Learning human motion: gestures, activities, pose, identity

Christian Wolf
INRIA-Chroma, CITI, LIRIS, INSA-Lyon
Gestures

Articulated pose

Activities

Face-to-face interactions
Trainable multi-modal models

This talk will focus on two main points:

1. How can we train deep complex models?
2. How can we deal with multi-modality?
Learning to predict

We would like to predict a value \( t \) from an observed input

\[
y = h(x, \theta)
\]

Parameters \( \theta \) are learned from training data.
Deep neural networks
From the standard toolbox

Convolutions introduce an inductive bias for imaging/vision applications.

Recurrent networks allow to model sequences.

[LeCun et al., 1998]

[Hocheiter and Schmidhuber, 1997]
Some gentle words on learning theory

\[ L_{\mathcal{D}, f}(h) \overset{\text{def}}{=} \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] \]

Expected error over (unknown) data distribution \( \mathcal{D} \) and (groundtruth) labeling function \( f \)

[Shai Shalev-Shwartz et al. 2014]
Empirical Risk Minimization

\[ L_S(h) \overset{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}, \]

Empirical error over Samples \( y_i \) from training set S

How different is the empirical error from the expected error?

[Shai Shalev-Shwartz et al. 2014]
Learning by gradient descent

Iterative minimisation through gradient descent:

\[ \theta[t+1] = \theta[t] + \eta \nabla \mathcal{L}(h(x, \theta), y^*) \]

Learning rate

Can be blocked in a local minimum (not that it matters much …)
Sources of error

\[ L_D, f(h) \overset{\text{def}}{=} \mathbb{P}_{x \sim D} [h(x) \neq f(x)] \]

1. Lack of generalization (shift between empirical error and expected error on the target domain)

2. Optimization problem (the solution of the ERM problem is not optimal)
Illustration: model fitting and generalization

How do we chose model complexity?

![Polynomial Curve Fitting Example](Image)

Figure 1.4 Plots of polynomials having various orders $M$, shown as red curves, fitted to the data set shown in Figure 1.2.

The root mean square (RMS) error defined by

$$E_{\text{RMS}} = \sqrt{\frac{E(w^\star)}{N}}$$

in which the division by $N$ allows us to compare different sizes of data sets on an equal footing, and the square root ensures that $E_{\text{RMS}}$ is measured on the same scale (and in the same units) as the target variable $t$. Graphs of the training and test set RMS errors are shown, for various values of $M$, in Figure 1.5. The test set error is a measure of how well we are doing in predicting the values of $t$ for new data observations of $x$. We note from Figure 1.5 that small values of $M$ give relatively large values of the test set error, and this can be attributed to the fact that the corresponding polynomials are rather inflexible and are incapable of capturing the oscillations in the function $\sin(2\pi x)$. Values of $M$ in the range $3 \leq M \leq 8$ give small values for the test set error, and these also give reasonable representations of the generating function $\sin(2\pi x)$, as can be seen, for the case of $M = 3$, from Figure 1.4.

[C. Bishop, Pattern recognition and Machine learning, 2006]
Big Data!

Overfitting decreases with increasing amount of data

\[ M = 9 \]

\[ N = 15 \]

\[ N = 100 \]

[C. Bishop, Pattern recognition and Machine learning, 2006]
[Yosinski, Clune, Bengio, Lipson, "How transferable are features in deep neural networks?", 2014]
Numerical/optimization issues (Positive or negative)

[Yosinski, Clune, Bengio, Lipson, "How transferable are features in deep neural networks?", 2014]
Difficult problems

Example: Visual Question Answering

“What is the moustache made of?”

Huge Models
Huge *and* complex models

[Fourure et al., BMVC 2017]
How can we still train all this?

- Regularization, normalization
- Data augmentation
- Throw large/insane amounts of data at the problem
  - Simulation, rendering
  - Complex and tiring acquisitions
- Increase the amount of information used from the data
  - Weakly supervised learning
  - Semi supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Create « smarter » models (inductive bias)
- Learn to focus on the relevant parts the data
  - Attention mechanisms
Throw insane amounts of data at the problem

CMU Panoptical dataset
[Neverova, Wolf, Taylor, Nebout. CVIU 2017]
How can we still train all this?

