Vision-based Hand Gesture Recognition

Sebastian Springenberg

Universität Hamburg
Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
Technische Aspekte Multimodaler Systeme

December 14, 2015

Outline

Motivation

- Hand Gestures are natural and intuitive
- Little cognitive load on the user when interacting with the system
- Various applications in HRI and HCI: Sign language interpretation, virtual reality, gesture based interaction with software / robots ...
- Cost-efficient and portable digital cameras can be used

Problems / Challenges

- System has to cope with complex scenes:
  Different backgrounds, variable lightning conditions, different gesture positions / orientations, occlusion

How to deal with it?

- Hand-segmentation vs. minimal preprocessing
- CNNs for robust real-time processing
- Static or continuous gestures?
- RGB or RGB-D information?
  RGB-D more precise but computationally more expensive

Conventional Approaches

Hand gesture recognition: A classification task

- Extract relevant features from images
- Pass features on to a classifier (e.g. SVM)

- Conventionally, hand tailored features are extracted

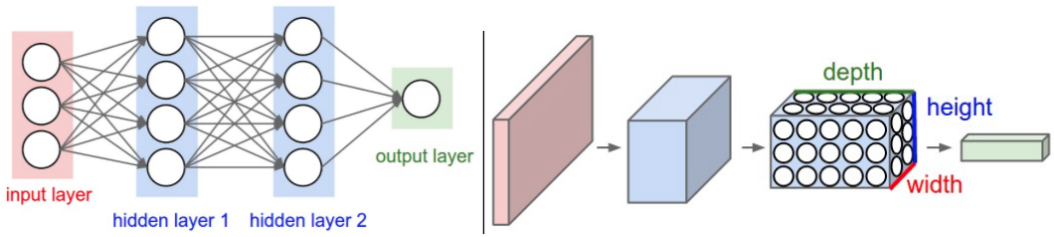
→ Classification performance highly dependent on task specific preprocessing

- Conventional feature detection methods include:
  HOG (histogram of oriented gradients), FFT, Hu Invariant Moments

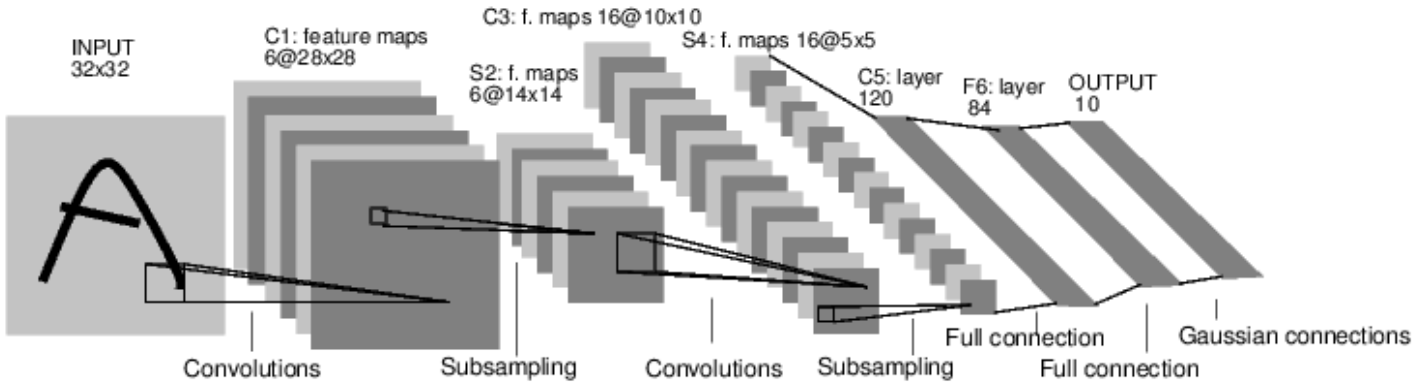
CNNs

- Biologically inspired approach on image processing
- Instead of exhaustive preprocessing, let system extract features on its own
- Similar to basic neural networks but with additional properties:
  - Weight sharing
  - Receptive fields
  - Subsampling
- Layers involved:
  - Input layer
  - Convolutional layer
  - Subsampling layer (i.e. max-pooling)
  - Output layer (usually fully connected)

