Deep Learning of Invariant Spatiotemporal Features from Video

Bo Chen, Jo-Anne Ting, Ben Marlin, Nando de Freitas

University of British Columbia

Introduction

- Focus: Unsupervised feature extraction from video for low and high-level vision tasks
- Desirable Properties:
  - Invariant features (for selectivity, robustness to input transformations)
  - Distributed representations (for memory & generalization efficiency)
  - Generative properties for inference (de-noising, reconstruction, prediction, analogy-making)
  - Good discriminative performance

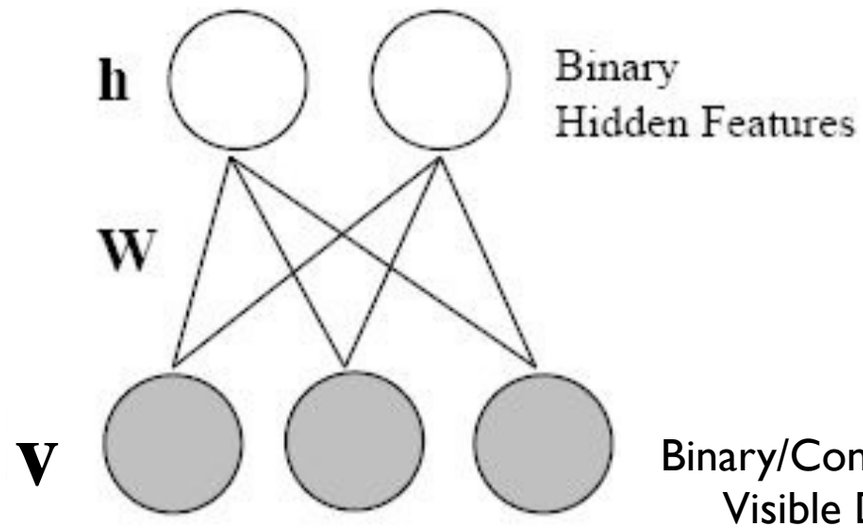
Feature extraction from video

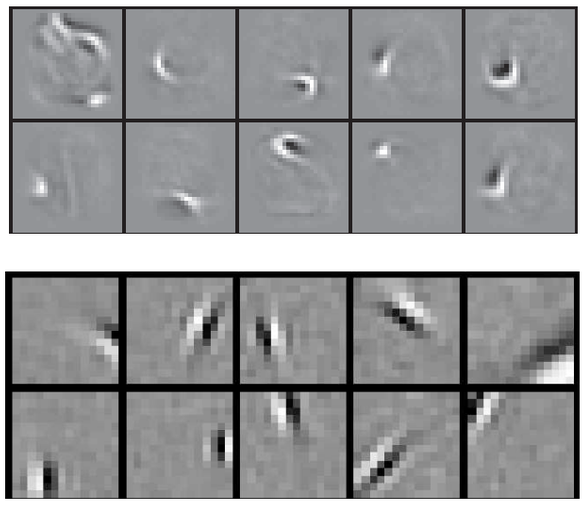
- Spatiotemporal feature detectors & descriptors:
  - 3D HMAX, 3D Hessian detector, cuboid detector, space-time interest points
  - HOG/HOF, 3D SIFT descriptor
  - recursive sparse coding (Dean et al., 2009), factored RBMs + sparse coding (Taylor et al., 2010), "factorized" sparse coding (Cadieu & Olshausen, 2008), deconvolutional network (Zeiler et al., 2010), recurrent temporal RBMs, gated RBMs, factored RBMs, mcRBMs, ...

Overview

- Background:
  - Restricted Boltzmann Machines (RBMs)
  - Convolutional RBMs
- Proposed model:
  - Spatio-temporal Deep Belief Networks
- Experiments:
  - Measuring Invariance
  - Recognizing actions (KTH dataset)
  - De-noising
  - Filling-in from sequence of "gazes"

