DEEP LEARNING FOR ACTIVITY RECOGNITION
(A BRIEF AND INCOMPLETE SURVEY)

GRAHAM TAYLOR

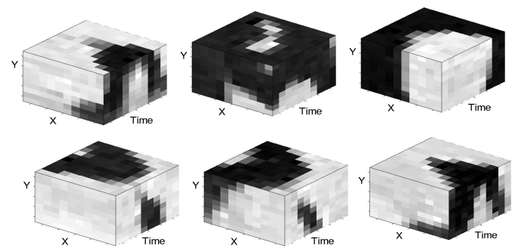
VISION, LEARNING AND GRAPHICS GROUP & MOVEMENT GROUP
COURANT INSTITUTE OF MATHEMATICAL SCIENCES
NEW YORK UNIVERSITY
NEW YORK, NY USA

Papers and software available at: http://www.cs.nyu.edu/~gwtaylor

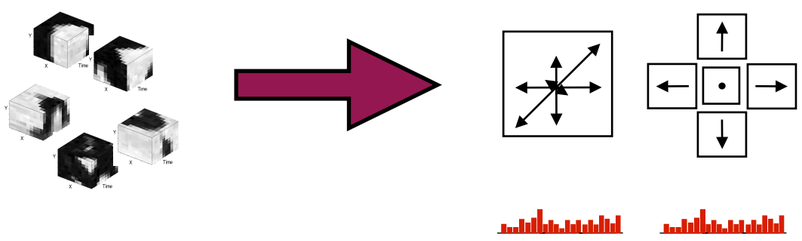
EXISTING PIPELINE FOR ACTIVITY RECOGNITION

Interest points

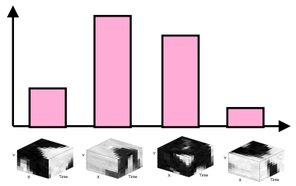
Collection of space-time patches



Cleverly engineered descriptors



Histogram of visual words

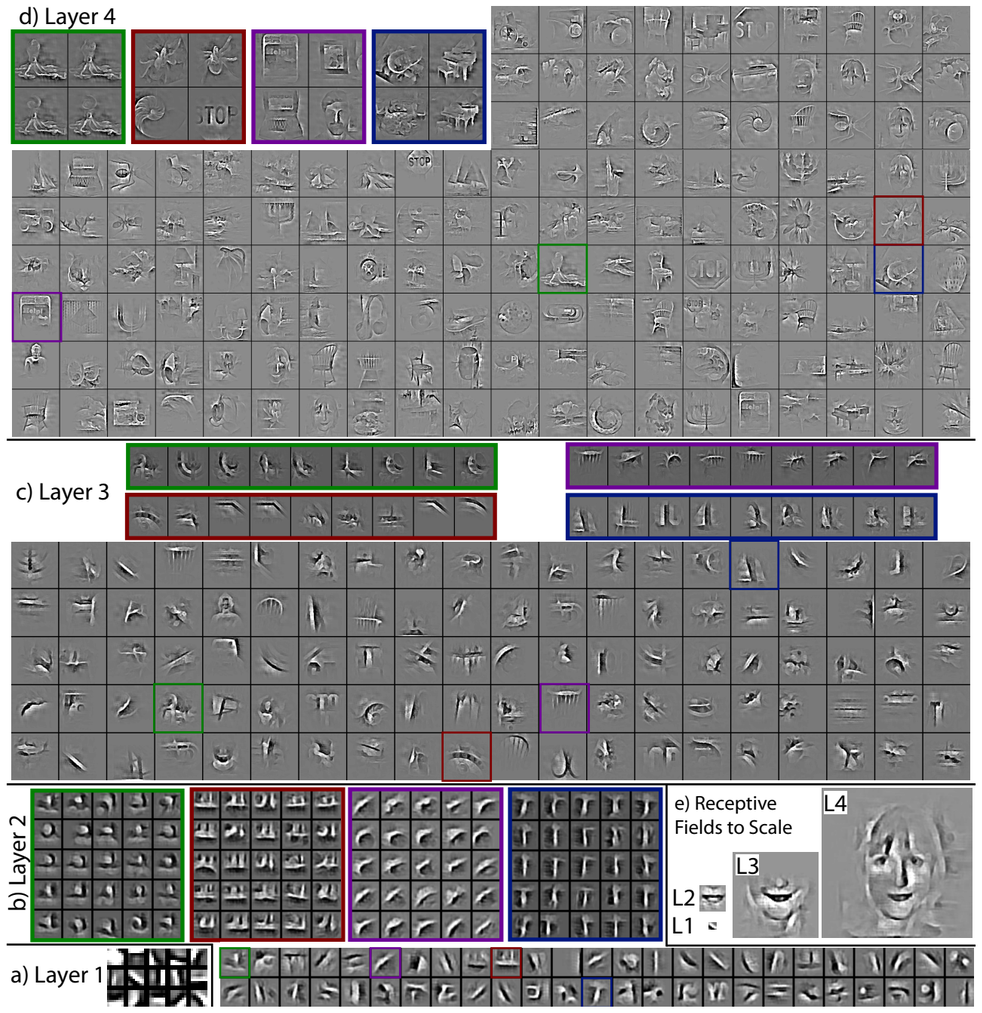


SVM classifier

(Images/videos from Ivan Laptev)

DEEP LEARNING

- Learning hierarchical data representations that are salient for high-level understanding
- Most often one layer at a time, building more abstract higher-level abstractions by composing lower-level representations
- Typically unsupervised
- Learned representations often used as input to classifiers



Deconvolutional Networks
(Zeiler, Taylor, and Fergus ICCV 2011)

MOTIVATIONS

- Representationally efficient (Bengio 2009)
- Produce hierarchical representations
  - Intuitive (humans organize their ideas hierarchically)
  - Permit non-local generalization
- Biologically motivated
  - brains use unsupervised learning
  - brains use distributed representations

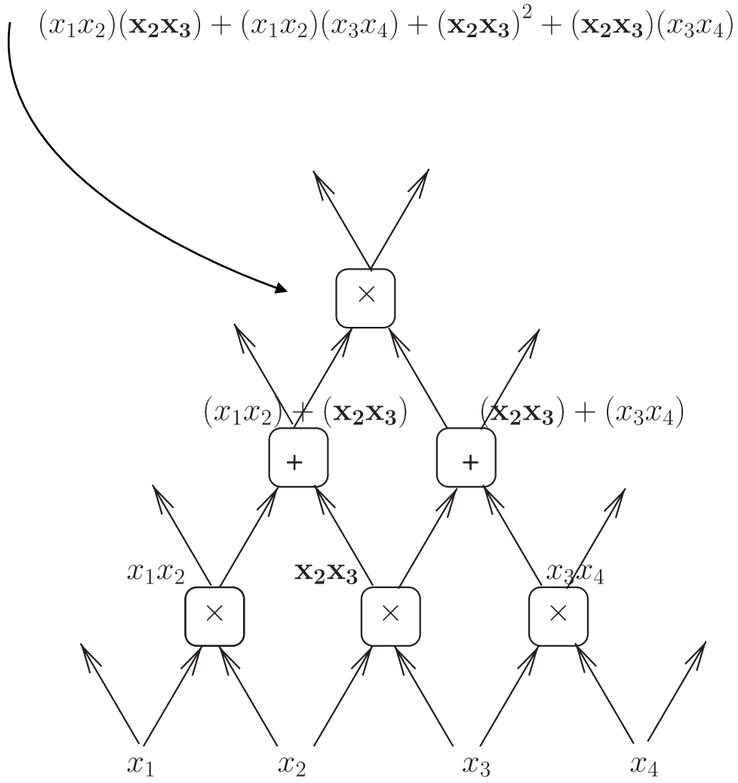


Image from Yoshua Bengio

POPULAR DEEP LEARNING ARCHITECTURES

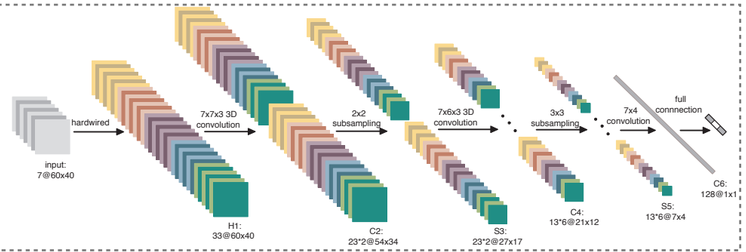
Name	Examples	Type
Deep Neural Networks	Rumelhart et al. 1986	S
Deep Belief Networks	Hinton et al. 2006, Lee et al. 2009, Norouzi et al. 2009	U*
Convolutional Networks	LeCun et al. 1998, Le et al. 2010	S
Stacked Denoising Autoencoders	Vincent et al. 2008	U*
Hierarchical Sparse Coding	Ranzato et al. 2007, Raina et al. 2007, Cadieu and Olshausen 2009, Yu et al. 2010	U
(De)Convolutional Sparse Coding	Kavacoglu et al. 2008, Zeiler et al. 2010, Chen et al. 2010, Masci et al. 2010	U
Deep Boltzmann Machines	Salakutdinov et al. 2009	U*

S - Supervised, U - Unsupervised, U* - Unsupervised but often fine-tuned discriminatively

OUTLINE

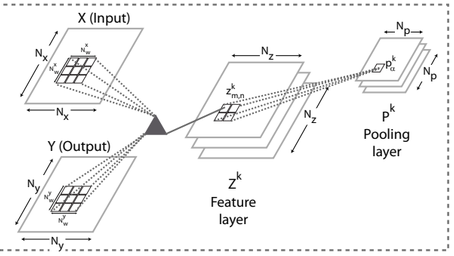
3D convolutional neural networks

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu (2010)