2D CNNs



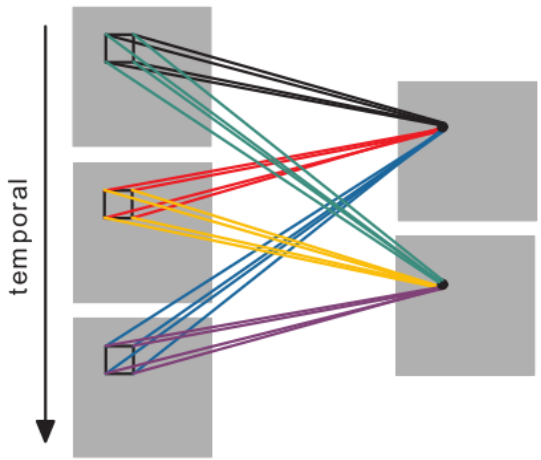
http://cs231n.github.io/convolutional-networks/



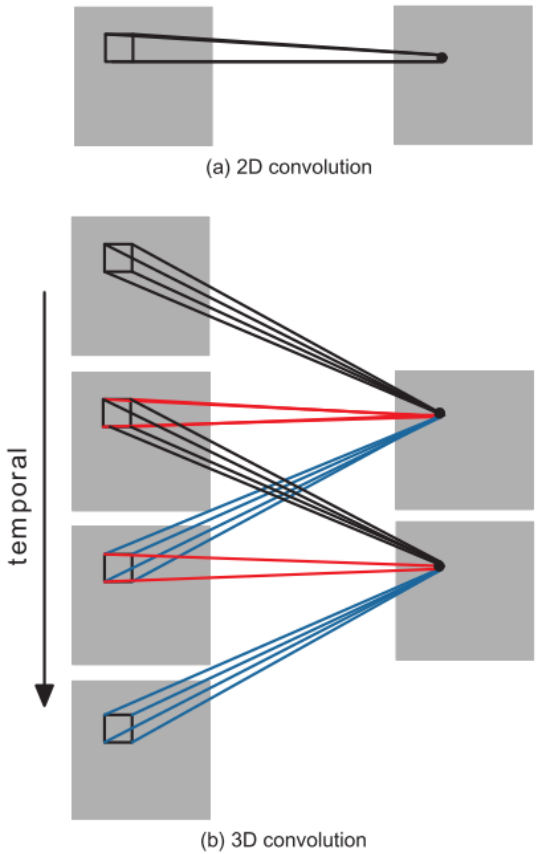
[2]

3D convolution

- Convolution both spatially and temporally for action recognition



[1]



RGB CNN

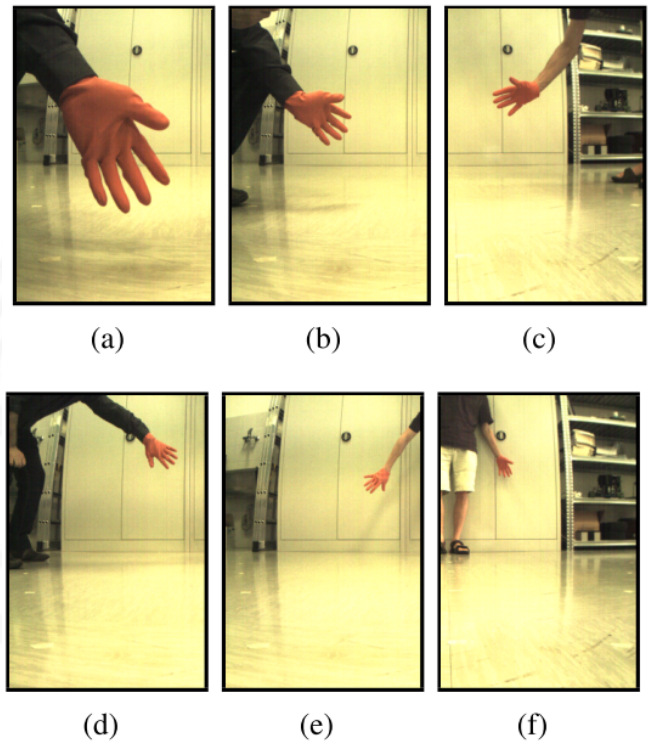
Nagi et al. use a CNN to classify hand gestures

- Gestures presented with coloured gloves.
- RGB images obtained by foot-bot robots as input to the network

[5]

RGB CNN cont.