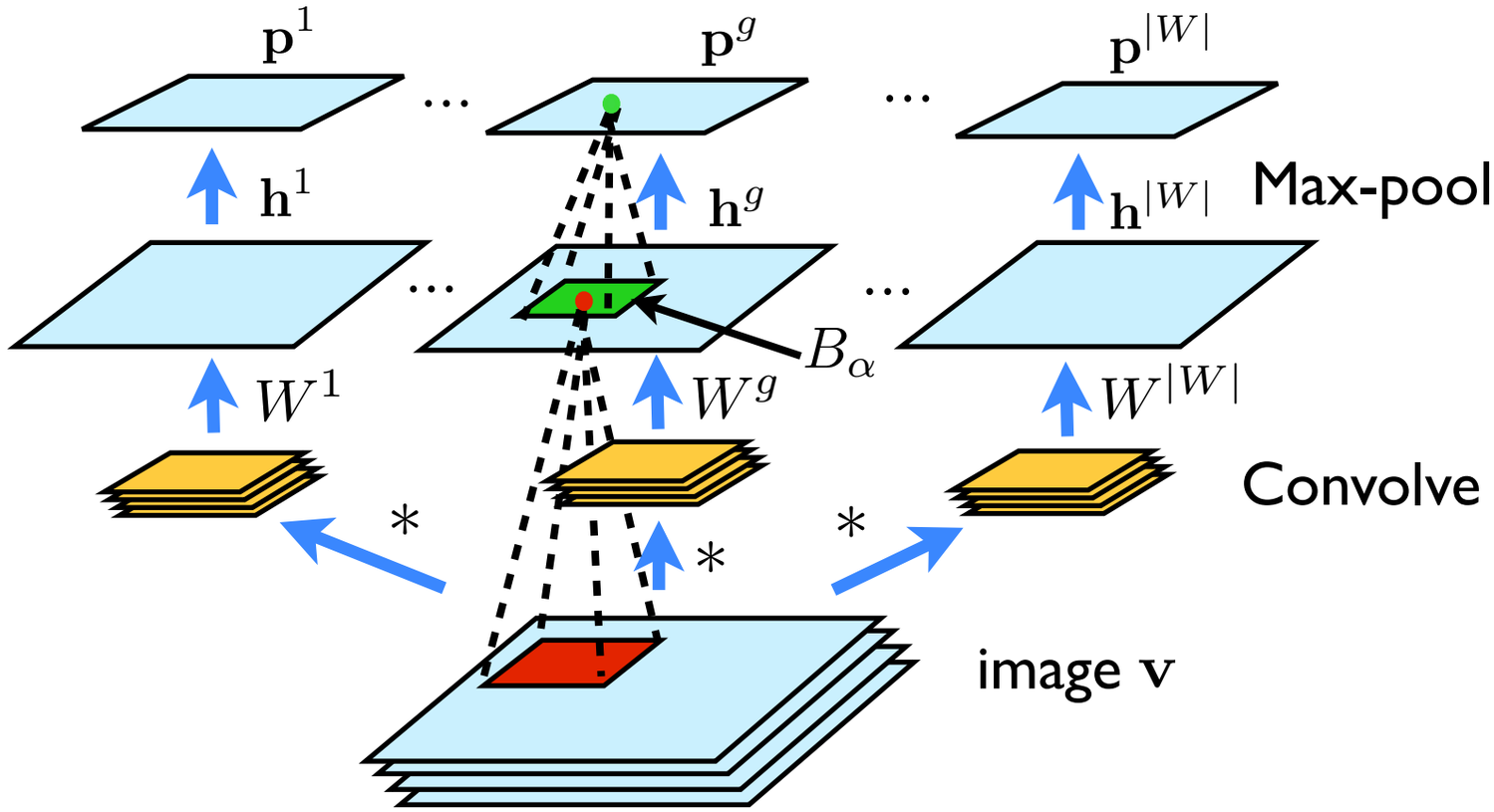
Background: RBMs





Hinton & Salakhutdinov (2006),
Lee et al. (2008)

Background: Convolutional RBMs

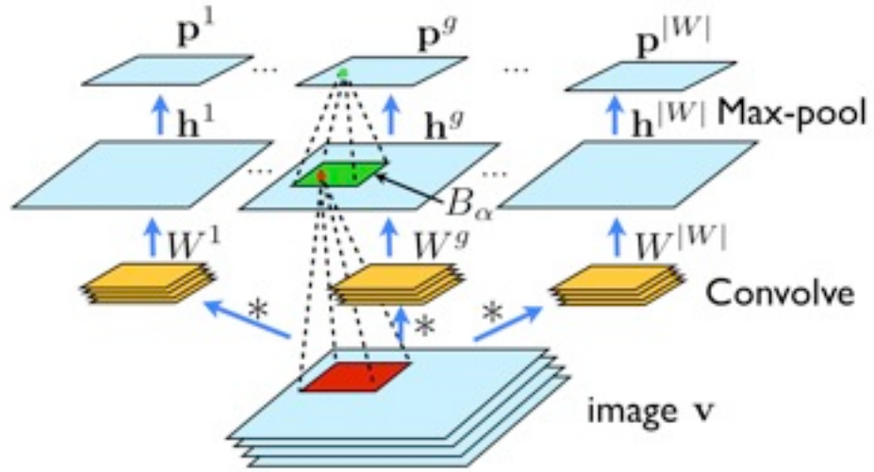


Desjardins & Bengio (2008), Lee, Grosse, Ranganath & Ng (2009),
Norouzi, Ranjbar & Mori (2009)

Convolutional RBMs

Standard RBMs:

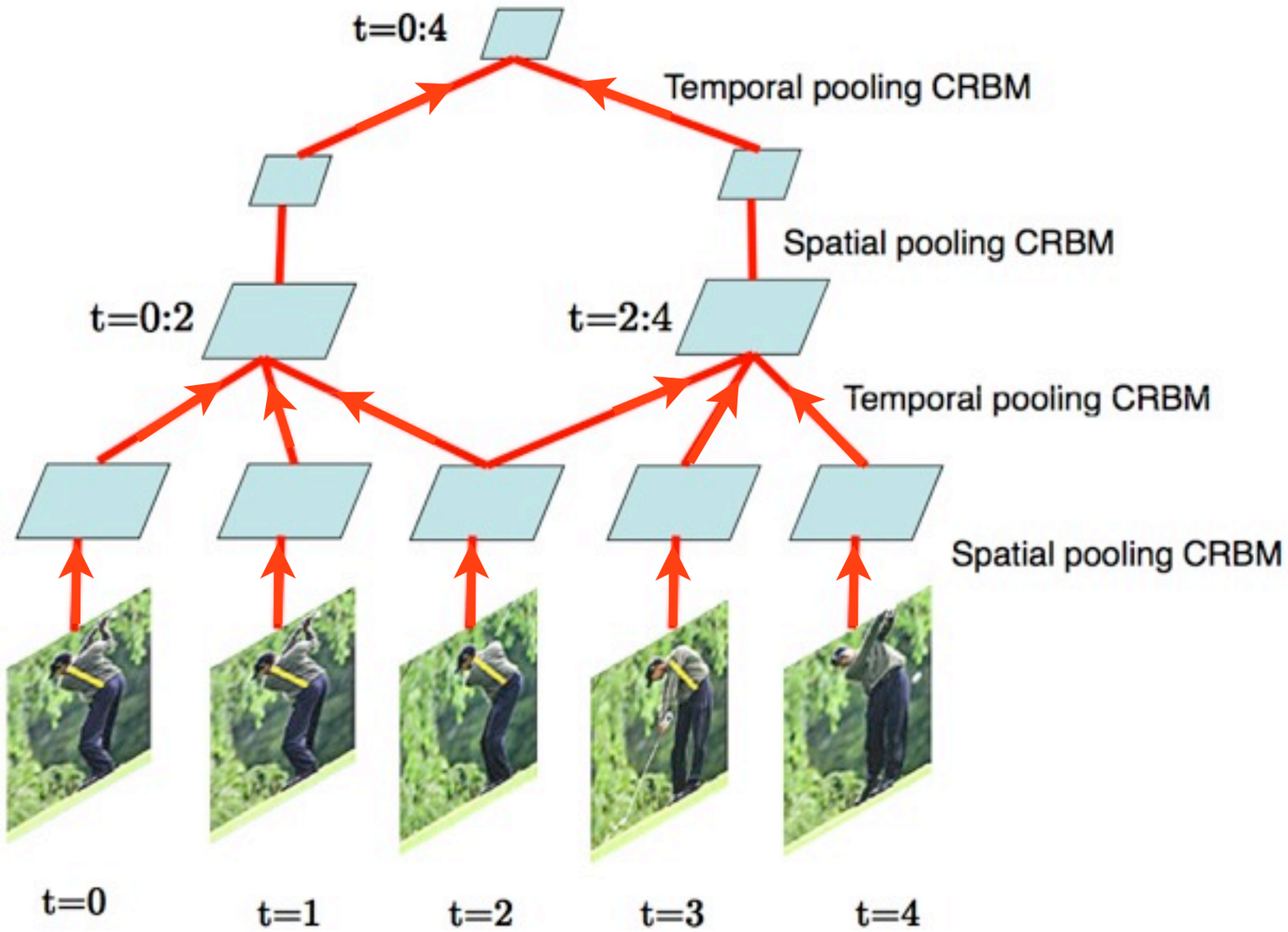
$$P(h_i = 1|\mathbf{v}) = logistic\left((W^j)^T \mathbf{v}\right)$$
$$P(v_i = 1|\mathbf{h}) = logistic\left((W_i)^T \mathbf{h}\right)$$