- Regularization, normalization
- Data augmentation
- Throw large/insane amounts of data at the problem
  - Simulation, rendering
  - Complex and tiring acquisitions
- Increase the amount of information used from the data
  - Weakly supervised learning
  - Semi supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Create « smarter » models (inductive bias)
- Learn to focus on the relevant parts the data
  - Attention mechanisms
Yann LeCun’s cake

- **“Pure” Reinforcement Learning** *(cherry)*
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**

- **Supervised Learning** *(icing)*
  - The machine predicts a category or a few numbers for each input.
  - Predicting human-supplied data.
  - **10→10,000 bits per sample**

- **Unsupervised/Predictive Learning** *(cake)*
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos.
  - **Millions of bits per sample**

Slide: Y. LeCun
Unsupervised Learning

« The brain has about $10^{14}$ synapses and we only live for about $10^9$ seconds. So we have more parameters than data. This motivates the idea that we must do a lot of unsupervised learning since the perceptual input is the only place we can get $10^5$ dimensions of constraint per second. »

Geoffrey Hinton
Combining real and simulated data

Joint positions (NYU Dataset)    Synthetic data (part segmentation)

Work of Natalia Neverova
Phd @ LIRIS, Now at Facebook AI

With Graham W. Taylor,
University of Guelph, Canada

Florian Nebout
AwAbot
A complementary representation

A dense finger part segmentation is complementary to joint coordinates

- Voronoi boundaries

[Neverova, Wolf, Taylor, Nebout. CVIU 2017]
Weakly/semi-supervised learning

- Synthetic data: segmentation maps
- Real data labels: joint coordinates
- Regression learner $f_r$
- Segmentation learner $f_s$
- Region localization
- Patchwise restoration
- Dictionary of synthetic patches
- Error segmentation vs joints

[Neverova, Wolf, Taylor, Nebout. CVIU 2017]
[Neverova, Wolf, Taylor, Nebout. CVIU 2017]
How can we still train all this?

- Regularization, normalization
- Data augmentation
- Throw large/insane amounts of data at the problem
  - Simulation, rendering
  - Complex and tiring acquisitions
- Increase the amount of information used from the data
  - Weakly supervised learning
  - Semi supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Create « smarter » models (inductive bias)
- Learn to focus on the relevant parts the data
  - Attention mechanisms
Gesture recognition

Project Interabot
Awabot / LIRIS / LIG / Voxler
Gestures have spatial and temporal scales
Multi-modal input

Skeleton descriptor

Depth and intensity images

Audio stream
Path V1:
- depth video, right hand

Path V1:
- intensity video, right hand

Path V2:
- depth video, left hand

Path V2:
- intensity video, left hand

Path M:
- mocap stream
- pose feature extractor

Path A:
- audio stream
- mel frequency histograms

[Neverova, Wolf, Taylor, Nebout, IEEE-T-PAMI 2016]
Learning modality fusion
Dropout

- Introduced in 2012 for Imagenet

- During training, for each training sample, 50% of the units are disabled.
- Punishes co-adaptation of units
- Large performance gains
Moddrop: modality-wise dropout

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

\[
\tilde{h}_{ij}^{(k)} = \sigma \left[ \sum_{i=1}^{F_k} w_{i,j}^{(k,k)} x_i^{(k)} + \delta(k) \sum_{n=1}^{K} \sum_{i=1}^{F_n} w_{i,j}^{(n,k)} x_i^{(n)} + b_j^{(k)} \right]
\]

Bernoulli selector
\[
P(\delta^{(k)} = 1) = p^{(k)}
\]
ECCV 2014 Challenge on Looking at People (gesture recognition track)

<table>
<thead>
<tr>
<th>#</th>
<th>Team</th>
<th>Score</th>
<th>#</th>
<th>Team</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Ours</strong> [47]</td>
<td>0.850</td>
<td>7</td>
<td>Camgoz et al. [61]</td>
<td>0.747</td>
</tr>
<tr>
<td>2</td>
<td>Monnier et al. [29]</td>
<td>0.834</td>
<td>8</td>
<td>Evangelidis et al. [62]</td>
<td>0.745</td>
</tr>
<tr>
<td>3</td>
<td>Chang [30]</td>
<td>0.827</td>
<td>9</td>
<td>Undisclosed authors</td>
<td>0.689</td>
</tr>
<tr>
<td>4</td>
<td>Peng et al. [63]</td>
<td>0.792</td>
<td>10</td>
<td>Chen et al. [64]</td>
<td>0.649</td>
</tr>
<tr>
<td>5</td>
<td>Pigou et al. [36]</td>
<td>0.789</td>
<td></td>
<td><strong>...</strong></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Wu [37]</td>
<td>0.787</td>
<td>17</td>
<td>Undisclosed authors</td>
<td>0.271</td>
</tr>
</tbody>
</table>

