Class 24-26

Learning Data Representations: beyond DeepLearning: the Magic Theory

Tomaso Poggio

Connection with the topic of learning theory

Notices of the American Mathematical Society (AMS), Vol. 50, No. 5, 537-544, 2003.

The Mathematics of Learning: Dealing with Data

Tomaso Poggio and Steve Smale

How then do the learning machines described in the theory compare with brains?

□ One of the most obvious differences is the ability of people and animals to learn from very few examples. The algorithms we have described can learn an object recognition task from a few thousand labeled images but a child, or even a monkey, can learn the same task from just a few examples. Thus an important area for future theoretical and experimental work is learning from partially labeled examples

□ A comparison with real brains offers another, related, challenge to learning theory. The “learning algorithms” we have described in this paper correspond to one-layer architectures. Are hierarchical architectures with more layers justifiable in terms of learning theory? It seems that the learning theory of the type we have outlined does not offer any general argument in favor of hierarchical learning machines for regression or classification.

□ Why hierarchies? There may be reasons of *efficiency* – computational speed and use of computational resources. For instance, the lowest levels of the hierarchy may represent a dictionary of features that can be shared across multiple classification tasks.

□ There may also be the more fundamental issue of *sample complexity*. Learning theory shows that the difficulty of a learning task depends on the size of the required hypothesis space. This complexity determines in turn how many training examples are needed to achieve a given level of generalization error. Thus our ability of learning from just a few examples, and its limitations, may be related to the hierarchical architecture of cortex.

Classical learning theory and Kernel Machines (Regularization in RKHS)

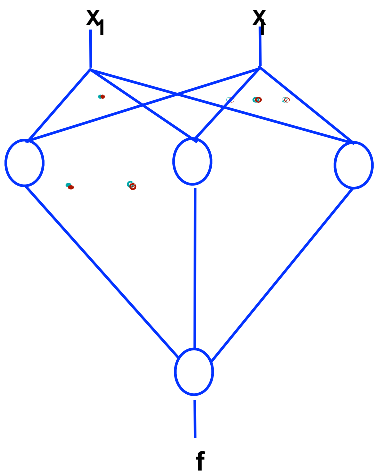
$$\min_{f\in H}\left[\frac{1}{n}\sum_{i=1}^{n} V(f(x_i)-y_i)+\lambda\ \|f\|_K^2\right]$$

implies

$$f(\mathbf{x})=\sum_i^n \alpha_i K(\mathbf{x},\mathbf{x}_i)$$

Remark:

Kernel machines correspond to *shallow* networks



M-theory:

unsupervised learning of hierarchical invariant representations

Plan

1. Motivation: models of cortex (and deep convolutional networks)
2. Core theory
   - the basic invariance module
   - the hierarchy
3. Computational performance
4. Biological predictions
5. Theorems and remarks
   - $n \to 1$
   - invariance and sample complexity
   - connections with scattering transform
   - invariances and beyond perception
   - ...

Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)

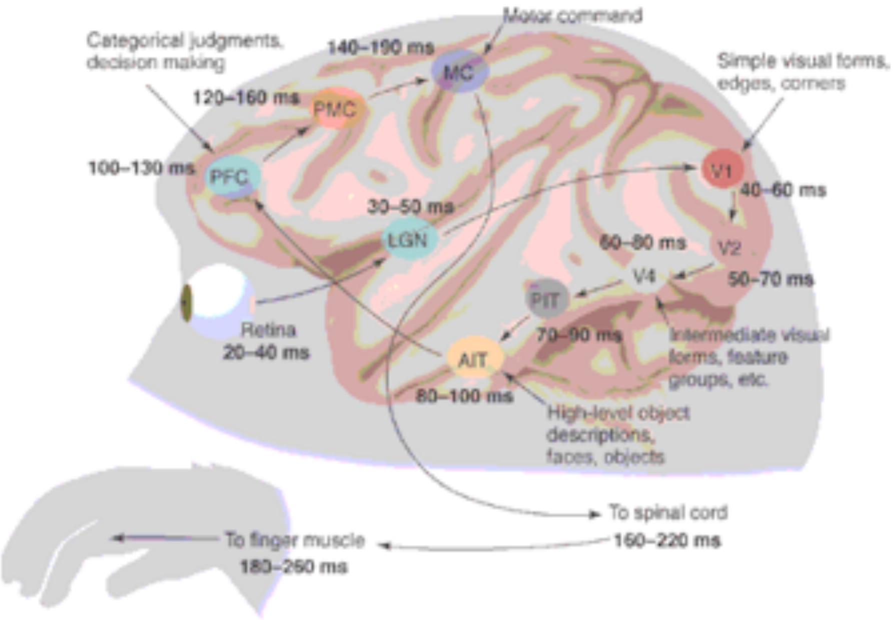
[software available online
with CNS (for GPUs)]

Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu
Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)



[software available online
with CNS (for GPUs)]

Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)

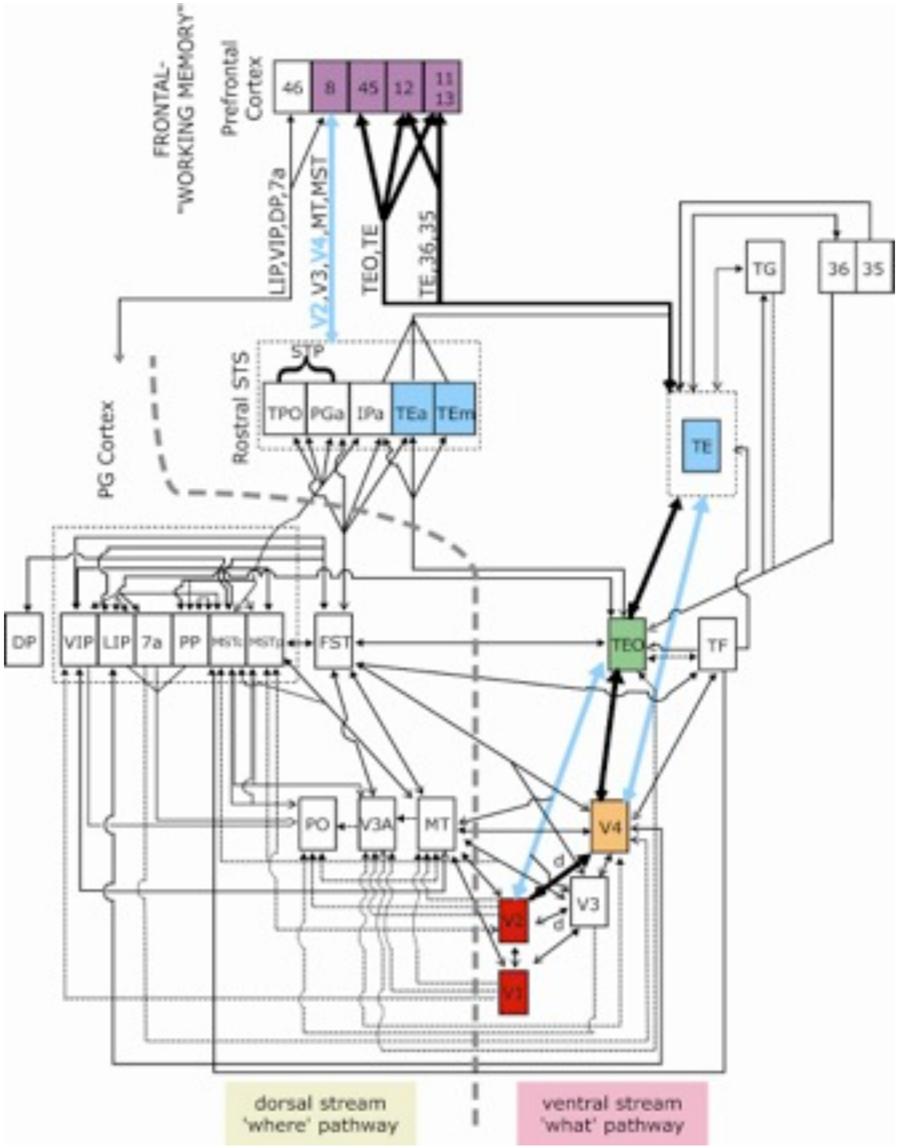
[software available online
with CNS (for GPUs)]

Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu
Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)



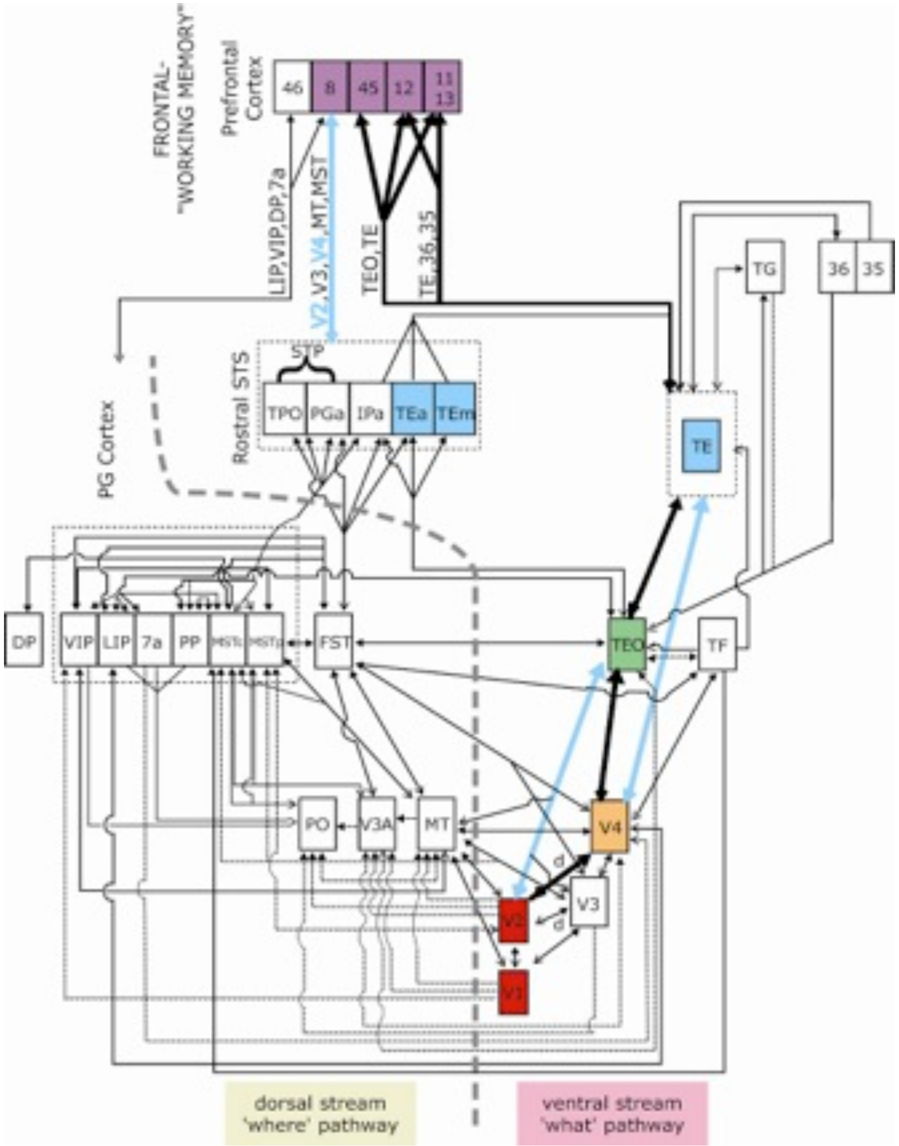
[software available online
with CNS (for GPUs)]

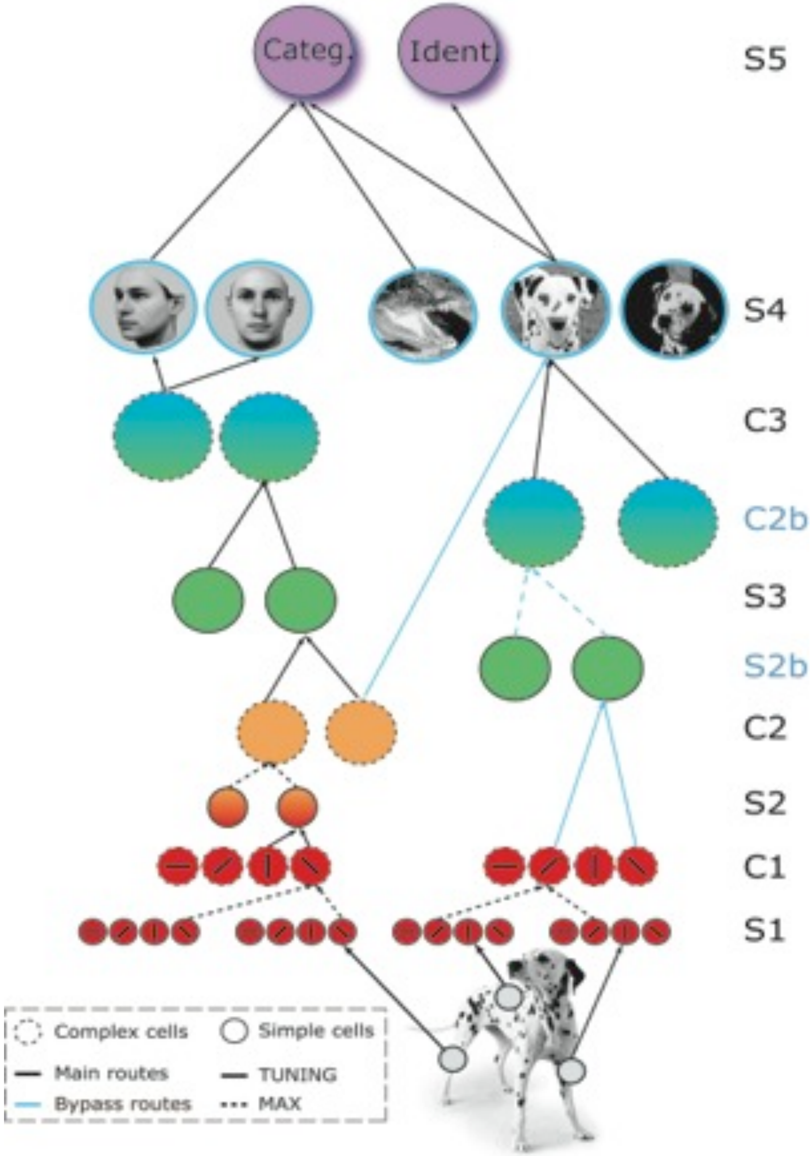
Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu
Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)

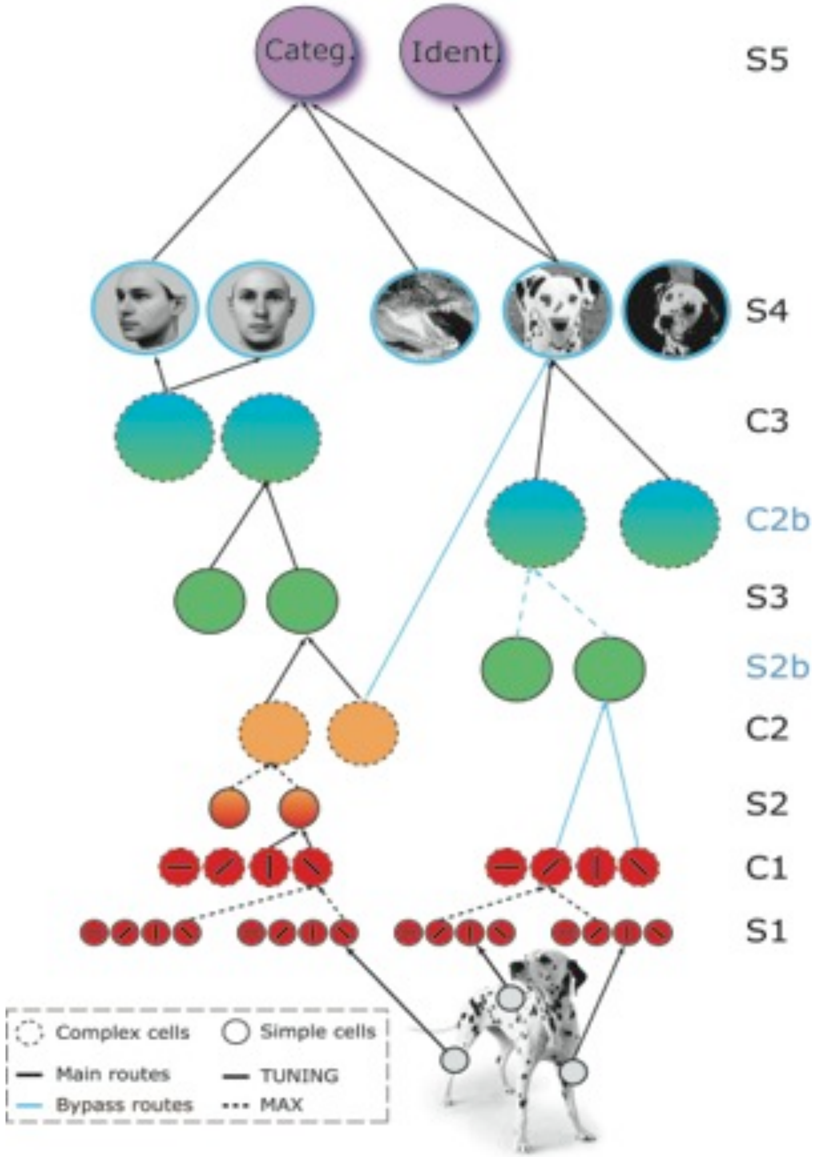




[software available online with CNS (for GPUs)]

Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Recognition in Visual Cortex: "classical model", selective and invariant



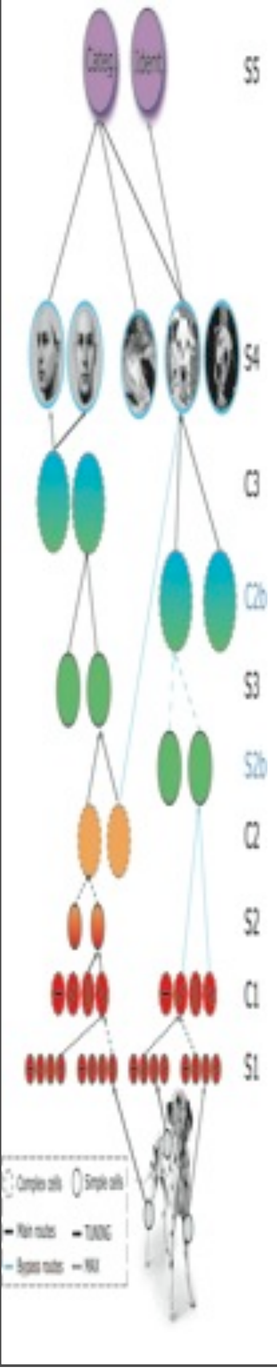
- It is in the family of "Hubel-Wiesel" models (Hubel & Wiesel, 1959: *qual.* **Fukushima**, 1980: *quant*; Oram & Perrett, 1993: *qual*; Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe, 2002; Ullman et al., 2002; Mel, 1997; Wersing and Koerner, 2003; LeCun et al 1998: *not-bio*; Amit & Mascaro, 2003: *not-bio*; Hinton, LeCun, Bengio *not-bio;* Deco & Rolls 2006…)

- As a biological model of object recognition in the ventral stream – from V1 to PFC -- it is *perhaps* the most quantitatively faithful to known neuroscience data

[software available online]

Riesenhuber & Poggio 1999, 2000; Serre Kouh Cadieu Knoblich Kreiman & Poggio 2005; Serre Oliva Poggio 2007

Model "works": it accounts for physiology



Hierarchical Feedforward Models: is consistent with or predict neural data

V1:

Simple and complex cells tuning (Schiller et al 1976; Hubel & Wiesel 1965; Devalois et al 1982)

MAX-like operation in subset of complex cells (Lampl et al 2004)

V2:

Subunits and their tuning (Anzai, Peng, Van Essen 2007)

V4:

Tuning for two-bar stimuli (Reynolds Chelazzi & Desimone 1999)

MAX-like operation (Gawne et al 2002)

Two-spot interaction (Freiwald et al 2005)

Tuning for boundary conformation (Pasupathy & Connor 2001, Cadieu, Kouh, Connor et al., 2007)

Tuning for Cartesian and non-Cartesian gratings (Gallant et al 1996)

IT:

Tuning and invariance properties (Logothetis et al 1995, paperclip objects)

Differential role of IT and PFC in categorization (Freedman et al 2001, 2002, 2003)

Read out results (Hung Kreiman Poggio & DiCarlo 2005)

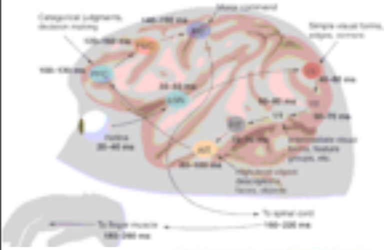
Pseudo-average effect in IT (Zoccolan Cox & DiCarlo 2005; Zoccolan Kouh Poggio & DiCarlo 2007)

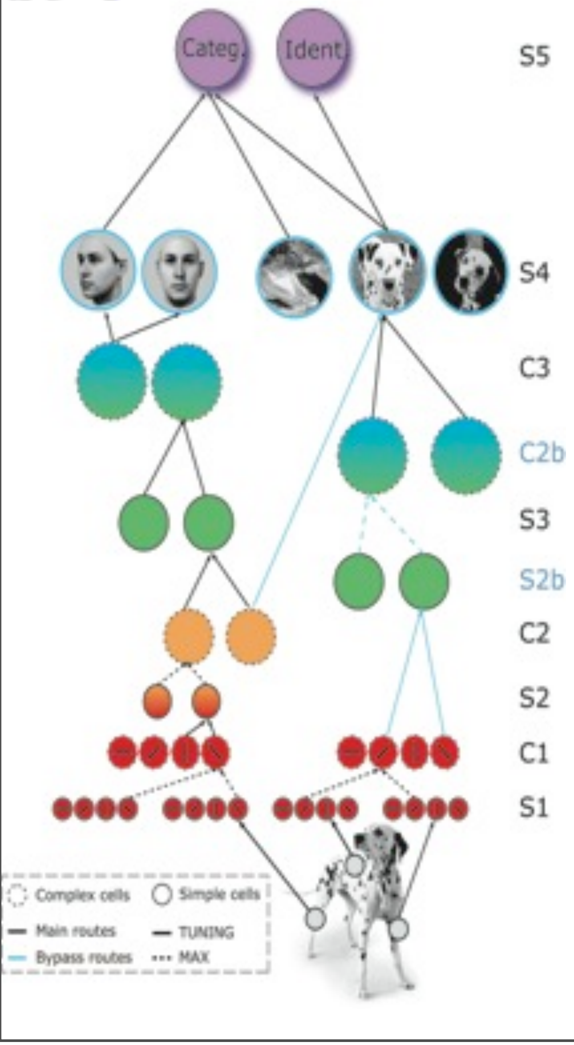
Human:

Rapid categorization (Serre Oliva Poggio 2007)

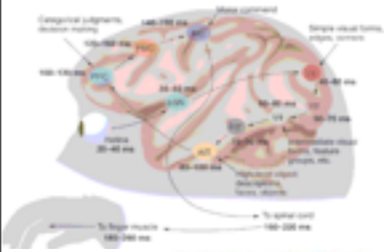
Face processing (fMRI + psychophysics) (Riesenhuber et al 2004; Jiang et al 2006)

