Pose and gestures: spatial deep learning

Mingyuan Jiu
Natalia Neverova
Christian Wolf
Graham W. Taylor
Atilla Baskurt

February 6th, 2014
Human motion analysis

Motion analysis from multimodal signals using deep learning

- Multi-modal gesture recognition, integrating
  - Articulated pose
  - Raw video signal
  - Audio

- Human body part segmentation
Multi-modal system

Articulated pose:
Full body skeleton provided by the sensor.

Depth video:
Cropped around right and left hands.

Audio signal:
Automatic speech recognition.
Two spatial scales

Full body motion and local movements of fingers

Communicative gestures:
- people communication (conversational gestures);
- natural human-robot interaction.
Proposed learning architecture
Video stream: convolutional neural network (ConvNet)

Hand positions are stabilized between frames within each short spatio-temporal block.
Skeleton stream: pose descriptor

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

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

3. Size normalization by the distance between HipCenter and ShoulderCenter.

4. Calculation of basis vectors (shown in blue) by applying PCA on 6 torso joints (shown in white).
Audio stream: two different methods

Official features (used for participation at ChaLearn challenge)
1. MFCC-based recognition of pronounced phonemes (Julius LVCSR engine).
2. Voice Activity Detection (VAD) + Automatic Speech Recognition (ASR)
3. Word-spotting strategy

New, learned representation (better; used after challenge)
1. Audio signal detection (thresholding over amplitude and duration).
2. Mel-frequency scaled spectrograms of the audio signal (40 filter banks).
3. 2-layer ConvNet + 2 hidden layers for individual predictions over short blocks.

(1) Raw audio signal  (2) Spectrogram  (3) Mel scale
Proposed learning architecture

- **Video stream (depth): hand 1**
  - 4-layer ConvNet
  - MLP
  - Output: 0...N

- **Video stream (depth): hand 2**
  - 4-layer ConvNet
  - MLP
  - Output: 0...N

- **Skeleton stream**
  - 139$L_1$-dimensional pose descriptor
  - MLP
  - Output: 0...N

- **Audio stream**
  - Features: MFCC
  - BoW
  - Output: 0...N

- **RNN**
  - Output: 0...N

**Long sequences (gestures)**

**Short spatio-temporal blocks (dynamic poses)**
Experimental results

2013 ChaLearn Gesture Challenge on multi-modal gesture recognition

RGB-D video, skeleton, audio
20 gesture classes
13,858 annotated samples
2013 ChaLearn Gesture Challenge on multi-modal gesture recognition

For detailed results on validation and test datasets see:

<table>
<thead>
<tr>
<th>Team</th>
<th>ED</th>
<th>Rank</th>
<th>Team</th>
<th>ED</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team 1</td>
<td>0.1276</td>
<td>1</td>
<td>Our team*</td>
<td>0.1773</td>
<td>6</td>
</tr>
<tr>
<td>Team 2</td>
<td>0.1539</td>
<td>2</td>
<td>Team 7</td>
<td>0.2445</td>
<td>7</td>
</tr>
<tr>
<td>Team 3</td>
<td>0.1711</td>
<td>3</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Team 4</td>
<td>0.1722</td>
<td>4</td>
<td>Team 16</td>
<td>0.8746</td>
<td>16</td>
</tr>
<tr>
<td>Team 5</td>
<td>0.1733</td>
<td>5</td>
<td>Team 17</td>
<td>0.9207</td>
<td>17</td>
</tr>
</tbody>
</table>

*Improved score with new audio features: 0.1136
Comparison of different modalities
(no long-term temporal dependencies, voting per short-time block)

<table>
<thead>
<tr>
<th>Modalities used</th>
<th>Recall</th>
<th>Precision</th>
<th>Edit distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth video only</td>
<td>0.5433</td>
<td>0.5494</td>
<td>0.6613</td>
</tr>
<tr>
<td>Articulated pose</td>
<td>0.7298</td>
<td>0.7420</td>
<td>0.4250</td>
</tr>
<tr>
<td>Audio stream</td>
<td>0.6754</td>
<td>0.6590</td>
<td>0.4966</td>
</tr>
<tr>
<td>Depth + pose</td>
<td>0.7618</td>
<td>0.7739</td>
<td>0.3809</td>
</tr>
<tr>
<td>Depth + audio</td>
<td>0.7669</td>
<td>0.7796</td>
<td>0.3742</td>
</tr>
<tr>
<td>Pose + audio</td>
<td>0.8765</td>
<td>0.8885</td>
<td>0.2125</td>
</tr>
<tr>
<td><strong>Depth + pose + audio</strong>*</td>
<td><strong>0.8784</strong></td>
<td><strong>0.8920</strong></td>
<td><strong>0.2091</strong></td>
</tr>
<tr>
<td>Random predictions</td>
<td></td>
<td></td>
<td>1.4747</td>
</tr>
</tbody>
</table>