**Ours, improved results after the competition** 0.870

<table>
<thead>
<tr>
<th>Model</th>
<th>Pose (mocap)</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours, submitted entry</strong> [47]</td>
<td>0.808</td>
<td>0.810</td>
</tr>
<tr>
<td>Peng et al. [63]</td>
<td>–</td>
<td>0.792</td>
</tr>
<tr>
<td>Pigou et al. [36]</td>
<td>–</td>
<td>0.789</td>
</tr>
<tr>
<td>Chang [30]</td>
<td>0.795</td>
<td>–</td>
</tr>
<tr>
<td>Monnier et al. [29] (validation set)</td>
<td>0.791</td>
<td>–</td>
</tr>
<tr>
<td>Wu and Shao [37]</td>
<td>0.787</td>
<td>0.637</td>
</tr>
<tr>
<td>Evangelidis et al. [62], after competition</td>
<td>0.768</td>
<td>–</td>
</tr>
<tr>
<td>Camgoz et al. [61]</td>
<td>0.747</td>
<td>–</td>
</tr>
<tr>
<td>Evangelidis et al. [62], submitted entry</td>
<td>0.745</td>
<td>–</td>
</tr>
</tbody>
</table>

**Ours, after competition** 0.831 0.836

[Neverova, Wolf, Taylor, Nebout, IEEE-T-PAMI 2016]
Results
Results w.r.t. to signal losses

Classification accuracy on the validation set (dynamic poses)

<table>
<thead>
<tr>
<th>Modalities</th>
<th>Dropout</th>
<th>Dropout+Moddrop</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>96.77</td>
<td>96.81</td>
</tr>
<tr>
<td>Mocap missing</td>
<td>38.41</td>
<td>92.82</td>
</tr>
<tr>
<td>Audio missing</td>
<td>84.10</td>
<td>92.59</td>
</tr>
<tr>
<td>Hands missing</td>
<td>53.13</td>
<td>73.28</td>
</tr>
</tbody>
</table>

Jacquard index on test set (full gestures)

<table>
<thead>
<tr>
<th>Modalities</th>
<th>Dropout</th>
<th>Dropout+Moddrop</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.875</td>
<td>0.875</td>
</tr>
<tr>
<td>Mocap missing</td>
<td>0.306</td>
<td>0.859</td>
</tr>
<tr>
<td>Audio missing</td>
<td>0.789</td>
<td>0.854</td>
</tr>
<tr>
<td>Hands missing</td>
<td>0.466</td>
<td>0.680</td>
</tr>
</tbody>
</table>

[Neverova, Wolf, Taylor, Nebout, IEEE-T-PAMI 2016]
Can we learn modality fusion?

- Path V1: depth video, right hand
- Path V2: depth video, left hand
- Path V1: intensity video, right hand
- Path V2: intensity video, left hand
- Path M: mocap stream
- Path A: audio stream

Diagram showing layers and connections for different modalities.
Method 1: Modout

Automatically learning multi-modal architectures with stochastic regularization.

<table>
<thead>
<tr>
<th>Method</th>
<th>BackProp</th>
<th>Dropout</th>
<th>Blockout</th>
<th>ModDrop</th>
<th>Modout</th>
<th>Modout +Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation accuracy</td>
<td>91.1</td>
<td>91.5</td>
<td>91.7</td>
<td>92.1</td>
<td>91.6</td>
<td><strong>92.9</strong></td>
</tr>
<tr>
<td>Test accuracy</td>
<td>92.0</td>
<td>92.5</td>
<td>92.6</td>
<td>92.4</td>
<td>93.6</td>
<td><strong>93.8</strong></td>
</tr>
</tbody>
</table>

[Li, Taylor, Neverova, Wolf, Face and Gesture recognition, 2017]
Method 2: Graph induced kernels

- Fusion strategies are described by graphs.
- Definition of graph distances
- Bayesian optimization

[Ramachandram, Lisicki, Shields, Amer, Taylor, arxiv 7/2017]
How can we still train all this?