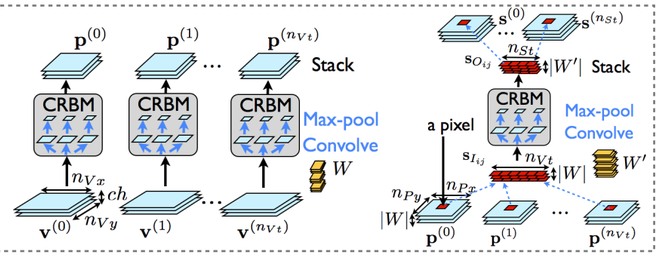
Convolutional gated restricted Boltzmann machines

Graham Taylor, Rob Fergus, Yann LeCun, and Chris Bregler (2010)



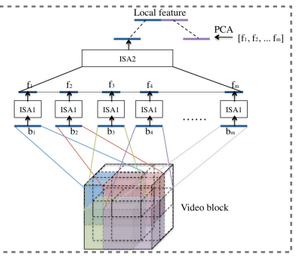
Space-time deep belief networks

Bo Chen, Jo-Anne Ting, Ben Marlin, and Nando de Freitas (2010)



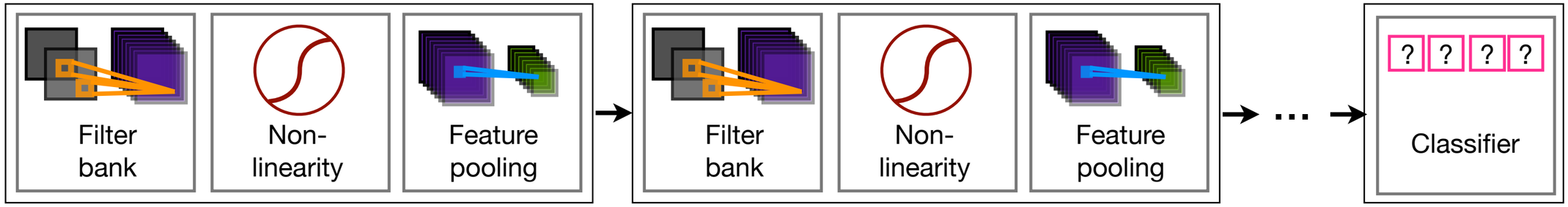
Stacked convolutional independent subspace analysis

Quoc Le Will Zou, Serena Yeung, and Andrew Ng (2011)



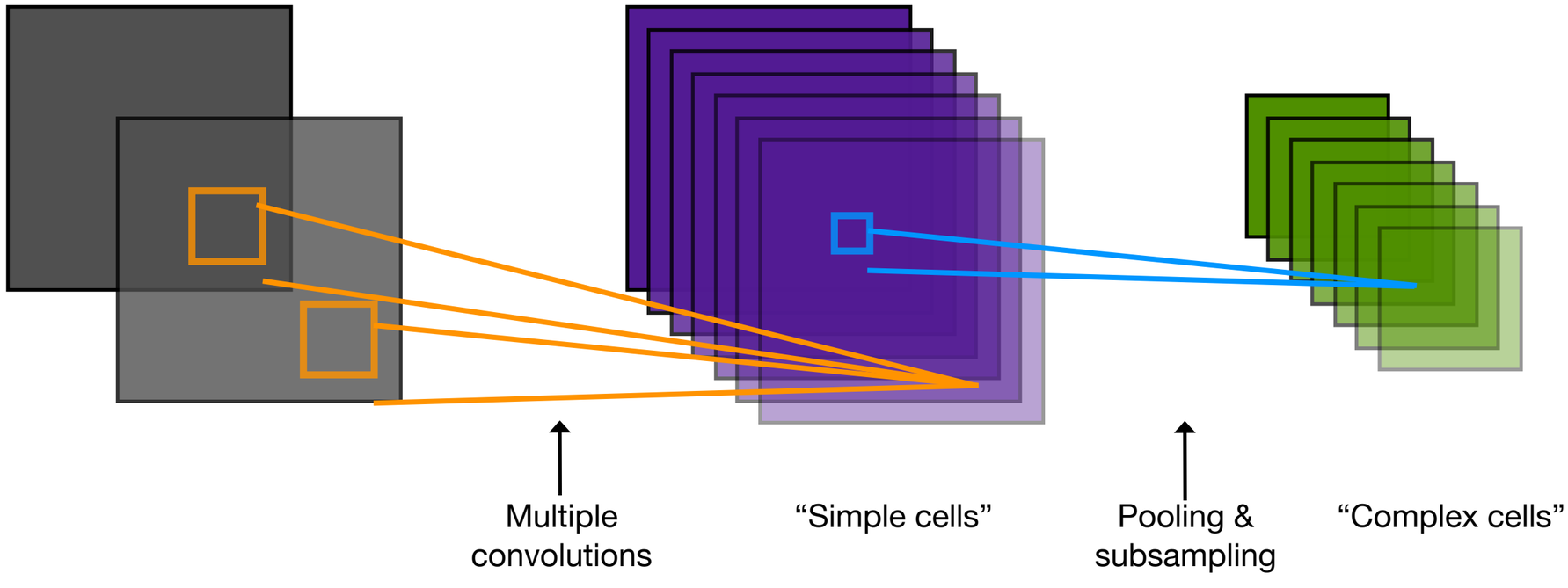
CONVOLUTIONAL NETWORKS

- Stacking multiple stages of Filter Bank + Non-Linearity + Pooling
- Shared with other approaches (SIFT, GIST, HOG)
- Main difference: Learn the filter banks at every layer



BIOLOGICALLY-INSPIRED

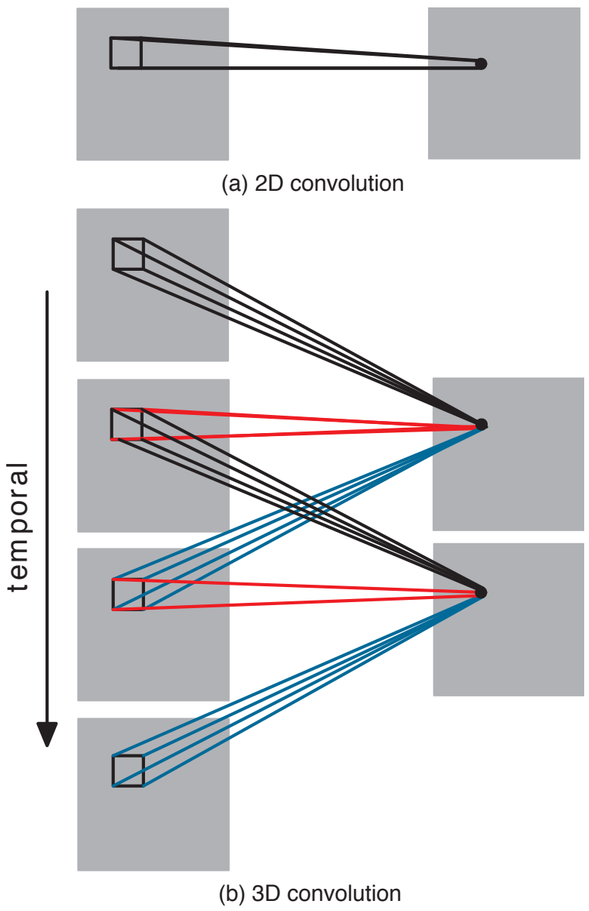
- Low-level features -> mid-level features -> high-level features -> categories
- Representations are increasingly abstract, global and invariant
- Inspired by Hubel & Wiesel (1962)
  - Simple cells detect local features
  - Complex cells pool the outputs of simple cells within a local neighborhood

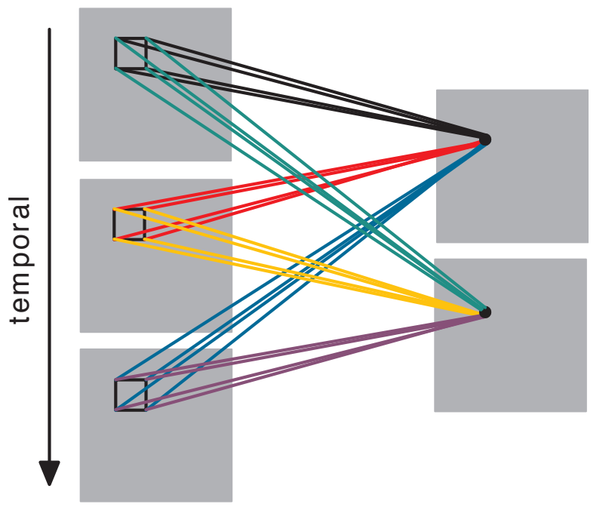


3D CONVNETS FOR ACTIVITY RECOGNITION

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu (ICML 2010)

