Mocap missing	30.6	85.9
Audio missing	78.9	85.4
Hands missing	46.6	68.0

Summary

Pose Estimation

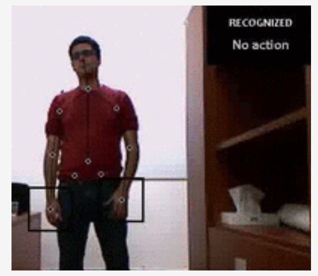
- Extreme variability
- Small # pixels
- Occlusions
- Dominated by convnets
- Structured output

Tracking



- Pose estimation + Dynamical models
- Still difficult outside of controlled environments

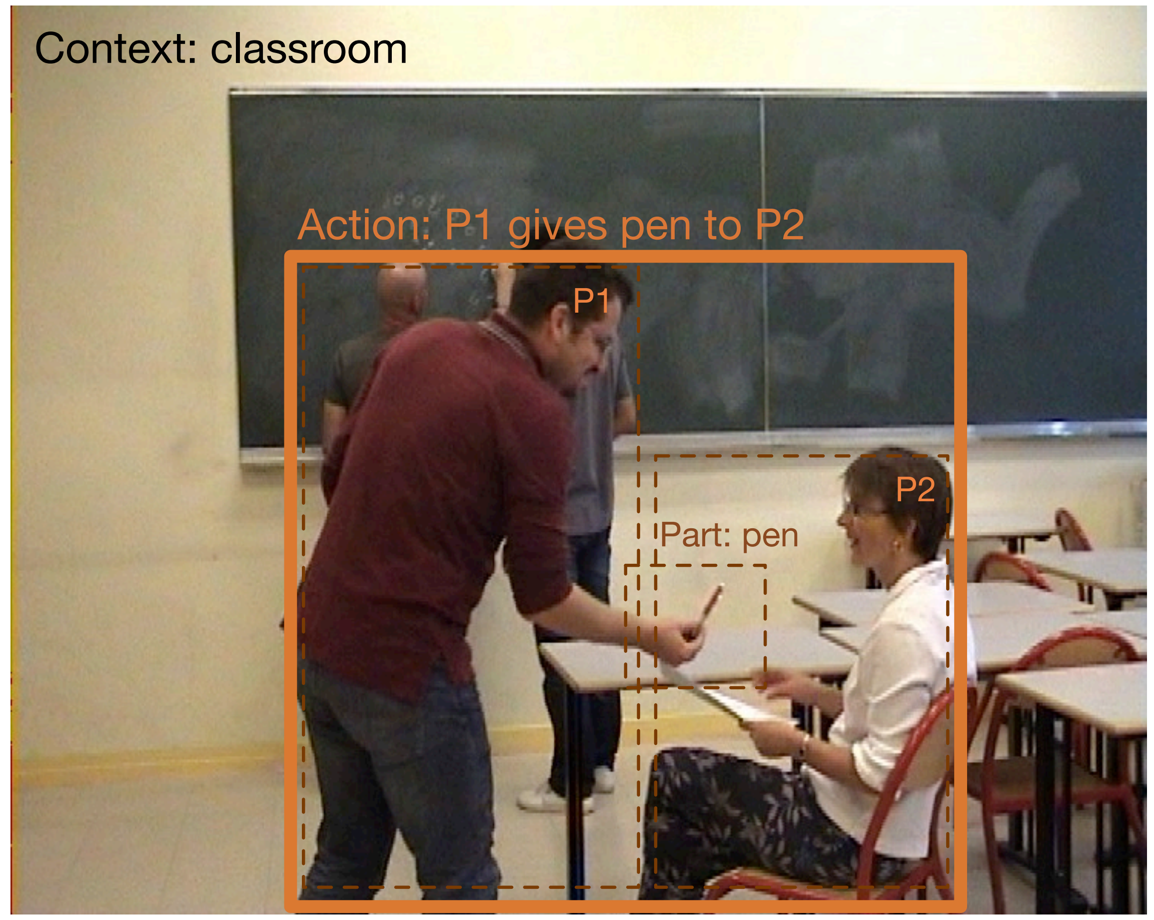
Activity /Gesture

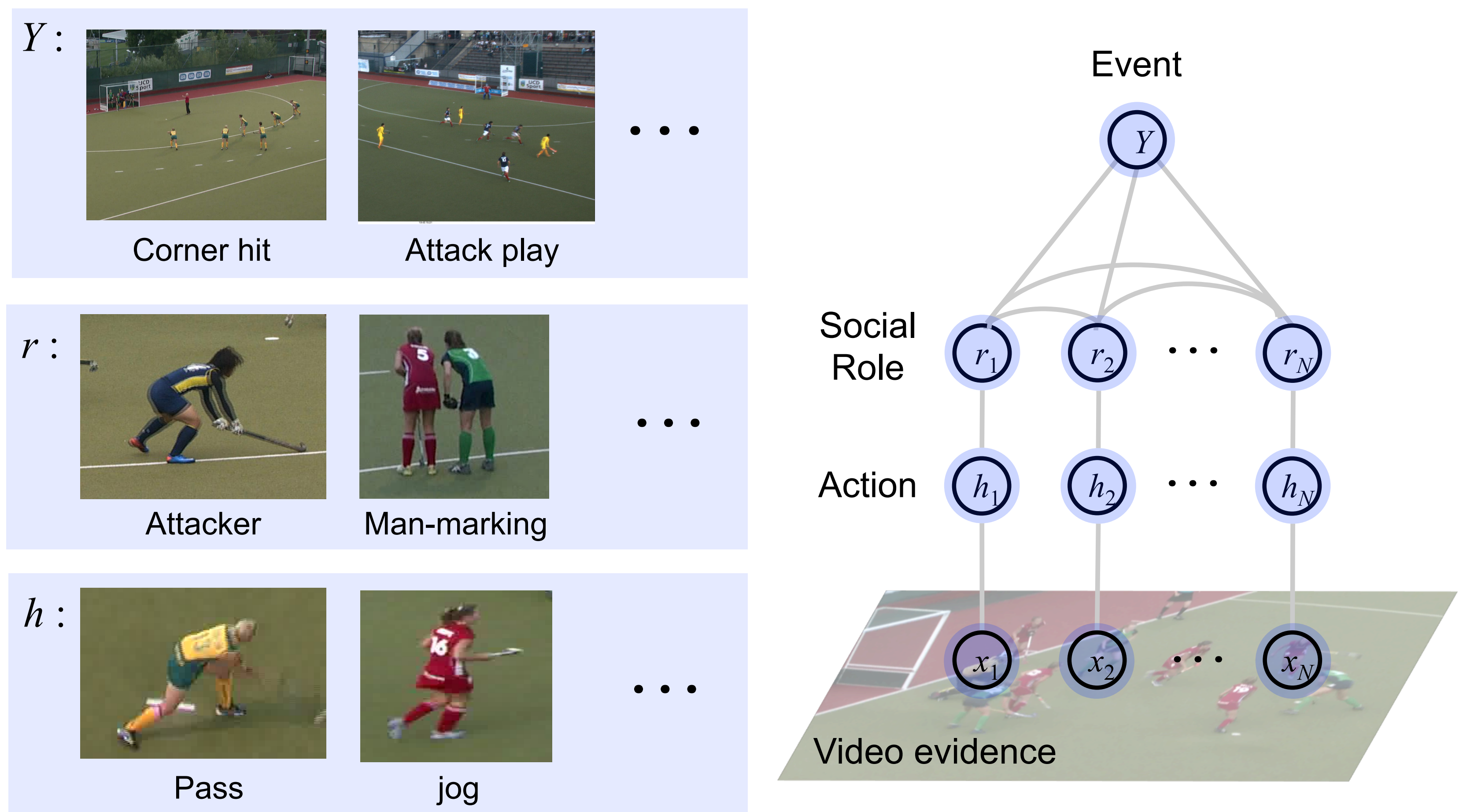


- Two families:
  - unsupervised feature extraction + pipeline
  - convnets (supervised)
- Potential for multi-modal data

Where to go from here?

- Limited labeled data
  - Unsupervised, weakly supervised learning?
- Going beyond classification of short, simple activities or gestures
  - Capture structural relationships w/ structured models: less flexible and efficient than DL models