- Regularization, normalization
- Data augmentation
- Throw large/insane amounts of data at the problem
  - Simulation, rendering
  - Complex and tiring acquisitions
- Increase the amount of information used from the data
  - Weakly supervised learning
  - Semi supervised learning
  - Unsupervised learning
  - Reinforcement learning
- Create « smarter » models (inductive bias)
- Learn to focus on the relevant parts the data
  - Attention mechanisms
Human activity recognition

(NTU)
Pose stream: state of the art
Articulated pose alone is not sufficient

Same class?!

Reading

Writing
Articulated pose alone is not sufficient

RGB is helpful...

Reading

Writing
Attention on relevant parts

Joint important for activity high attention

Joint wrongly located low attention

Work of Fabien Baradel, Phd @ LIRIS

With Julien Mille (INSA Val de Loire)
Teaser 1: how?

[Baradel, Wolf, Mille, arxiv 2017]
RGB raw sequence

Visualization of the attention process of our model

‘Handshaking’
Body motion of the full sub-sequence
Body motion of the full sub-sequence
$S$

Body motion of the full sub-sequence
# Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>Pose</th>
<th>RGB</th>
<th>CS</th>
<th>CV</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lie Group [37]</td>
<td>X</td>
<td>-</td>
<td>50.1</td>
<td>52.8</td>
<td>51.5</td>
</tr>
<tr>
<td>Skeleton Quads [9]</td>
<td>X</td>
<td>-</td>
<td>38.6</td>
<td>41.4</td>
<td>40.0</td>
</tr>
<tr>
<td>Dynamic Skeletons [13]</td>
<td>X</td>
<td>-</td>
<td>60.2</td>
<td>65.2</td>
<td>62.7</td>
</tr>
<tr>
<td>HBRNN [8]</td>
<td>X</td>
<td>-</td>
<td>59.1</td>
<td>64.0</td>
<td>61.6</td>
</tr>
<tr>
<td>Deep LSTM [30]</td>
<td>X</td>
<td>-</td>
<td>60.7</td>
<td>67.3</td>
<td>64.0</td>
</tr>
<tr>
<td>Part-aware LSTM [30]</td>
<td>X</td>
<td>-</td>
<td>62.9</td>
<td>70.3</td>
<td>66.6</td>
</tr>
<tr>
<td>ST-LSTM + TrustG. [23]</td>
<td>X</td>
<td>-</td>
<td>69.2</td>
<td>77.7</td>
<td>73.5</td>
</tr>
<tr>
<td>STA-LSTM [34]</td>
<td>X</td>
<td>-</td>
<td>73.2</td>
<td>81.2</td>
<td>77.2</td>
</tr>
<tr>
<td>JTM [39]</td>
<td>X</td>
<td>-</td>
<td>76.3</td>
<td>81.1</td>
<td>78.7</td>
</tr>
<tr>
<td>DSSCA - SSLM [31]</td>
<td>X</td>
<td>X</td>
<td>74.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Ours (pose only)</strong></td>
<td>X</td>
<td>-</td>
<td><strong>77.1</strong></td>
<td><strong>84.5</strong></td>
<td><strong>80.8</strong></td>
</tr>
<tr>
<td><strong>Ours (RGB only)</strong></td>
<td>-</td>
<td>X</td>
<td><strong>75.6</strong></td>
<td><strong>80.5</strong></td>
<td><strong>78.1</strong></td>
</tr>
<tr>
<td><strong>Ours (pose + RGB)</strong></td>
<td>X</td>
<td>X</td>
<td><strong>84.8</strong></td>
<td><strong>90.6</strong></td>
<td><strong>87.7</strong></td>
</tr>
</tbody>
</table>

Table 1: Results on the NTU RGB+D dataset with Cross-Subject (CS) and Cross-View (CV) settings (accuracies in %)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Pose</th>
<th>RGB</th>
<th>Depth</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw skeleton [45]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>49.7</td>
</tr>
<tr>
<td>Joint feature [45]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>80.3</td>
</tr>
<tr>
<td>Raw skeleton [46]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>79.4</td>
</tr>
<tr>
<td>Joint feature [46]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>86.9</td>
</tr>
<tr>
<td>HBRNN [8]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>80.35</td>
</tr>
<tr>
<td>Co-occurrence RNN [47]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>90.4</td>
</tr>
<tr>
<td>STA-LSTM [34]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>91.5</td>
</tr>
<tr>
<td>ST-LSTM + Trust Gate [23]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>93.3</td>
</tr>
<tr>
<td>DSPM [22]</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>93.4</td>
</tr>
<tr>
<td><strong>Ours (Pose only)</strong></td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>90.5</td>
</tr>
<tr>
<td><strong>Ours (RGB only)</strong></td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>72.0</td>
</tr>
<tr>
<td><strong>Ours (Pose + RGB)</strong></td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>94.1</td>
</tr>
</tbody>
</table>