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





Multiple convolutions applied to contiguous frames to extract multiple features

Images from Ji et al. 2010

3D CNN ARCHITECTURE

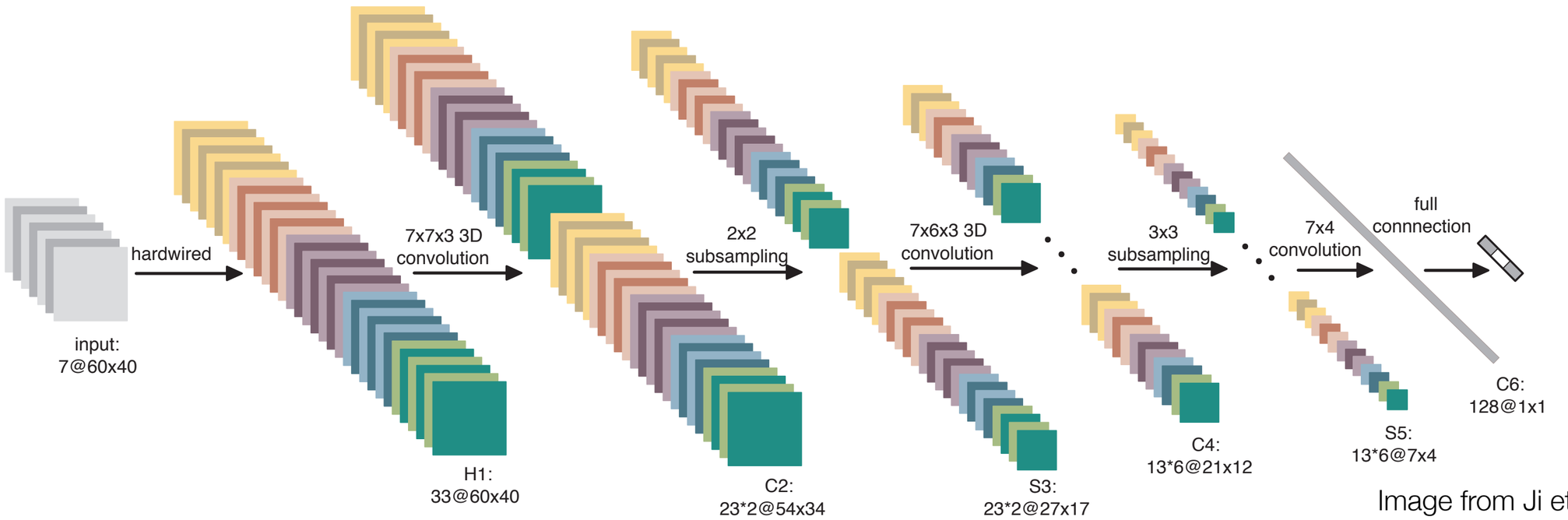


Image from Ji et al. 2010

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

Action units

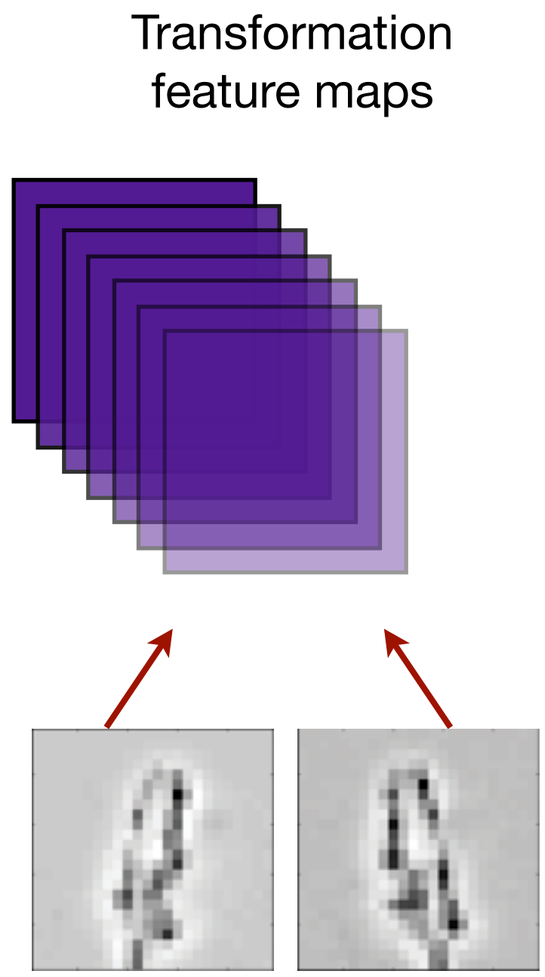
3D CONVNET: DISCUSSION

- Good performance on TRECVID surveillance data (*CellToEar, ObjectPut, Pointing*)
- Good performance on KTH actions (*box, handwave, handclap, jog, run, walk*)
- Still a fair amount of engineering: person detection (TRECVID), foreground extraction (KTH), hard-coded first layer

Image from Ji et al. 2010

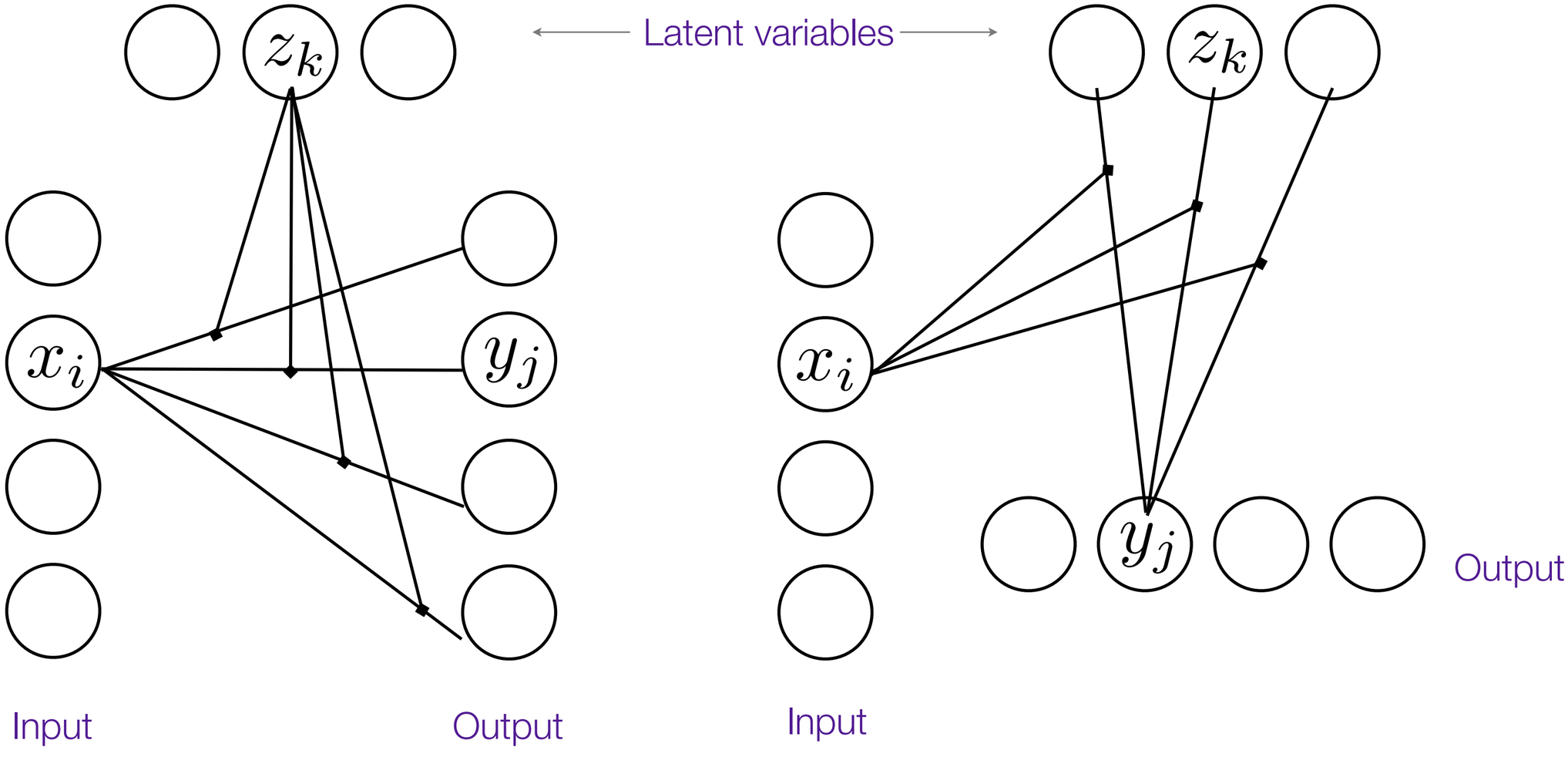
LEARNING FEATURES FOR VIDEO UNDERSTANDING

- Most work on unsupervised feature extraction has concentrated on *static images*
- We propose a model that extracts motion-sensitive features from *pairs of images*
- Existing attempts (e.g. Memisevic & Hinton 2007, Cadieu & Olshausen 2009) ignore the *pictorial* structure of the input
- Thus limited to modeling small image patches



GATED RESTRICTED BOLTZMANN MACHINES

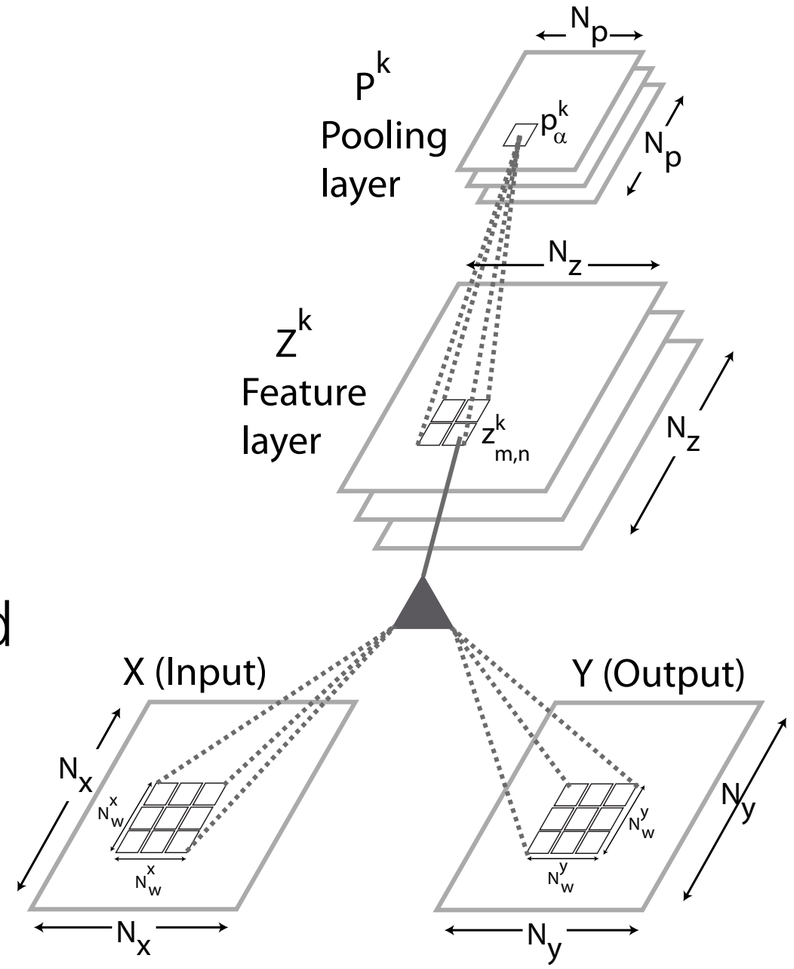
- Two views (Memisevic and Hinton 2007):



