Multi-scale deep learning for gesture detection and localization

Natalia Neverova, Christian Wolf,
Graham Taylor, Florian Nebout

ECCV 2014 Chalearn Workshop on Looking at People :
Pose recoverty, action/interaction, gesture recognition

Team LIRIS / Univ. Guelph / Awabot

Natalia Neverova

Christian Wolf

Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205
(Lyon, France)

Graham Taylor

University of Guelph,
School of Engineering
(Guelph, Canada)

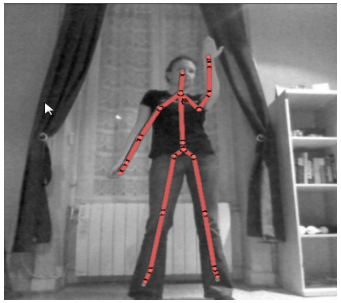
Florian Nebout

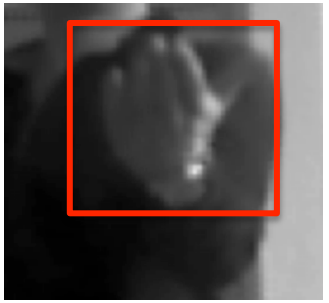
Awabot SAS
(Lyon, France)

Problem and its context

Communicative gestures:

- communication between people;
- natural human-robot interaction.





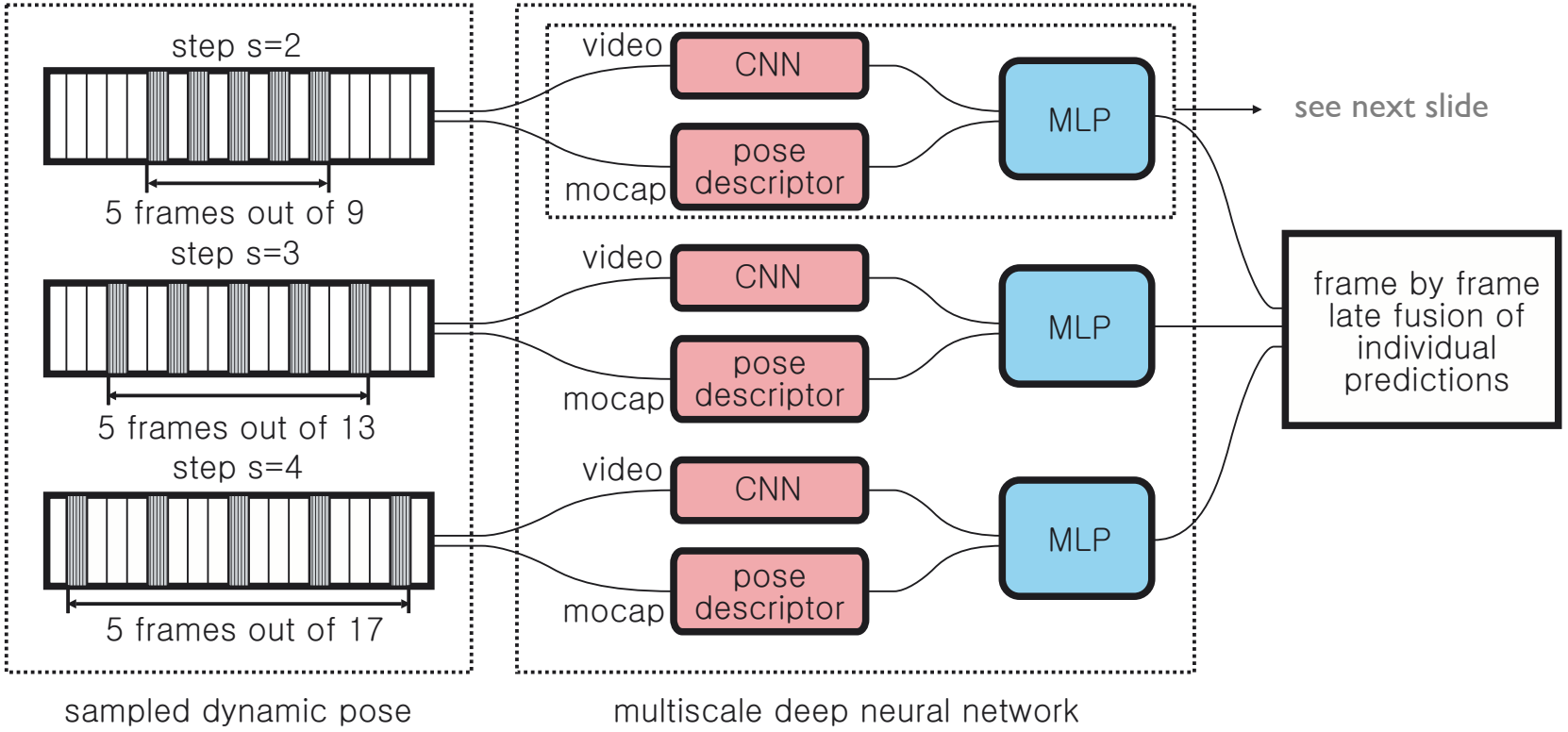
Multiple modalities:

- color and depth video signals;
- skeleton stream (articulated pose).

Multiple spatial and temporal scales:

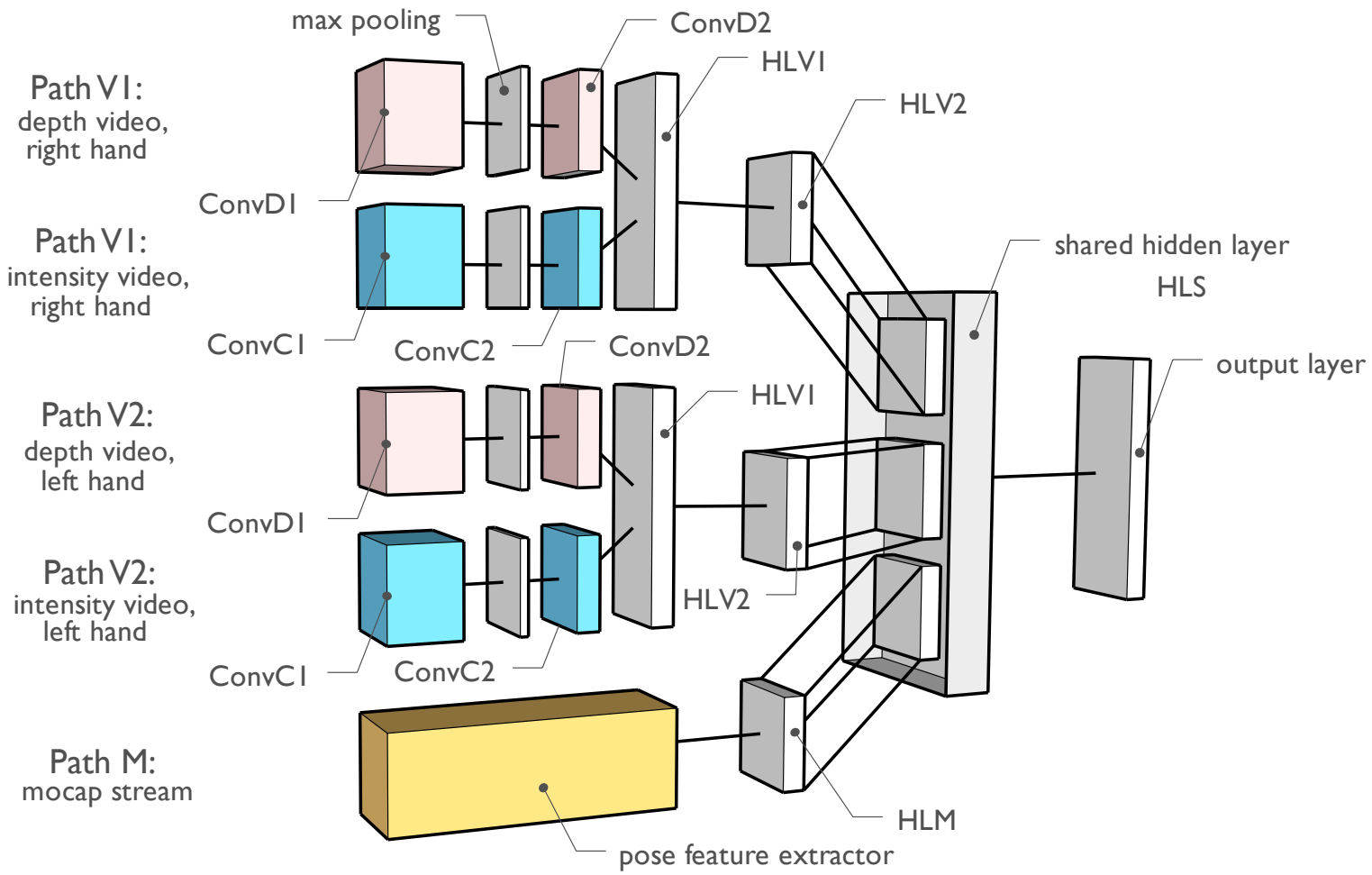
- full upper-body motion and fine hand articulation,
- short and long-term temporal dependencies.