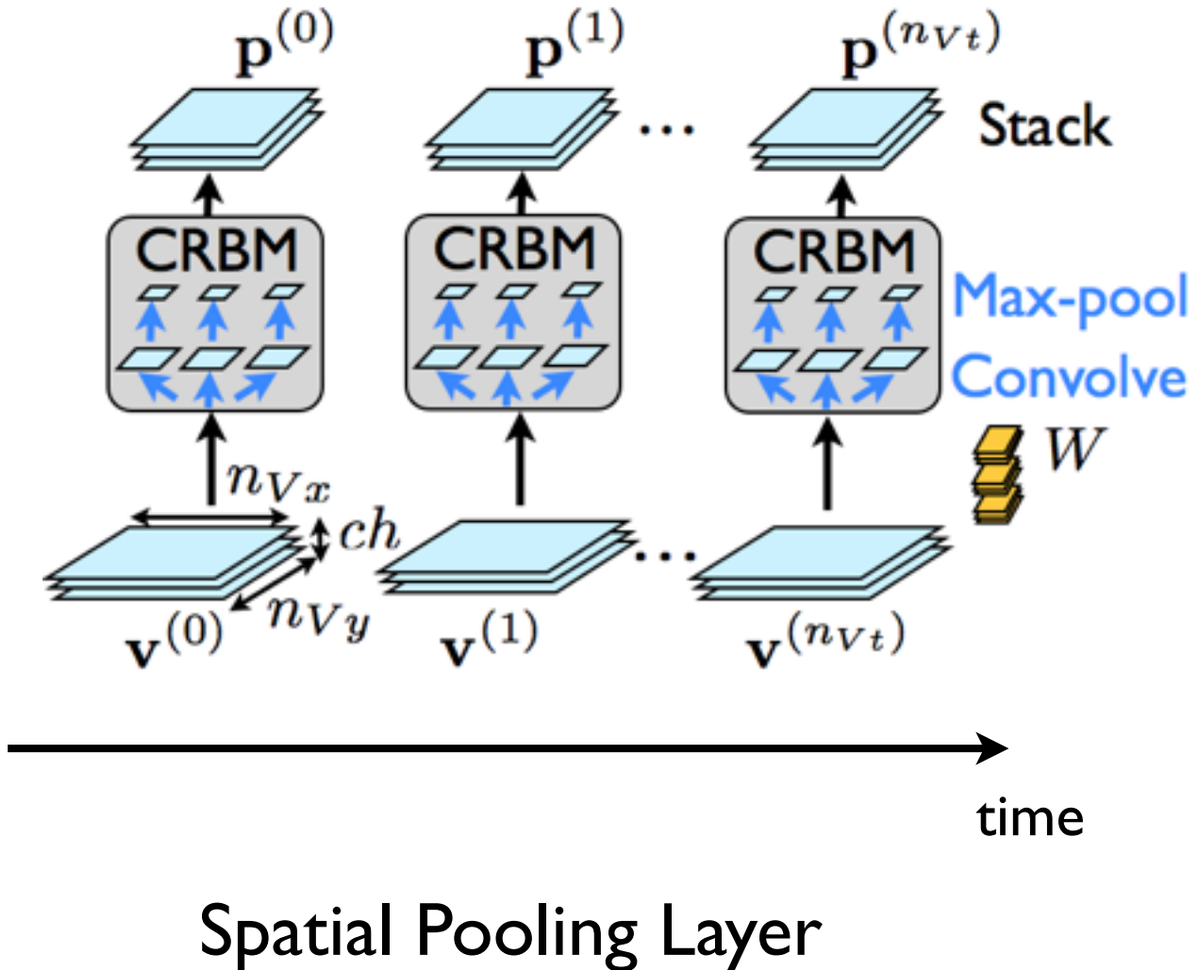
Convolutional RBMs:

$$P(h^g, p^g|\mathbf{v}) = ProbMaxPool(W^g * v^g)$$
$$P(\mathbf{v}|\mathbf{h}) = logistic\left(\sum_{g=1}^{|W|} W^g \star h^g\right)$$