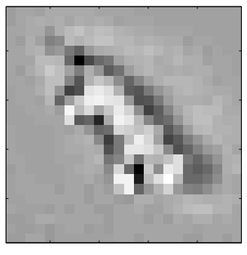
CONVOLUTIONAL GRBM

Graham Taylor, Rob Fergus, Yann LeCun, and Chris Bregler (ECCV 2010)

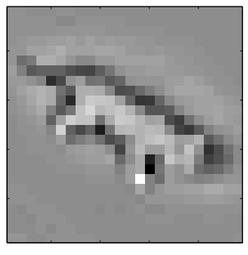
- Like the GRBM, captures third-order interactions
- Shares weights at all locations in an image
- As in a standard RBM, exact inference is efficient
- Inference and reconstruction are performed through convolution operations



VISUALIZING FEATURES THROUGH ANALOGY



Input



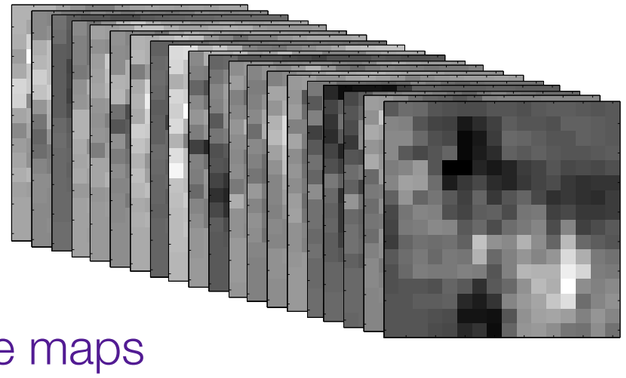
Output

VISUALIZING FEATURES THROUGH ANALOGY

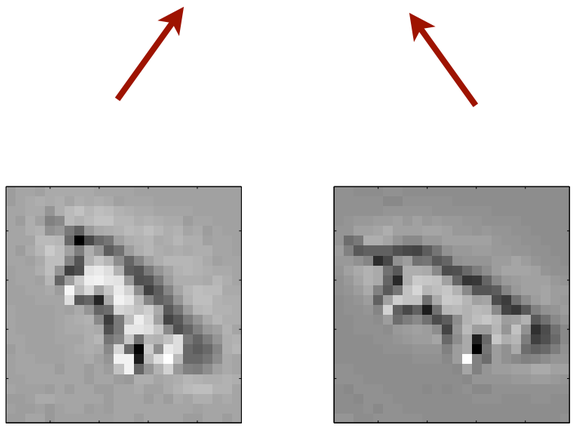
Feature maps

Input Output

VISUALIZING FEATURES THROUGH ANALOGY



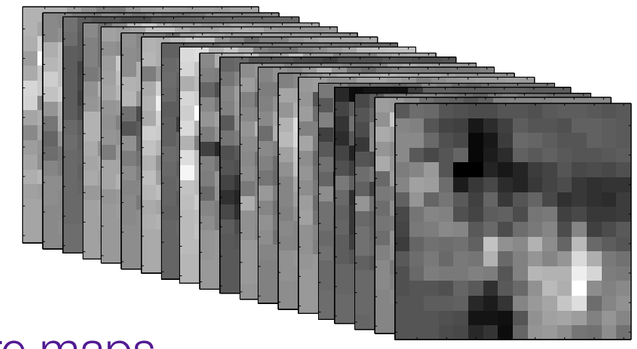
Feature maps



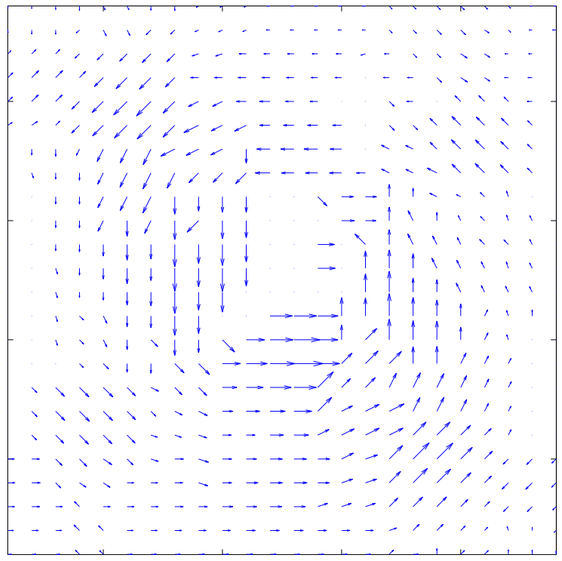
Input

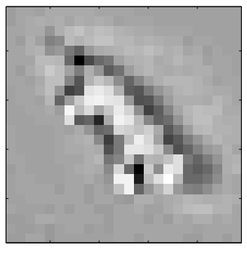
Output

VISUALIZING FEATURES THROUGH ANALOGY

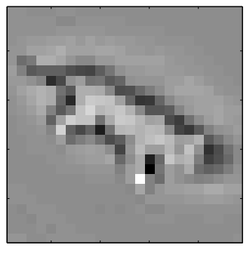


Feature maps

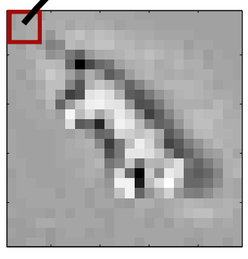




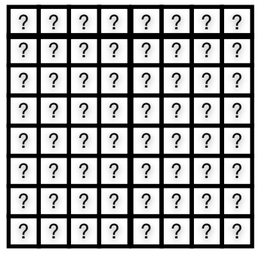
Input



Output

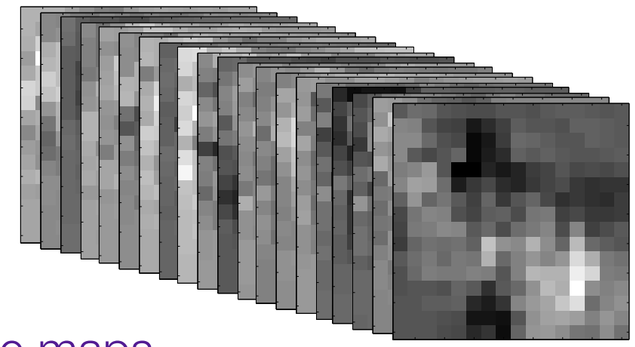


Input

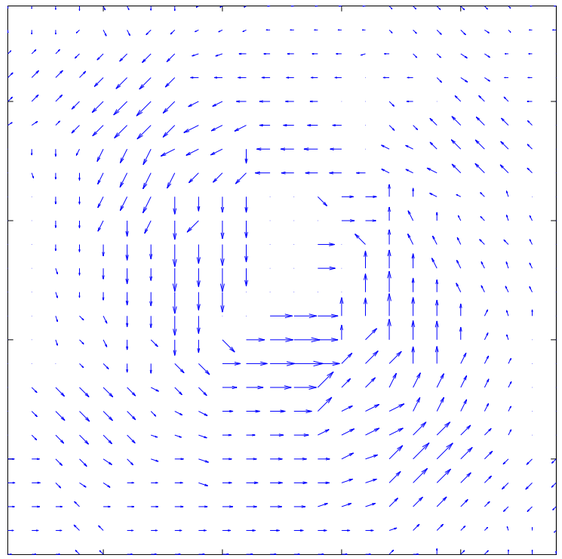


Output

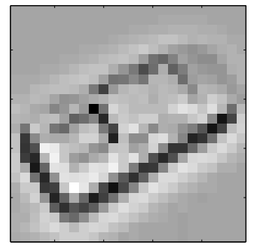
VISUALIZING FEATURES THROUGH ANALOGY

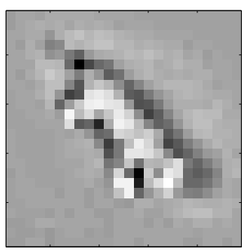


Feature maps

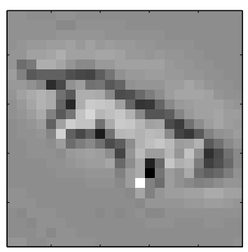


Ground truth

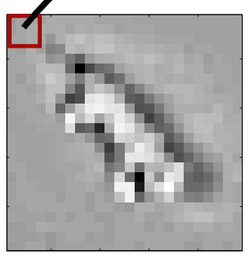




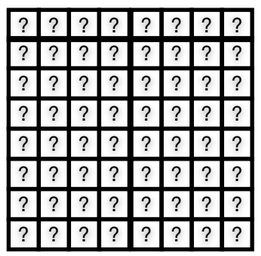
Input



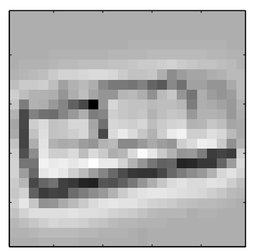
Output



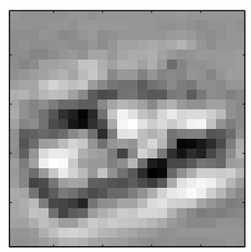
Input



Output



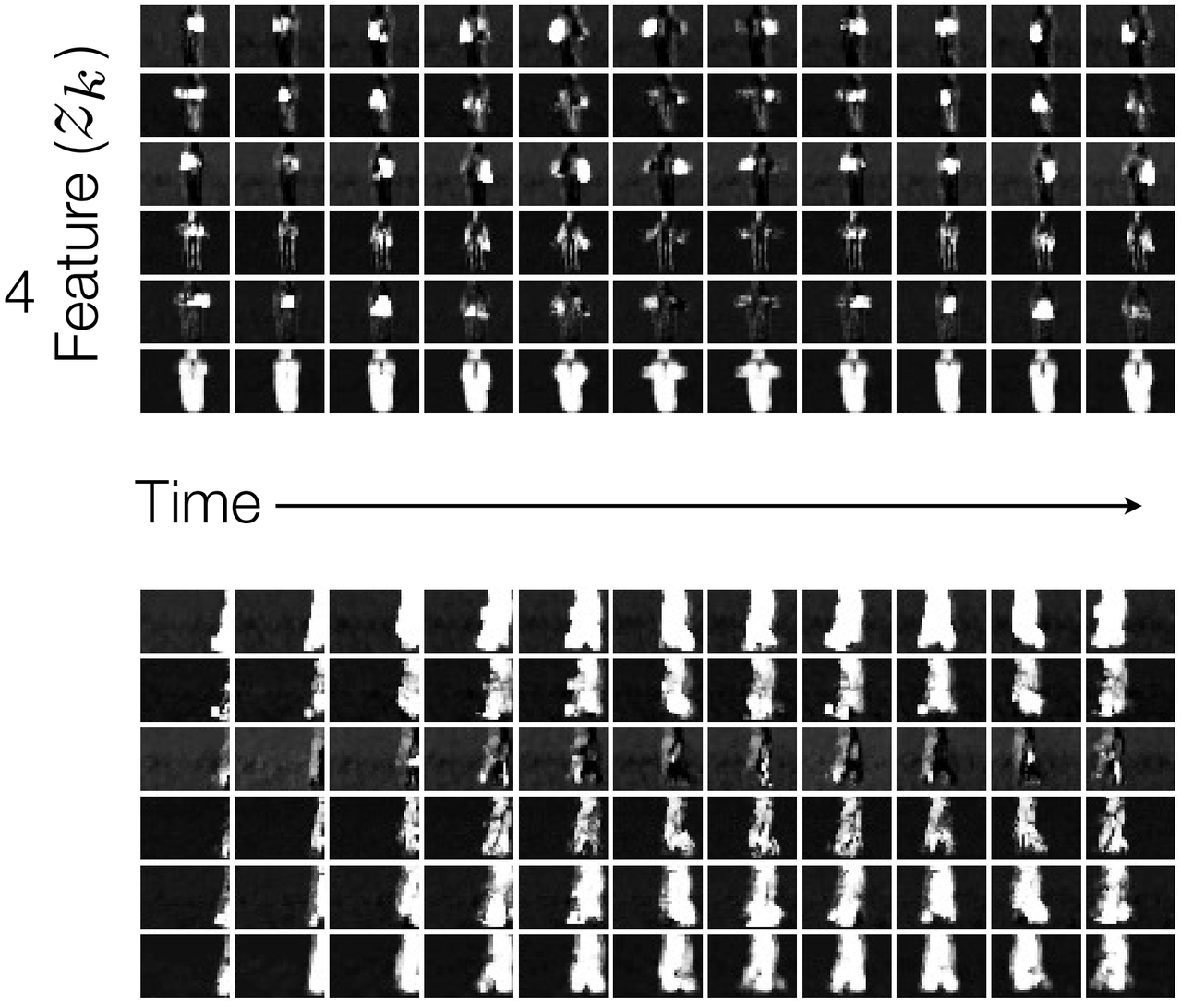
Novel input



Transformation (model)

HUMAN ACTIVITY: KTH ACTIONS DATASET

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