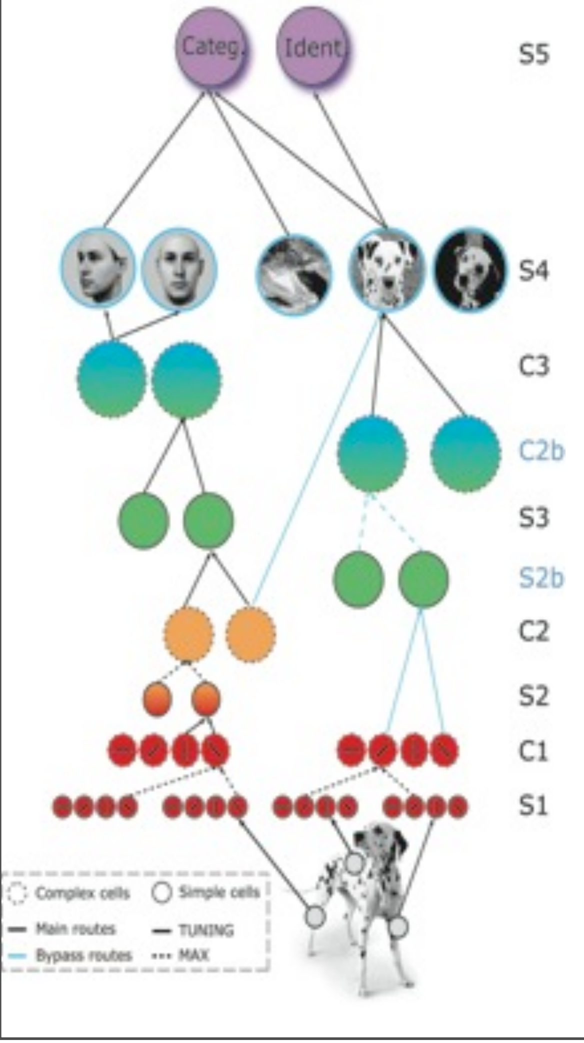
Model "works": it accounts for psychophysics


Categorical judgments, decision making
Simple visual forms, edges, corners
PFC
V1
V2
V4
AIT
Retina
20–40 ms
Intermediate visual forms, feature groups, etc.
High-level object descriptions, faces, objects
To spinal cord
To finger muscle

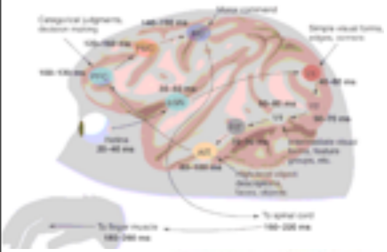


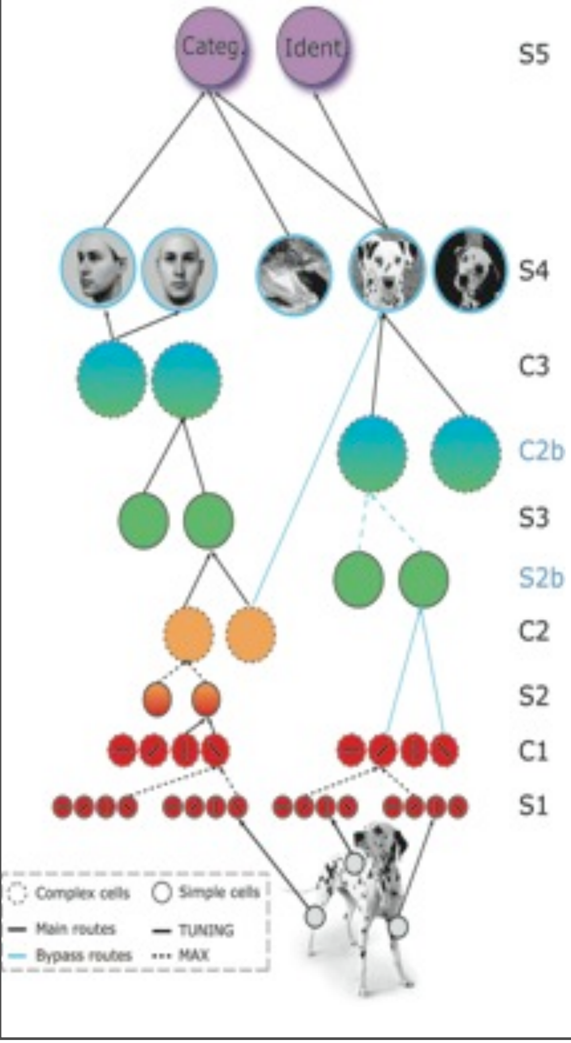
Model "works":
it accounts for psychophysics


Categorical judgments, decision making
Motor command
Simple visual forms, edges, corners
Retina
20–40 ms
Intermediate visual forms, feature groups, etc.
High-level object descriptions, faces, objects
To spinal cord
To finger muscle



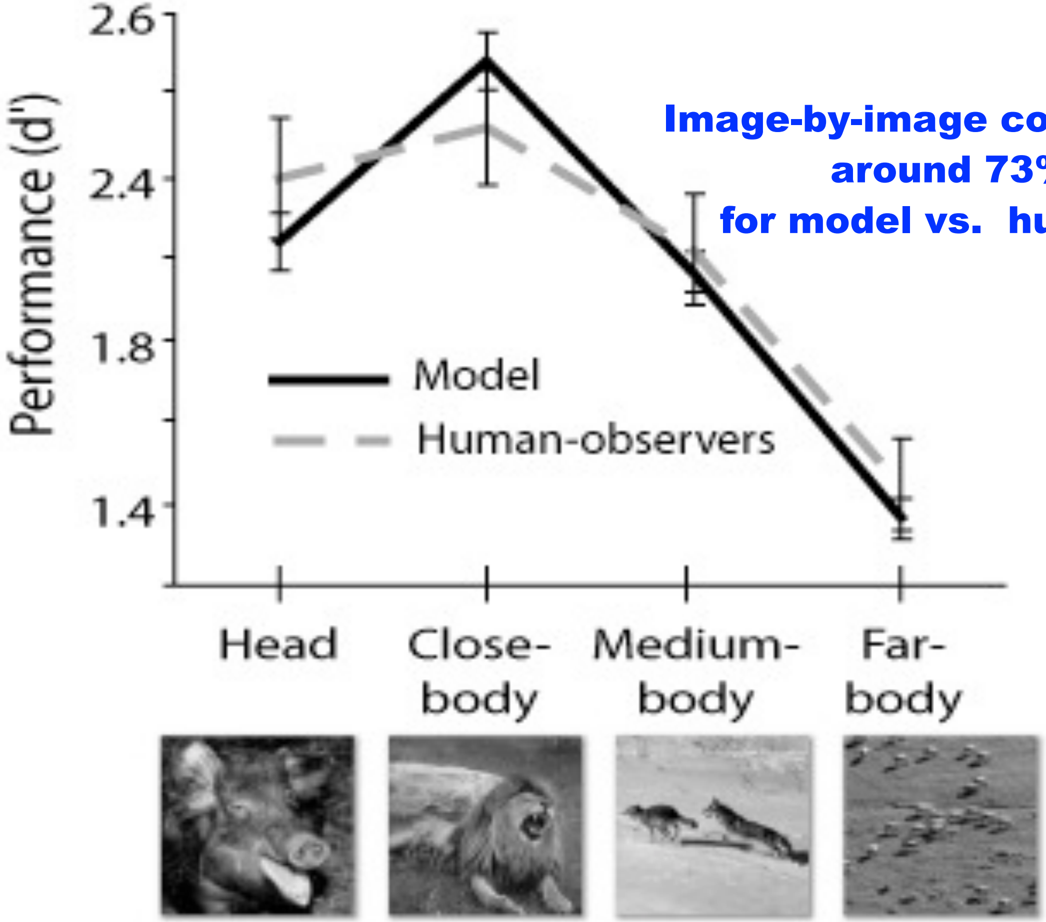
Model "works": it accounts for psychophysics


Categorical judgments, decision making
Motor command
Simple visual forms, edges, corners
PFC
PMC
MC
V1
V2
V4
PIT
AIT
Retina
Intermediate visual forms, feature groups, etc.
High-level object descriptions, faces, objects
To spinal cord
To finger muscle

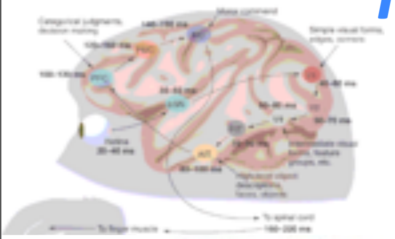


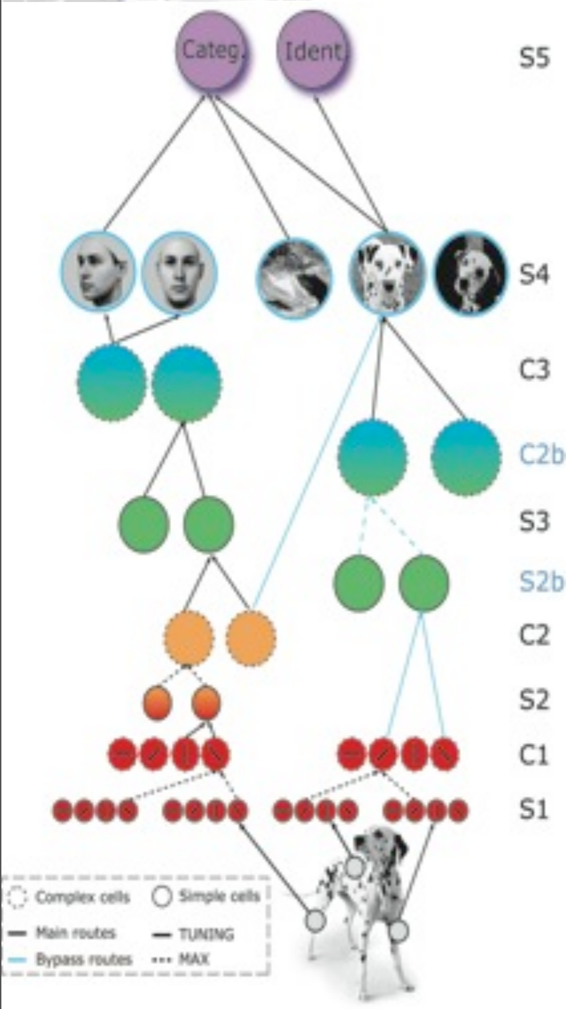
Feedforward Models: "predict" rapid categorization (82% model vs. 80% humans)

Image-by-image correlation: around 73% for model vs. humans)



Model “works”:
it performs well at computational level

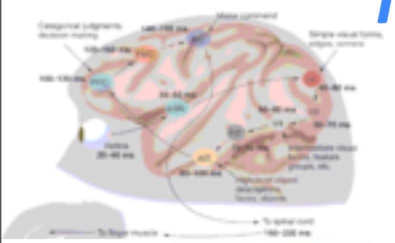


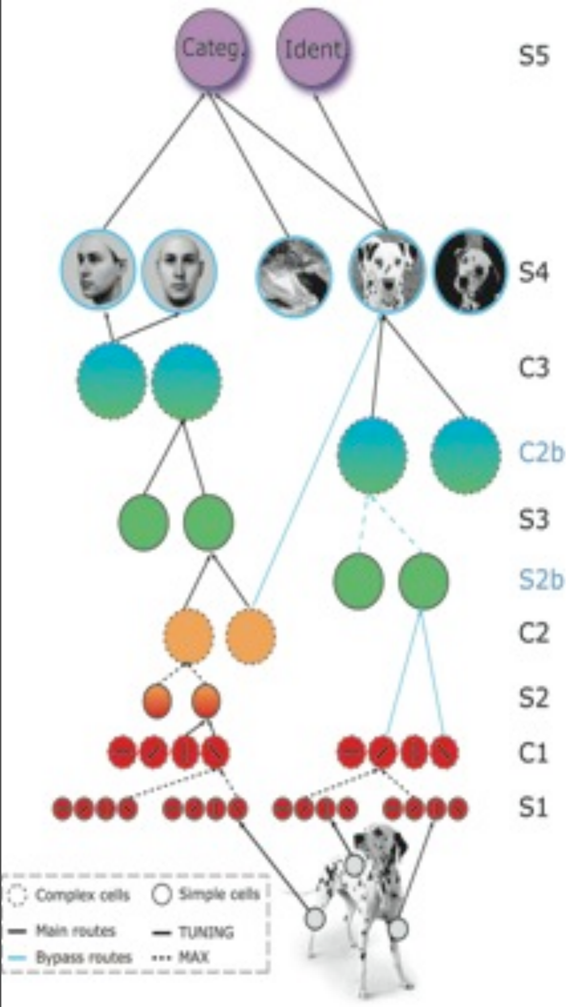


Models of the *ventral stream* in cortex perform well compared to engineered computer vision systems (in 2006) on several databases

Bileschi, Wolf, Serre, Poggio, 2007

Model "works":
it performs well at computational level

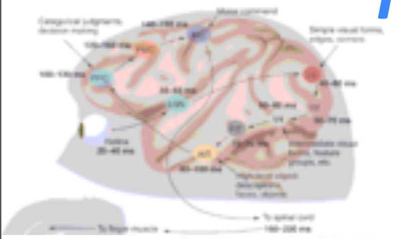


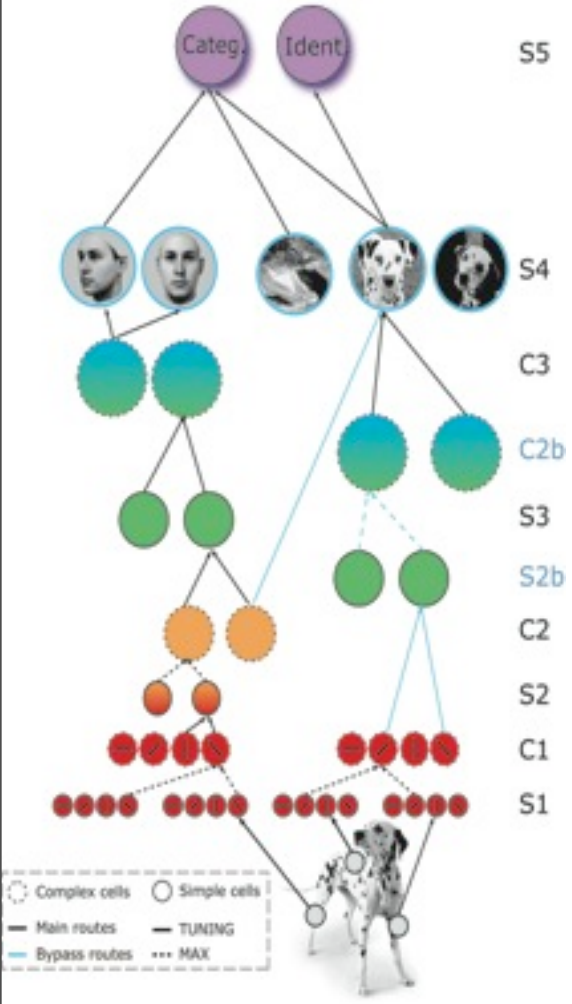


Models of the *ventral stream* in cortex perform well compared to engineered computer vision systems (in 2006) on several databases

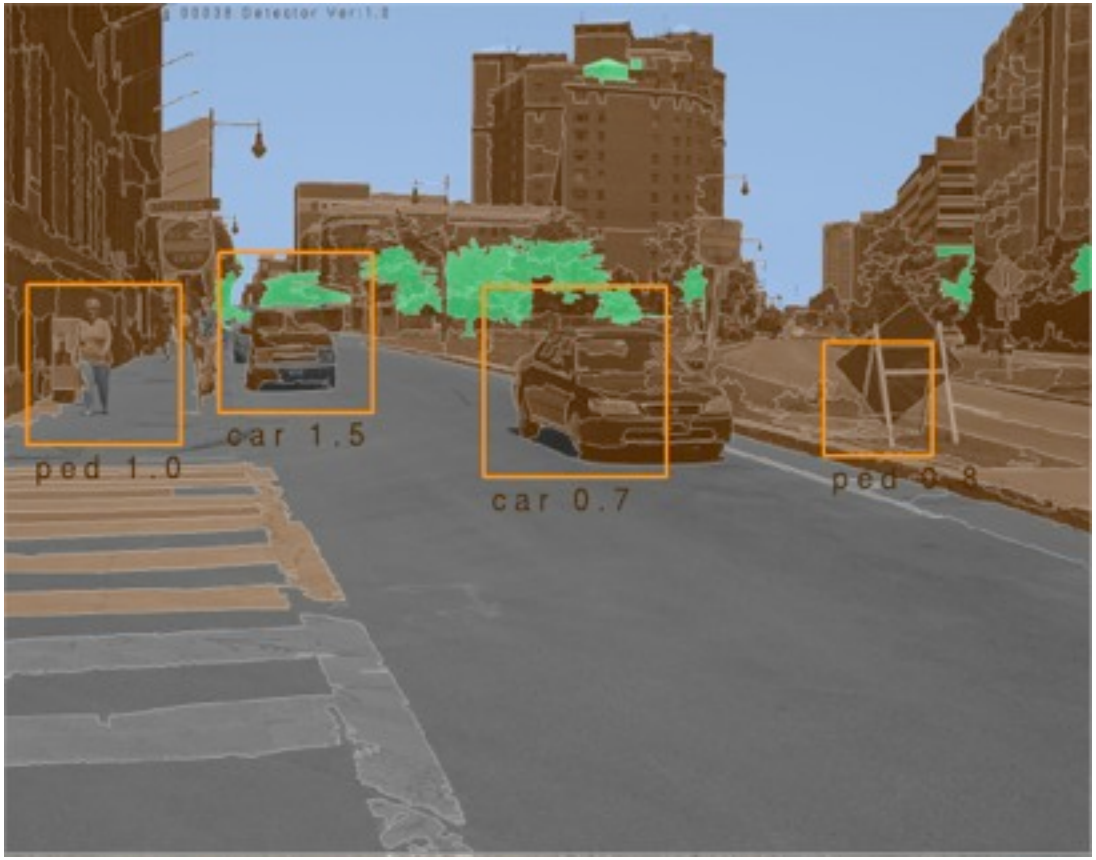
Bileschi, Wolf, Serre, Poggio, 2007

Model "works":
it performs well at computational level





Models of the *ventral stream* in cortex
perform well compared to
engineered computer vision systems (in 2006)
on several databases



Bileschi, Wolf, Serre, Poggio, 2007

Model "works":
it performs well at computational level

Performance

human agreement	72%
proposed system	77%
commercial system	61%
chance	12%

Models of cortex lead to better systems for action recognition in videos: automatic phenotyping of mice

Jhuang , Garrote, Yu, Khilnani, Poggio, Mutch Steele, Serre, Nature Communicatons, 2010

Model “works”:
it performs well at computational level

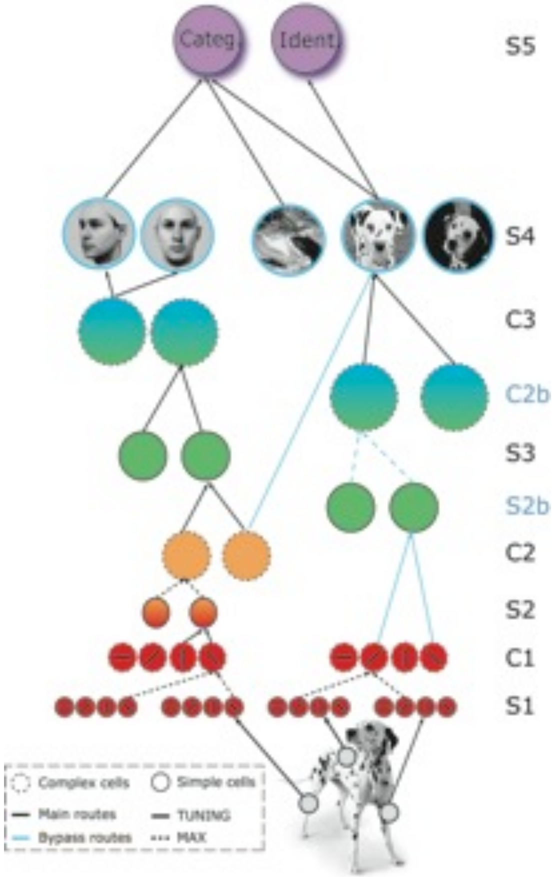
Performance

human agreement	72%
proposed system	77%
commercial system	61%
chance	12%

Models of cortex lead to better systems for action recognition in videos: automatic phenotyping of mice

Jhuang , Garrote, Yu, Khilnani, Poggio, Mutch Steele, Serre, Nature Communicatons, 2010

Motivation:
theory is needed!

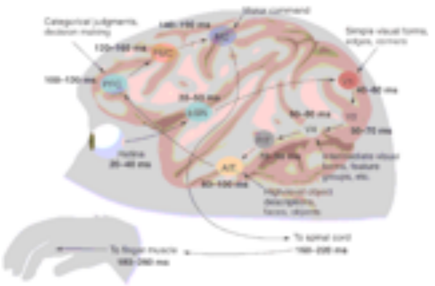


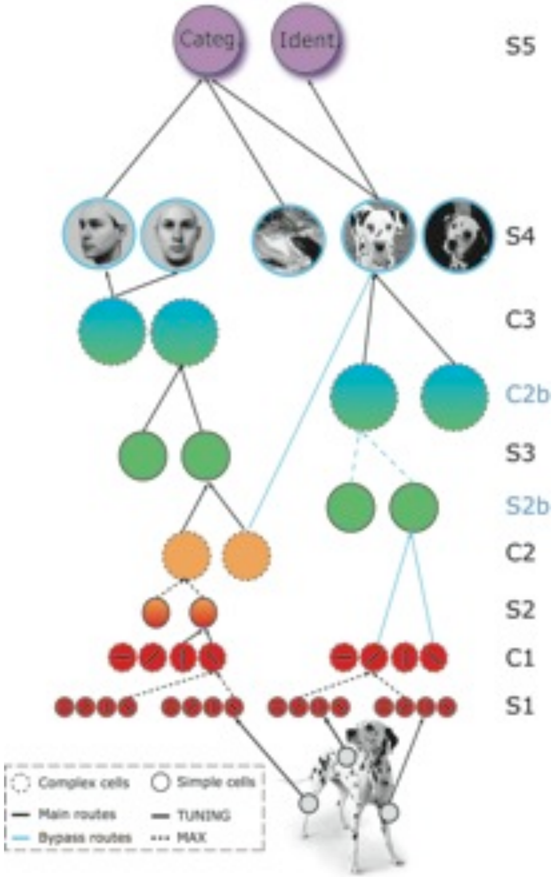
Hierarchical, Hubel and Wiesel (HMAX-type)
models
work well, as model of cortex and as
computer vision systems
but...why? and how can we improve them?

Similar convolutional networks
called deep learning networks
(LeCun, Hinton,...)
are
unreasonably successful
in vision and speech (ImageNet+Timit)...

why?

Motivation:
theory is needed!



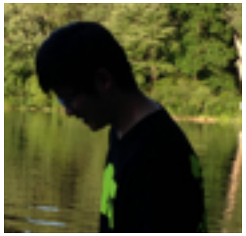


Hierarchical, Hubel and Wiesel (HMAX-type)
models
work well, as model of cortex and as
computer vision systems
but...why? and how can we improve them?

Similar convolutional networks
called deep learning networks
(LeCun, Hinton,...)
are
unreasonably successful
in vision and speech (ImageNet+Timit)...

why?

Collaborators (MIT-IIT, LCSL) in recent work



F. Anselmi, J. Mutch , J. Leibo, L. Rosasco, A. Tacchetti, Q. Liao

+ +

Evangelopoulos, Zhang, Voinea

Also: L. Isik, S. Ullman, S. Smale, C. Tan, M. Riesenhuber, T. Serre, G. Kreiman, S. Chikkerur, A. Wibisono, J. Bouvrie, M. Kouh, J. DiCarlo, C. Cadieu, S. Bileschi, L. Wolf, D. Ferster, I. Lampl, N. Logothetis, H. Buelthoff

Plan

1. Motivation: models of cortex (and deep convolutional networks)
2. Core theory
    - the basic invariance module
    - the hierarchy
3. Computational performance
4. Biological predictions
5. Theorems and remarks
    - $n \to 1$
    - invariance and sample complexity
    - connections with scattering transform
    - invariances and beyond perception
    - ...

Theory: underlying hypothesis

The main computational goal of the *feedforward* ventral stream hierarchy is to compute a representation for each incoming image which is invariant to transformations previously experienced in the visual environment.

Remarks:

- A *theorem* shows that invariant representations may reduce by orders of magnitude the sample complexity of a classifier at the top of the hierarchy
- Empirical evidence also supports the claim

Theory: underlying hypothesis

The main computational goal of the *feedforward* ventral stream hierarchy is to compute a representation for each incoming image which is invariant to transformations previously experienced in the visual environment.