Proposed model: Spatiotemporal DBNs

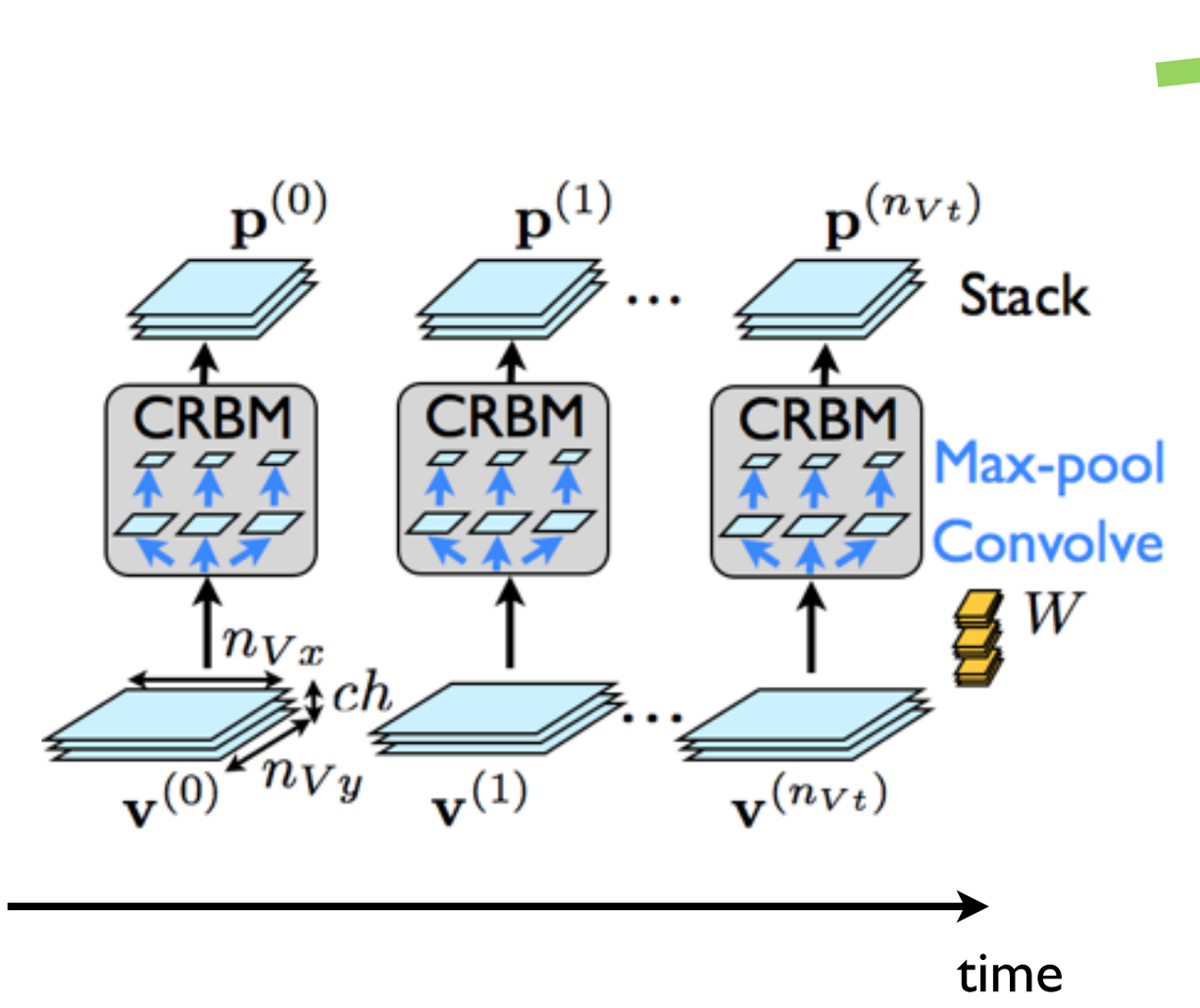


Proposed model: Spatiotemporal DBNs

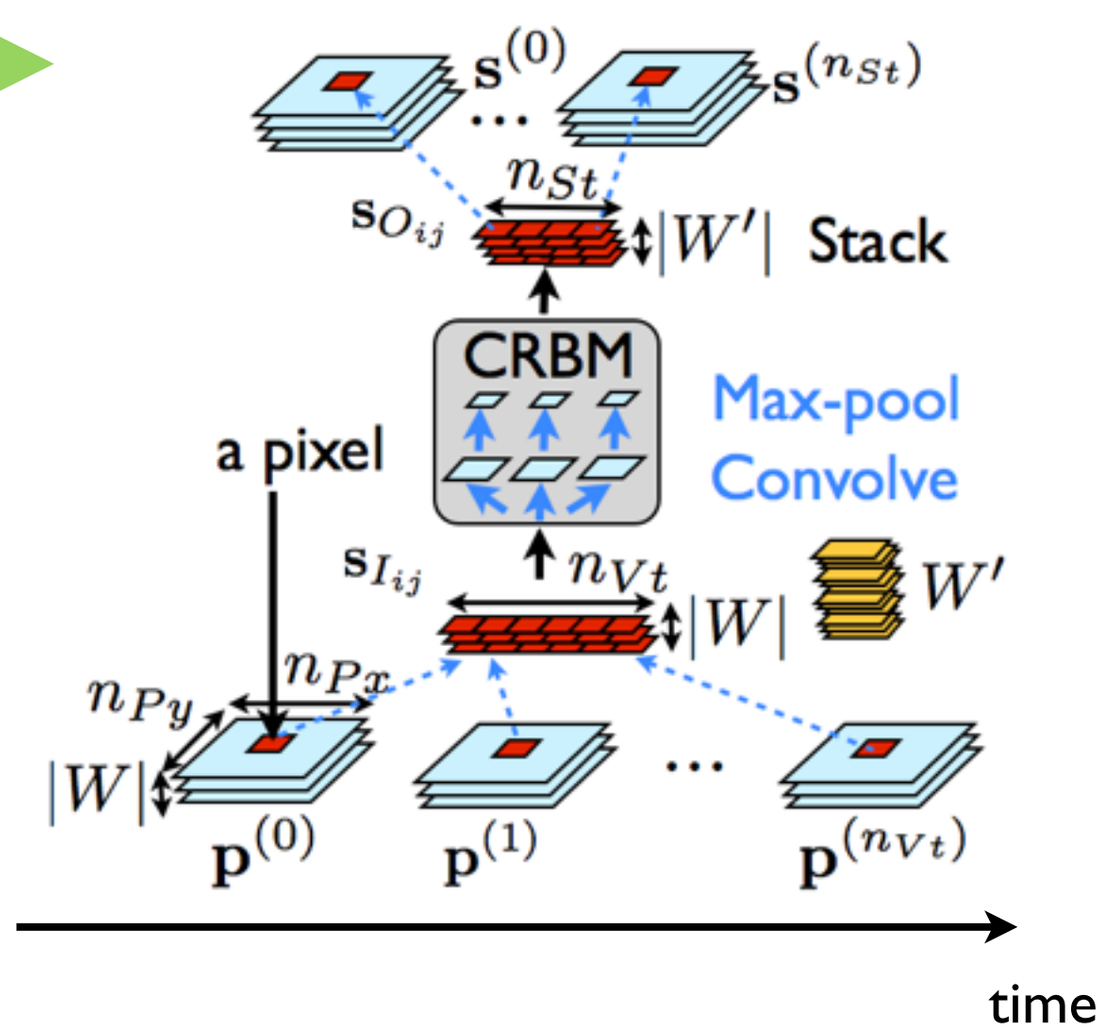


Spatial Pooling Layer

Proposed model: Spatiotemporal DBNs

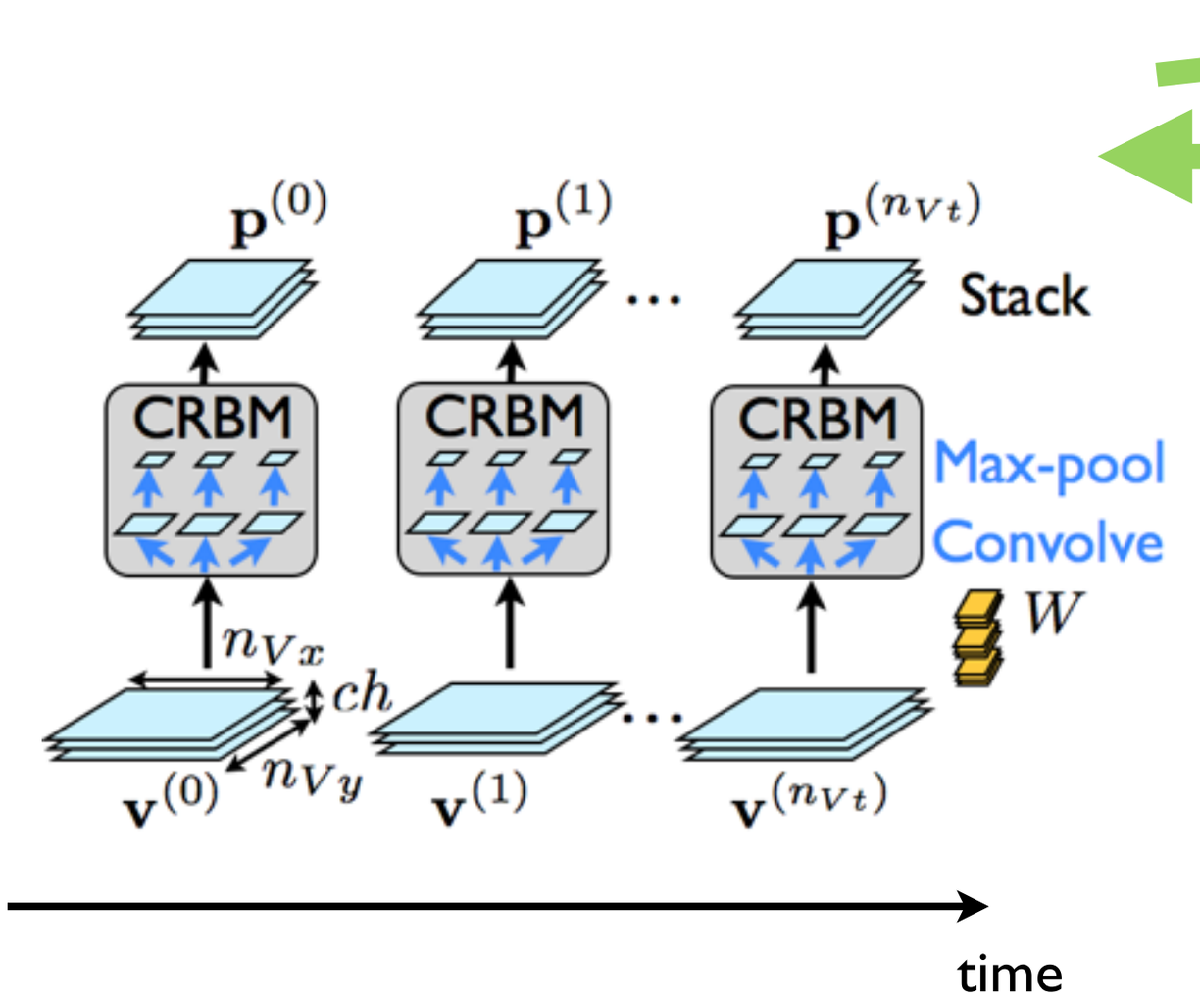