Gesture detection and classification: a multi-scale architecture



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

Single-scale deep architecture



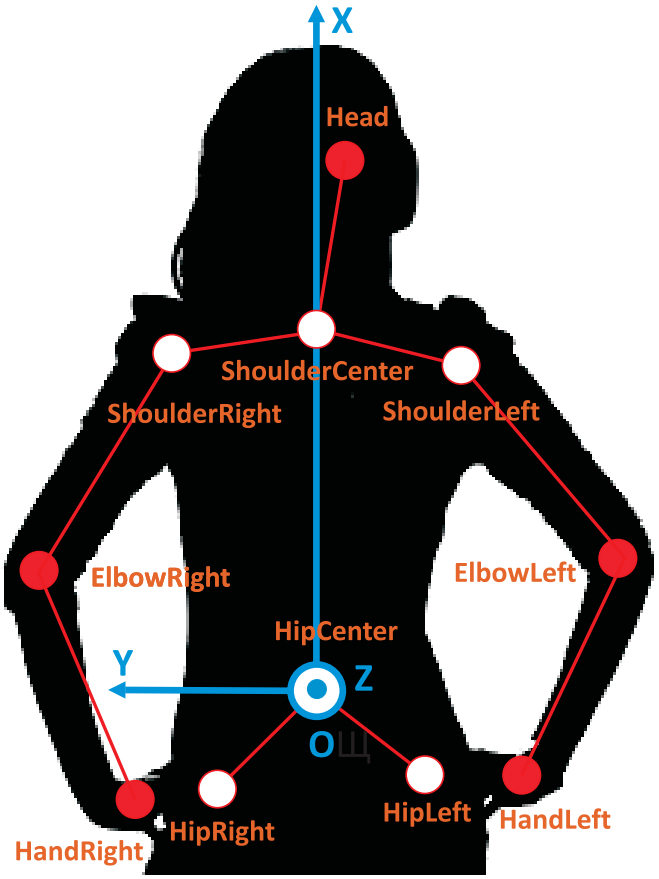
Articulated pose descriptor

1. Based on 11 upper body joints (excludes unstable wrist joints).

2. Position normalization: HipCenter is an origin of a new coordinate system.

3. Size normalization by the mean distance between each pair of joints.[1]

4. Calculation of basis vectors for each frame (shown in blue) by applying PCA on 6 torso joints (shown in white).



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

Articulated pose descriptor

1. Normalized positions of joints.[1]

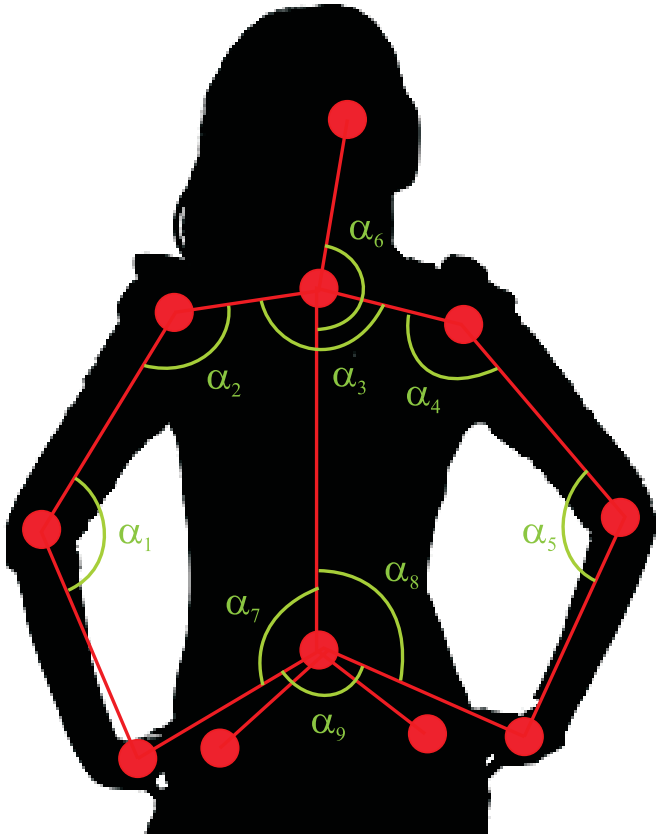
2. Velocity and acceleration of each joint.[1]