- 6 gestures based on count of fingers
- 6000 images of size 512×384 at 6 different distances to robot



[5]

RGB CNN cont.

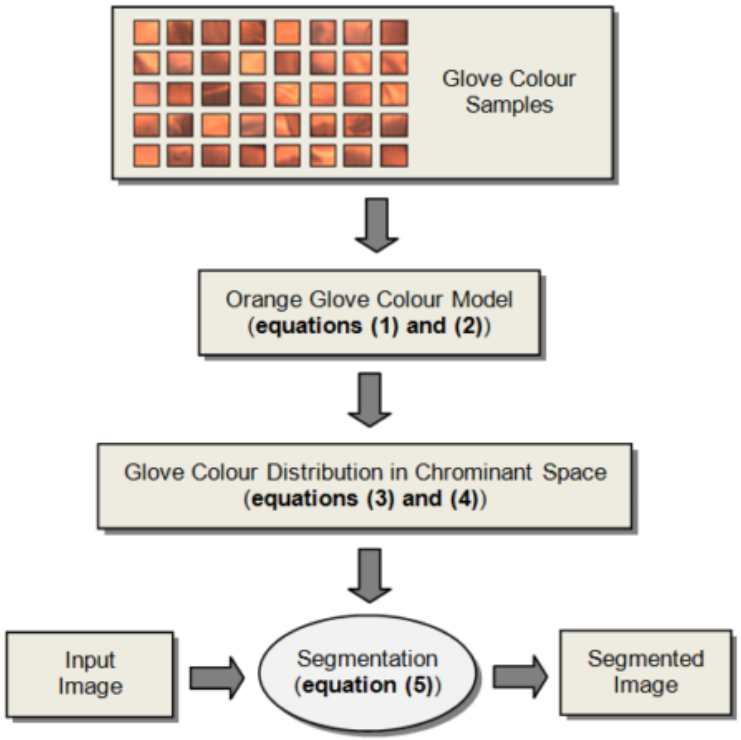
Preprocessing

- Transform RGB image to YCbCr (chrominance) color space
- Segment Gloves using a Single Gaussian Model:
  Model glove color using mean and covariance of chrominant color with bivariate Gaussian

[5]

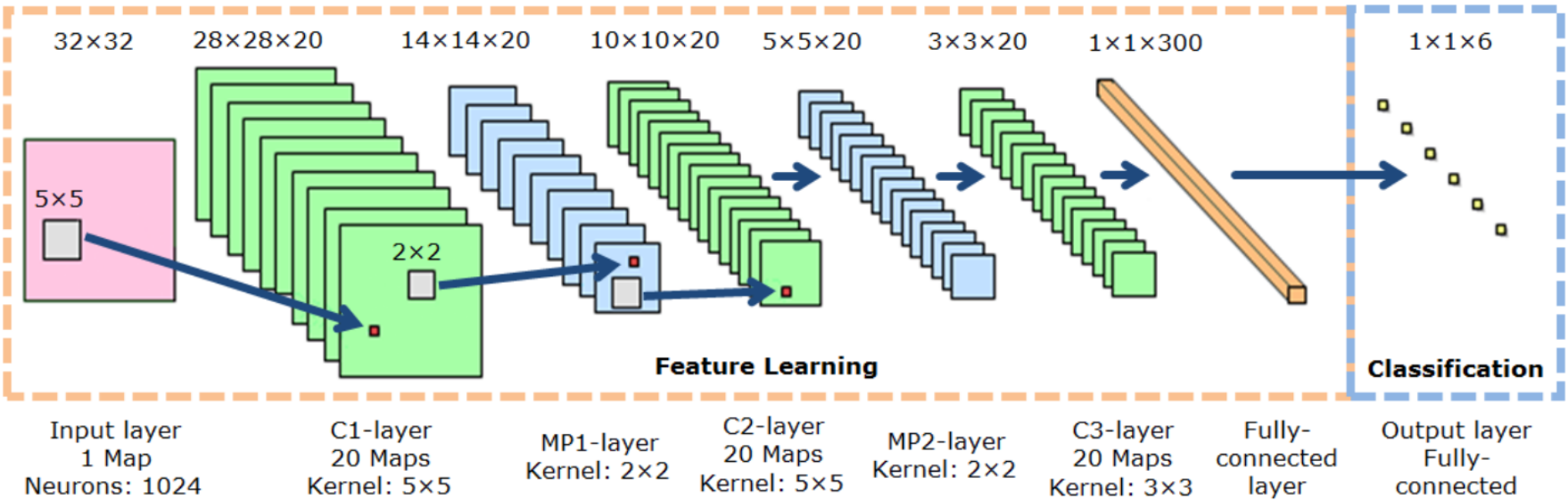
RGB CNN cont.