Spatial Pooling Layer

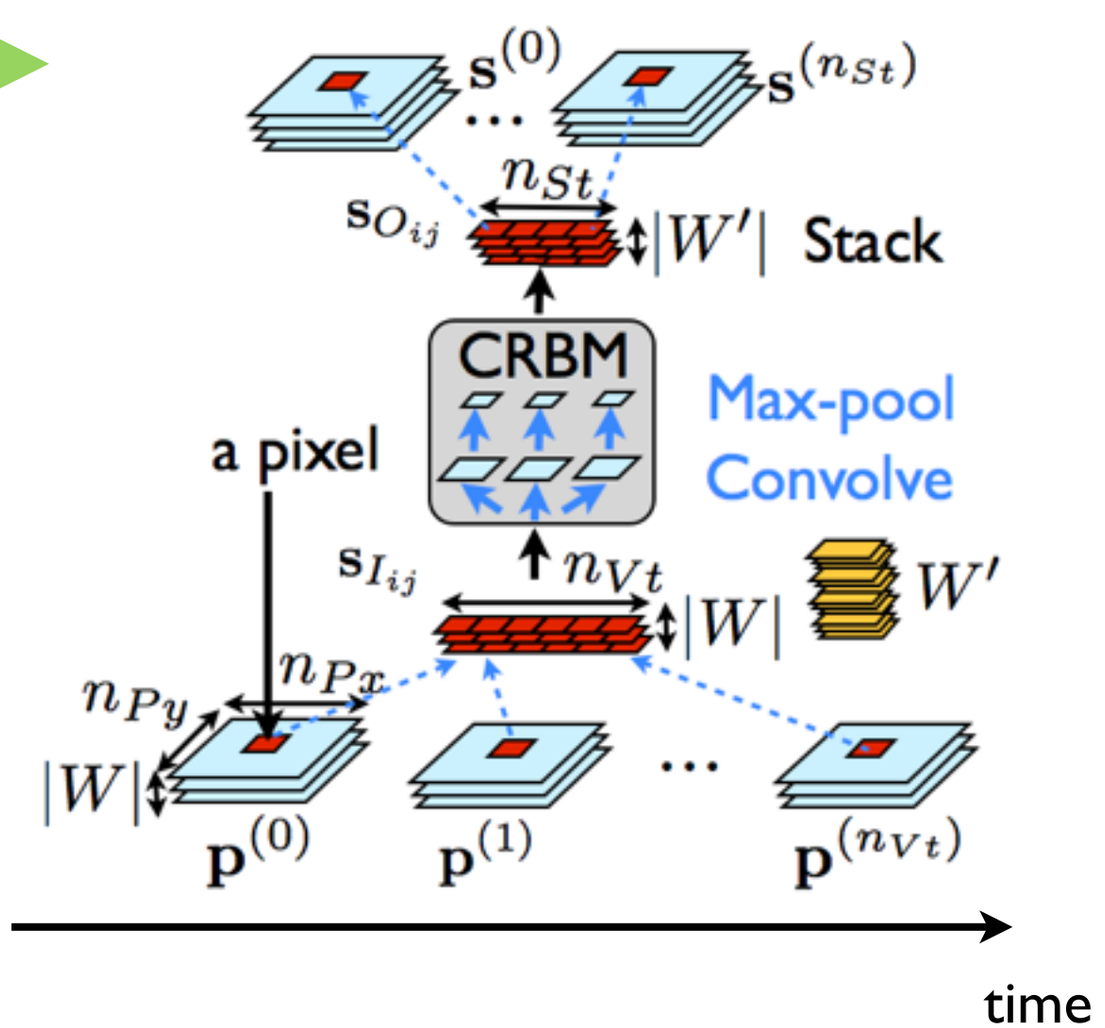


Temporal Pooling Layer

Proposed model: Spatiotemporal DBNs



Spatial Pooling Layer

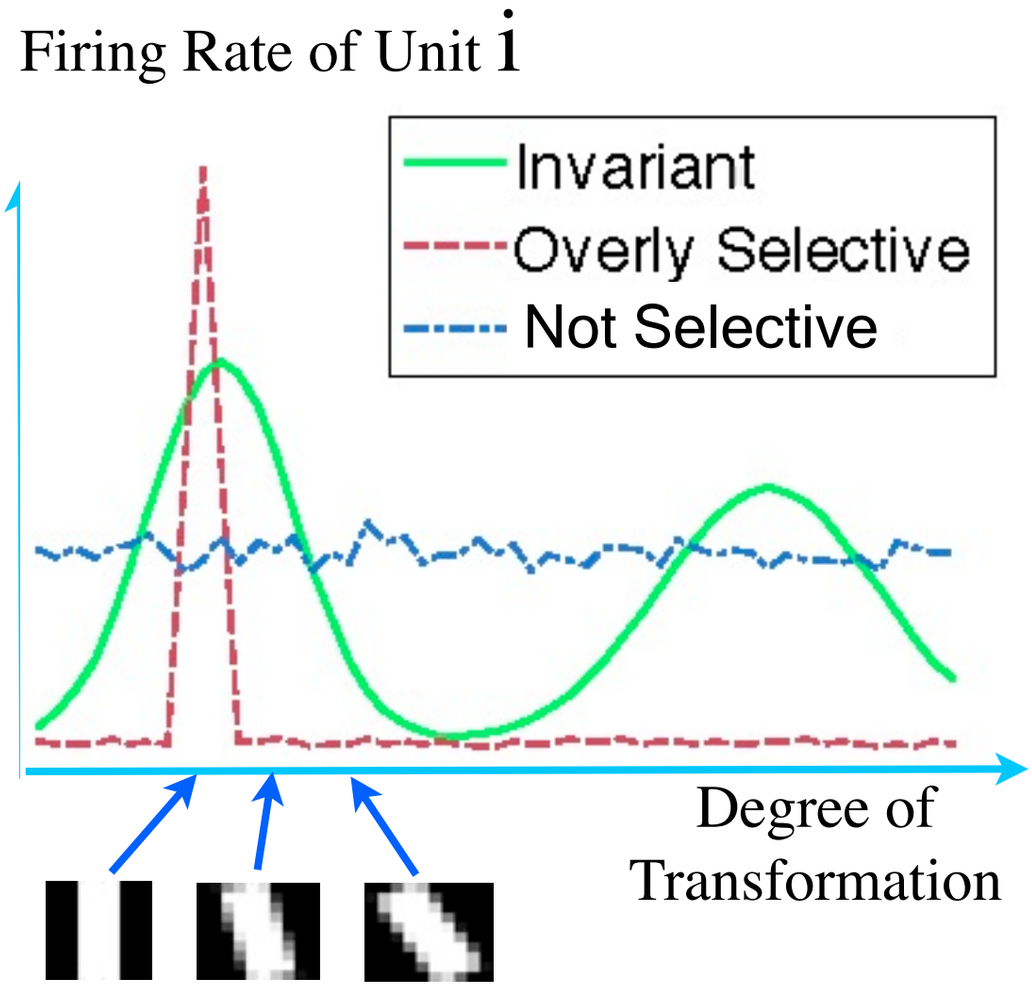


Temporal Pooling Layer