Hand clapping (above); Walking (below)

ACTIVITY RECOGNITION: KTH

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	**90.0**
HOG+KM+SVM	79.0	3D convnet+3D convnet+logistic reg.	79.4
HOF+KM+SVM	88.0	3D convnet+3D convnet+MLP	79.5

- Compared to methods that do not use explicit interest point detection
- State of the art: 92.1% (Laptev et al. 2008) 93.9% (Le et al. 2011)
- Other reported result on 3D convnets uses a different evaluation scheme

ACTIVITY RECOGNITION: HOLLYWOOD 2

- 12 classes of human action extracted from 69 movies (20 hours)
- Much more realistic and challenging than KTH (changing scenes, zoom, etc.)
- Performance is evaluated by mean average precision over classes

Method	Average Prec.
Prior Art (Wang et al. survey 2009):	
HOG3D+KM+SVM	45.3
HOG/HOF+KM+SVM	**47.4**
HOG+KM+SVM	39.4
HOF+KM+SVM	45.5
Our method:	
GRBM+SC+SVM	**46.8**

SPACE-TIME DEEP BELIEF NETWORKS

Bo Chen, Jo-Anne Ting, Ben Marlin, and Nando de Freitas (NIPS Deep Learning Workshop 2010)

- Two previous approaches we saw used <u>discriminative</u> learning
- We now look at a <u>generative</u> method, opening up more applications
  - e.g. in-painting, denoising
- Another key aspect of this work is demonstrated learned <u>invariance</u>
- Basic module: Convolutional Restricted Boltzmann Machine (Lee et al. 2009)

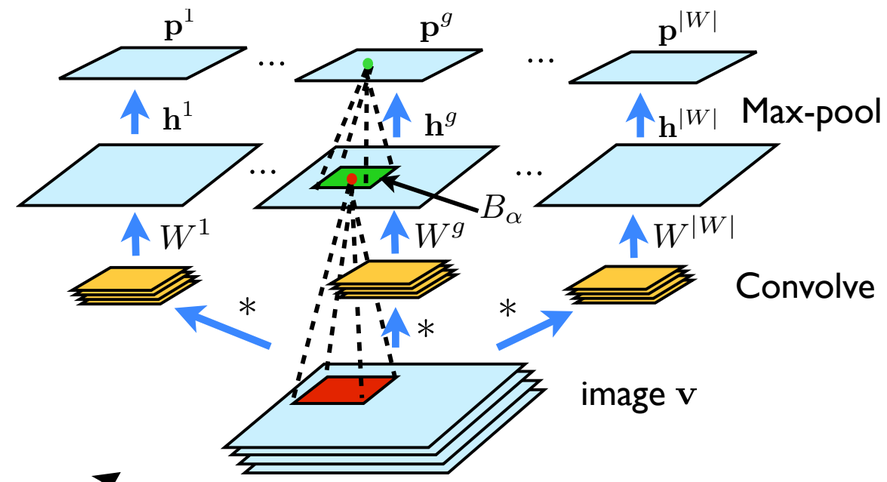
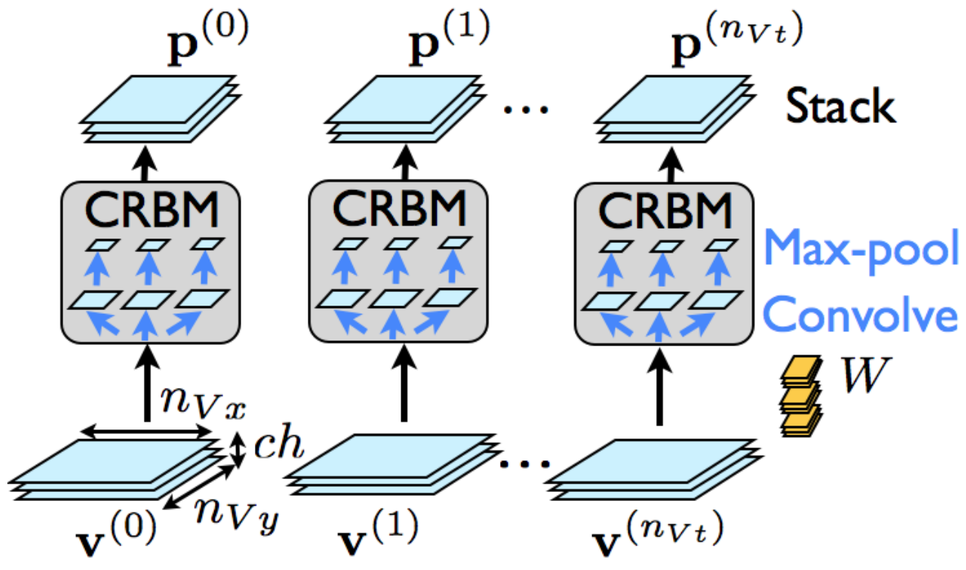


Image from Chen al. 2010

ST-DBN

- Key idea: alternate layers of spatial and temporal Convolutional RBMs
- Weight sharing across all CRBMs in a layer
- Highly overcomplete: use sparsity on activations of max-pooling units