* Improved score: 0.93 recall / precision
Human motion analysis

Motion analysis from multimodal signals using deep learning

Multi-modal gesture recognition, integrating
• Articulated pose
• Raw video signal
• Audio

Human body part segmentation
Segmentation through classification

Application:
- Calculate human pose: set of joint positions
- Use an intermediate representation: body part segmentation

Figure: Shotton et al., CVPR 2011

Jiu, Wolf, Baskurt, 2013
Segmentation through classification

Goal:
- Extremely fast segmentation: real time
- Models which do not require solving any combinatorial problems
- Training of classifier; independent classification, pixel by pixel

Figure: Shotton et al., CVPR 2011
Spatial relationships: (1) pixels

- Classical solutions for the integration of spatial relationships:
  - Variational methods (level sets, active contours etc.);
  - Probabilistic graphical models (MRF);
  - Heuristics and context based methods (filtering);
- Example for regularization terms: neighboring pixels are likely to have similar labels
- Example: energy function / MRF + discrete optimization

\[ E(l_1, \ldots, l_N) = \sum_i U(l_i, Z_i) + \alpha \sum_{(i,j) \in \mathcal{E}} D(l_i, l_j) \]
Spatial relationships: (2) labels

**Additional information:** neighboring pixels are likely
- to have similar labels, or
- to have labels which are adjacent in the object layout (!!)

Could also be solved by MRF + discrete optimization

\[
E(l_1, \ldots, l_N) = \sum_i U(l_i, Z_i) + \alpha \sum_{(i,j) \in E} D(l_i, l_j)
\]
Spatial relationships

Possible models differing in integration of spatial relationships:

---

No
Independent classification, pixel by pixel.

Yes
MRF with pairwise terms. Global minimization, high computational complexity.

Yes
Independent classification, pixel by pixel. The prior improves the classifier.
Structured models ... w/o structure

- It is **not** possible to include pairwise terms into a classifier which classifies pixels independently.
- Pairwise terms lead to combinatorial problems.
- Alternative strategy:
  - do **not** proceed by pairs
  - change the loss function for pixelwise classification
  - punish errors (classically), but:
    - punish errors **less**, if the misclassified label is a **neighbor** of the groundtruth label
- It will be shown that this strategy decreases "pure" **classical (!!) classification error."
the data, although embedded in a high dimensional space, can be learned. In most cases only few data are complex enough. In reality, the available amount of training images or objects into parts through pixelwise classification, which integrates the spatial layout of the part labels.

Classifying a prediction machine. Our objective is to learn features in a prediction machine. Our objective is to learn features in their design mostly relies on human intelligence. Recently, sophisticated learning architectures. It may also exploit the much larger amount of data, although embedded in a high dimensional space, can be learned. In most cases only few data are complex enough. In reality, the available amount of training.

Classification by random forests
Integration into an entropy based algorithm

Feature learning
Integration into non-linear dimensionality reduction

Classification by MLP or log. regr.
Integration into a loss function; label ranking strategy

[Pattern Recognition Letters 2014]
Spatial deep learning

M images \{X_1^m, \ldots, X_M^m\}

- A parametric function maps pixels \(i\) (and their receptive fields) to a feature representation

\[ Z_i^m \in \mathbb{R}^Q \]

\[ Z_i^m = f(X_i^m | \theta_f) \]

- A classifier predicts part labels

\[ \hat{l}_i = g(Z_i^m | \theta_g) \]
Classical supervised learning

Stimulated network output:

\[ \hat{l}_i = g(Z_i^m | \theta_g) \]

Target output (groundtruth):

\[ \bar{l}_1, \bar{l}_2, \bar{l}_3 \]

Classical loss function: cross entropy

\[ E(w) = -\sum_{n} \left\{ \bar{l}_n \ln \hat{l}_n + (1 - \bar{l}_n) \ln(1 - \hat{l}_n) \right\} \]
Learning to rank class labels

- The groundtruth class label is supposed to be ranked first (highest classifier response)
- The neighboring class labels are supposed to ranked next
- The non-neighboring class labels are ranked last
- The rankings inside the groups (gt, nb, non-nb) are irrelevant
Learning to rank class labels

Similar to (Burges, NIPS 2006), the loss function is decomposed into terms over pairs. For each pair, differences in network output are mapped to probabilities:

\[ o_{uv} = g(Z_{i,u}) - g(Z_{i,v}) \]

\[ P_{uv} = \frac{e^{o_{ij}}}{1 + e^{o_{ij}}} \]

A target probability is defined according to desired ranking: \( \bar{P}_{uv} \) is set to \( \lambda > 0.5 \) if \( u \) is ranked higher than \( v \), and \( 1 - \lambda \) otherwise.