3. 9 inclination angles between selected virtual bones (shown).

4. 9 azimuth angles between projections of virtual bones relative to body.

5. 11 angles relative to the camera sensor.

6. 55 pairwise distances between all joints.



[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 and color video streams

1. Interested in capturing fine movements of palms and fingers.

2. Hand positions provided by the skeleton stream are stabilized inside a short spatio-temporal block.

3. Extract a bounding box around right and left hands centered at hand positions provided by skeleton.

4. Subtract background by thresholding along depth axis.

5. Apply local contrast normalization.

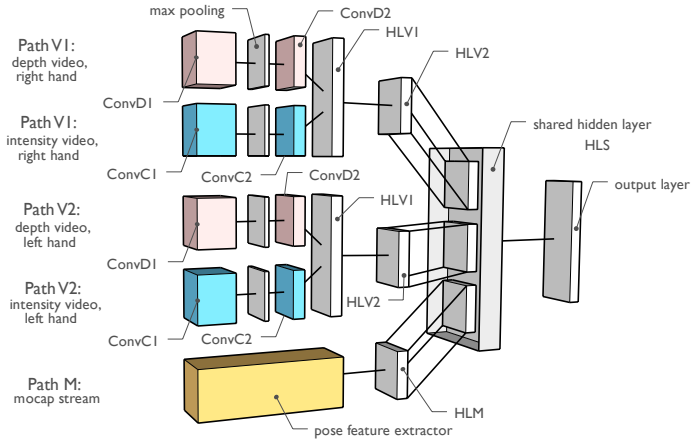
Training algorithm

Difficulties:

Number of parameters to learn: ~6 200 000.

Number of training gestures: ~10 000.

Result: poor convergence.