Table 2: Results on SBU Kinect Interaction dataset (accuracies in %)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Pose</th>
<th>RGB</th>
<th>Depth</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action Ensemble [38]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>68.0</td>
</tr>
<tr>
<td>Efficient Pose-Based [10]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>73.1</td>
</tr>
<tr>
<td>Moving Pose [47]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>73.8</td>
</tr>
<tr>
<td>Moving Poselets [36]</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>74.5</td>
</tr>
<tr>
<td>Depth Fusion [48]</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>88.8</td>
</tr>
<tr>
<td>MMMP [32]</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>91.3</td>
</tr>
<tr>
<td>DL-GSGC [24]</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>95.0</td>
</tr>
<tr>
<td>DSSCA - SSLM [31]</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>97.5</td>
</tr>
<tr>
<td><strong>Ours (Pose only)</strong></td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>74.6</td>
</tr>
<tr>
<td><strong>Ours (RGB only)</strong></td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>75.3</td>
</tr>
<tr>
<td><strong>Ours (Pose + RGB)</strong></td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Table 3: Results on MSR Daily Activity 3D dataset (accuracies in %)
## Results: ablation study 1

<table>
<thead>
<tr>
<th>Methods</th>
<th>Pose</th>
<th>RGB</th>
<th>Attention Spatial</th>
<th>Attention Temporal</th>
<th>CS</th>
<th>CV</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Pose only</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>77.1</td>
<td>84.5</td>
<td>80.8</td>
</tr>
<tr>
<td>B RGB only, no attention (sum of features)</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>61.5</td>
<td>65.9</td>
<td>63.7</td>
</tr>
<tr>
<td>C RGB only, no attention (concat of features)</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>63.2</td>
<td>67.2</td>
<td>65.2</td>
</tr>
<tr>
<td>E RGB only + spatial attention</td>
<td>o</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>67.4</td>
<td>71.2</td>
<td>69.3</td>
</tr>
<tr>
<td>G RGB only + spatio-temporal attention</td>
<td>o</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>75.6</td>
<td>80.5</td>
<td>78.1</td>
</tr>
<tr>
<td>H Multi-modal, no attention (A+B)</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>83.0</td>
<td>88.5</td>
<td>85.3</td>
</tr>
<tr>
<td>I Multi-modal, spatial attention (A+E)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>84.1</td>
<td>90.0</td>
<td>87.1</td>
</tr>
<tr>
<td>K Multi-modal, spatio-temporal attention (A+G)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>84.8</td>
<td>90.6</td>
<td>87.7</td>
</tr>
</tbody>
</table>
Results: ablation study 2

<table>
<thead>
<tr>
<th>Methods</th>
<th>Attention</th>
<th>CS</th>
<th>CV</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conditional to pose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGB only</td>
<td>-</td>
<td>67.4</td>
<td>71.2</td>
<td>69.3</td>
</tr>
<tr>
<td>RGB only</td>
<td>X</td>
<td>75.6</td>
<td>80.5</td>
<td>78.1</td>
</tr>
<tr>
<td>Multi-modal</td>
<td>-</td>
<td>83.9</td>
<td>90.0</td>
<td>87.0</td>
</tr>
<tr>
<td>Multi-modal</td>
<td>X</td>
<td>84.8</td>
<td>90.6</td>
<td>87.7</td>
</tr>
</tbody>
</table>
Conclusions

- Goal: use multiple modalities where they are available and where they can increase performance.

- Goal: get rid of requirements on modalities where they are not needed.
  - E.g. Get of the requirement on articulated pose (skeleton) w/o any performance drop

- Current tendency: replace structured models by models of (visual) attention.

- Active vision vs. robotics: robot navigation for human activity recognition.
Outlook

Progress in Computer Vision and AI (Deep Learning) will bring realistic service robotics soon into our reach.

We are hiring a post-doc!
Robot compagnon

We are hiring a post-doc!