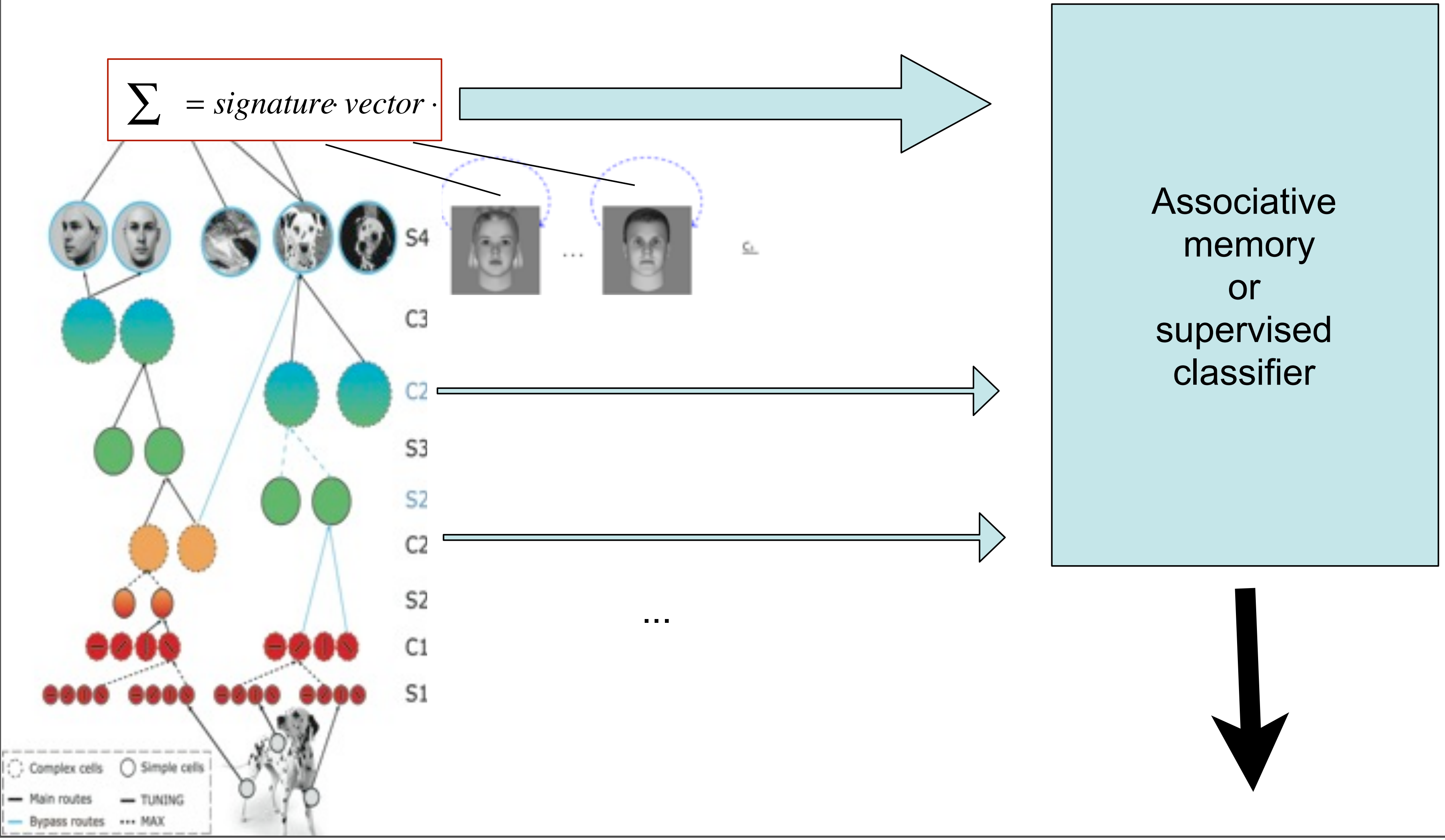
Features do not matter!

Remarks:

- A *theorem* shows that invariant representations may reduce by orders of magnitude the sample complexity of a classifier at the top of the hierarchy

- Empirical evidence also supports the claim

Use of invariant representation ---> signature vectors for memory access at several levels of the hierarchy



$\sum \; = signature\cdot vector\cdot$

Associative memory or supervised classifier

Neuroscience constraints on image representations

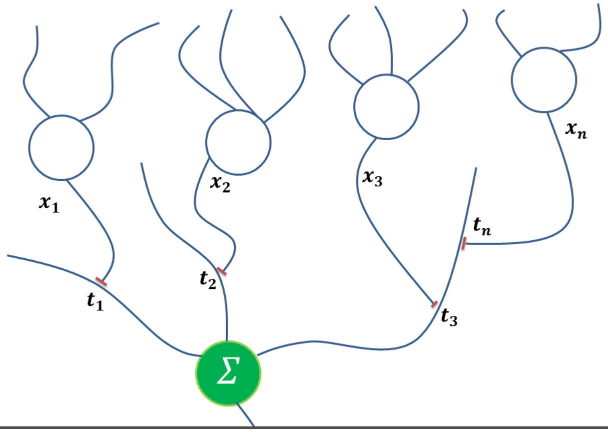
Remarks:

- Images can be represented by a set of functionals on the image, e.g. a set of measurements

- Neuroscience suggests that natural functionals for a neuron to compute is a high-dimensional dot product between an “image patch” and another image patch (called template) which is stored in terms of synaptic weights (synapses per neuron $\sim 10^2 - 10^5$)

- Projections via dot products are natural for neurons: here simple cells

Neuroscience definition of dot product!



Neuroscience constraints on image representations

Remarks:

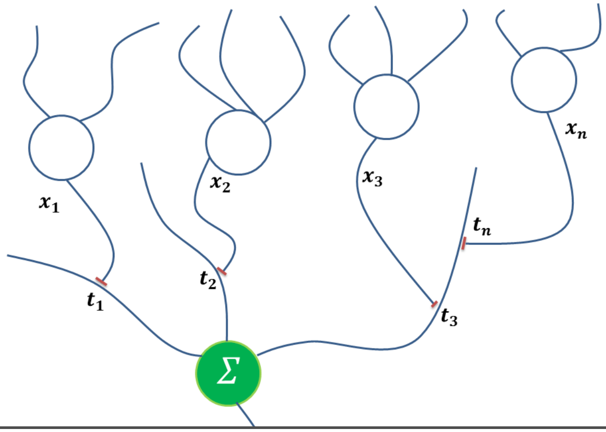
- Images can be represented by a set of functionals on the image, e.g. a set of measurements

- Neuroscience suggests that natural functionals for a neuron to compute is a high-dimensional dot product between an "image patch" and another image patch (called template) which is stored in terms of synaptic weights (synapses per neuron $\sim 10^2 - 10^5$)

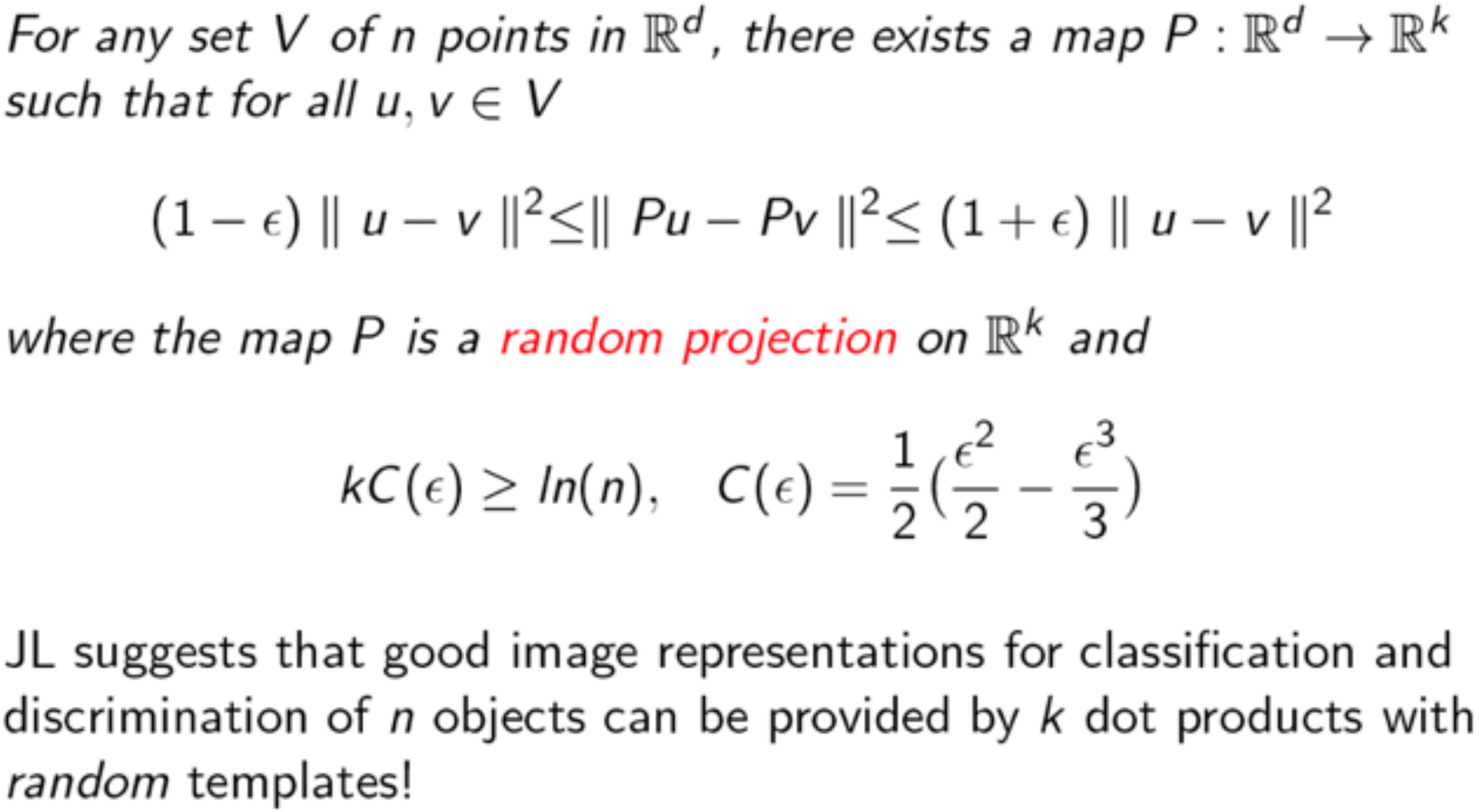
- Projections via dot products are natural for neurons: here simple cells

$$< x, t > \longleftrightarrow$$

Neuroscience definition of dot product!



Signatures: the Johnson-Lindenstrauss theorem (features do not matter much!)



For any set V of n points in $\mathbb{R}^d$, there exists a map $P : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$

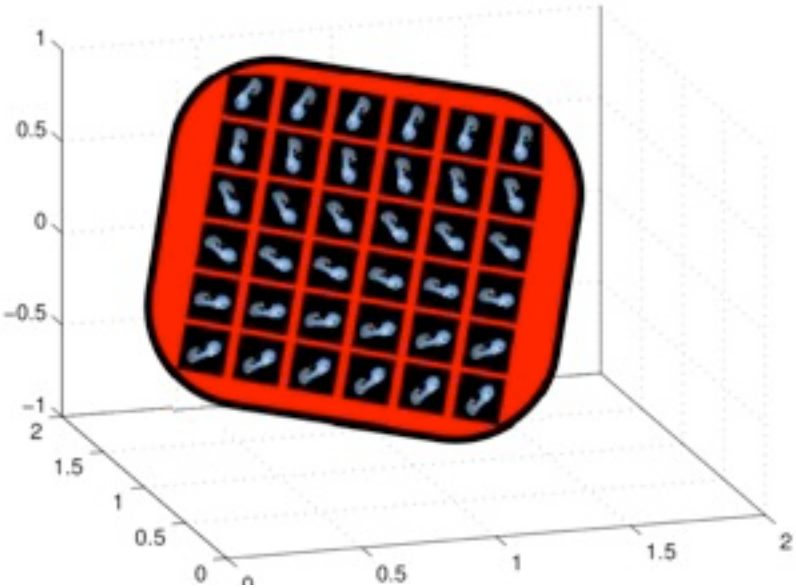
$$(1-\epsilon) \parallel u - v \parallel^2 \leq \parallel Pu - Pv \parallel^2 \leq (1+\epsilon) \parallel u - v \parallel^2$$

where the map P is a random projection on $\mathbb{R}^k$ and

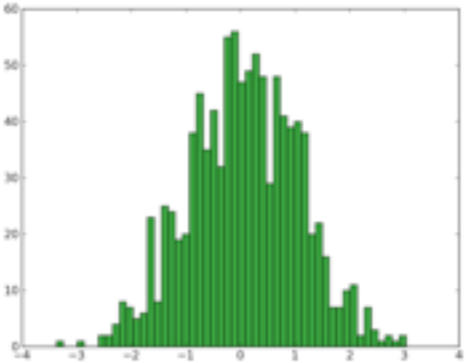
$$kC(\epsilon) \geq ln(n), \quad C(\epsilon) = \frac{1}{2}\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)$$

JL suggests that good image representations for classification and discrimination of n objects can be provided by k dot products with *random* templates!

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



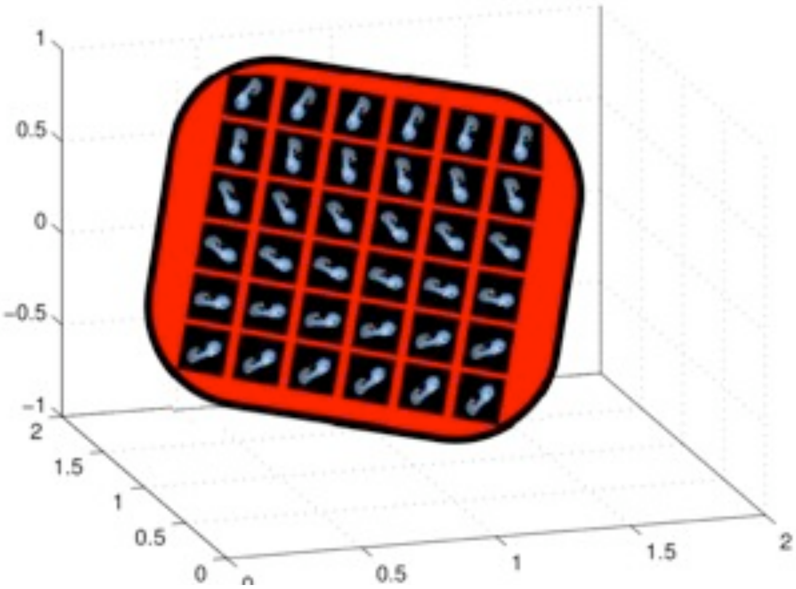
A template (e.g. a car,)
undergoes all in plane rotations



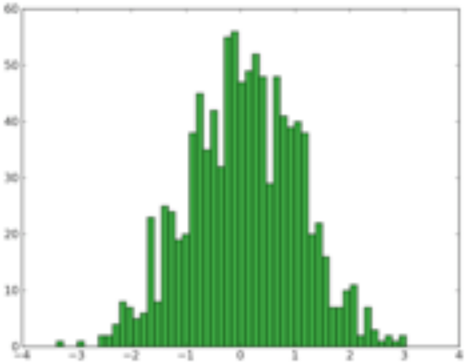
An histogram of the values of the
dot products of with the image
(e.g. a face) is computed.
Histogram gives a unique and
invariant image signature

poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



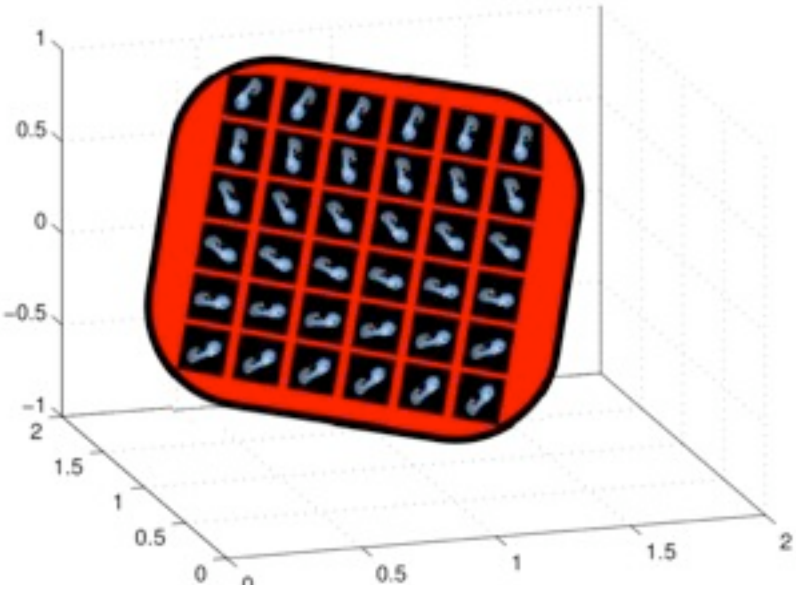
A template (e.g. a car, t)
undergoes all in plane rotations



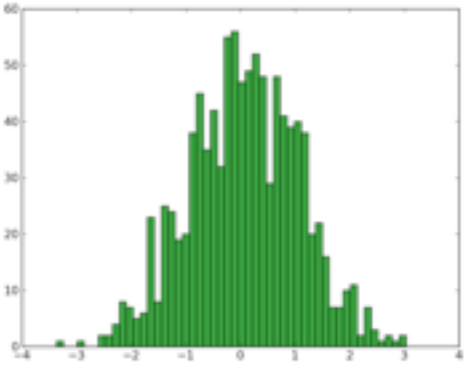
An histogram of the values of the
dot products of   with the image
(e.g. a face) is computed.
Histogram gives a unique and
invariant image signature

poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



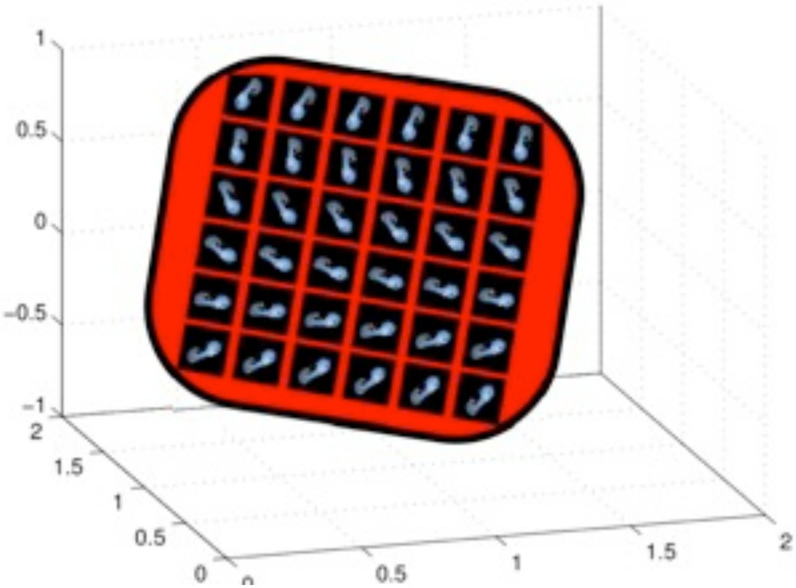
A template (e.g. a car, t)
undergoes all in plane rotations gt



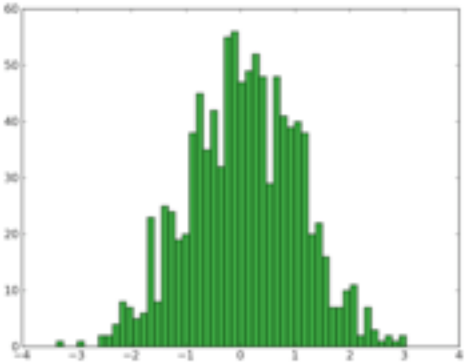
An histogram of the values of the
dot products of with the image
(e.g. a face) is computed.
Histogram gives a unique and
invariant image signature

poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



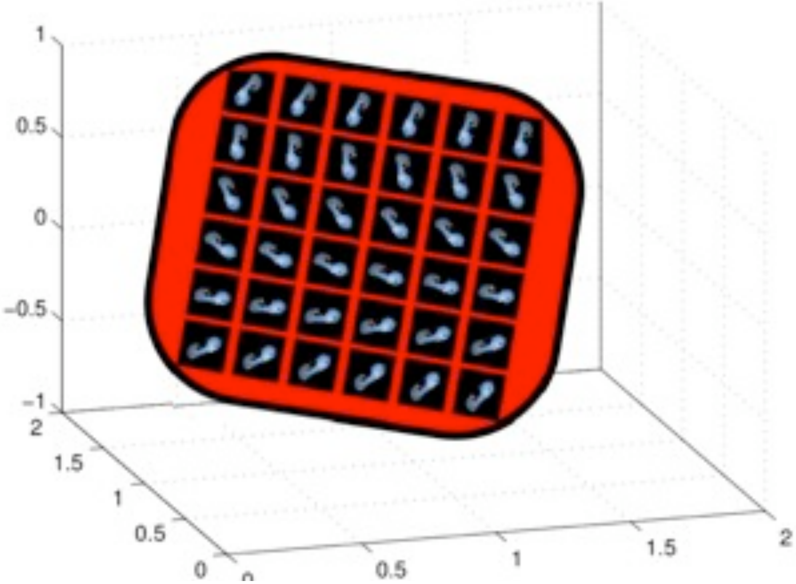
A template (e.g. a car, t) undergoes all in plane rotations gt



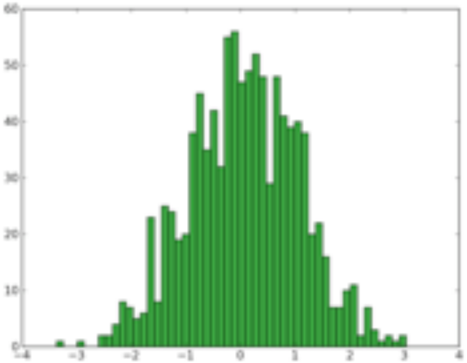
An histogram of the values of the dot products of gt with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature

poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



A template (e.g. a car, t)
undergoes all in plane rotations gt

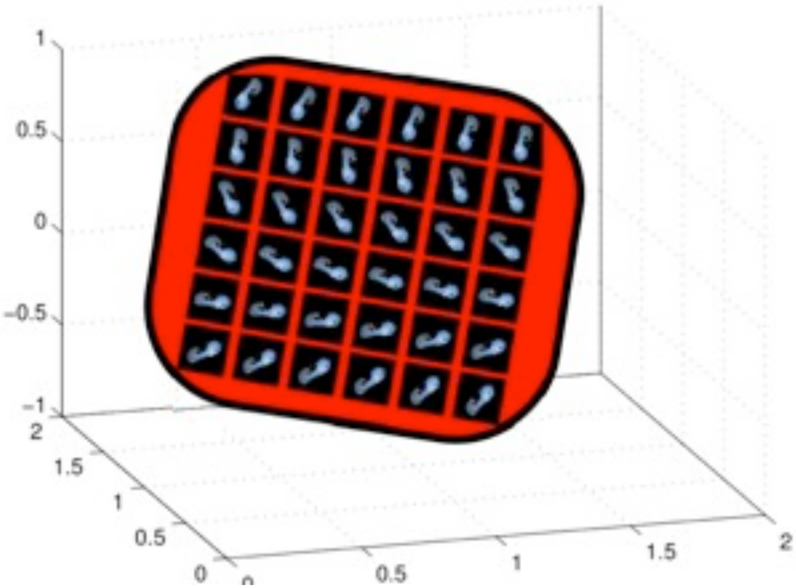


An histogram of the values of the
dot products of gt with the image
(e.g. a face) is computed.
Histogram gives a unique and
invariant image signature