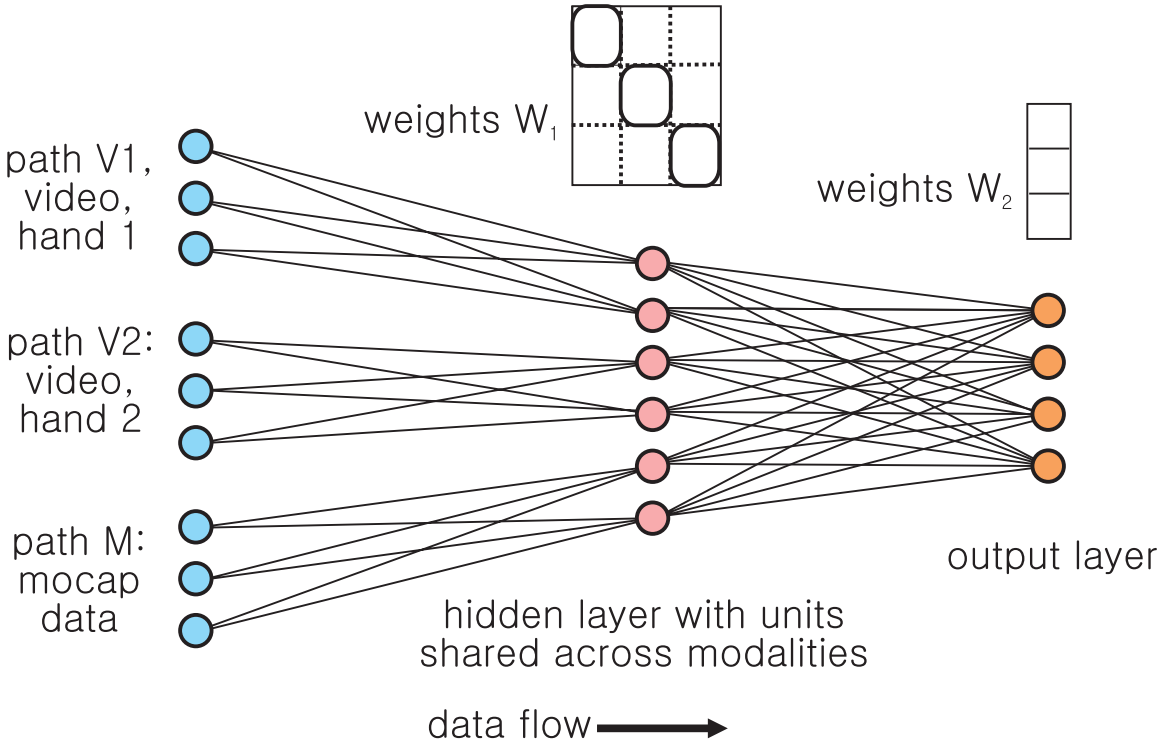


Proposed solution:

- pretraining of individual channels separately;
- “clever” initialization of shared layers;
- an iterative training algorithm gradually increasing the number of parameters to learn.

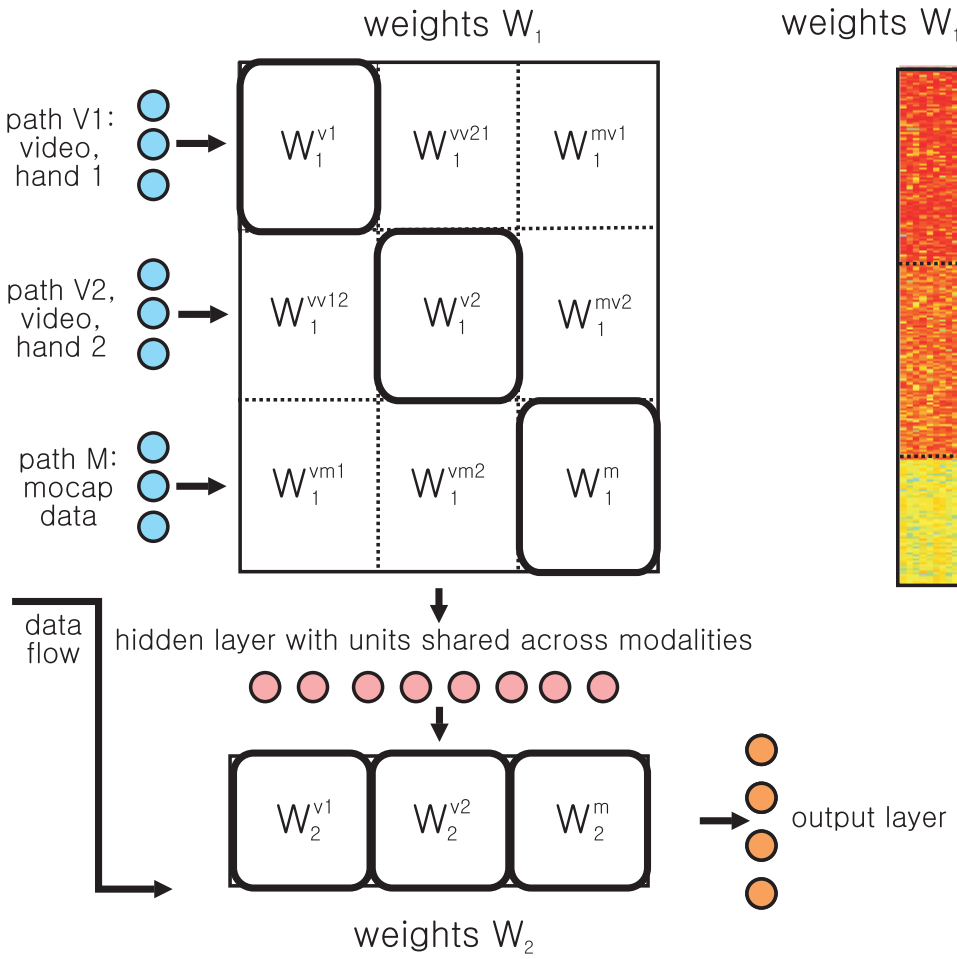
Initialization of the shared layer: structured weight matrices