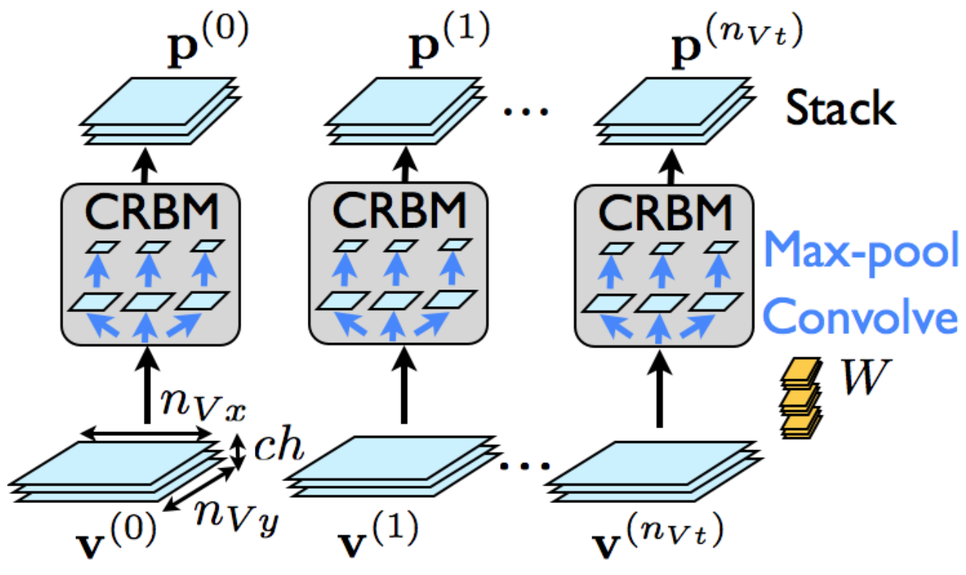


Spatial pooling layer

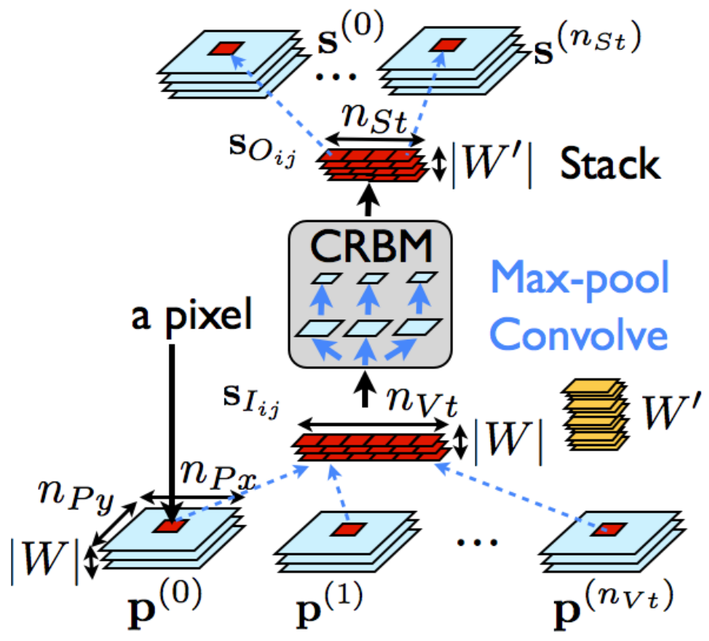
Images from Chen al. 2010

ST-DBN

- Key idea: alternate layers of spatial and temporal Convolutional RBMs
- Weight sharing across all CRBMs in a layer
- Highly overcomplete: use sparsity on activations of max-pooling units



Spatial pooling layer

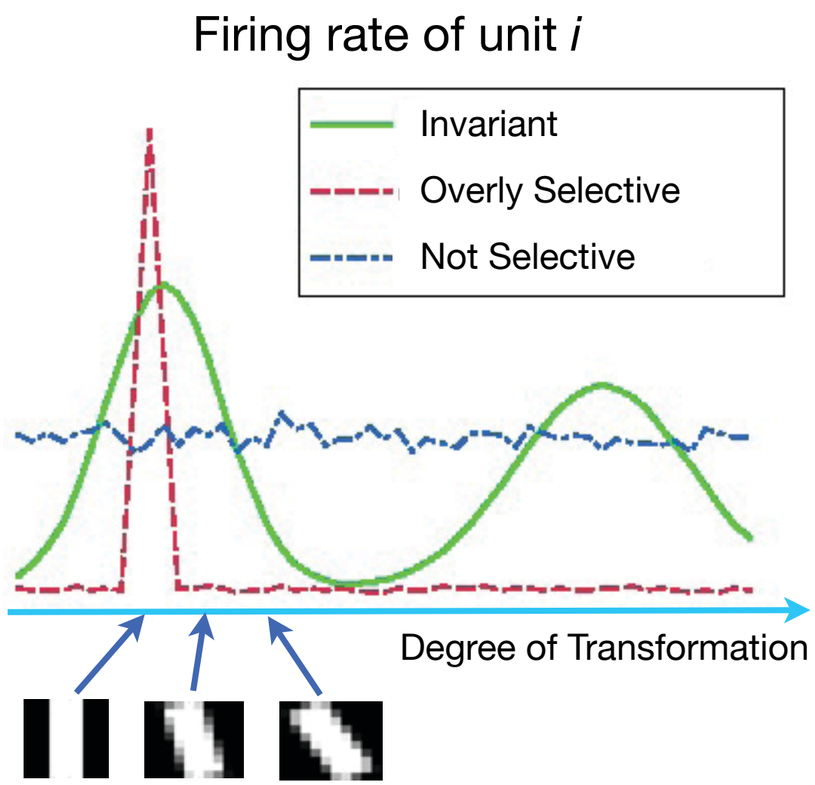


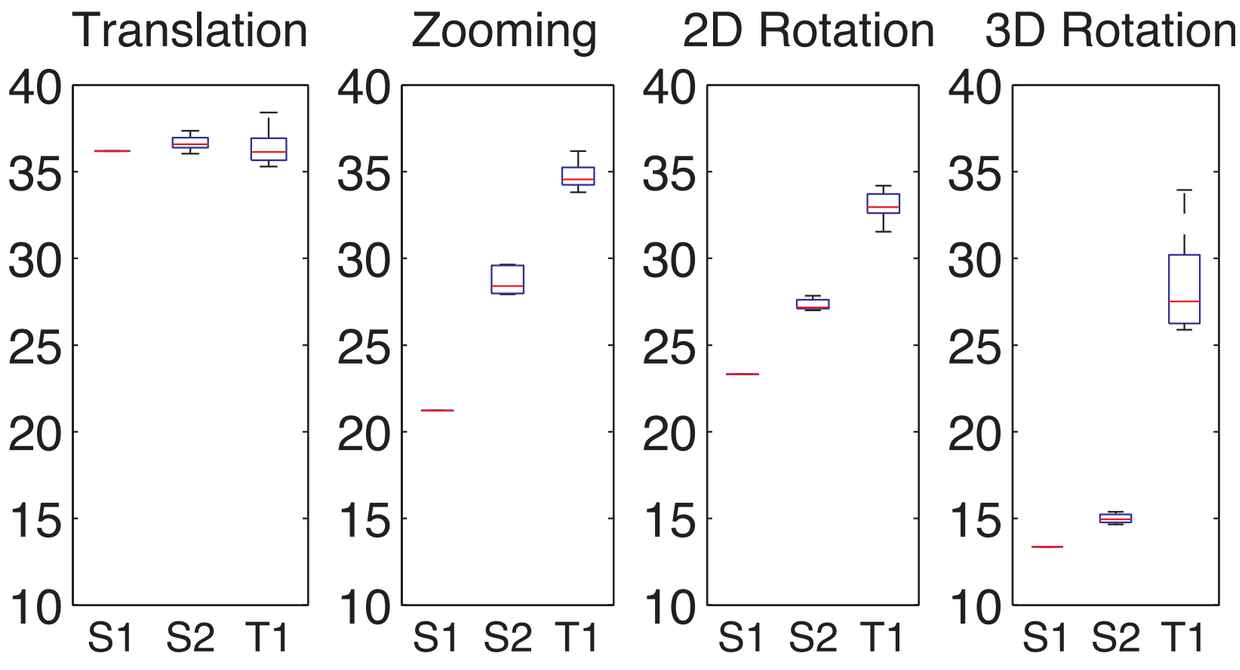
Temporal pooling layer

Images from Chen al. 2010

MEASURING INVARIANCE

- Measure invariance at each layer for various transformations of the input
- Use measure proposed by Goodfellow et al. (2009)



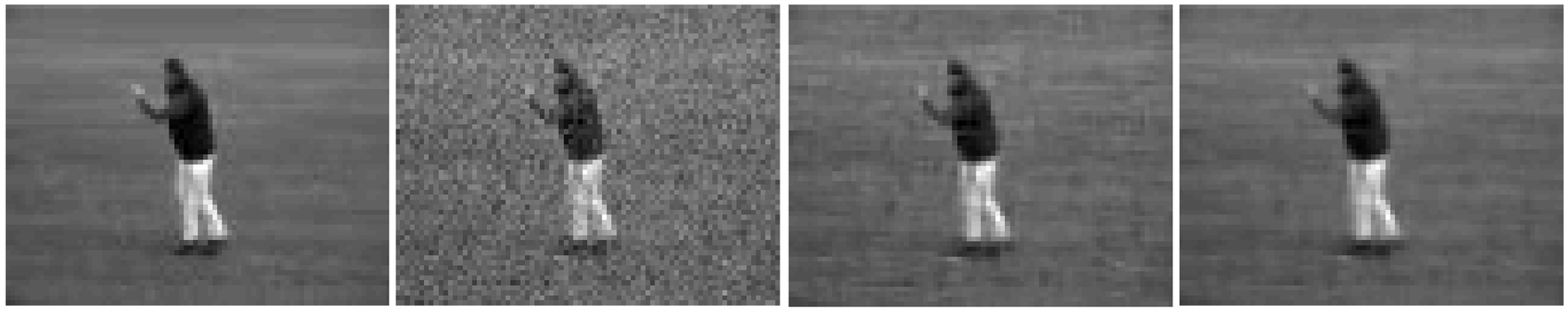


Invariance scores computed for Spatial Pooling Layer 1 (S1), Spatial Pooling Layer 2 (S2) and Temporal Pooling Layer 1 (T1).
Higher is better.