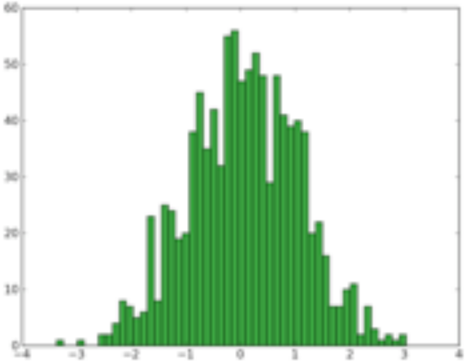
$,gt\rangle$

poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



A template (e.g. a car, t)
undergoes all in plane rotations gt

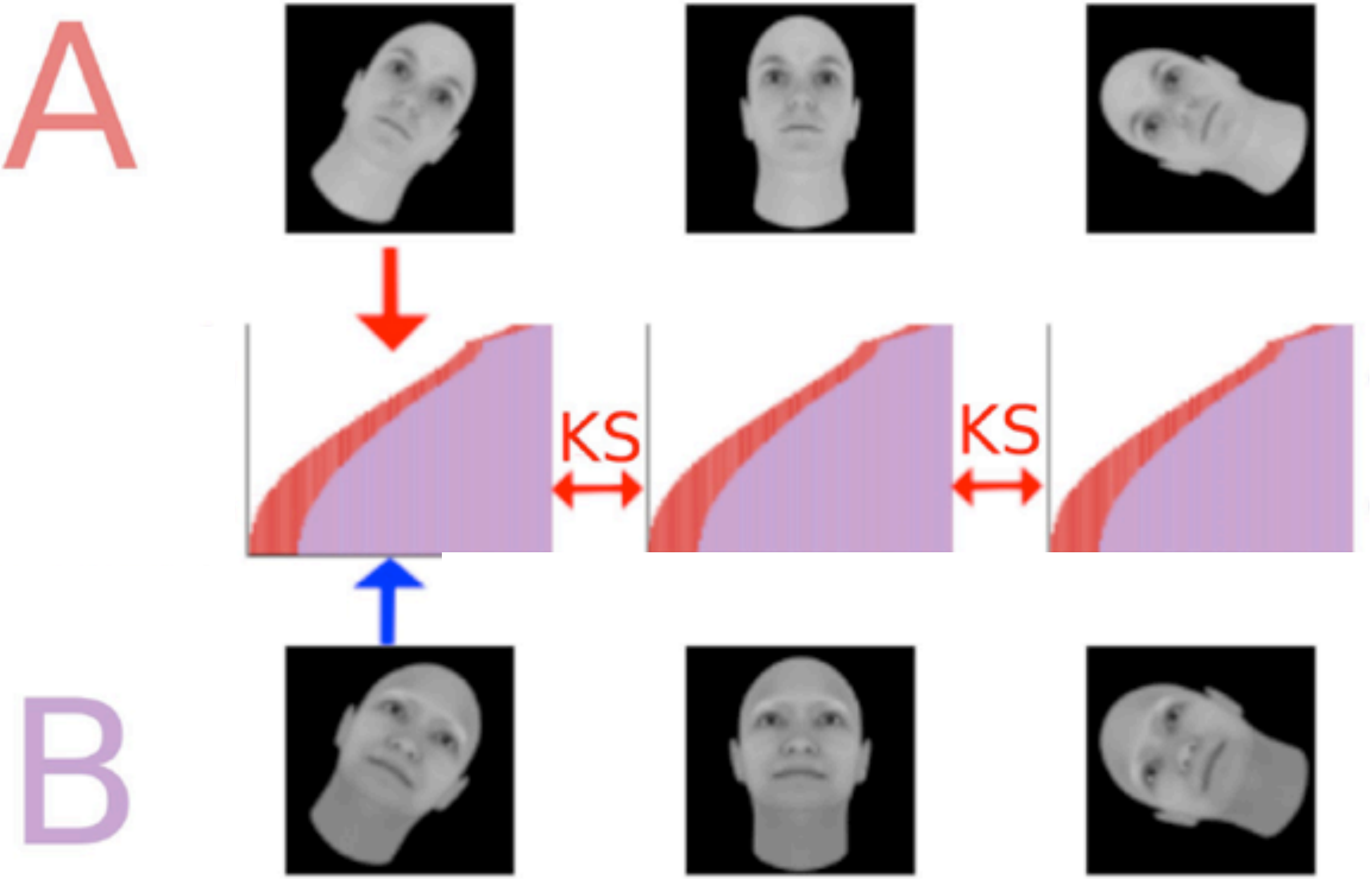


An histogram of the values of the
dot products of gt with the image
(e.g. a face) is computed.
Histogram gives a unique and
invariant image signature

$Hist\langle$ $,gt\rangle$

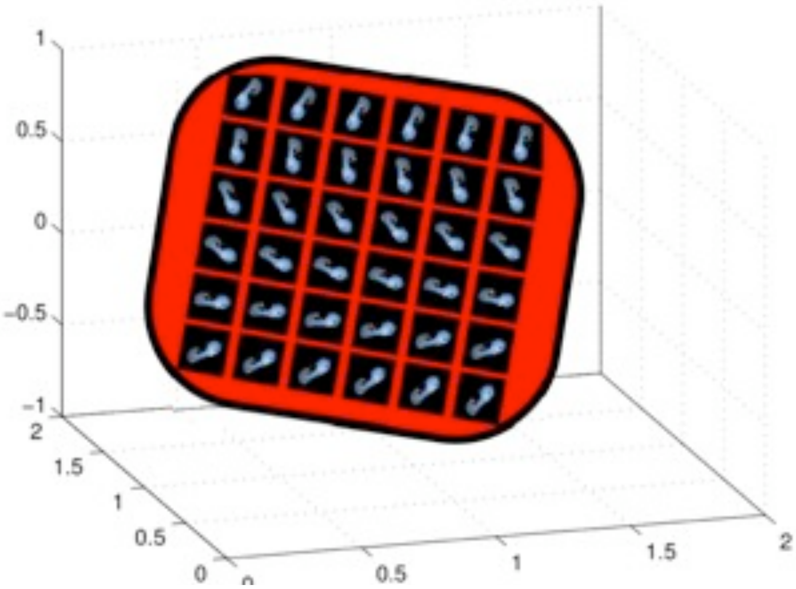
poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



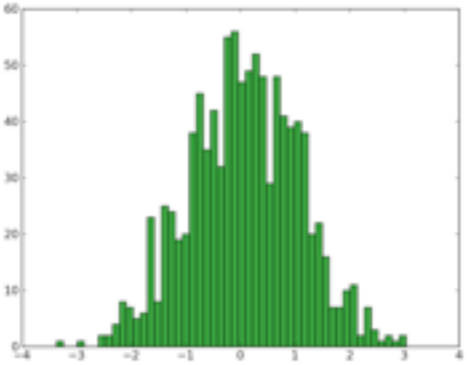
poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



A template (e.g. a car,   ) undergoes all in plane rotations

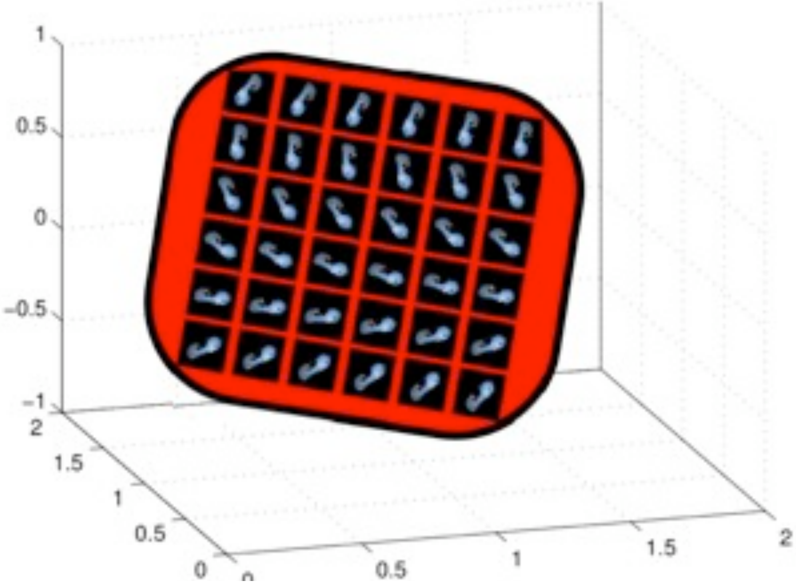
Random template could be used instead of car



An histogram of the values of the dot products of   with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature

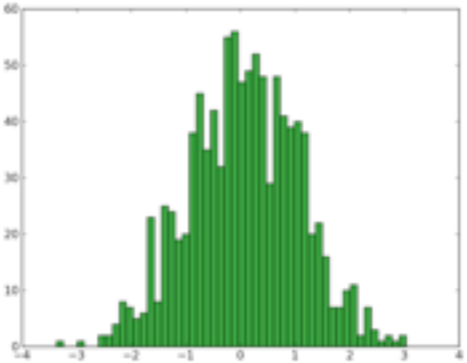
poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



A template (e.g. a car, t)
undergoes all in plane rotations

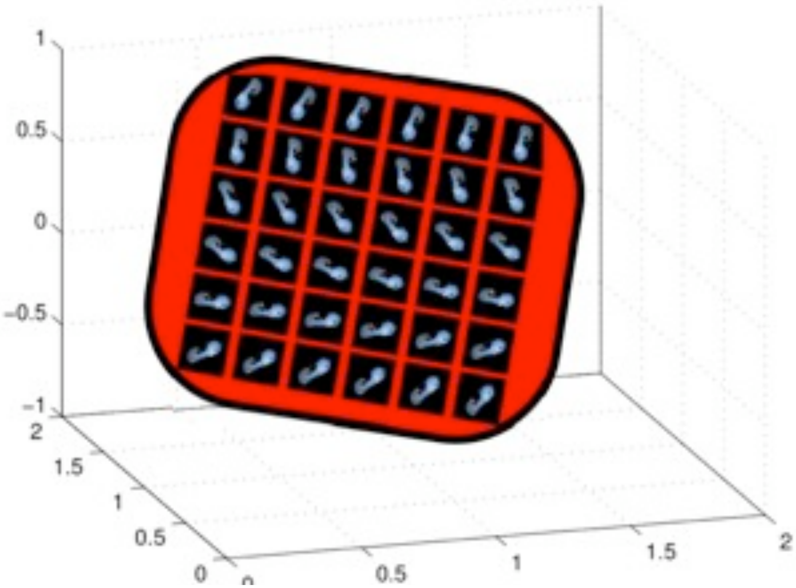
Random template could be
used instead of car



An histogram of the values of the
dot products of   with the image
(e.g. a face) is computed.
Histogram gives a unique and
invariant image signature

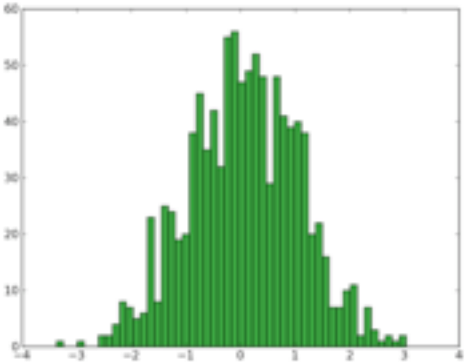
poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



A template (e.g. a car, t) undergoes all in plane rotations gt

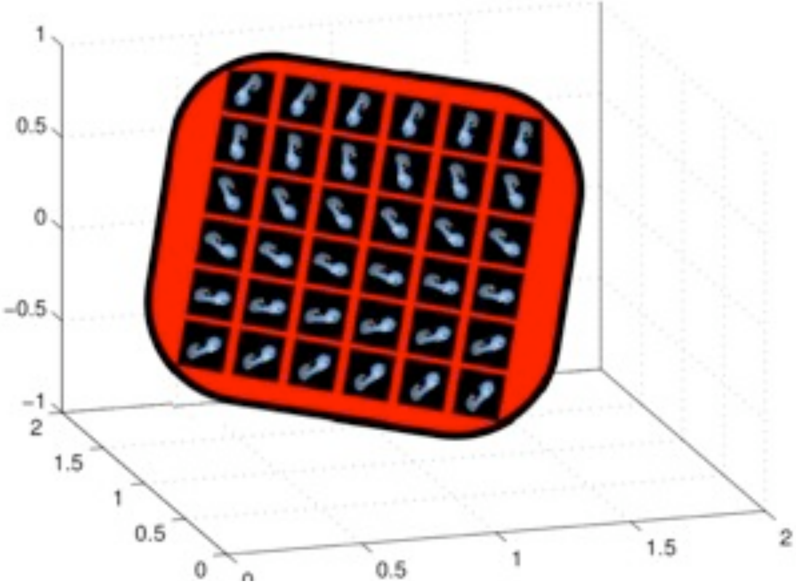
Random template could be used instead of car



An histogram of the values of the dot products of   with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature

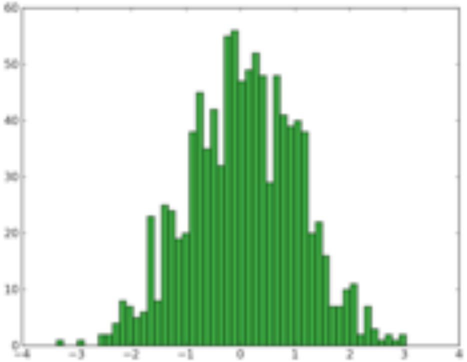
poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



A template (e.g. a car, t)
undergoes all in plane rotations gt

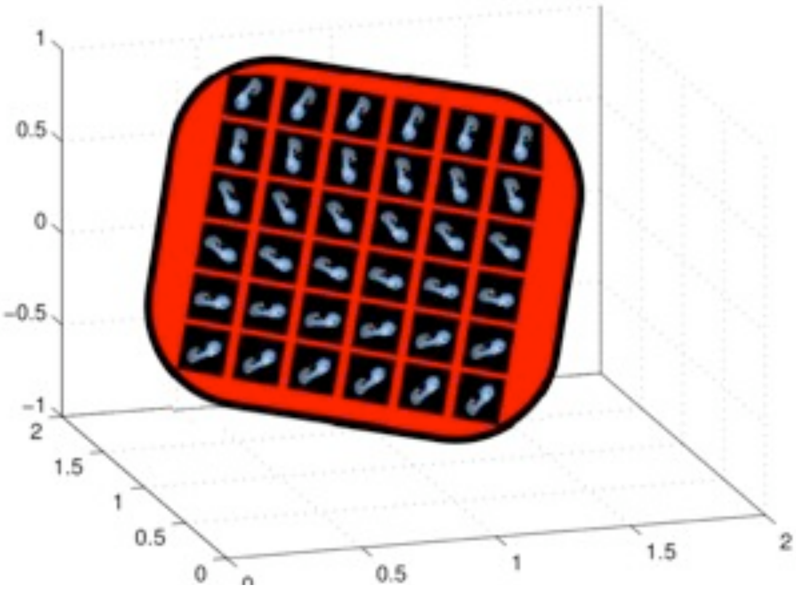
Random template could be
used instead of car



An histogram of the values of the
dot products of gt with the image
(e.g. a face) is computed.
Histogram gives a unique and
invariant image signature

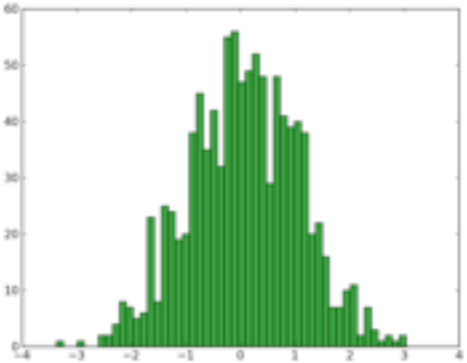
poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



A template (e.g. a car, t) undergoes all in plane rotations gt

Random template could be used instead of car

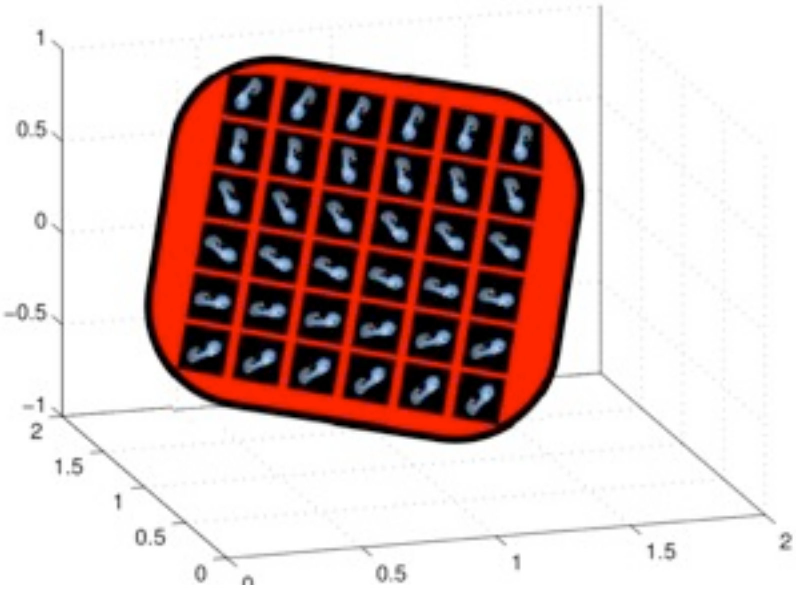


An histogram of the values of the dot products of gt with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature

$, gt\rangle$

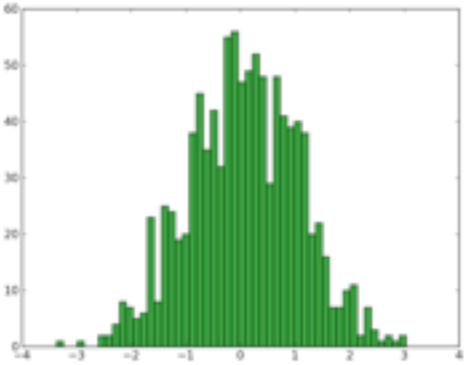
poggio, anselmi, rosasco, tacchetti, leibo, liao

Computing an invariant signature with the HW module (dot products and histograms of an image in a window)



A template (e.g. a car, t) undergoes all in plane rotations gt

Random template could be used instead of car



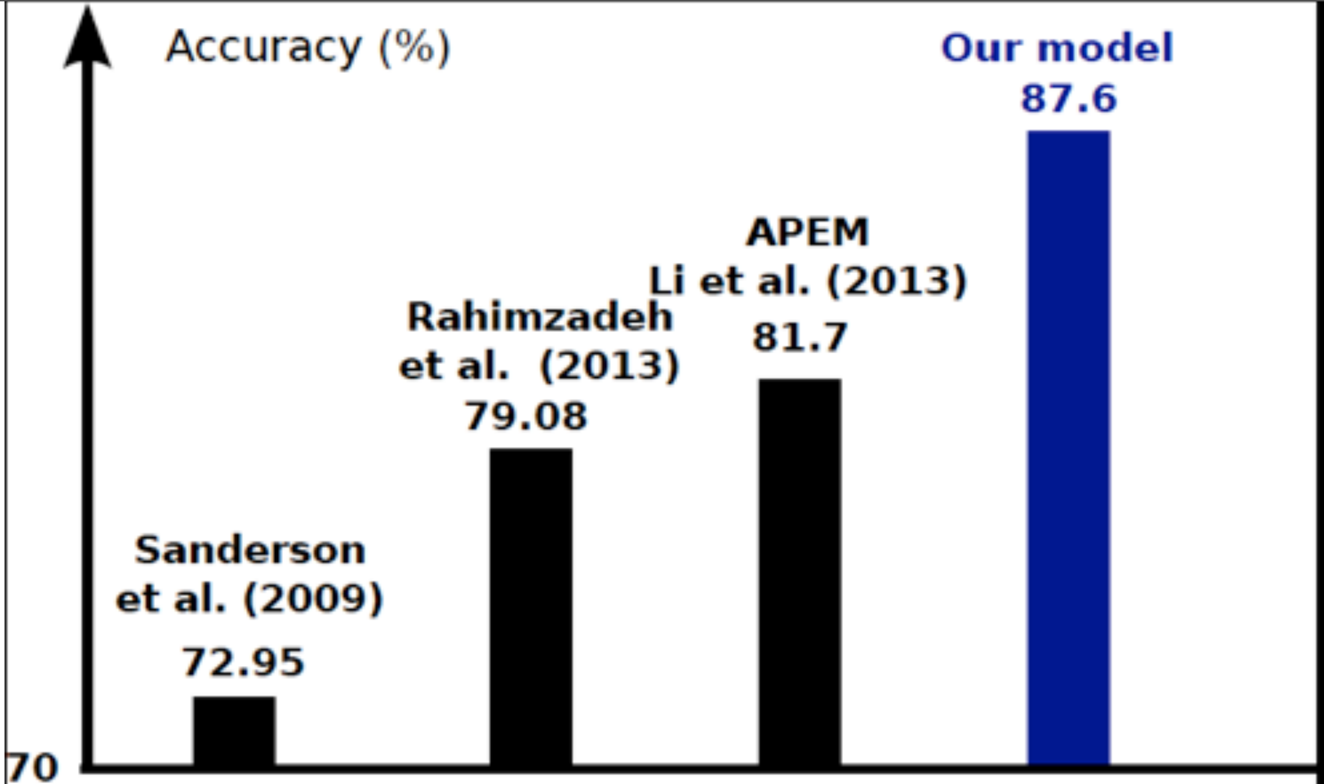
An histogram of the values of the dot products of gt with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature

$Hist\langle$ $,gt\rangle$

poggio, anselmi, rosasco, tacchetti, leibo, liao

This is it

- The basic HW module works for all transformations (no need to know anything about it, just collect unlabeled videos)
- Recipe:
  - memorize a set of images/objects called templates
  - for each template memorize observed transformations
  - to generate an representation/signature invariant to those transformation for each template
    - compute dot products of its transformations with image
    - compute histogram of the resulting values
- The same rule works on many types of transformations:
    - affine in 2D, image blur, image undersampling,...
    - 3D pose for faces, pose for bodies, perspective deformations, color constancy, aging, face expressions,...



LFW - no outside data used & no alignment

Q. Liao, J. Leibo

I want to get into more detail of two points here:

1. invariant representations are good because they reduce sample complexity

2. theorems on the magic of computing a good representation

Motivation

Cardinality of the universe of possible images generated by an object:

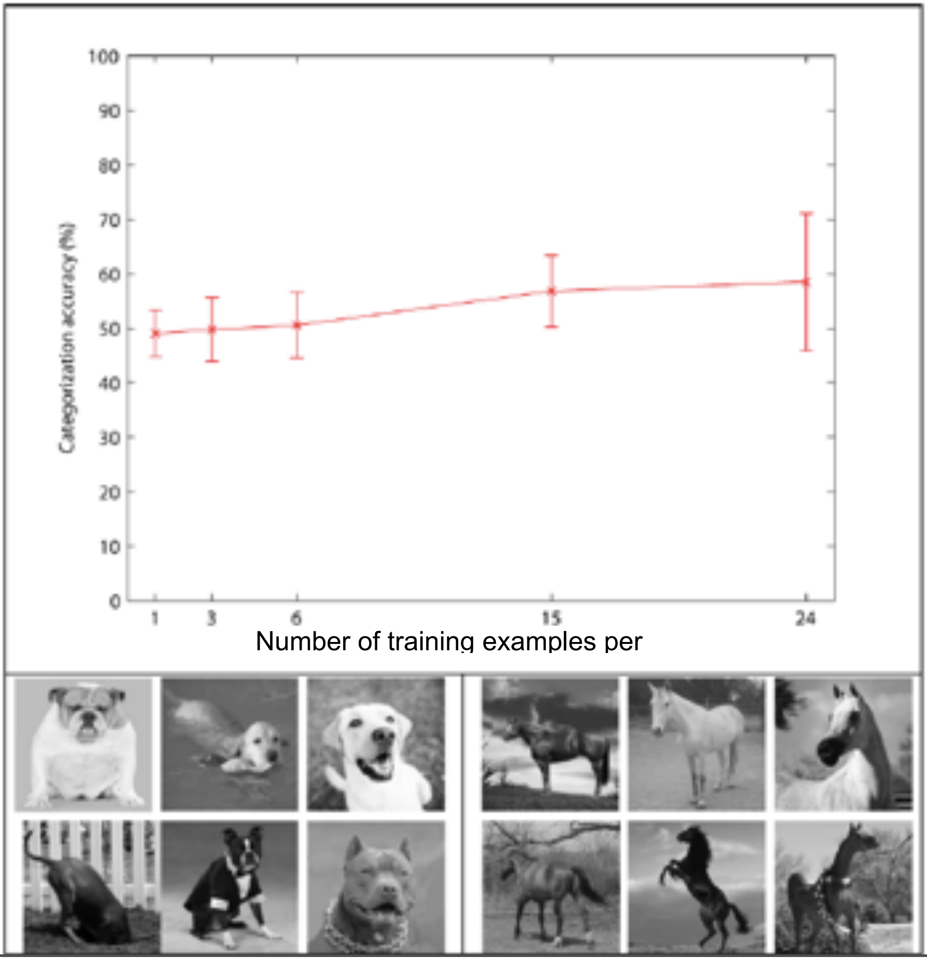
- Assuming: a granularity of a few minutes of arc + a visual field of say 10 degrees
- then
  - $10^3 - 10^5$ different images of the same object from x, y translations
  - $10^3 - 10^5$ from rotations in depth
  - a factor of $10 - 10^2$ from rotations in the image plane
  - another factor of $10 - 10^2$ from scaling.