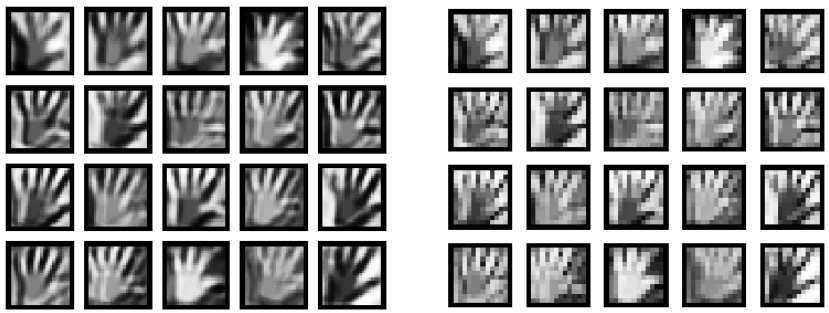
Colour Segmentation



[5]

MPCNN


32×32
28×28×20
14×14×20
10×10×20
5×5×20
3×3×20
1×1×300
1×1×6
5×5
2×2
Feature Learning
Classification
Input layer
1 Map
Neurons: 1024
C1-layer
20 Maps
Kernel: 5×5
MP1-layer
Kernel: 2×2
C2-layer
20 Maps
Kernel: 5×5
MP2-layer
Kernel: 2×2
C3-layer
20 Maps
Kernel: 3×3
Fully-
connected
layer
Output layer
Fully-
connected



[5]

RGB CNN cont.

Convolutional layer:

- M maps of equal size (M_x, My)
- Kernel of size (K_x, K_y) shifted over image
- Neurons in a map share weights but have different input fields

Max-pooling layer:

- Maximum activation over non-overlapping rectangular regions of size (K_x, K_y)
- Method of downsampling for more invariant features in higher layers

Classification layer:

- MLP for final classification
- One neuron per class with softmax activation function

RGB CNN cont.

Results

Feature Learner	Classifier	Reference	Error Rate
PHOG	SVM	[37]	27.04%
FFT	SVM	[39]	25.32%
Skeletonization	SVM	[35] [36]	21.55%
Hu Invariant Moments	SVM	[30] [31] [32]	20.34%
Shape Properties	SVM	[31] [34]	17.91%
Spatial Pyramid (BoW)	SVM	[33]	15.68%
Tiled CNN	NN	[40]	9.52%
Big and Deep MPCNN	NN	Proposed	3.23%

[5]