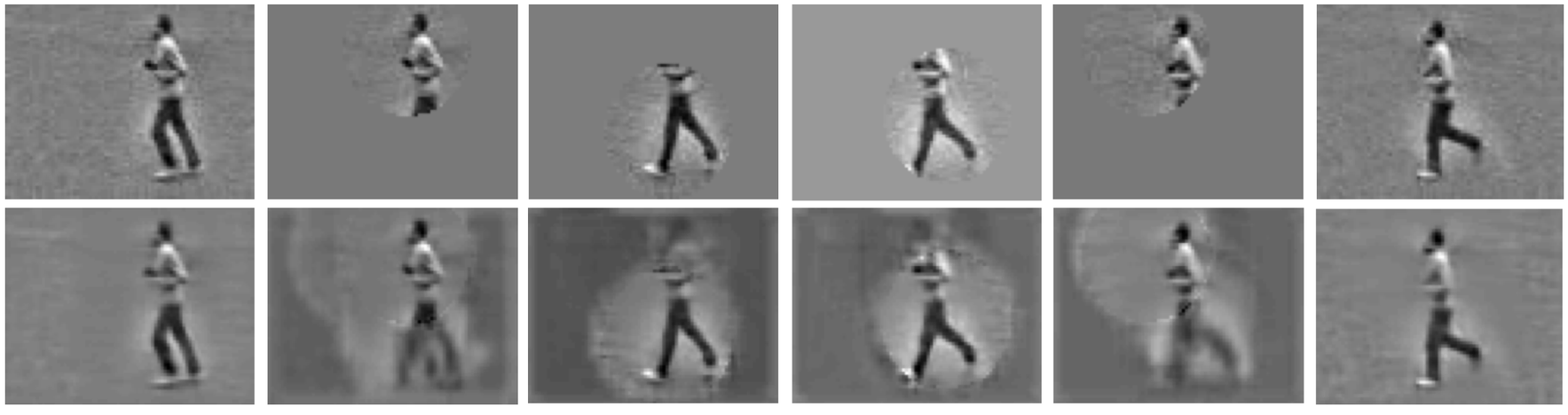
Images from Chen al. 2010

DENOISING AND RECONSTRUCTION

•Operations not possible with a discriminative approach



Test frame | Corrupted test frame | Reconstruction: 1 layer ST-DBN | Reconstruction: 2 layer ST-DBN



Observed gazes

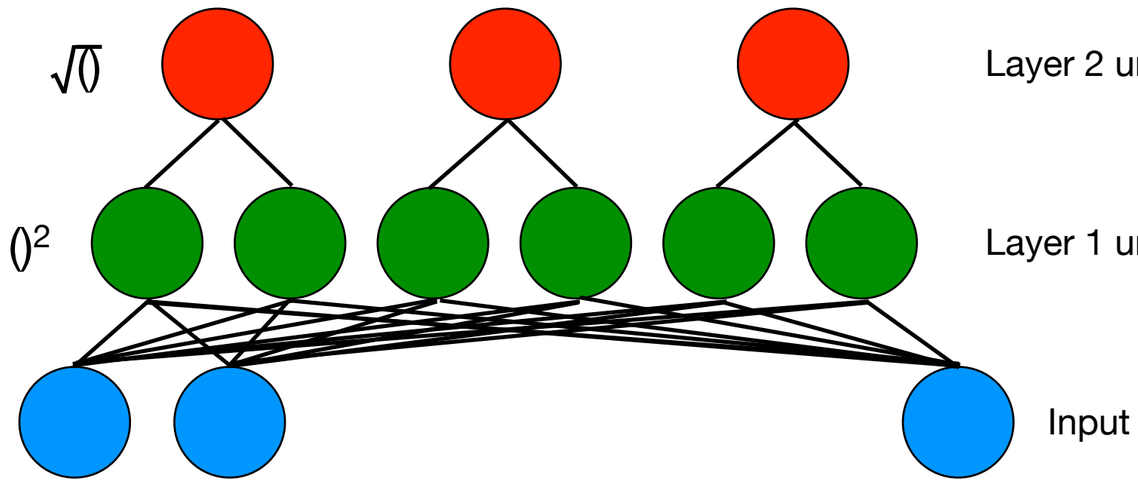
Reconstructions

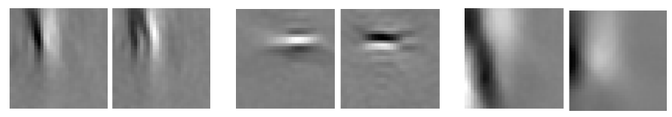
Images from Chen al. 2010

STACKED CONVOLUTIONAL INDEPENDENT SUBSPACE ANALYSIS (ISA)

Quoc Le Will Zou, Serena Yeung, and Andrew Ng (CVPR 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 from Le et al. 2010

SCALING UP: CONVOLUTION AND STACKING

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

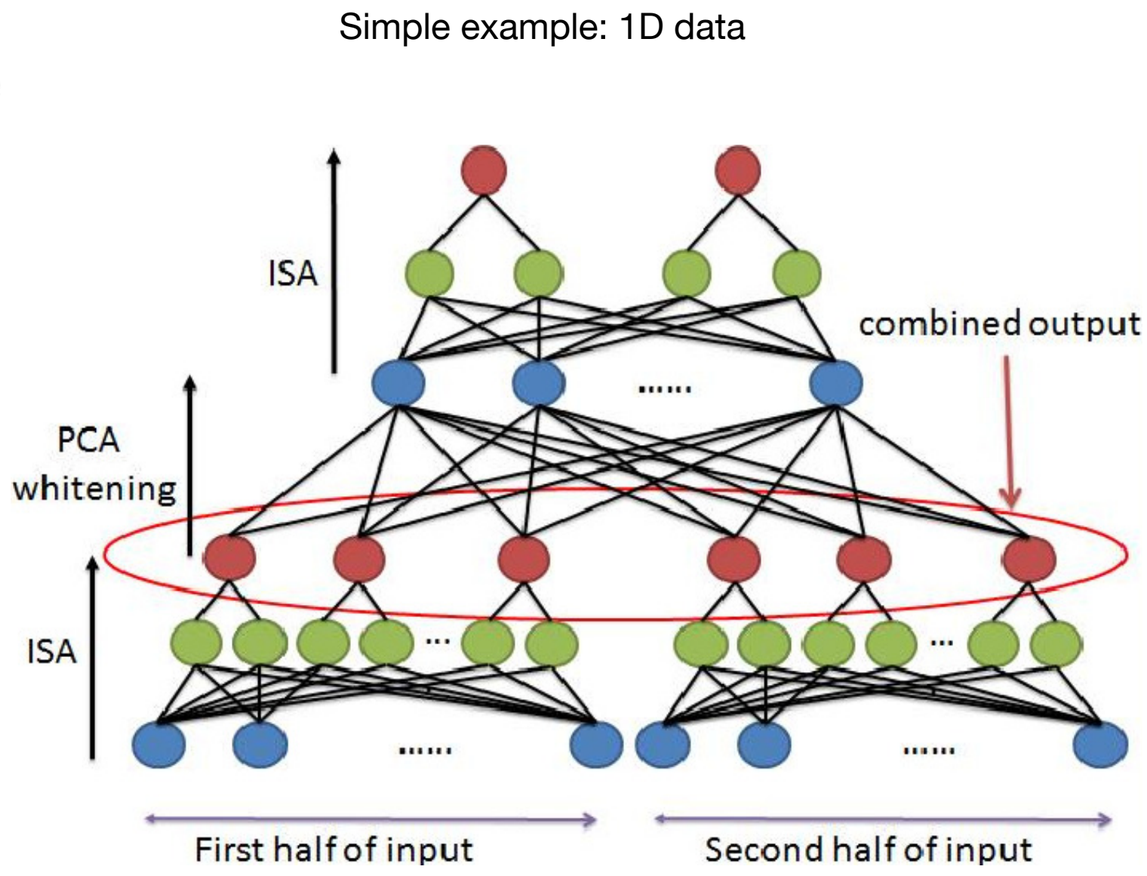
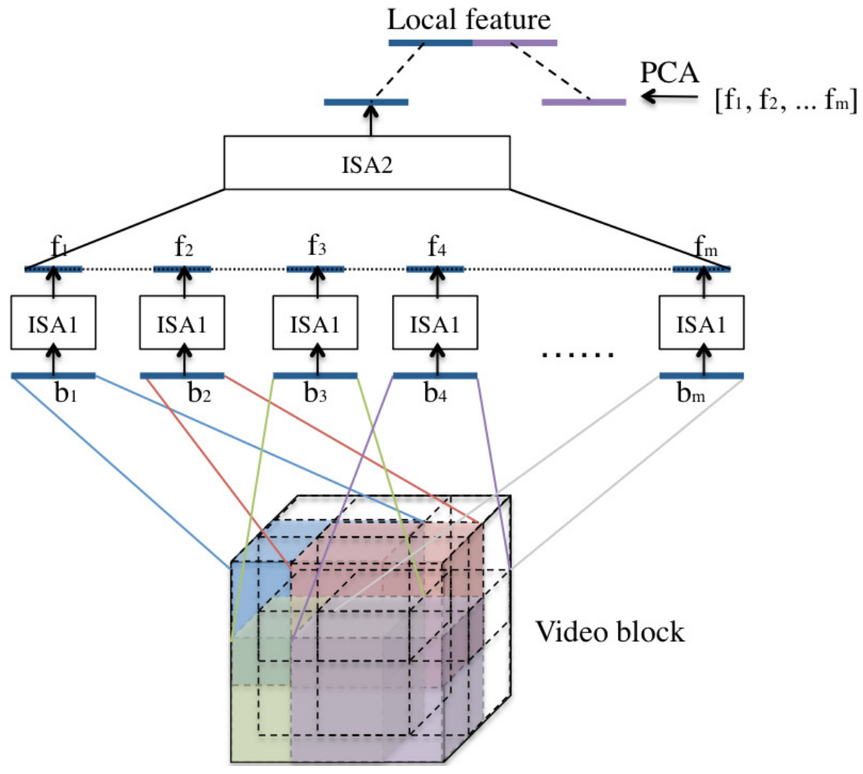