Output and target probability are compared with cross-entropy loss:

\[ C_{uv} = -\bar{P}_{uv} \log P_{uv} - (1 - \bar{P}_{uv}) \log (1 - P_{uv}) \]
Learning the feature mapping

Learn a mapping from pixels (+receptive field) to features

\[ Z_{i}^{m} = f(X_{i}^{m}|\theta_{f}) \]

\( f \) can be modeled as a ConvNet (here shown for classification):

\[ \text{Image d'entrée} \]

\[ 32 \times 32 \]

\[ \text{Convolutions 2D} \]

\[ 5 \times 5 \]

\[ \text{Convolutions 2D} \]

\[ 5 \times 5 \]

\[ \ldots \]

\[ \ldots \]

\[ \ldots \]

\[ \ldots \]

\[ 120xN1 \]

\[ 84xN2 \]

\[ 10xN3 \]

\[ \ldots \]
Learning the feature mapping

Consider differences in mapped features over pairs of pixels:

\[ d_{ij}^m = \| Z_i^m - Z_j^m \|_2 \]

Minimize energy function:

\[ E = \sum_{i,j} \delta_{l_i,l_j} L_S(X_i^m, X_j^m) + \sum_{i,j} \nu_{l_i,l_j} L_D(X_i^m, X_j^m) \]

Label relationships:

- \( \delta_{a,b} = 1 \) if \( a = b \) and 0 otherwise;
- \( \nu_{a,b} \) is defined as \( \nu_{a,b} = 1 \) if parts \( a \) and \( b \) are not neighbors in the corpus,

Loss functions:

\[ L_S(X_i^m, X_j^m) = \frac{1}{2} (d_{ij}^m)^2, \]
\[ L_D(X_i^m, X_j^m) = \frac{1}{2} \left[ \max(0, \alpha - d_{ij}^m) \right]^2, \quad \alpha > 0 \text{ is a margin.} \]
Results

Figure 4: Classification examples from the CDC4CV dataset. (a) input depth image; (b) groundtruth segmentation; (c) appropriate baseline: randomized forest for CDC4CV; (d) DrLIM+LR without spatial learning; (e) our method (spatial pre-training and spatial LR learning).

Input  Groundtruth  Random forest (Shotton et al., CVPR 2011)  ConvNet w/ DrLIM pretraining (Hadsell/Chopra/Lecun, CVPR 2006) + classical backprop  ConvNet w/ spatial pretraining + spatial backprop (Our method)

CDC4CV Poselets dataset (Holt et al., 2011)
## Experimental results: accuracy

### Table 1: Evaluation of different baselines on the CDC4CV dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized forest (Shotton et al., 2011)</td>
<td>60.30%</td>
</tr>
<tr>
<td>Spatial Randomized forest (Jiu et al., 2013)</td>
<td>61.05%</td>
</tr>
<tr>
<td>Single-scale (vanilla) ConvNet (LeCun et al., 1998)</td>
<td>47.17%</td>
</tr>
<tr>
<td>Multi-scale ConvNet (Farabet et al., 2012)</td>
<td>62.54%</td>
</tr>
</tbody>
</table>

### Table 2: Results of different combinations of classical and spatial learning on the CDC4CV dataset.

<table>
<thead>
<tr>
<th>Convolutional layers</th>
<th>LR</th>
<th>Fine-tuning</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrLIM (Hadsell et al., 2006)</td>
<td>classical</td>
<td>no</td>
<td>35.10%</td>
</tr>
<tr>
<td>DrLIM (Hadsell et al., 2006)</td>
<td>spatial</td>
<td>no</td>
<td>41.05%</td>
</tr>
<tr>
<td>DrLIM (Hadsell et al., 2006)</td>
<td>spatial</td>
<td>no</td>
<td>41.65%</td>
</tr>
<tr>
<td>DrLIM (Hadsell et al., 2006)</td>
<td>spatial</td>
<td>no</td>
<td>41.65%</td>
</tr>
<tr>
<td>DrLIM (Hadsell et al., 2006)</td>
<td>spatial</td>
<td>yes</td>
<td>64.39%</td>
</tr>
<tr>
<td>DrLIM (Hadsell et al., 2006)</td>
<td>spatial</td>
<td>yes</td>
<td>65.12%</td>
</tr>
<tr>
<td>DrLIM (Hadsell et al., 2006)</td>
<td>spatial</td>
<td>yes</td>
<td>65.18%</td>
</tr>
<tr>
<td>DrLIM (Hadsell et al., 2006)</td>
<td>spatial</td>
<td>yes</td>
<td>66.92%</td>
</tr>
</tbody>
</table>

 CDC4CV Poselets dataset (Holt et al., 2011)
Spatial error distribution

All pixels

Difference between:
- Proposed > baseline
- Baseline > proposed
Normalized by all pixels
Conclusion

Learning representations can be as successful on depth images as it is on RGB images.

Learning representations can lead to significant advantages over handcrafted representations.

Learning allows to design generic frameworks which can be applied to different modalities (RGB, D, Audio etc.) and which automatically adapt to the data.

Combining all modalities in a single framework allows the model to compensate for major weaknesses of individual models and reduce the negative influence of noise.