3D RGB-D CNN

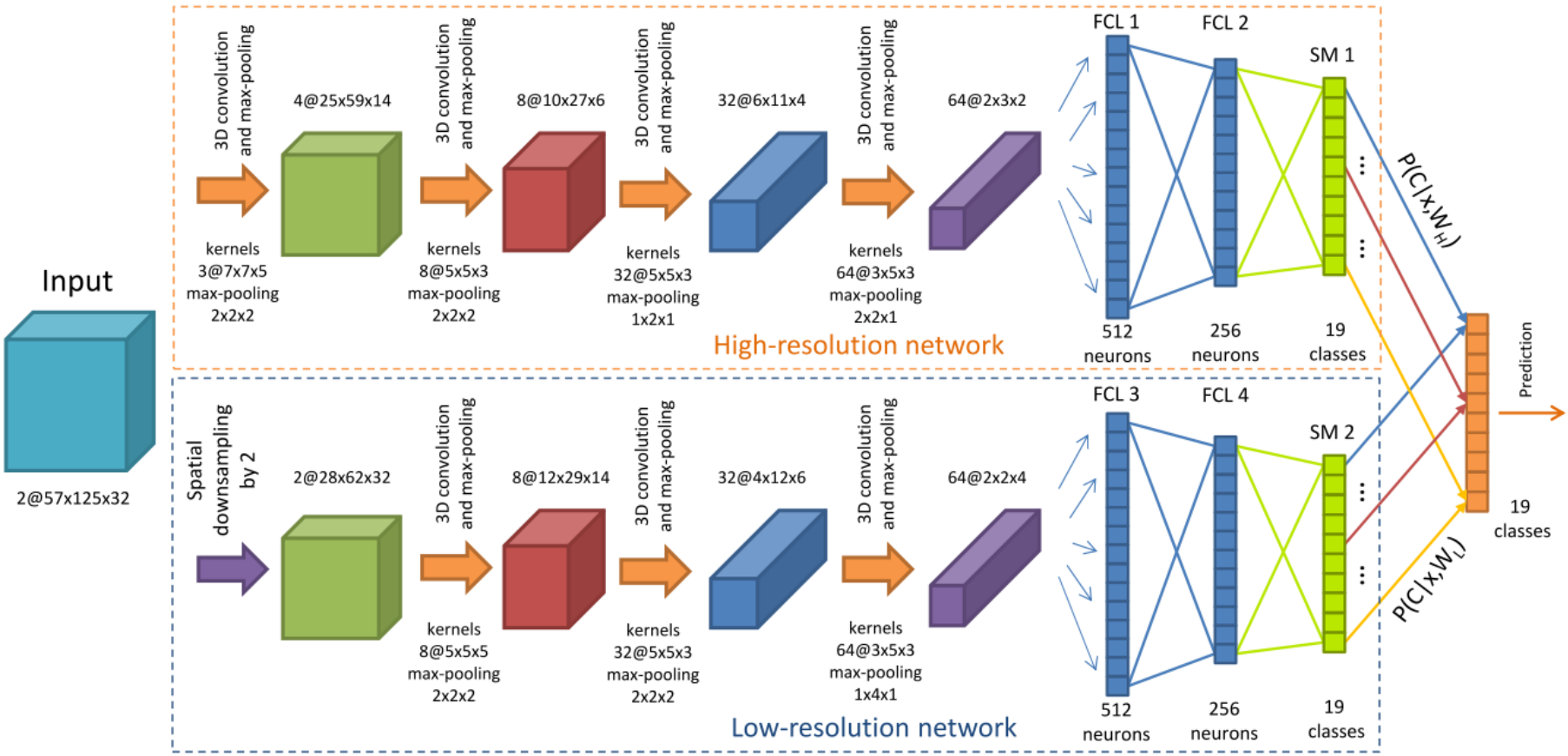
Molchanov et al. use interleaved RGB-D information as input for robust gesture recognition

- 3D CNN to process multiple frames at once
- Two subnetworks that give a fused prediction of gesture classification:
  High-resolution network (HRN) and low-resolution network (LRN)
- Spatio-temporal data augmentation to omit overfitting on the training data

3D RGB-D CNN

The VIVA challenge Dataset

- 885 intensity and depth video sequences recorded with Kinect camera
- 19 different gestures to be classified
- 8 subjects, varying lightning conditions