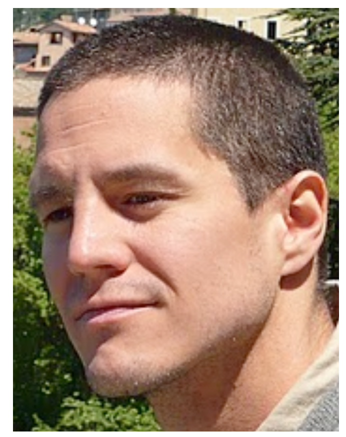




Images: Greg Mori

Acknowledgements

- Much of the background was developed in collaboration with a larger research team:
  - Christian Wolf and Julien Mille (INSA-Lyon)
  - Greg Mori (SFU)
  - Matthieu Cord and Nicolas Thome (UPMC-Paris 6)



Thank You!


Montreal
Guelph
Toronto
New York
Quebec City
Sudbury
North Bay
Trois-Rivières
Victoriaville
Drummondville
Algonquin Provincial Park
Ottawa
Granby
Sherbrooke
Saint-Jean-sur-Richelieu
Maine
Bangor
Kingston
Vermont
Markham
Mississauga
Rochester
Portland
New Hampshire
Hamilton
London
Buffalo
New York
Albany
Manchester
Chatham
Massachusetts
Boston
Springfield
Providence
Cleveland
Connecticut
Akron
New Haven
Bridgeport
Canton
Pennsylvania
Pittsburgh
Allentown
Altoona
Rhode Island
Trenton
Columbus
Philadelphia
Toms River
Presque Isle