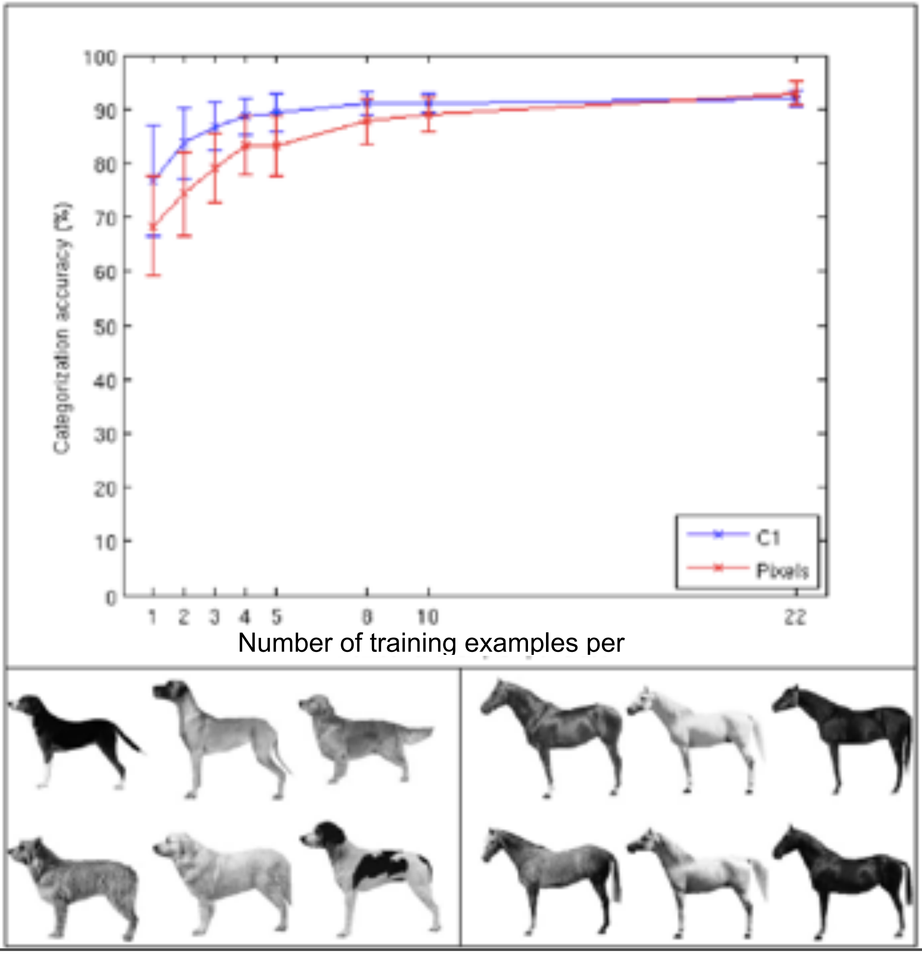
  for a total $10^8 - 10^{14}$ distinguishable images for a single object.

How many different types of dogs exist within the "dog" category? No more than, say, $10^2 - 10^3$. Thus it is greater win to be able to factor out the geometric transformations than the intracategory differences.

Learning how biology does learn - from very few labeled examples

Idea: unsupervised learning of invariant representations reduces number of labeled examples





Theory: underlying hypothesis

Invariance can significantly reduce sample complexity

Theorem *(translation case)* Consider a space of images of dimensions $d \times d$ pixels which may appear in any position within a window of size $rd \times rd$ pixels. The usual image representation yields a sample complexity (of a linear classifier) of order $m = O(r^2 d^2)$;the oracle representation (invariant) yields (because of much smaller covering numbers) a -- much better -- sample complexity of order

$$m_{oracle} = O(d^2) = \frac{m_{image}}{r^2}$$

A second phase in machine learning: a paradigm shift?

The first phase (and successes) of ML: supervised learning: $n \to \infty$

Siri
Use your voice to send messages, set reminders, search for information, and more.
Autonomous Driving

The next phase of ML: unsupervised learning of invariant representations for learning: $n \to 1$

Class 25

Learning Data Representations: beyond DeepLearning: the Magic Theory

Tomaso Poggio

I want to get into more detail of two points here:

1. invariant representations are good because they reduce sample complexity

2. theorems on the magic of computing a good representation

Overview of a “deep” theory

- Formal proofs --> *exact invariance* for generic images under group transformations using the basic HW module with generic templates (it is an *invariant Johnson-Lindenstrauss*-like embedding)

Transformation example: affine group

The action of a group transformation on an image I is defined as:

In the case of affine group:

Transformation example: affine group

The action of a group transformation g on an image I is defined as:

In the case of affine group:

Transformation example: affine group

The action of a group transformation g on an image I is defined as:

$$gI(\vec{x}) = I(g^{-1}\vec{x})$$

In the case of affine group:

Transformation example: affine group

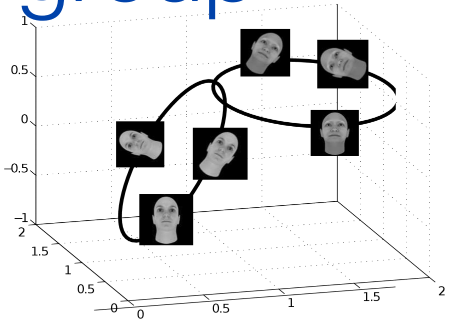
The action of a group transformation g on an image I is defined as:

$$gI(\vec{x}) = I(g^{-1}\vec{x})$$

In the case of affine group:

$$gI(\vec{x}) = I(A^{-1}\vec{x} - \vec{b}), \qquad A \in GL(2), \vec{b} \in R^2$$

Theorems for the compact group



The image orbit and its associated probability distribution is invariant and unique

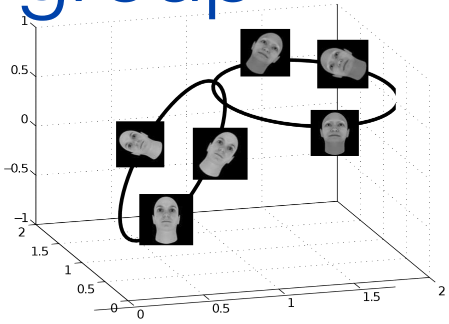
For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates

Theorems for the compact group

$$I \sim I' \Leftrightarrow O_I = O_{I'} \Leftrightarrow P_I = P_{I'}$$



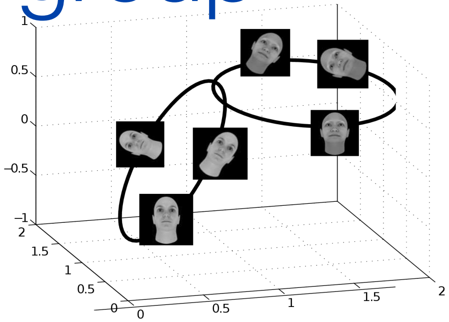
The image orbit and its associated probability distribution is invariant and unique

For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates

Theorems for the compact group



$$I \sim I' \Leftrightarrow O_I = O_{I'} \Leftrightarrow P_I = P_{I'}$$

The image orbit and its associated probability distribution is invariant and unique

$$P_{\langle I, t^k \rangle}$$

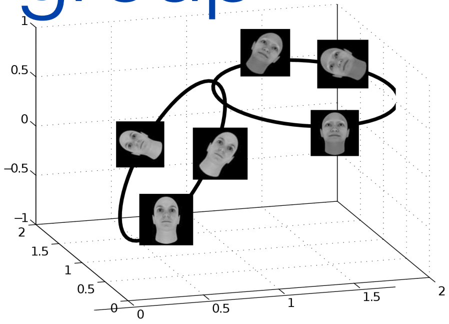
For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates

Theorems for the compact group

$$I \sim I' \Leftrightarrow O_I = O_{I'} \Leftrightarrow P_I = P_{I'}$$



The image orbit and its associated probability distribution is invariant and unique

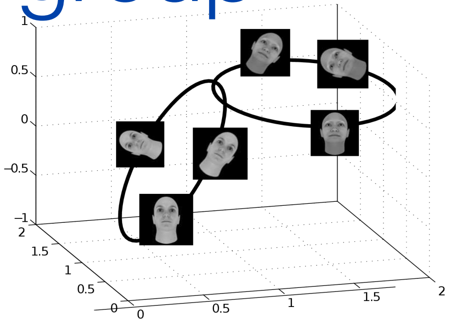
$$\begin{tikzcd}
P_I \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & P_{\langle I, t^k \rangle} \arrow[d, leftrightarrow] \\
{} & {}
\end{tikzcd}$$

For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates

Theorems for the compact group



$$I \sim I' \Leftrightarrow O_I = O_{I'} \Leftrightarrow P_I = P_{I'}$$

The image orbit and its associated probability distribution is invariant and unique

$$\begin{tikzcd} P_I \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & P_{\langle I,t^k \rangle} \arrow[d, leftrightarrow] \\ gI & {} \end{tikzcd}$$

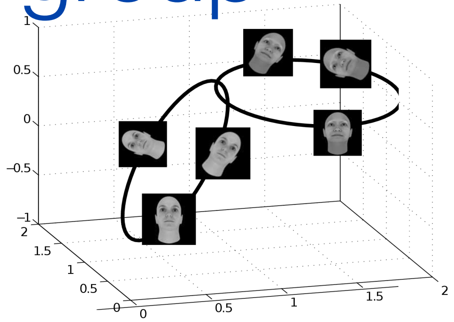
For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates

Theorems for the compact group

$$I \sim I' \Leftrightarrow O_I = O_{I'} \Leftrightarrow P_I = P_{I'}$$



The image orbit and its associated probability distribution is invariant and unique

$$\begin{tikzcd}
P_I \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & P_{\langle I, t^k \rangle} \arrow[d, leftrightarrow] \\
gI & \langle gI, t^k \rangle = \langle I, \boxed{g^{-1} t^k} \rangle
\end{tikzcd}$$

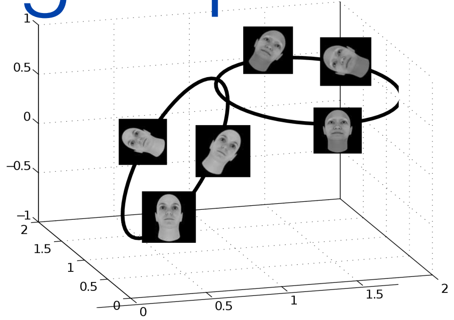
For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates

Theorems for the compact group

$$I \sim I' \Leftrightarrow O_I = O_{I'} \Leftrightarrow P_I = P_{I'}$$



The image orbit and its associated probability distribution is invariant and unique

$$\begin{tikzcd}
P_I \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & P_{\langle I,t^k\rangle} \arrow[d, leftrightarrow] \\
gI & \langle gI,t^k\rangle = \langle I,\boxed{g^{-1}t^k}\rangle
\end{tikzcd}$$

For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

$t^k, k = 1,\ldots,K$: set of templates

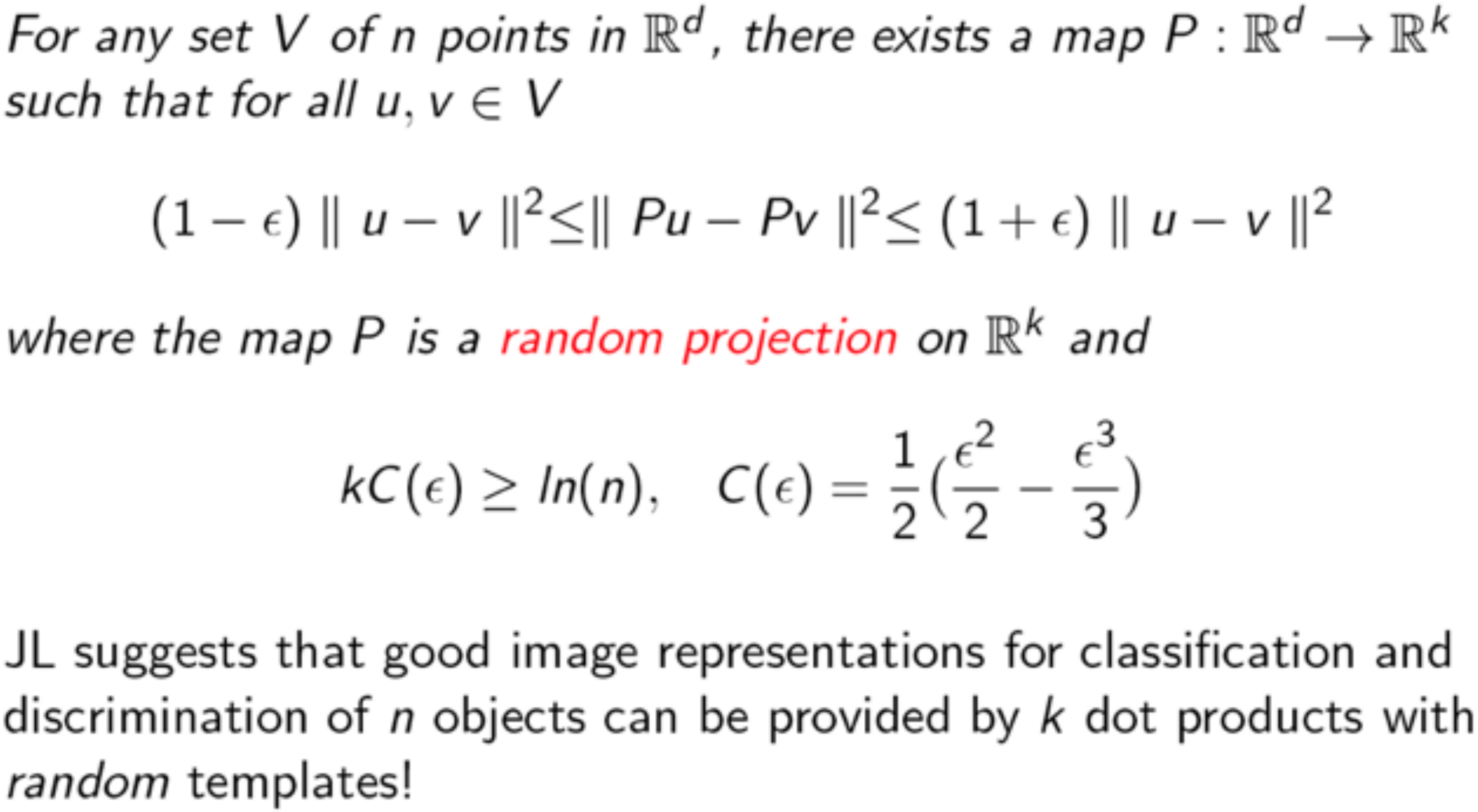
Probability distribution from finite projections

Theorem: Consider n images I_j in X_n. Let $K \geq \frac{c}{\varepsilon^2} \log \frac{n}{\delta}$ where c is a universal constant. Then

$$| d(P_I - P_{I'}) - \hat{d}_K (P_I - P_I ') | \leq \varepsilon$$

with probability $1 - \delta^2$, for all $I, I' \in \mathrm{X}_n$.

A motivation for signatures: the Johnson-Lindenstrauss theorem (features do not matter much!)



For any set V of n points in $\mathbb{R}^d$, there exists a map $P : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$

$$(1-\epsilon) \parallel u - v \parallel^2 \leq \parallel Pu - Pv \parallel^2 \leq (1+\epsilon) \parallel u - v \parallel^2$$

where the map P is a random projection on $\mathbb{R}^k$ and

$$kC(\epsilon) \geq ln(n), \quad C(\epsilon) = \frac{1}{2}(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3})$$

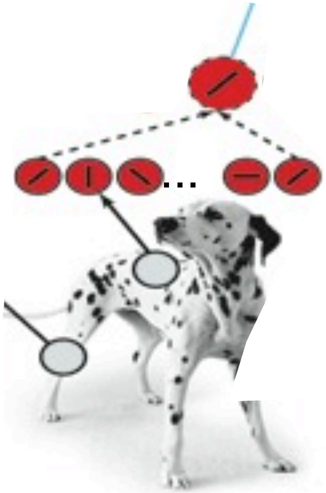
JL suggests that good image representations for classification and discrimination of n objects can be provided by k dot products with *random* templates!

Our basic machine: a HW module

(dot products and histograms for an image in a receptive field window)

- The signature provided by complex cells at each "position" is associated with histograms of the simple cells activities that is

$$\mu_n^k(I) = \frac{1}{|G|}\sum_{i=1}^{|G|}\sigma(\langle I, g_i t^k \rangle + n\Delta)$$

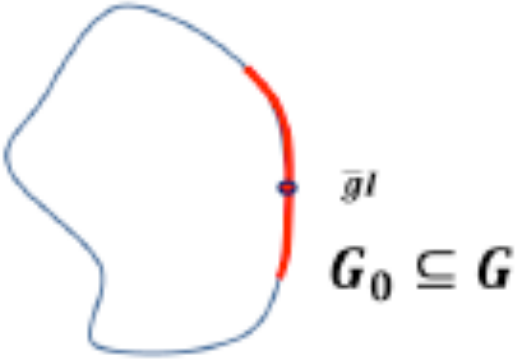


- Related quantities such as moments of the distributions are also invariant, for instance as computed by the energy model of complex cells or the max, related to the sup norm ---> we have a *full theory of pooling*

- Neural computation/represnetation of a histogram requires a set of complex cells -- neurons with different thresholds

- Histograms provide uniqueness independently of pooling range

Images, groups and orbit



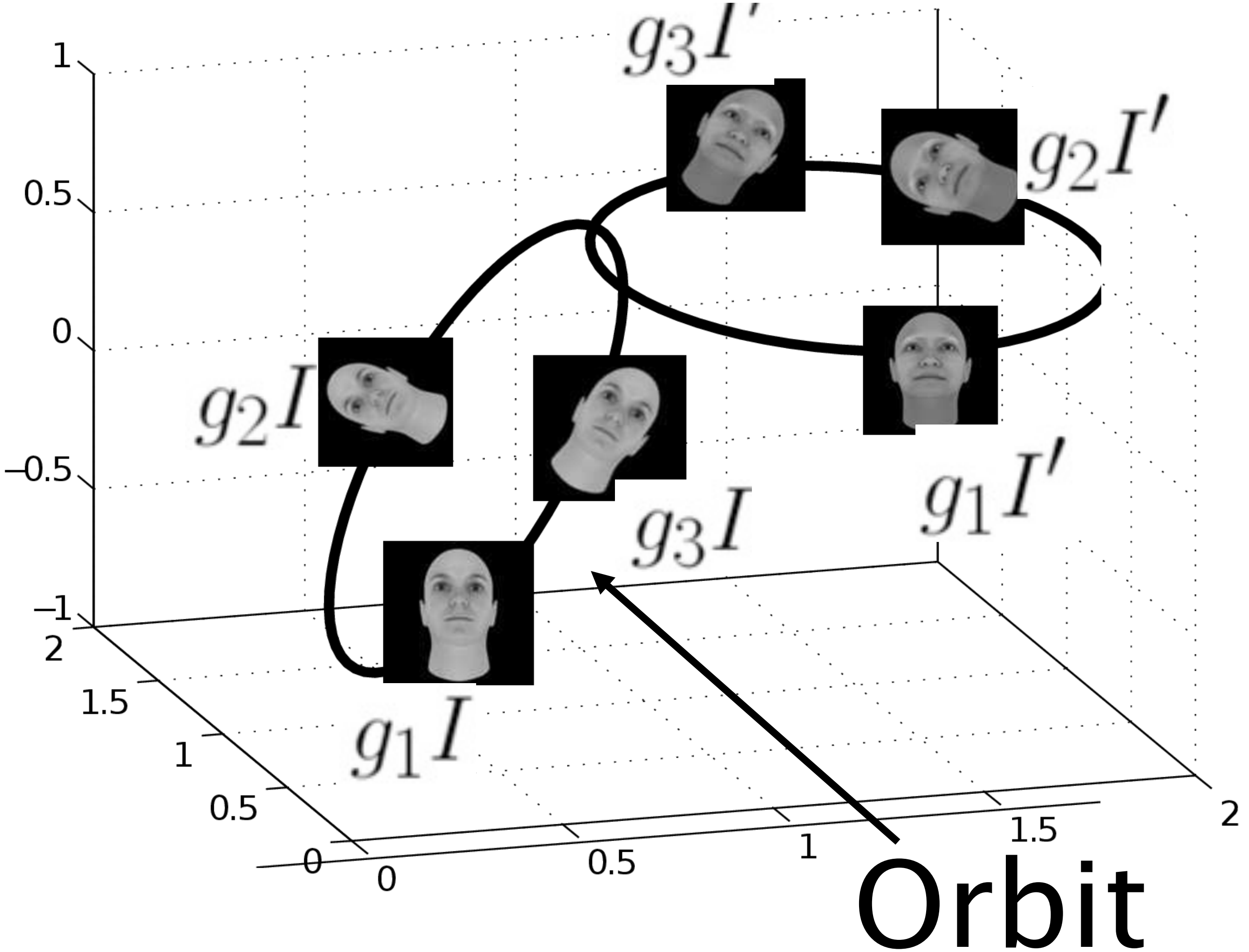
Orbit O_I

$I \sim I' if \quad \exists g \in G$ s.t. $I' = gI$

Orbit O_I can be proved to be *invariant* and *unique*

Orbit is unique and invariant

$$I \sim I' \iff O_I = O_{I'}$$



Orbit: set of images gI generated from a single image I under the action of the group

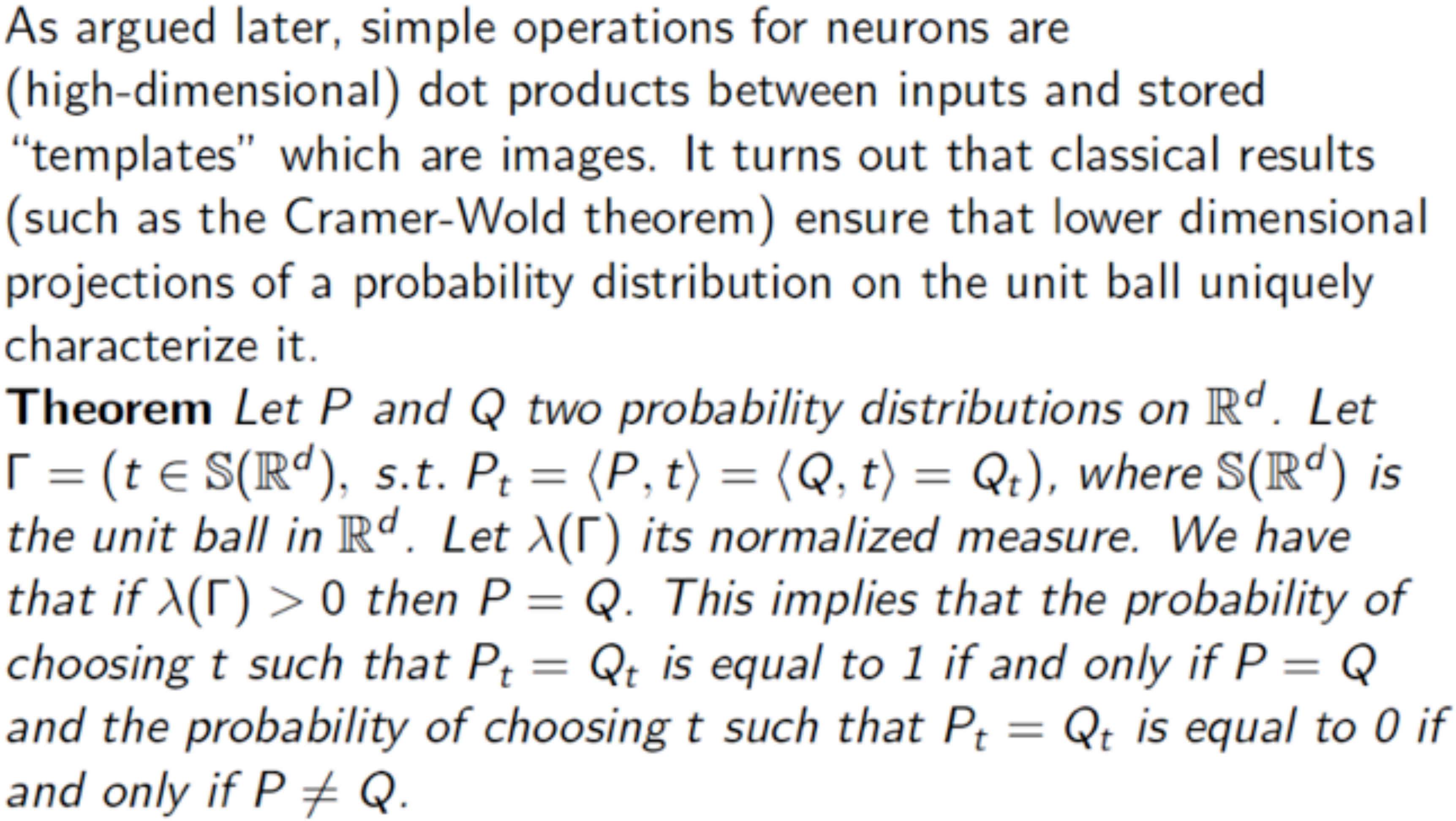
Preview: group invariance theorems

- An orbit is fully characterized by the probability density$P_G(gI)$

- An application of Cramer-Wold theorems suggests that that a proxy for $P_G(gI)$ is a set of K one-dimensional $P_G(< gI, t^k >)$

- Since $P_G(< gI, t^k >) = P_G(< I, g^{-1}t^k >)$ it is possible to get an invariant representation from a single image I if all transformations of t^k are stored.