Input
2@57x125x32
Spatial downsampling by 2
2@28x62x32
3D convolution and max-pooling
kernels 3@7x7x5 max-pooling 2x2x2
4@25x59x14
kernels 8@5x5x3 max-pooling 2x2x2
8@10x27x6
kernels 32@5x5x3 max-pooling 1x2x1
32@6x11x4
kernels 64@3x5x3 max-pooling 2x2x1
64@2x3x2
High-resolution network
kernels 8@5x5x3 max-pooling 2x2x2
8@12x29x14
kernels 32@5x5x3 max-pooling 2x2x2
32@4x12x6
kernels 64@3x5x3 max-pooling 1x4x1
64@2x2x4
Low-resolution network
FCL 1
FCL 2
SM 1
512 neurons
256 neurons
19 classes
FCL 3
FCL 4
SM 2
512 neurons
256 neurons
19 classes
P(C|x,W_H)
P(C|x,W_L)
Prediction
19 classes

[4]

3D RGB-D CNN cont.

Only little preprocessing:

- Align temporal length of video sequences with nearest neighbour interpolation
- Downsample the original image data by a factor of 2
- Compute gradients from intensity channel and interleave image gradients with depth values

Classification:

- Combine class member-ship probabilities $P(C|x, W_H)$ and $P(C|x, W_L)$ from HRN and LRN:
  $P(C|x) = P(C|x, W_H) * P(C|x, W_H)$

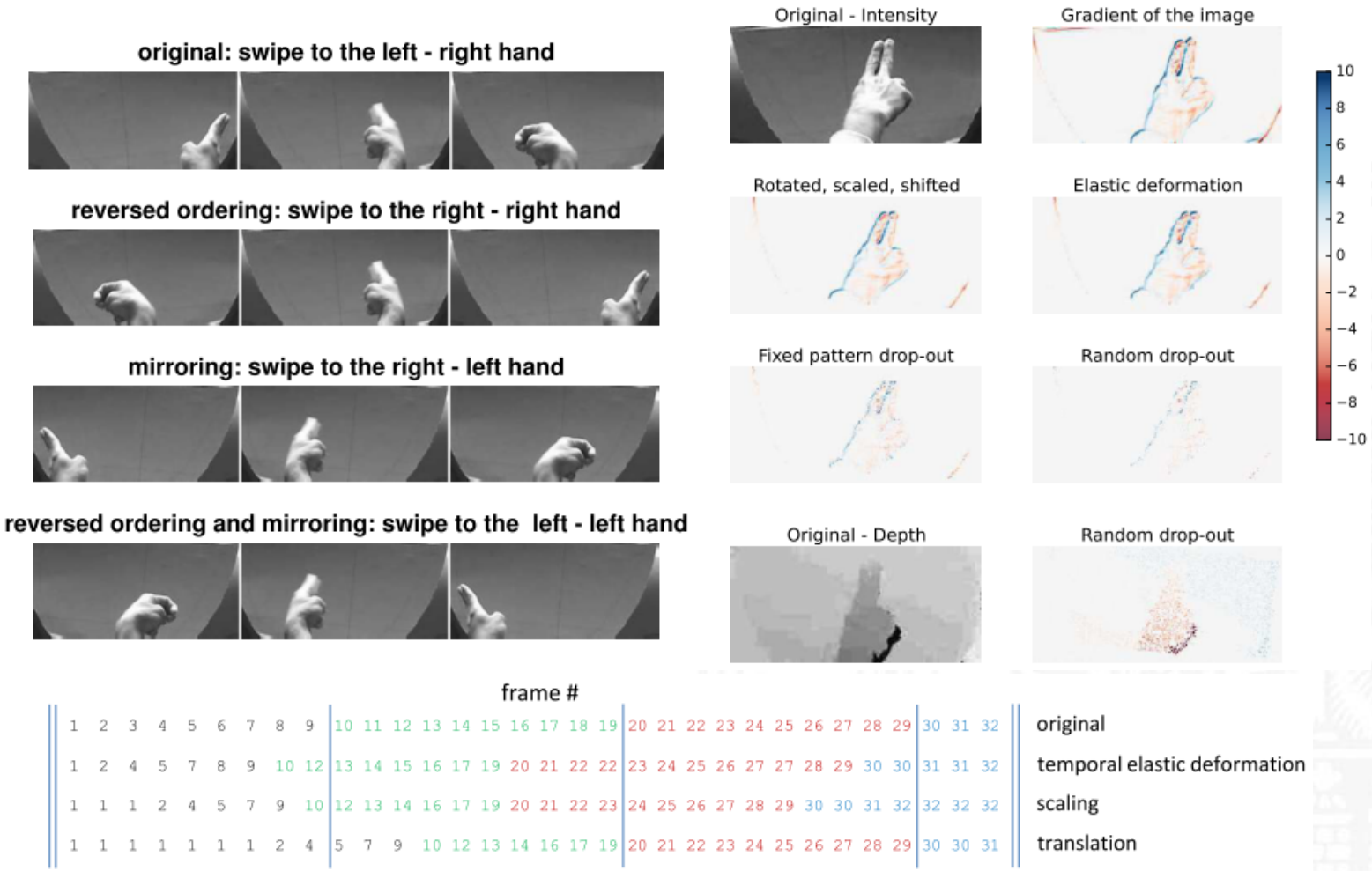
3D RGB-D CNN cont.