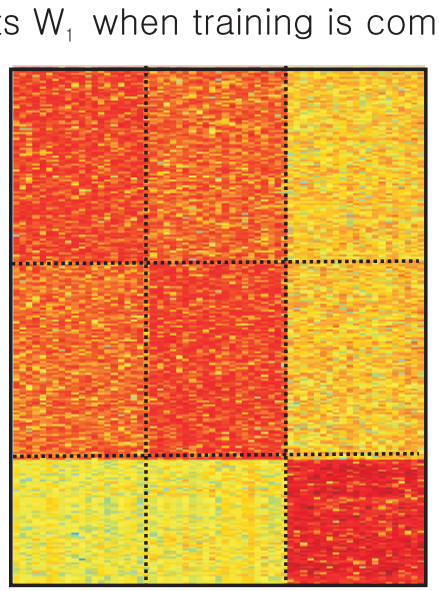
The top hidden layer from each path is initially wired to a subset of neurons in the shared layer.



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

A slightly different view

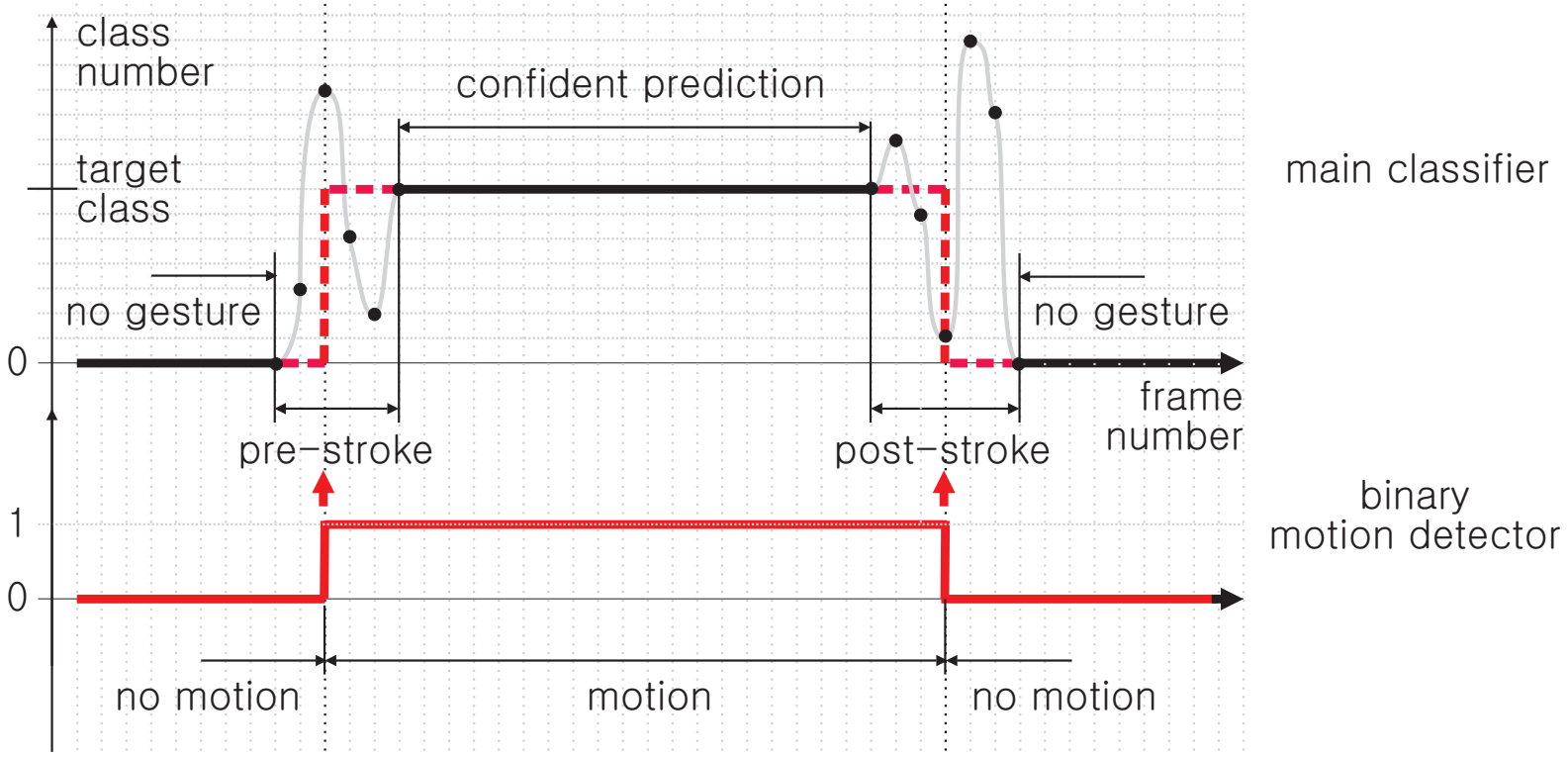




Blocks of the weight matrices are learned iteratively after proper initialization of the diagonal elements.

Gesture localization

The system outputs noisy predictions at pre-stroke and post-stroke phases due to high similarity between gesture classes at these time periods and temporal inertia of the classifier.



An additional binary classifier is employed for filtering and refinement of temporal position of each gesture.

2014 ChaLearn Looking at People Challenge[2] Track 3: Gesture recognition

Rank	Team	Score
1	**LIRIS (ours)**	**0.8500**
2	CraSPN	0.8339
3	JY	0.8268
4	CUHK-SWJTU	0.7919
5	Ipigou	0.7888
6	stevenwudi	0.7873
7	ismar	0.7466
8	Quads	0.7454
9	Telepoints	0.6888
10	TUM-fortiss	0.6490

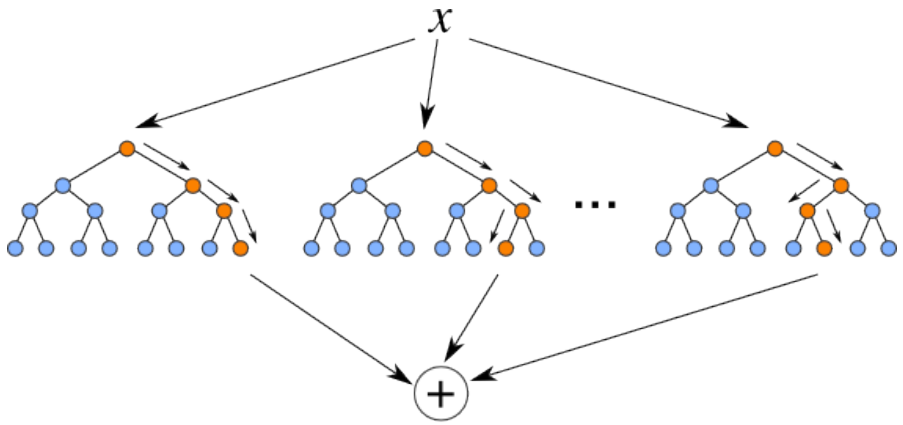
[2]Escalera et al, "ChaLearn Looking at People Challenge 2014: Dataset and Results", ECCV workshop, 2014

Performance of separate modalities at different temporal scales

Step	Articulated pose	Video	All
2	0.6938	0.7862	0.8188
3	0.7734	0.7926	0.8255
4	0.7891	0.7990	0.8449
All	**0.8080**	**0.8096**	**0.8488**

(0.8382)

Baseline model:
handcrafted descriptors + an ensemble classifier trained in a similar iterative fashion



1. Same 99-dimensional articulated pose descriptor.
2. HoG features from intensity images.
3. Histograms of depths.
4. Derivatives of histograms of depths and gradients.
5. Extremely randomized trees (ERT) classifier.

Performance of different architectures

Model	Without motion detector	With motion detector	Virtual rank
Deep learning (proposed)	0.8118	0.8488	(1)
ERT (baseline)	0.7278	0.7811	(6)
Deep learning + ERT (hybrid)	**0.8143**	**0.8500**	(1)

Conclusion

1. General method for gesture and near-range action recognition from a combination of color and depth video and articulated pose data.

2. Each channel captures a spatial scale, the system operates at three temporal scales.

3. An effective learning algorithm including pre-training with diagonal initialization of shared weights and iterative fusion.

4. Possible extensions:
   - additional modalities;
   - additional temporal scales;
   - feedback connections to handle noisy or missing channels.

ChaLearn Looking at People: pose recovery, action/interaction, gesture recognition. In conjunction with ECCV 2014
September 6th, 2014, Zurich, Switzerland