Projections of Probabilities: Cramer-Wold



As argued later, simple operations for neurons are (high-dimensional) dot products between inputs and stored "templates" which are images. It turns out that classical results (such as the Cramer-Wold theorem) ensure that lower dimensional projections of a probability distribution on the unit ball uniquely characterize it.

Theorem *Let P and Q two probability distributions on $\mathbb{R}^d$. Let $\Gamma = (t \in \mathbb{S}(\mathbb{R}^d),\ s.t.\ P_t = \langle P, t \rangle = \langle Q, t \rangle = Q_t)$, where $\mathbb{S}(\mathbb{R}^d)$ is the unit ball in $\mathbb{R}^d$. Let $\lambda(\Gamma)$ its normalized measure. We have that if $\lambda(\Gamma) > 0$ then $P = Q$. This implies that the probability of choosing t such that $P_t = Q_t$ is equal to 1 if and only if $P = Q$ and the probability of choosing t such that $P_t = Q_t$ is equal to 0 if and only if $P \neq Q$.*

Invariant projections theorem

Consider

$$d(P_I, P_{I'}) = \int d_0(P_{\langle I,t\rangle}, P_{\langle I',t\rangle}) d\lambda(t), \quad \forall I, I' \in \mathcal{X},$$

$$d(P_I, P_{I'}) \approx \int d_\mu(\mu^t(I), \mu^t(I')) d\lambda(t), \quad \forall I, I' \in \mathcal{X},$$

where d_μ is a metric on histograms induced by d_0.

$$d_\mu(\mu^k(I), \mu^k(I')) = \left\|\mu^k(I) - \mu^k(I)\right\|_{\mathbb{R}^N}$$

where $\|\cdot\|_{\mathbb{R}^N}$ is the Euclidean norm in $\mathbb{R}^N$

Theorem *Consider n images $\mathcal{X}_n$ in $\mathcal{X}$. Let $K \geq \frac{c}{\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant. Then*

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| \leq \epsilon,$$

with probability $1 - \delta^2$, for all $I, I' \in \mathcal{X}_n$.

Plan

1. Motivation: models of cortex (and deep convolutional networks)
2. Core theory
    - the basic invariance module
    - the hierarchy
3. Computational performance
4. Biological predictions
5. Theorems and remarks
    - $n \to 1$
    - invariance and sample complexity
    - connections with scattering transform
    - invariances and beyond perception
    - ...

Implementations/specific models: computational performance

- Deep convolutional networks (such as Lenet) as an architecture are special case of Mtheory (with just translation invariance and max/sigmoid pooling)
- HMAX as an architecture is a special case of Mtheory (with translation + scale invariance and max pooling) and used to work well

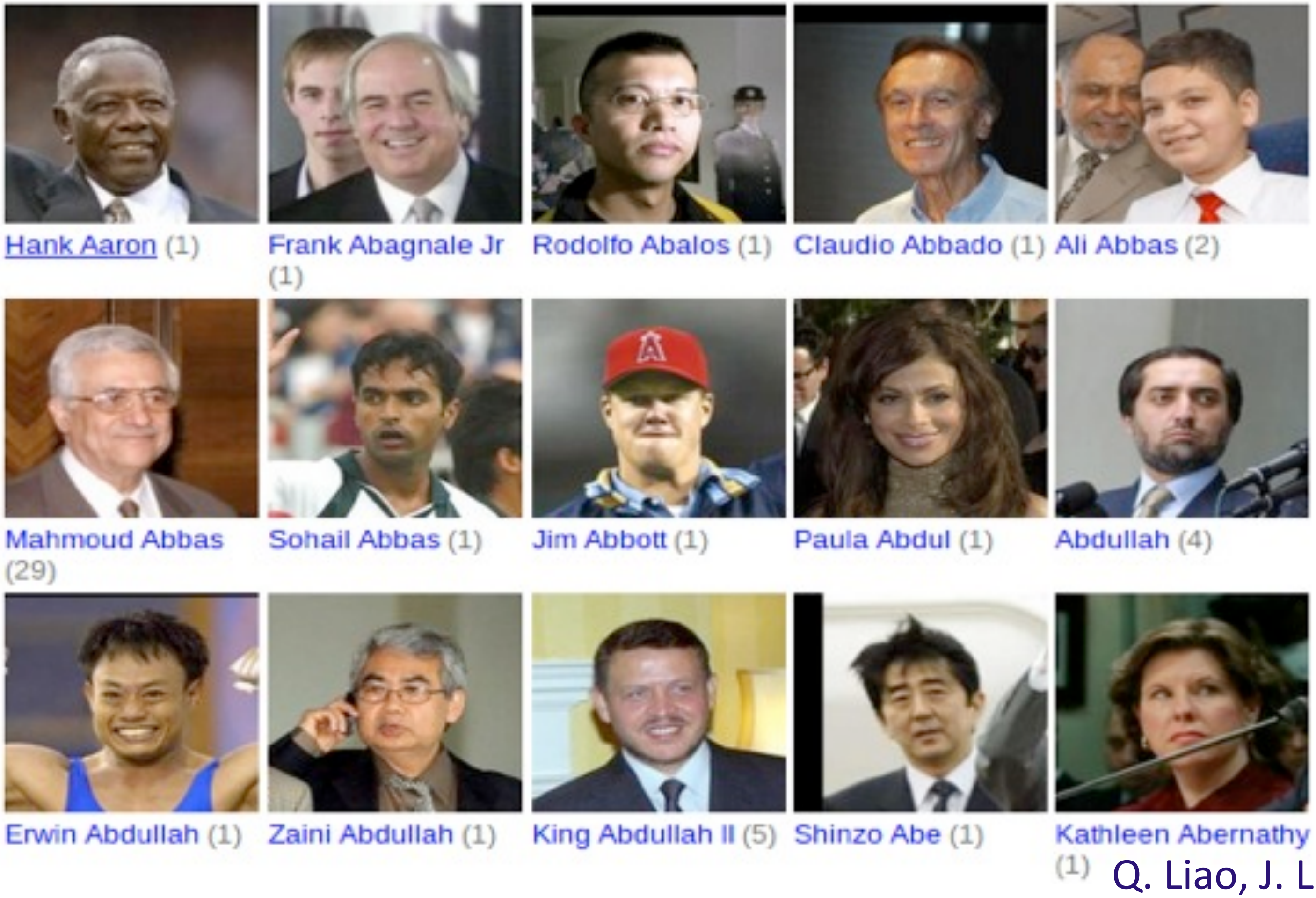
Models: computational performance

- Deep convolutional networks (such as Lenet) as an architecture are special case of Mtheory (with just translation invariance and max/sigmoid pooling)
- HMAX as an architecture is a special case of Mtheory (with translation + scale invariance and max pooling) and used to work well
- Encouraging initial results in speech and music classification (Evangelopoulos, Zhang, Voinea)
- Example in face identification (Liao, Leibo) --->

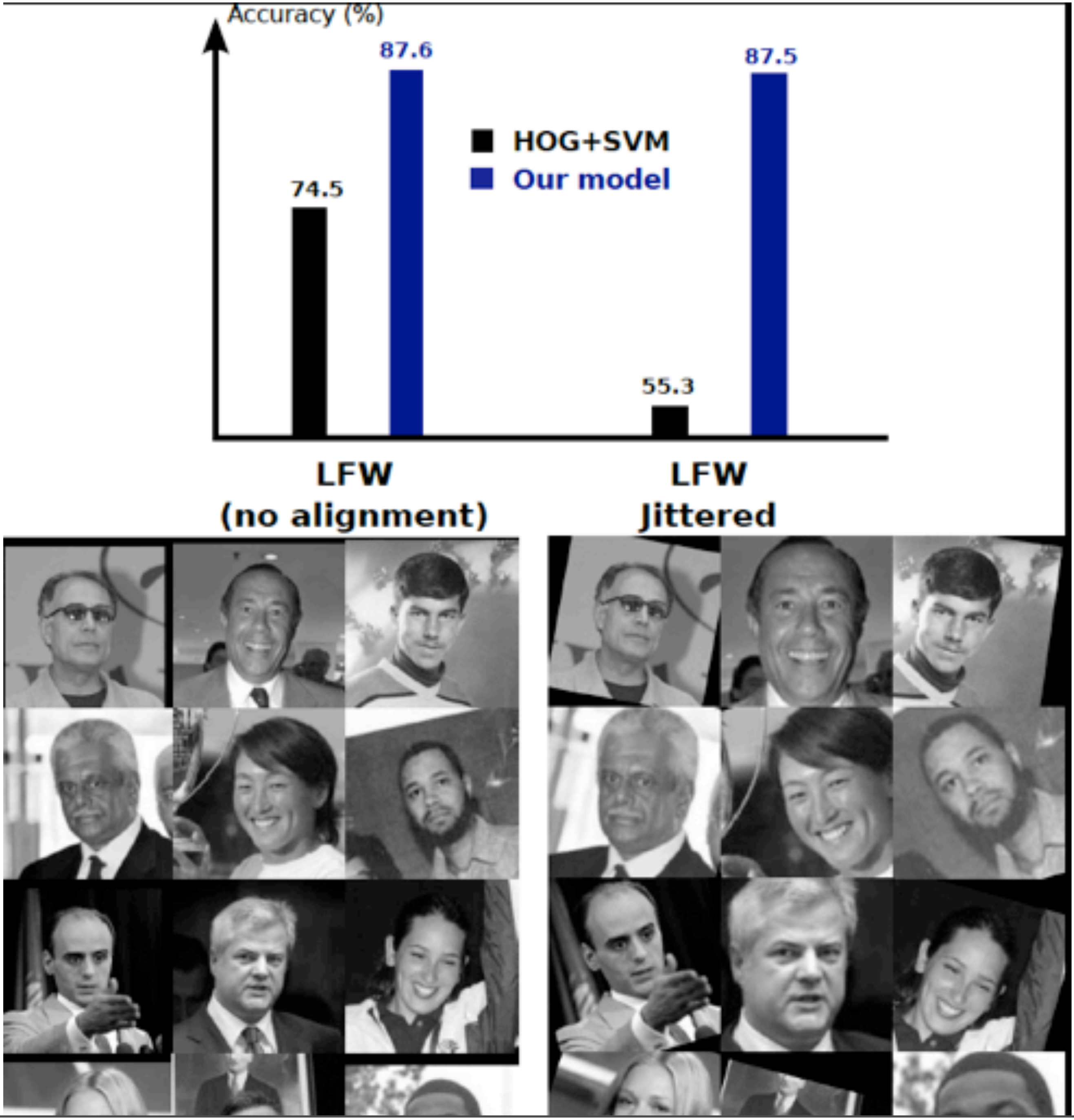
Computational performance: example faces

Labeled Faces in the Wild

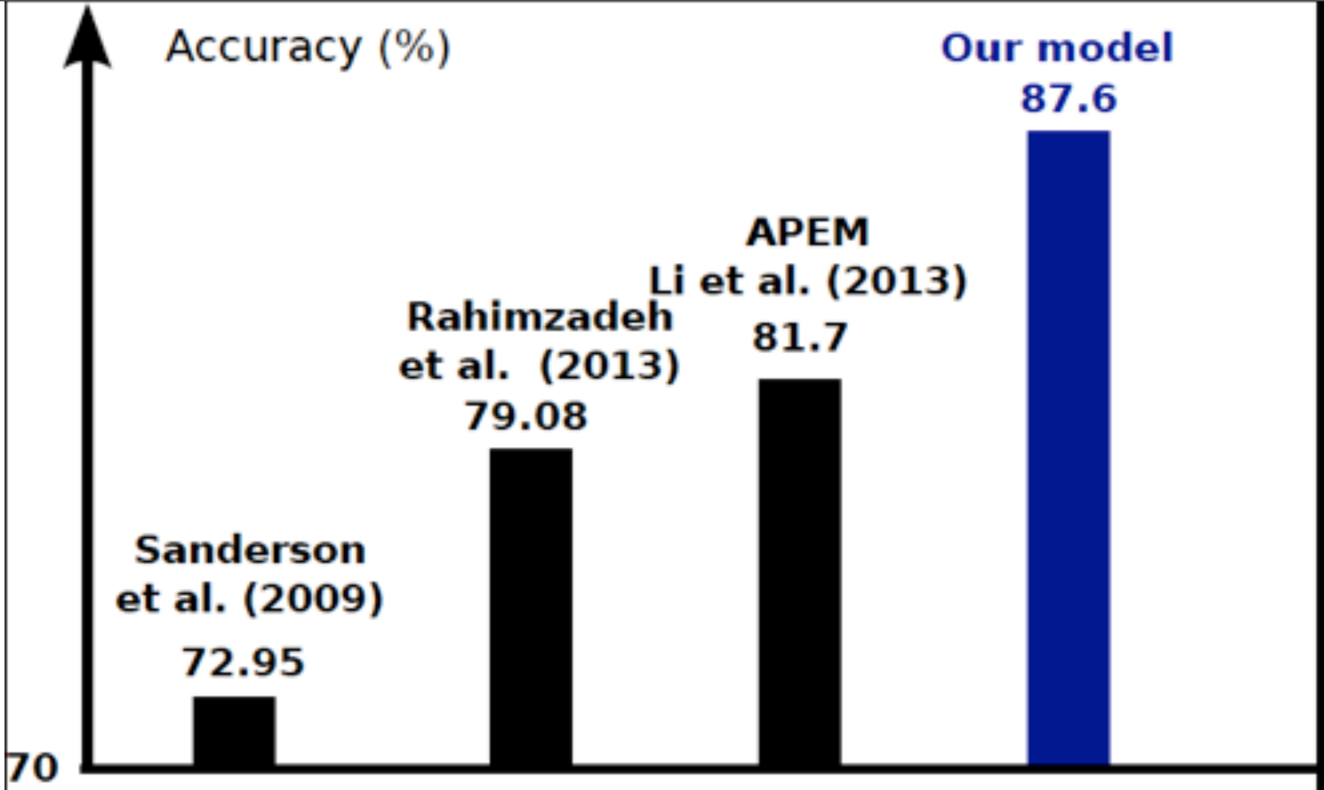
Contains 13,233 images of 5,749 people



Q. Liao, J. Leibo, NIPS 2013



Q. Liao, J. Leibo



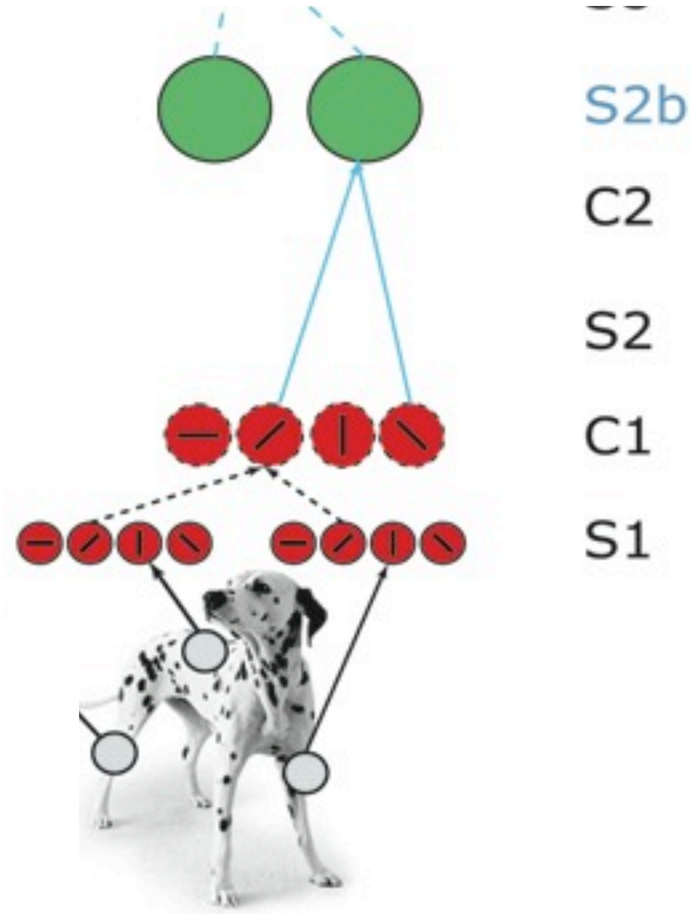
LFW - no outside data used & no alignment

Q. Liao, J. Leibo

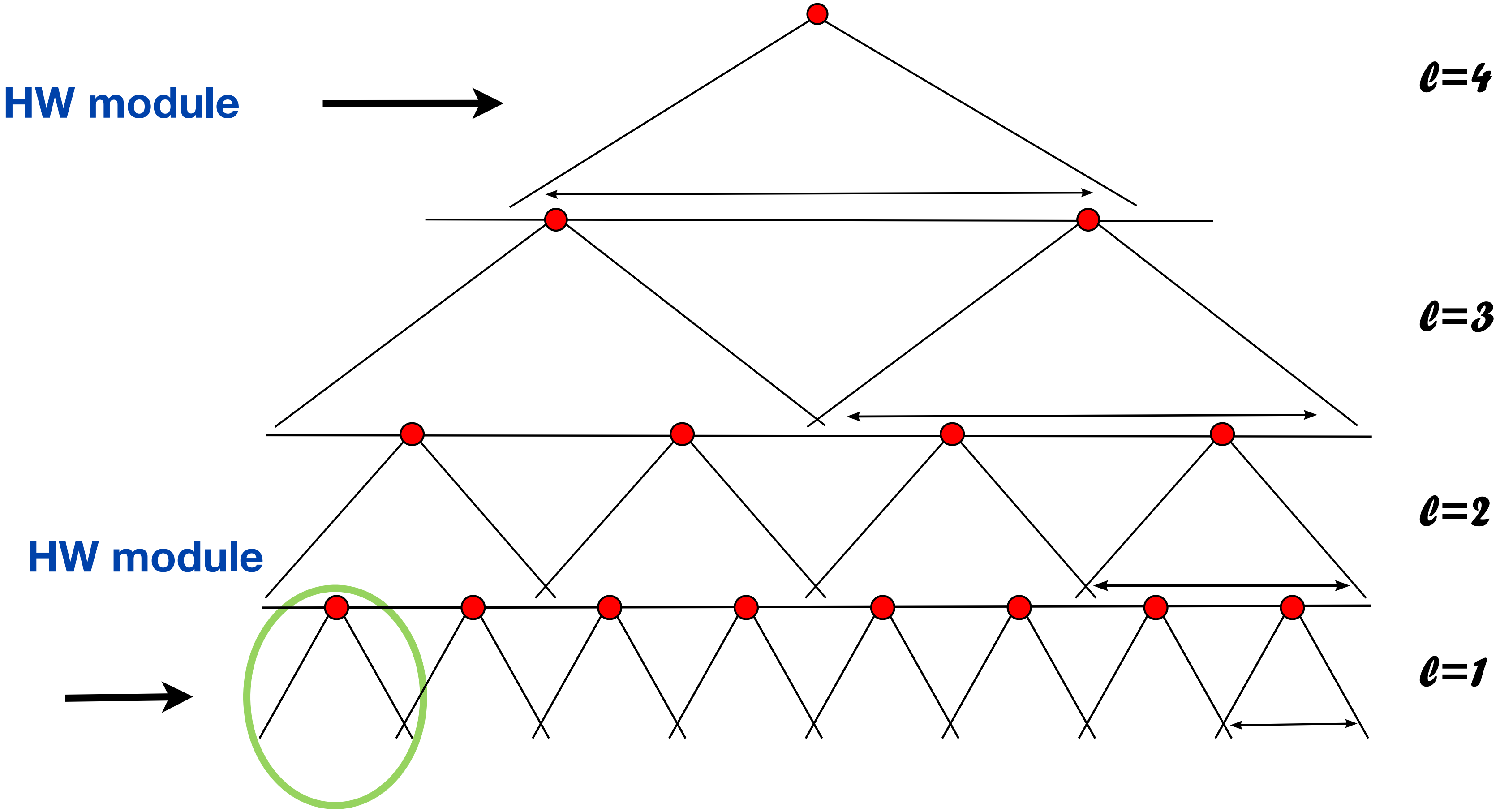
I want to go now into another part of the theory:

3. covariance allows to extend to hierarchies
4.