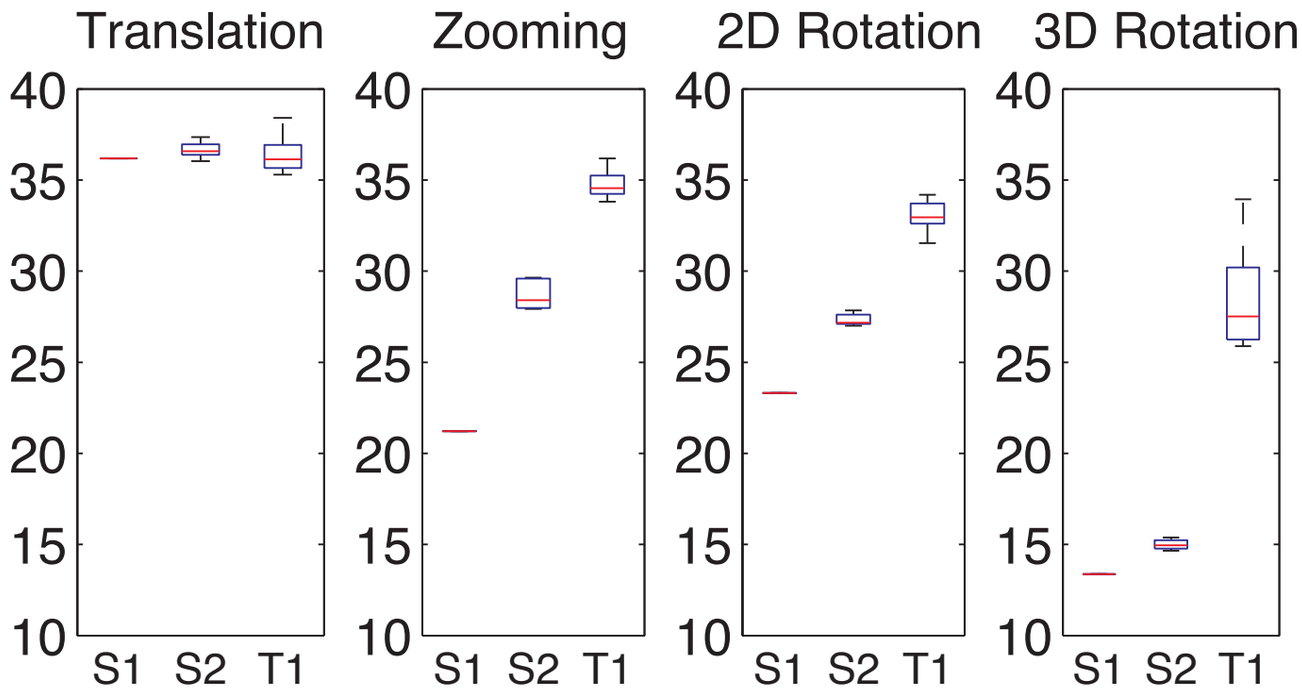
Training STDBNs

- Greedy layer-wise pre-training (Hinton, Osindero & Teh, 2006)
- Contrastive divergence for each layer (Carreira-Perpignan & Hinton, 2005)
- Sparsity regularization (e.g. Olshausen & Field, 1996)