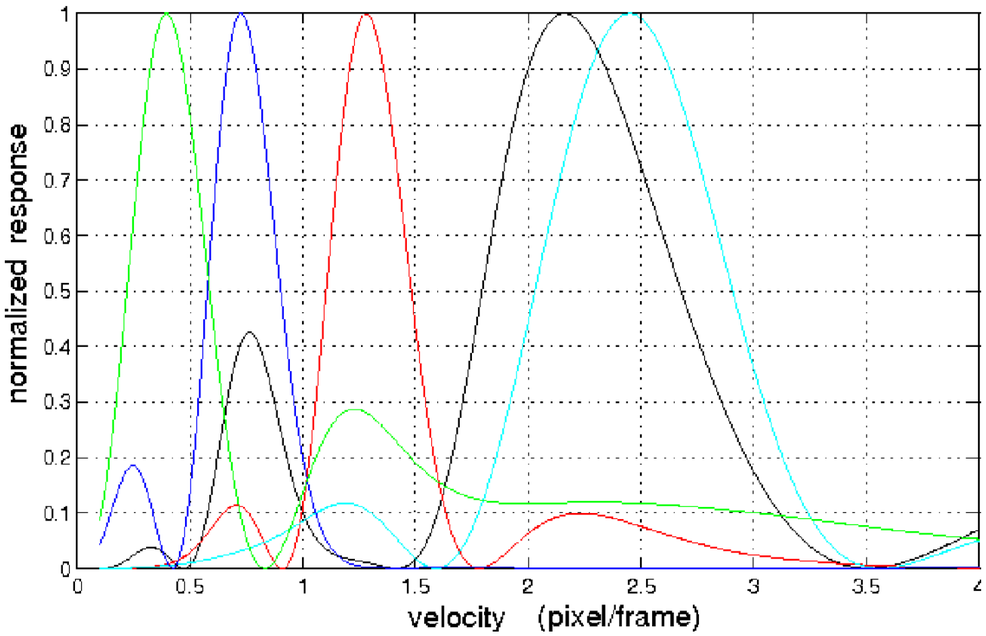
Image from Le et al. 2010

LEARNING SPATIO-TEMPORAL FEATURES

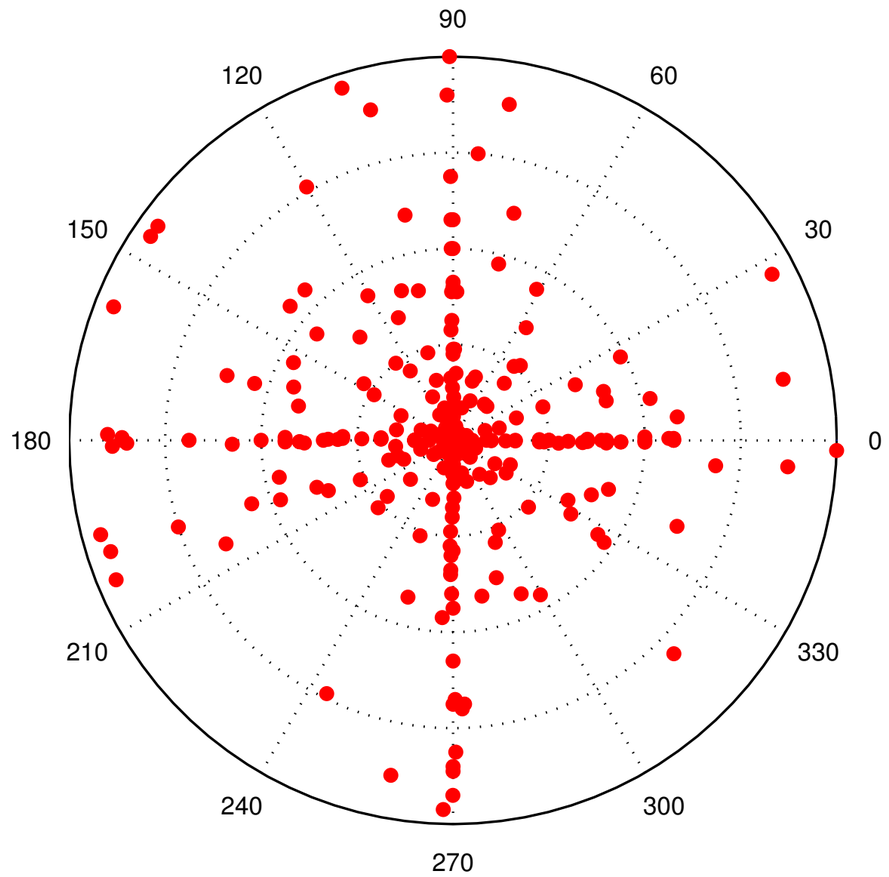
- Inputs to the network are blocks of video
- Each block is vectorized and processed by ISA
- Features from Layer 1 and Layer 2 are combined prior to classification



VELOCITY AND ORIENTATION SELECTIVITY



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

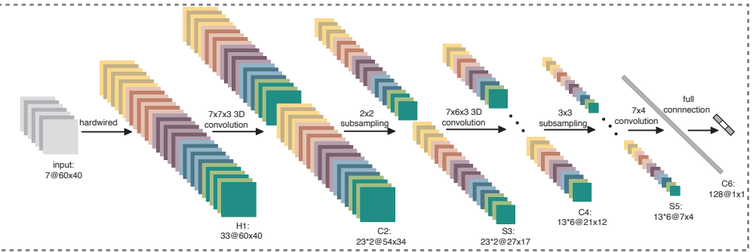


Edge velocities (radius) and orientations (angle) to which filters give maximum response
Outermost velocity: 4 pixels per frame

SUMMARY

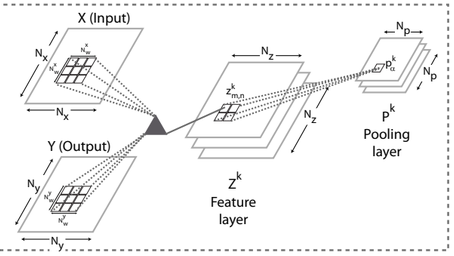
3D convolutional neural networks

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu (2010)



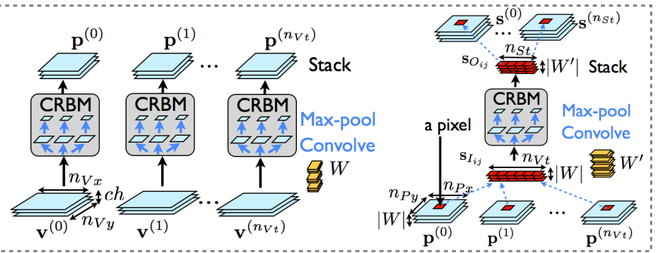
Convolutional gated restricted Boltzmann machines

Graham Taylor, Rob Fergus, Yann LeCun, and Chris Bregler (2010)



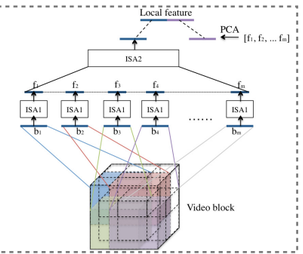
Space-time deep belief networks

Bo Chen, Jo-Anne Ting, Ben Marlin, and Nando de Freitas (2010)



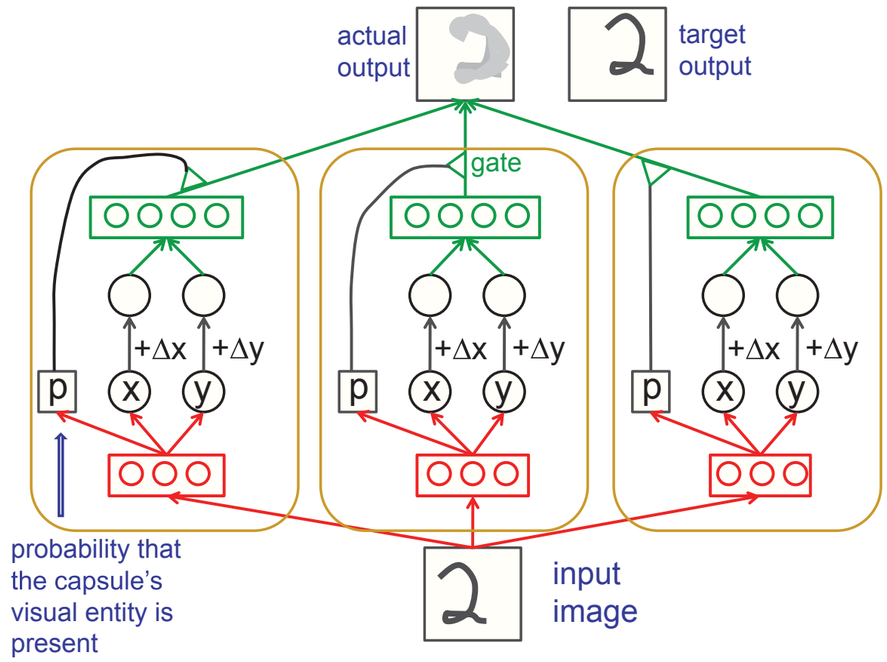
Stacked convolutional independent subspace analysis

Quoc Le Will Zou, Serena Yeung, and Andrew Ng (2011)



CONCLUSION

- Deep learning methods have already shown promise in the domain of activity recognition
- To this point, they are still neck-and-neck with more engineered systems
- Homogeneous network built by simple, trainable modules
- Future improvements in activity recognition will be driven by efficient and robust learning algorithms that build hierarchical representations (almost) entirely unsupervised
- Are we done with learning invariant representations?



Transforming Autoencoder
(Hinton, Krizhevsky, and Wang 2011)
Image from Geoff Hinton