Preview: from a HW module to a hierarchy via covariance

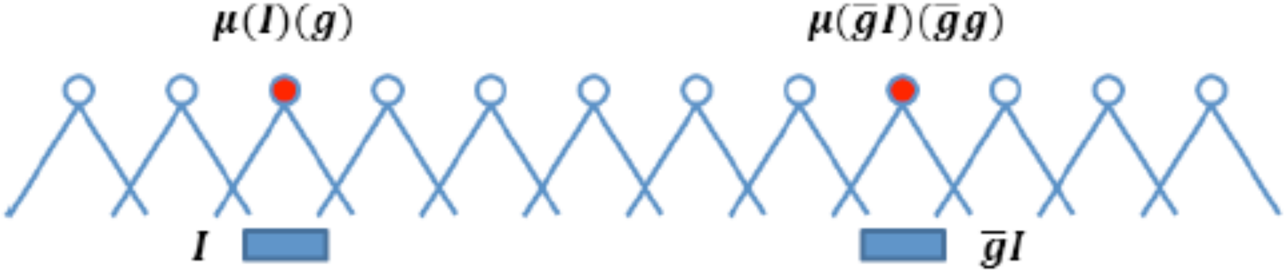


Preview: from a HW module to a hierarchy via covariance



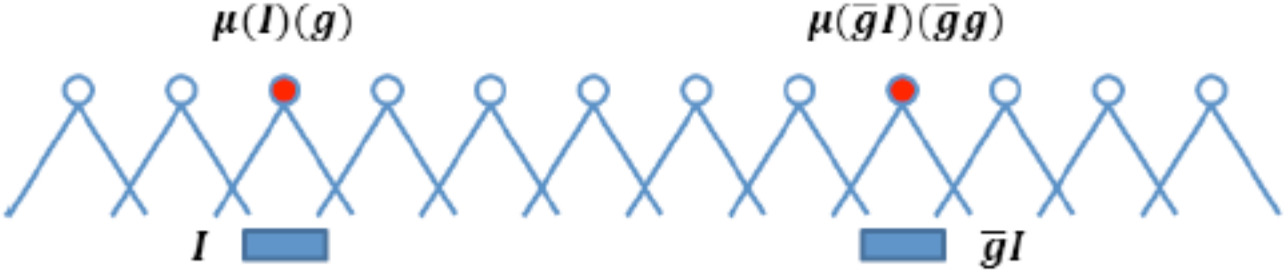
● complex cell node gives output of the HW module

Preview: from a HW module to a hierarchy via covariance



Covariance theorem (informal): *for isotropic networks the activity at a layer of "complex" cells for shifted an image at position g is equal to the activity induced by the group shifted image at the shifted position.*

Preview: from a HW module to a hierarchy via covariance

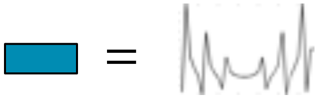


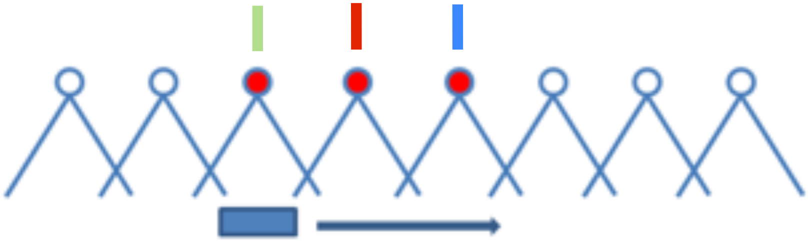
Covariance theorem (informal): *for isotropic networks the activity at a layer of "complex" cells for shifted an image at position g is equal to the activity induced by the group shifted image at the shifted position.*

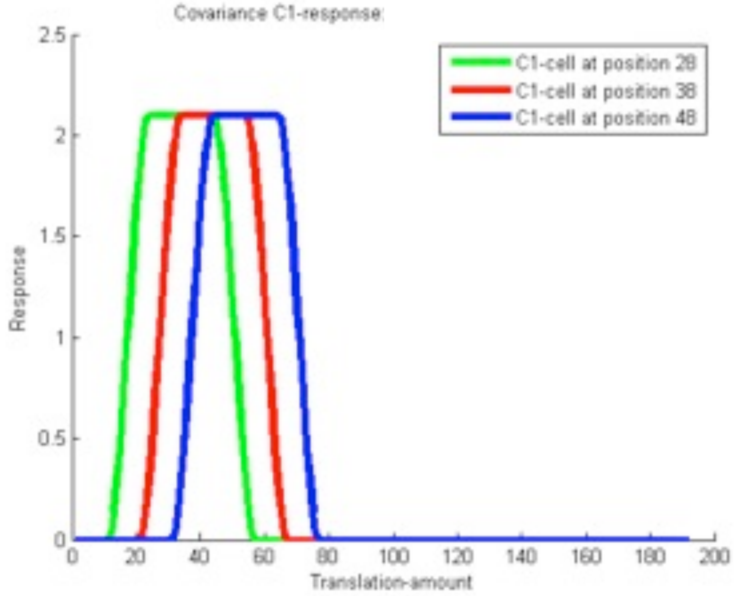
Remarks:

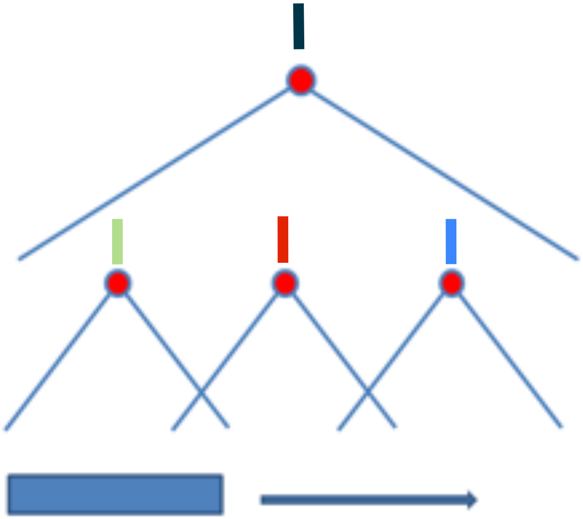
- Covariance allows to consider a higher level HW module, looking at the neural image at the lower layer and apply again the invariance/covariance arguments

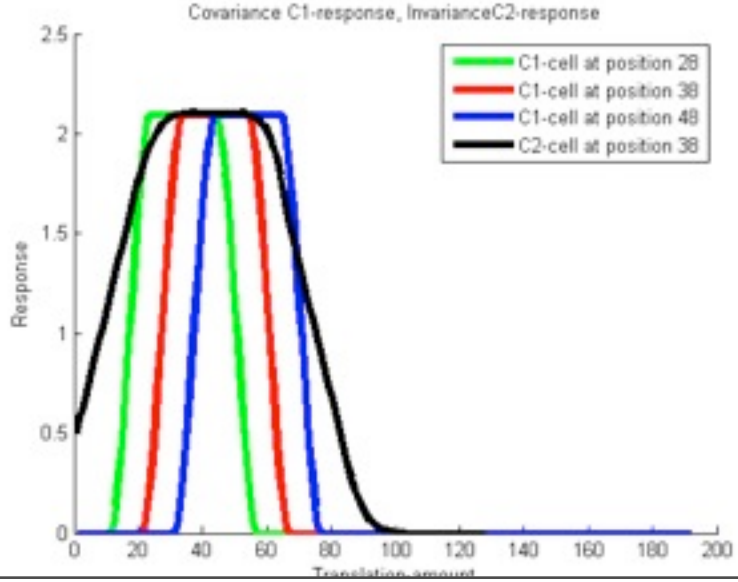
Toy example: 1D translation



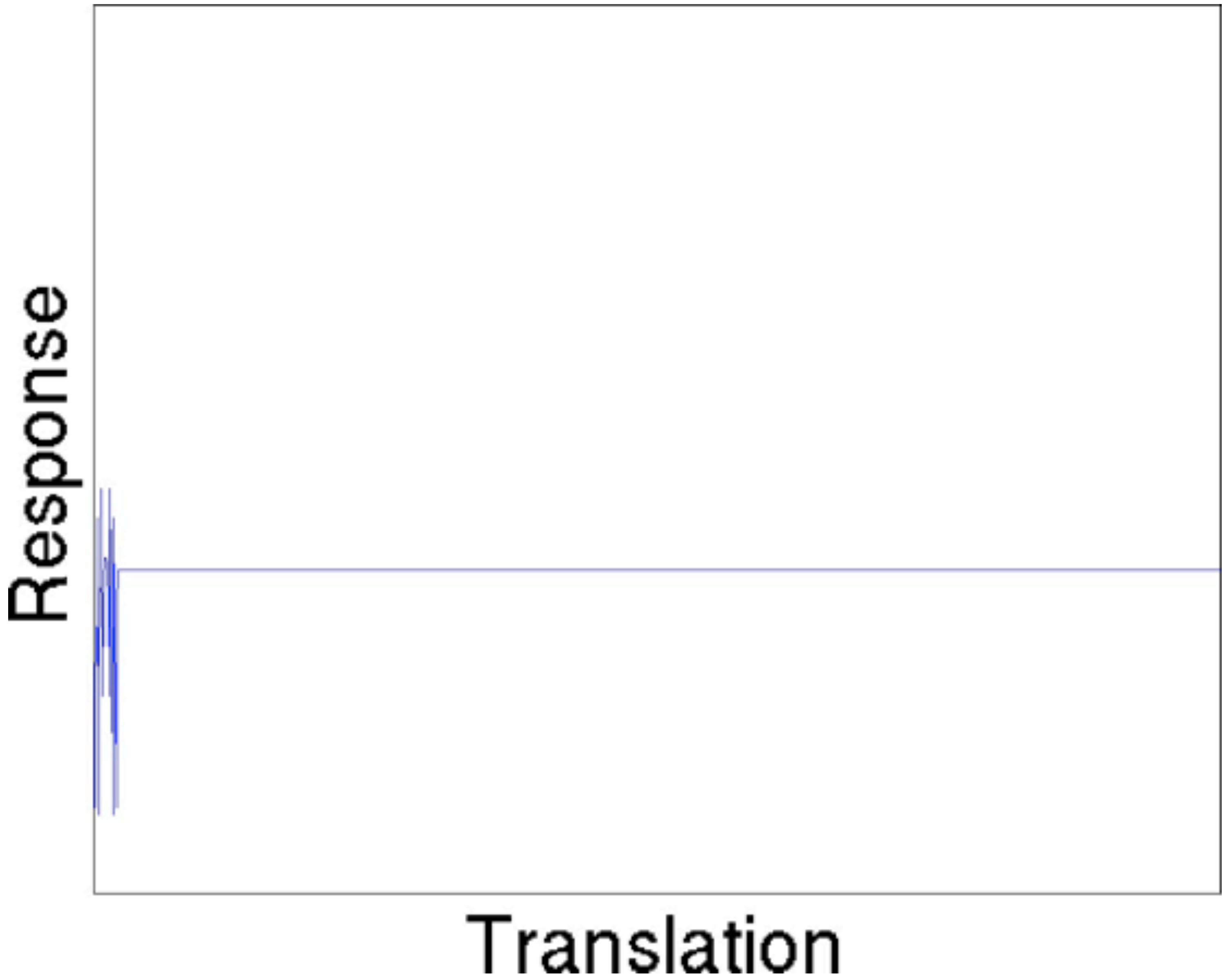








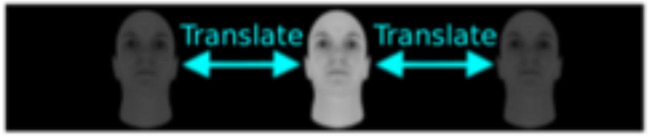
Toy example: 1D translation



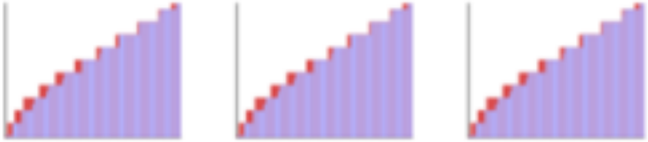
M-Theory

So far: compact groups in R^2

M-theory extend result to



- partially observable groups



- non-group transformations

- hierarchies of magic HW modules (multilayer)

Non compact groups

We assume that the dot products is “normalized": the signals x and t are zero-mean and norm = 1. Thus starting with x”, t”

$$x' = x'' - E(x''), \qquad x = \frac{x'}{|x'|};$$

$$\mathrm{t}' = t'' - E(t''), \qquad \mathrm{t} = \frac{t'}{|t'|}$$

We assume that the empty surround of an isolated image patch has value 0, being equal to the average value over the ensemble of images. In particular the dot product of a template and the region outside an isolated image patch is 0.

Partially Observable Groups

(includes non compact)

For a transformation observed via a “receptive field” there is only “partial invariance”

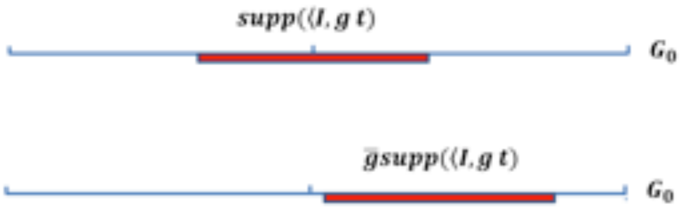
Let $I,t \in H$ with H Hilbert space, $\eta_n : R \to R^+$ a set of bijective (positive) functions and G a locally compact group. Let $G_0 \subseteq G$ and suppose $\operatorname{supp} \langle I, g_i t^k \rangle \subseteq G_0$ .Then for any $\bar{g} \in G, t^k, I$

$$\mu_n^k(I) = \mu_n^k(\bar{g}I) \Leftrightarrow \langle I, g_i t^k \rangle = 0, \forall g \in G_0 \cup \bar{g}G_0 \setminus G_0 \cap \bar{g}G_0$$

eg if $G_0 \cup \bar{g}G_0$ *is* our universe

then $\forall g \in G_0 \cup \bar{g}G_0 \setminus G_0 \cap \bar{g}G_0$

can be written as $\forall g \in (G_0 \cap \bar{g}G_0)^c$



Rosasco, poggio, Anselmi

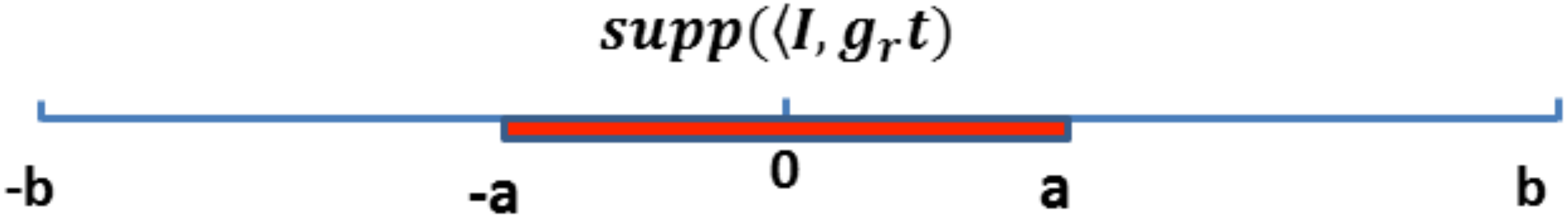
Partially Observable Groups

Invariance for POGs implies a localization property we call

sparsity of the image I *wrt the* t *dictionary under the set of transformations* G

Example: consider the case of a 1D parameter translation group: invariance of $\mu_n^k(I)$ with pooling region $[-b,b]$ is ensured if

$$\left\langle I, g_r t^k \right\rangle = 0, \qquad \text{for } |r| > b-a$$



Invariance, sparsity, wavelets

Thus sparsity implies, and is implied by, invariance.
Sparsity can be satisfied in two different regimes:

- exact sparsity for *generic* images holds for affine group.

- approximate sparsity of a subclass of I w.r.t. dictionary of transformed templates gt^k holds locally for any smooth transformation.

Invariance, sparsity, wavelets

Theorem: Sparsity is *necessary and sufficient* condition for translation and scale invariance. Sparsity for translation (respectively scale) invariance is equivalent to the support of the template being small in space or frequency.

Proposition: Maximum simultaneous invariance to translation and scale is achieved by Gabor templates:

$$t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}$$

M-Theory

M-theory extends result to

- non compact groups
- non-group transformations
- hierarchies of magic HW modules (multilayer)

Non-group transformations: approximate invariance in class-specific regime

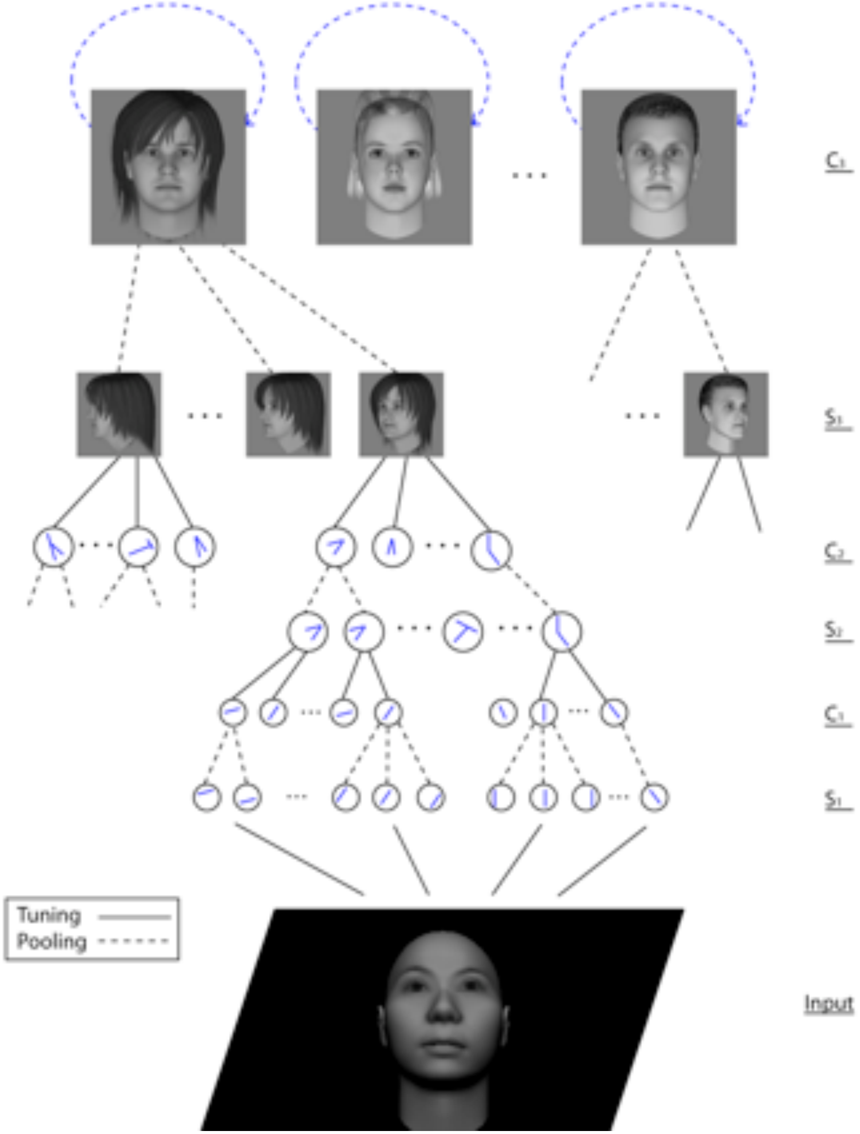
$\mu_n^k(I)$ is locally invariant if:

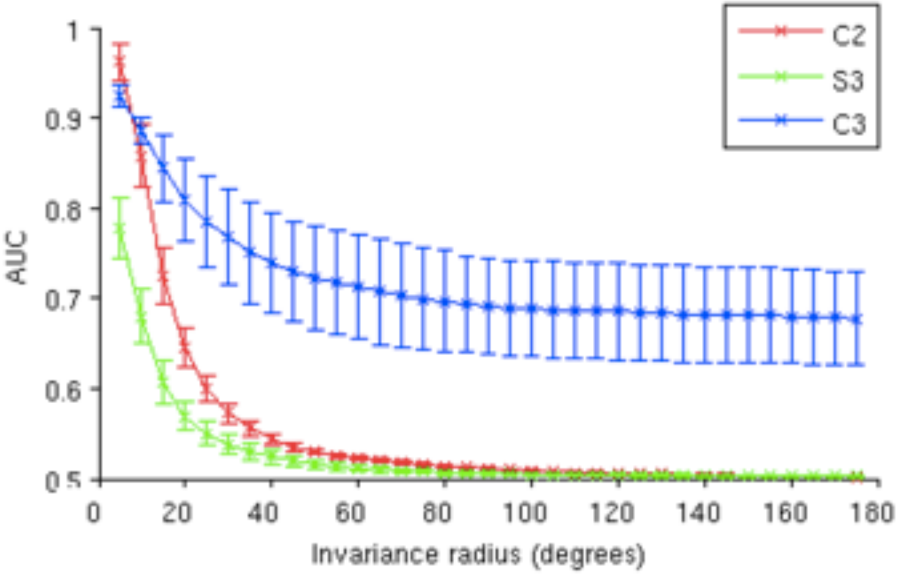
- I is sparse in the dictionary of t^k

- I transforms in the same way (belong to the same class) as t^k

- the transformation is sufficiently smooth

Class specific pose invariance for faces



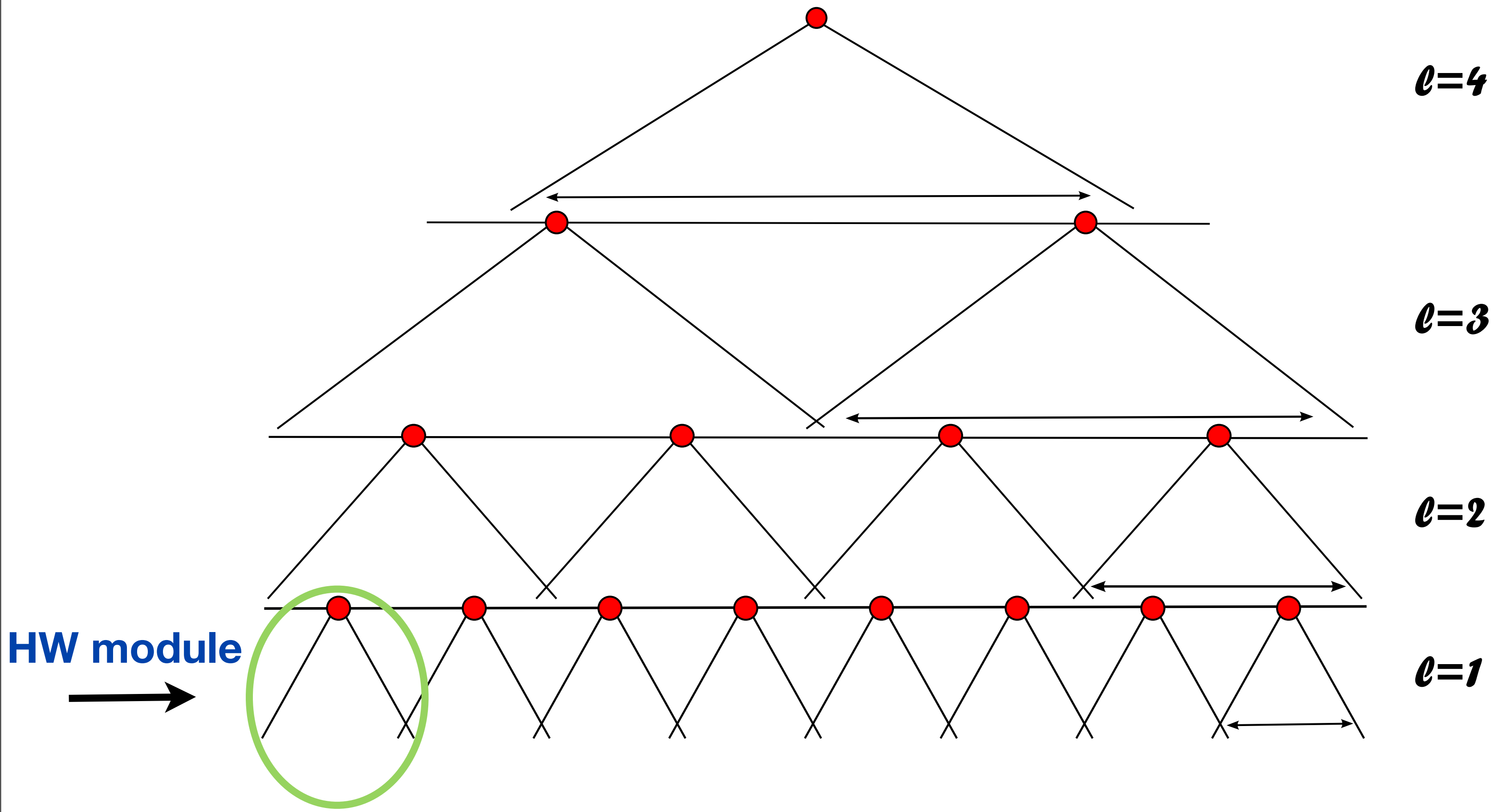


M-Theory

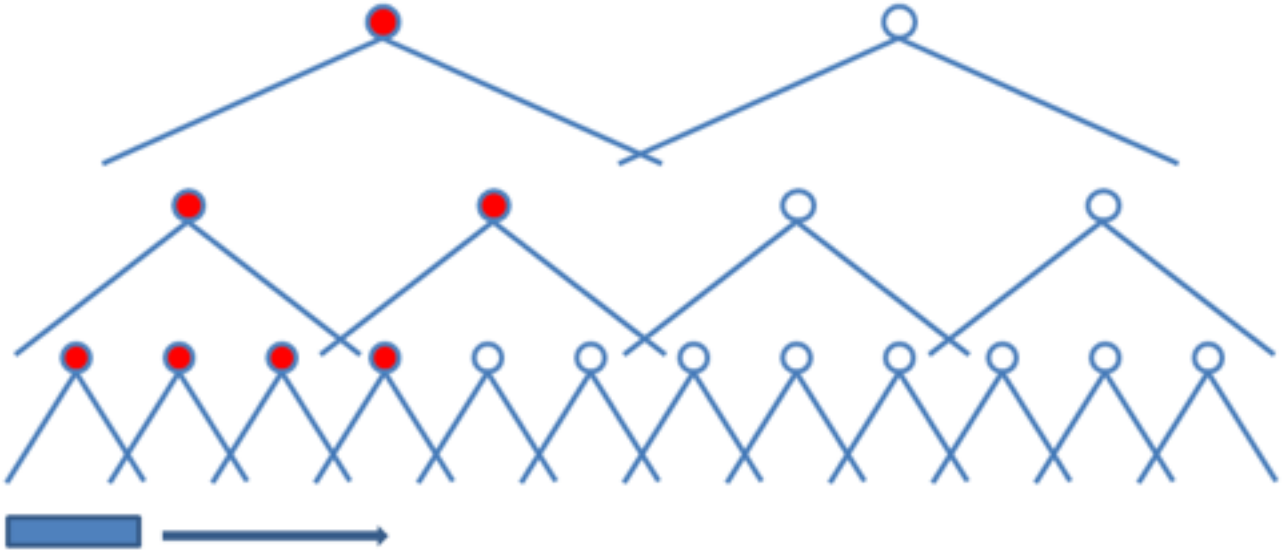
M-theory extend result to

- non compact groups
- non-group transformations
- hierarchies of magic HW modules (multilayer)