STDBN as a discriminative feature extractor: Measuring invariance

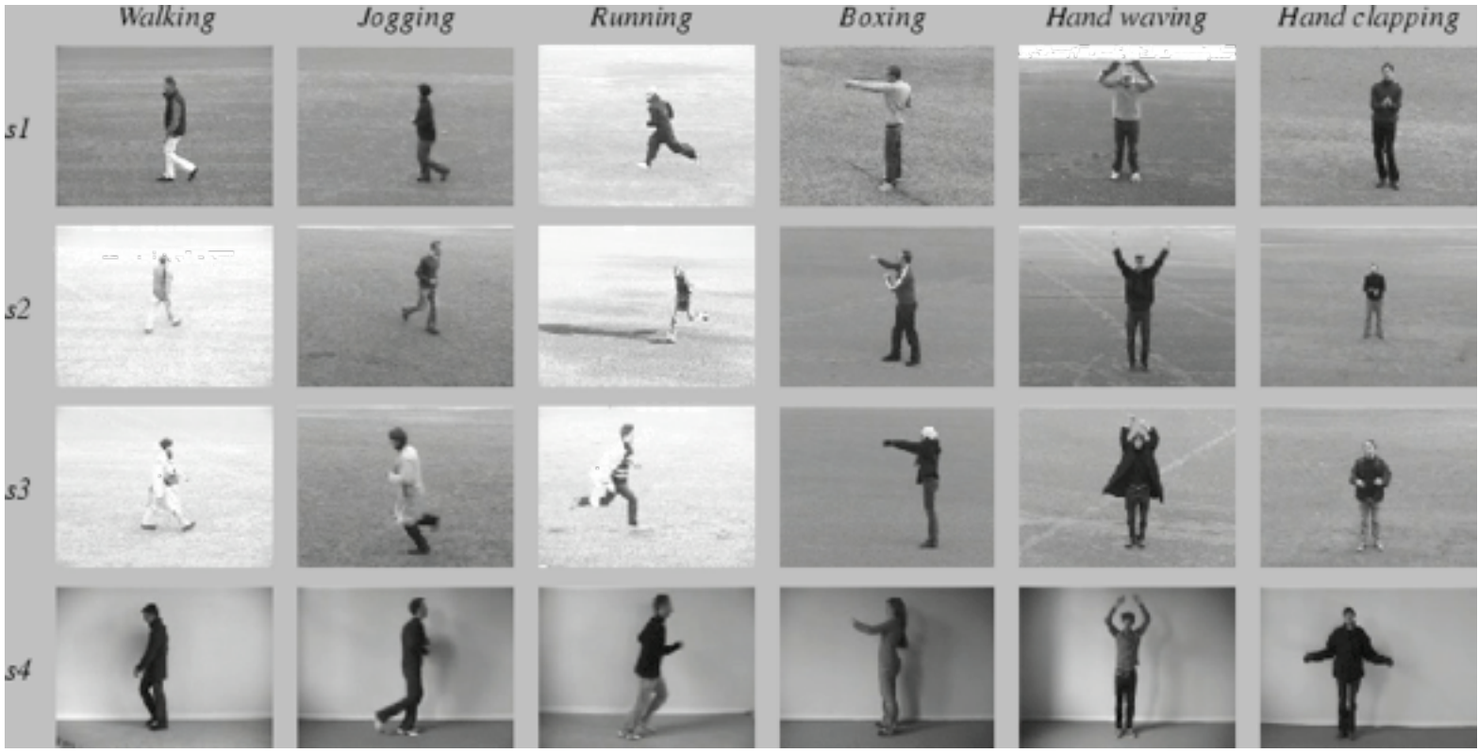


Goodfellow, Le, Saxe & Ng (2009)



Invariance scores for common transformations in natural videos, computed for layer 1 (S1) and layer 2 (S2) of a CDBN and layer 2 (T1) of STDBN. Higher is better.

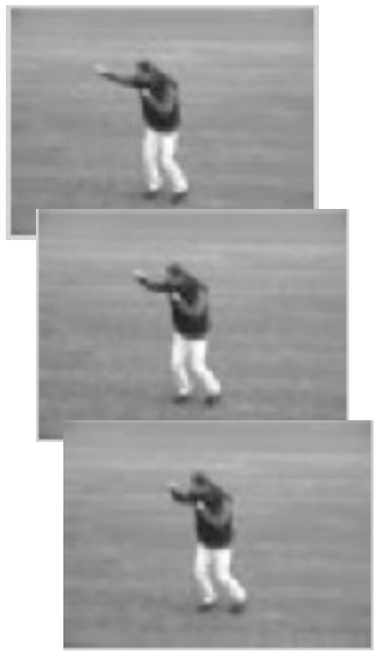
STDBN as a discriminative feature extractor: Recognizing actions



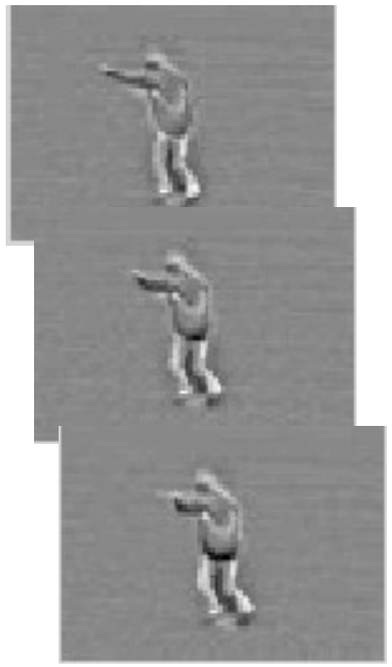
KTH Dataset: 2391 videos (~1GB)

Schuldt, Laptev & Caputo (2004), Dollar, Rabaud, Cottrell & Belongie (2005), Laptev, Marszalek, Schmid & Rozenfield (2008), , Liu & Shah (2008), Dean, Corrado & Washington (2009), Wang & Li (2009), Taylor & Bregler (2010), ...