Training:

- Minimize cost function:

$$L(W, D) = -\frac{1}{|D|} \sum_{i=0}^{|D|} \log(P(C^{(i)}|x^{(i)}, W))$$

Data Augmentation



[4]

3D RGB-D CNN cont.

Results

Method	HOG	LRN	HRN	LRN+HRN
Mean	64.5%	74.4%	70.0%	**77.5%**
Std	16.9%	8.9%	7.8%	7.9%

LRN_{RGB}	LRN_D	LRN_{RGB}+LRN_D	LRN_{RGBD}
57.0%	65.0%	70.4%	74.4%

	None	Offline	Online	Both
Train	99.9%	99.8%	93.0%	91.1%
Test	48.3%	56.2%	59.1%	74.4%

[4]

Comparison RGB / 3D RGB-D CNN

RGB CNN (Nagi et al.)

+ Fast and efficient gesture recognition
+ Runs on a mobile robot
- Requires some preprocessing / hand segmentation (see also [3])
- Gloves are used to facilitate hand segmentation
- Only possible to recognize static gestures

Comparison RGB / 3D RGB-D CNN cont.

3D RGB-D CNN (Molchanov et al.)

+ No need for extensive hand segmentation and preprocessing
+ Robust classification, no over-fitting (data augmentation)
+ Can recognize gestures involving motion
- Computationally more expensive than 2D RGB CNN

Evaluation: CNNs for HGR

+ Less preprocessing than hand-tailored feature extraction
+ Robust to distortion
+ Good generalization: Not particularly sensible to individual user properties (hand appearance, style of gesture execution), lighting conditions, background

- While application is possible in real-time, training can take long
- Training computationally expensive (parallelization on a fast GPU)
- Sufficient training data required for good generalization and to avoid overfitting

References

[1] S Ji, M Yang, and K Yu.
3D convolutional neural networks for human action recognition.
IEEE Transactions on Pattern Analysis & Machine Intelligence, 35(1):221–231, 2013.

[2] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner.
Gradient-based learning applied to document recognition.
Proceedings of the IEEE, 86(11):2278–2323, 1998.

[3] Hsien-I Lin, Ming-Hsiang Hsu, and Wei-Kai Chen.
Human hand gesture recognition using a convolution neural network.
Automation Science and Engineering (CASE), 2014 IEEE International Conference on, pages 1038–1043, 2014.

[4] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, Jan Kautz, and Santa Clara.
Hand Gesture Recognition with 3D Convolutional Neural Networks.
pages 1–7, 2015.

[5] Jawad Nagi and Frederick Ducatelle.
Max-pooling convolutional neural networks for vision-based hand gesture recognition.
2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pages 342–347, 2011.