Hierarchies of magic HW modules: key property is covariance



Local and global invariance: whole-parts theorem

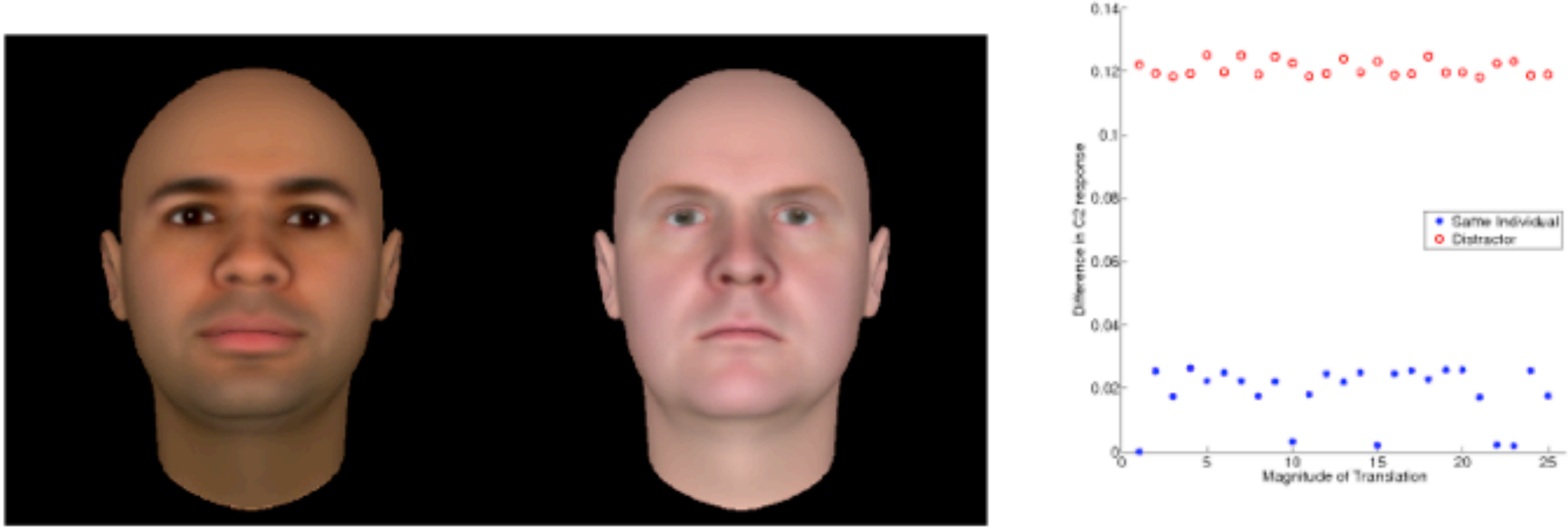


For any signal (image) there is a layer in the hierarchy such that the response is invariant w.r.t. the signal transformation.

Why multilayer architectures

- Compositionality: signatures for wholes and for parts of different size at different locations
- Minimizing clutter effects
- Invariance for certain non-global affine transformations
- Retina to V1 map

Invariance and uniqueness

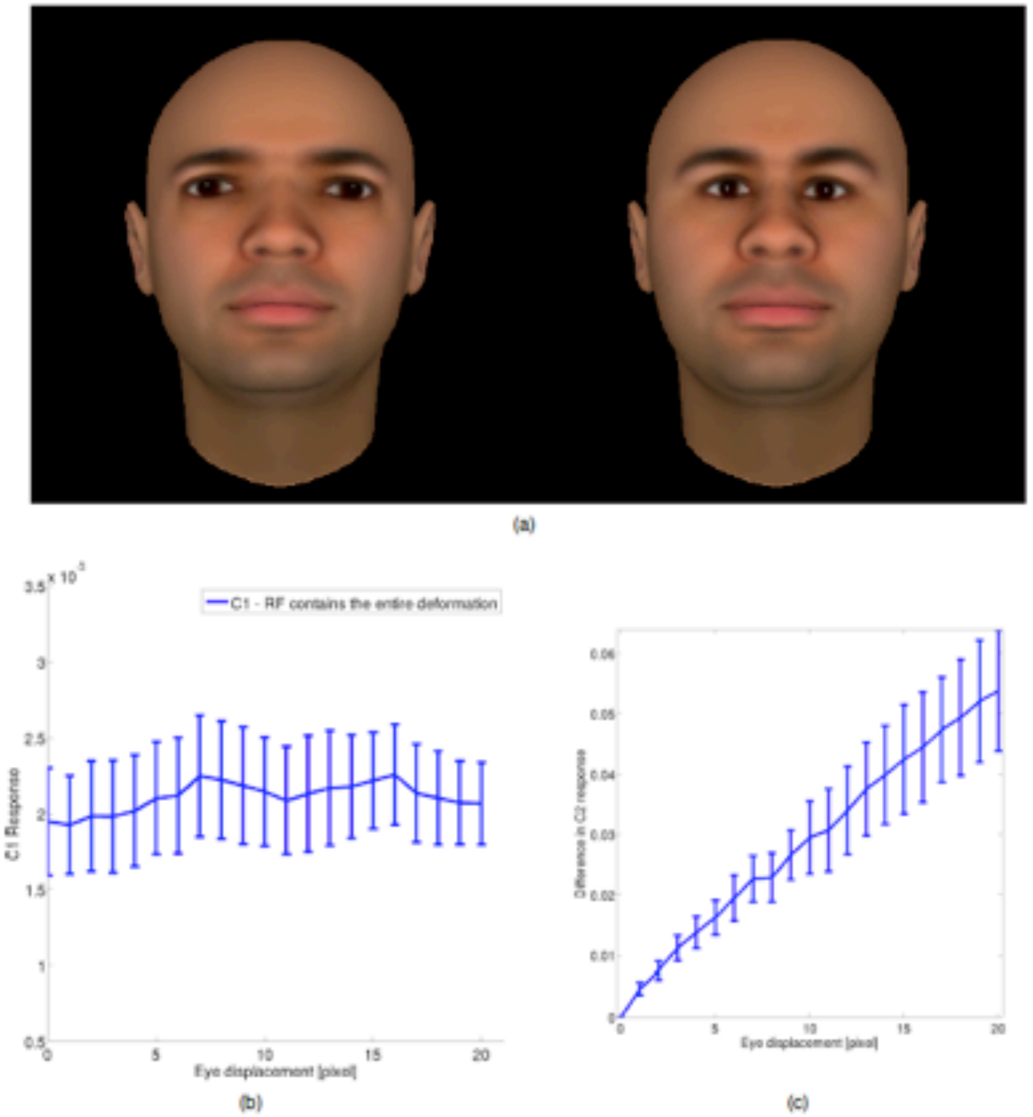


(a) *Reference input and distractor.*

(b)

Figure 3: *Two distinct stimuli (left) are presented at various location in the visual field. The Euclidean distance between C2 response vectors in HMAX is reported (right). It can be seen how the response are invariant to global translation and discriminative. The C2 units represent the top of a hierarchical, convolutional architecture.*

Invariance for parts and stability for wholes



(a)

(b)

(c)

Figure 4: *(a) shows the reference image on the left and a local deformation of it (the eyes are closer to each other); (b) shows that a C1 signature from complex cells whose receptive fields covers the left eye is invariant to the deformation; in (c) C2 cells whose receptive fields contain the whole face are (Lipschitz) stable with respect to the deformation. In all cases just the euclidean norm of the response is shown on the y axis.*

Plan

1. Motivation: models of cortex (and deep convolutional networks)
2. Core theory
    - the basic invariance module
    - the hierarchy
3. Computational performance
4. Biological predictions
5. Theorems and remarks
    - $n \to 1$
    - invariance and sample complexity
    - connections with scattering transform
    - invariances and beyond perception
    - ...

Implementations/specific models: computational performance

- Deep convolutional networks (such as Lenet) as an architecture are special case of Mtheory (with just translation invariance and max/sigmoid pooling)
- HMAX as an architecture is a special case of Mtheory (with translation + scale invariance and max pooling) and used to work well

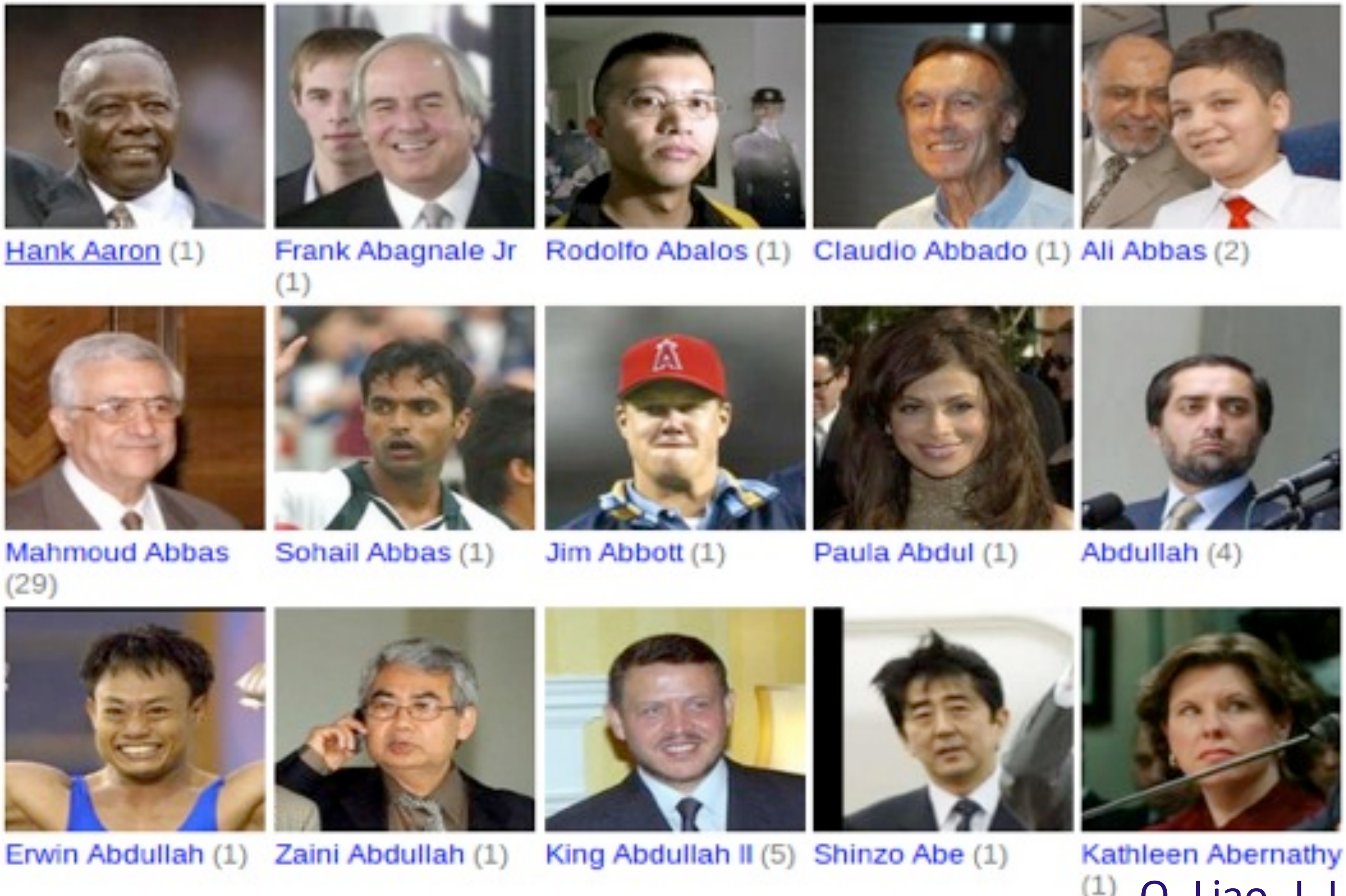
Models: computational performance

- Deep convolutional networks (such as Lenet) as an architecture are special case of Mtheory (with just translation invariance and max/sigmoid pooling)
- HMAX as an architecture is a special case of Mtheory (with translation + scale invariance and max pooling) and used to work well
- Encouraging initial results in speech and music classification (Evangelopoulos, Zhang, Voinea)
- Example in face identification (Liao, Leibo) --->

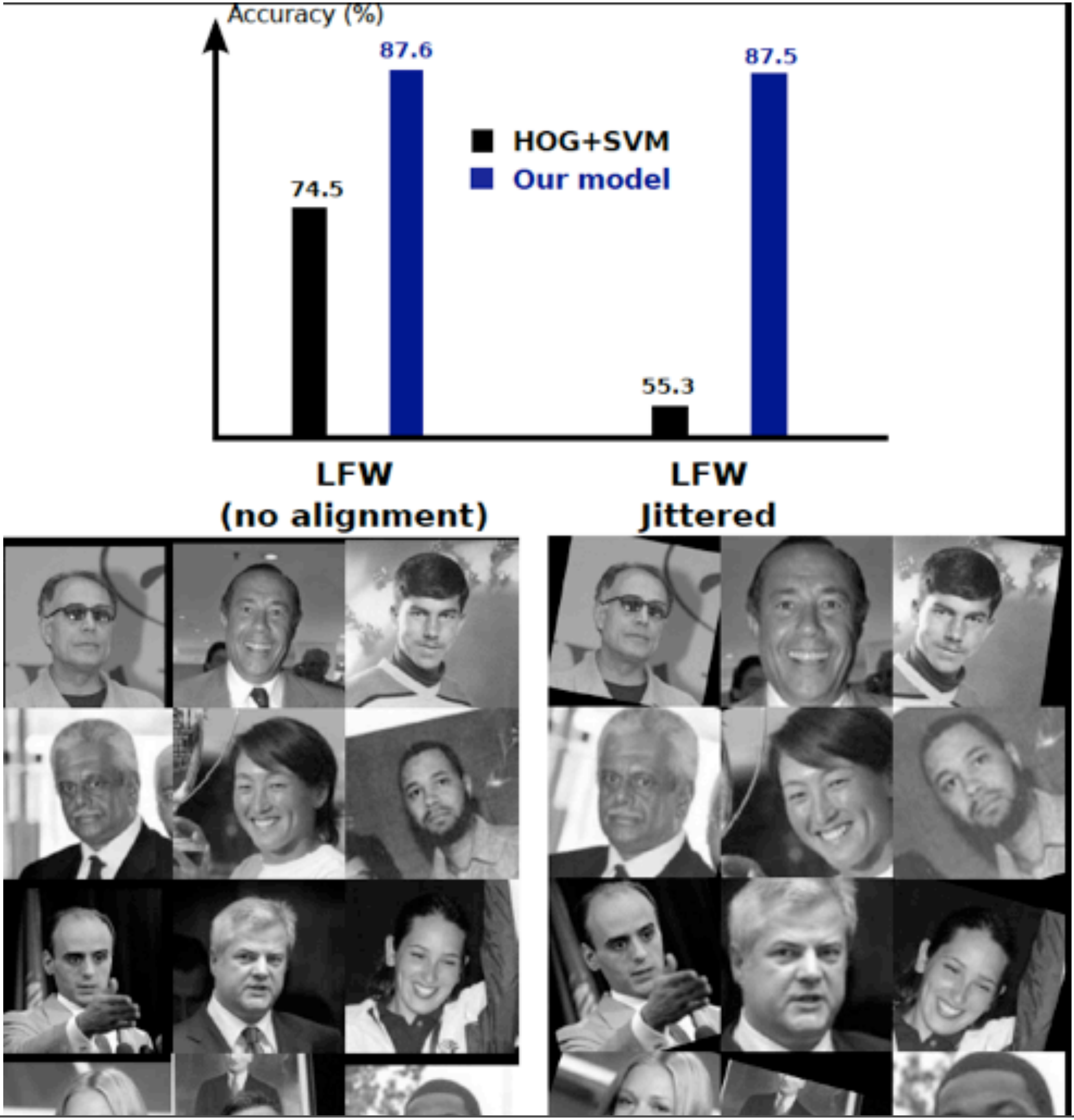
Computational performance: example faces

Labeled Faces in the Wild

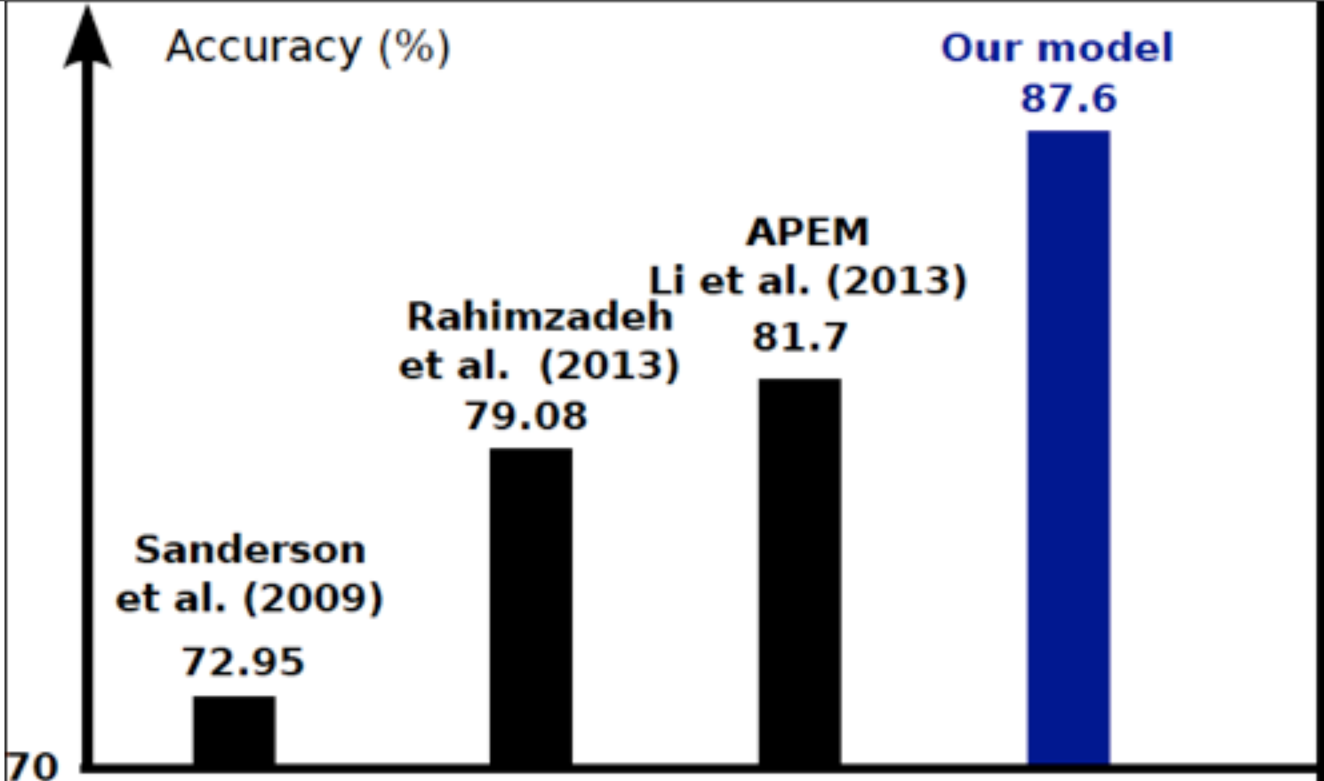
Contains 13,233 images of 5,749 people



Q. Liao, J. Leibo, NIPS 2013



Q. Liao, J. Leibo



LFW - no outside data used & no alignment

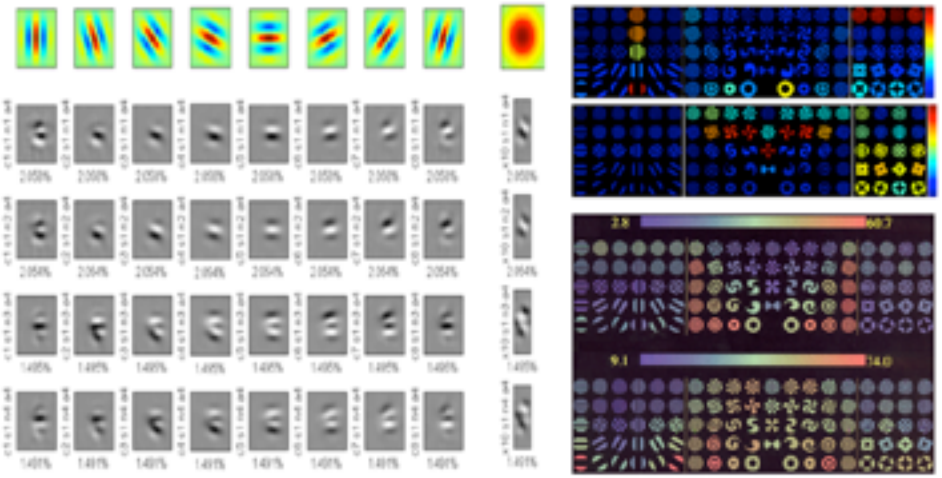
Q. Liao, J. Leibo

Plan

1. Motivation: models of cortex (and deep convolutional networks)
2. Core theory
    - the basic invariance module
    - the hierarchy
3. Computational performance
4. Biological predictions
5. Theorems and remarks
    - $n \to 1$
    - invariance and sample complexity
    - connections with scattering transform
    - invariances and beyond perception
    - ...

Theory of unsupervised invariance learning in hierarchical architectures

- neurally plausible: HW module of simple-complex cells
- says what simple-complex cells compute
- provides a theory of pooling: energy model, average, max...
- leads to a new characterization of complex cells
- provides a computational explanation of why Gabor tuning
- may explain tuning and functions of V1, V2, V4 and in face patches!
- suggests generic, Gabor-lik tuning in early areas <u>and</u> specific selective tuning higher up



poggio, anselmi, rosasco, tacchetti, leibo, liao

Plan

1. Motivation: models of cortex (and deep convolutional networks)
2. Core theory
    - the basic invariance module
    - the hierarchy
3. Computational performance
4. Biological predictions
5. Theorems and remarks
    - $n \to 1$
    - invariance and sample complexity
    - connections with scattering transform
    - invariances and beyond perception
    - ...

Musing on technology: a second phase in Machine Learning?

- The first phase -- from ~1980s -- led to a rather complete theory of supervised learning and to practical systems (MobilEye, Orcam,...) that need lots of examples for training: $n \to \infty$

- The second phase may be about unsupervised learning of (invariant) representations that make supervised learning possible with very few examples:

$$n \to 1$$