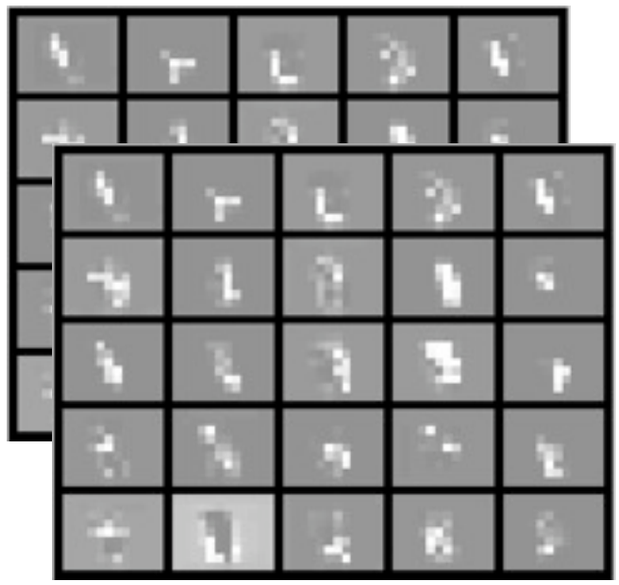
Recognizing actions: Pipeline



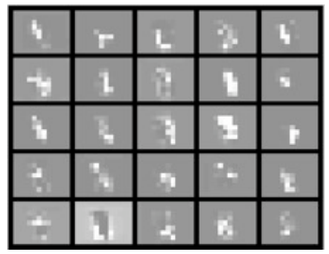
Input video



Input to STDBN



STDBN responses



Input to SVM

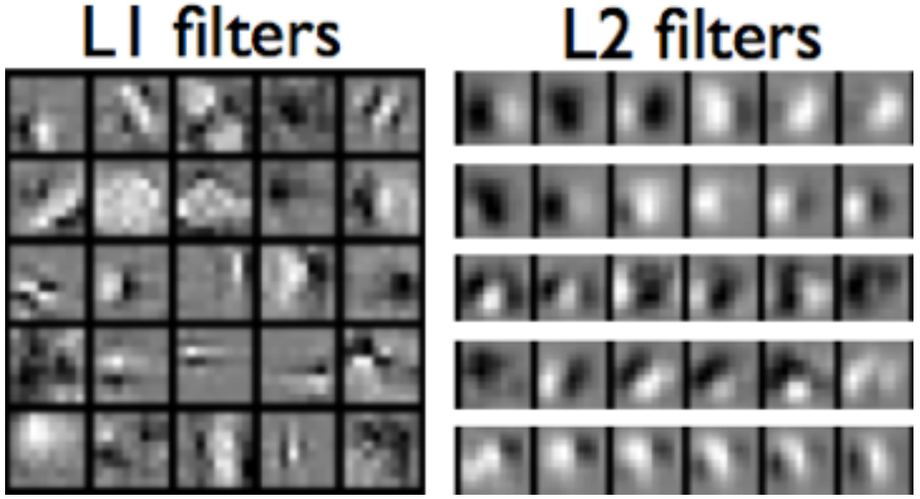
Local and global contrast normalization

Unsupervised feature extraction using STDBN

Dimensionality standardization (resolution reduction)

Recognizing actions: Model architecture

- 4-layers: 8x8 filters for spatial pooling layers, 6x1 filters for temporal pooling layers, pooling ratio of 3



Experiments: number of parameters ~300k, training time ~5 days

Recognizing actions: Classification accuracy

	Box	Clap	Wave	Jog	Run	Walk
Box	**89.4**	4.7	2.7	0.8	0	2.4
Clap	2.8	**89.7**	6.7	0	0.8	0
Wave	0.8	2.0	**96.9**	0	0.3	0
Jog	0	0	0	**68.0**	19.9	12.1
Run	0	0	0	16.4	**80.9**	2.7
Walk	0	0	0	10.5	1.2	**88.3**

Layer 1

	Box	Clap	Wave	Jog	Run	Walk
Box	**91.4**	4.7	1.2	0	0	2.7
Clap	3.1	**91.3**	5.6	0	0	0
Wave	0.8	1.2	**98**	0	0	0
Jog	0	0	0	**77.3**	13.7	9.0
Run	0	0	0	12.5	**85.6**	1.9
Walk	0	0	0	5.5	0	**94.5**

Layer 2

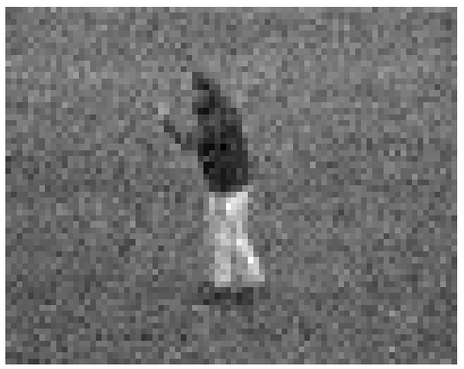
	Box	Clap	Wave	Jog	Run	Walk
Box	**94.5**	2.4	1.2	0	0	1.9
Clap	3.9	**93.7**	2.0	0	0.4	0
Wave	1.2	1.6	**96.9**	0	0	0
Jog	0.3	0	0	**80.5**	13.3	5.9
Run	0	0	0	10.5	**89.5**	0
Walk	0.8	0	0	3.1	0	**96.1**

Layer 3

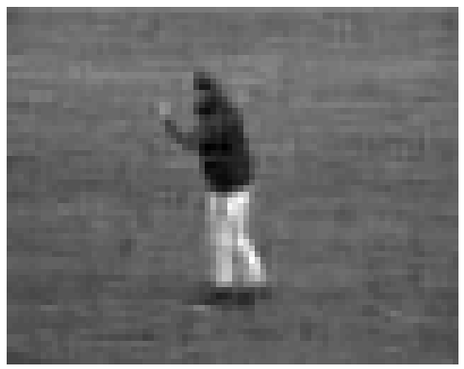
Method	Accuracy (%)
4-layer ST-DBN	90.3 ± 0.83
3-layer ST-DBN	$\mathbf{91.13} \pm 0.85$
2-layer ST-DBN	89.73 ± 0.18
1-layer ST-DBN	85.97 ± 0.94
Liu & Shah [21]	94.2
Wang & Li [31]	87.8
Dollár et al. [18]	81.2

STDBN as a generative model: De-noising

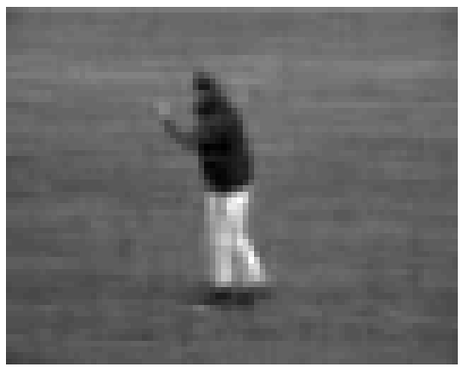
Clean Video



Noisy Video
NMSE = 1



Spatially
denoised video
NMSE=0.175



Spatiotemporally
denoised video
NMSE=0.155

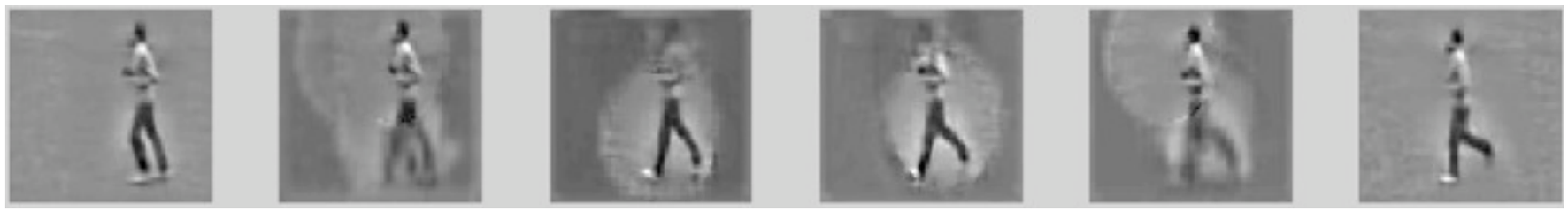
STDBN as a generative model: Filling-in from sequence of "gazes"

Missing

STDBN as a generative model: Filling-in from sequence of "gazes"

Missing

Prediction



STDBN as a generative model: Filling-in from sequence of "gazes"

Missing

Prediction

Truth (Fully Observed)

Discussion

- Limitations:
  - computation
  - invariance vs. reconstruction trade-off
  - alternating space-time vs. joint space-time convolution
- Extensions:
  - attentional mechanisms for gaze planning?
  - leveraging feature detection to reduce training data size