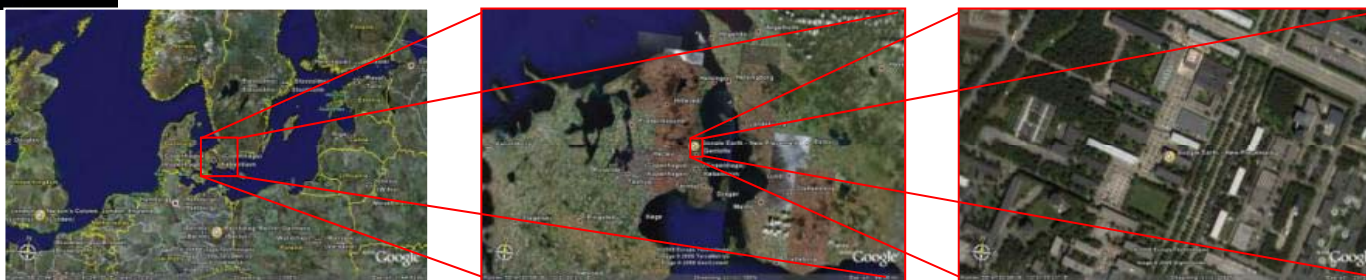


# Transformation Invariant Sparse Coding



**Morten Mørup & Mikkel N. Schmidt**  
 Section for Cognitive Systems  
 DTU Informatics, Technical University of Denmark

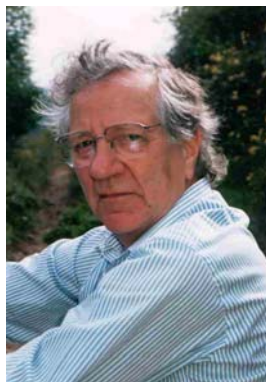


$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

$$\int_a^b \varepsilon \Theta^{\sqrt{17}} + \Omega \int \delta e^{i\pi} = \{2.7182818284\}$$

$$\chi^2 \sum \gg \Sigma !$$

# Redundancy Reduction



Horace Barlow  
(1921-)

## What is a pattern?

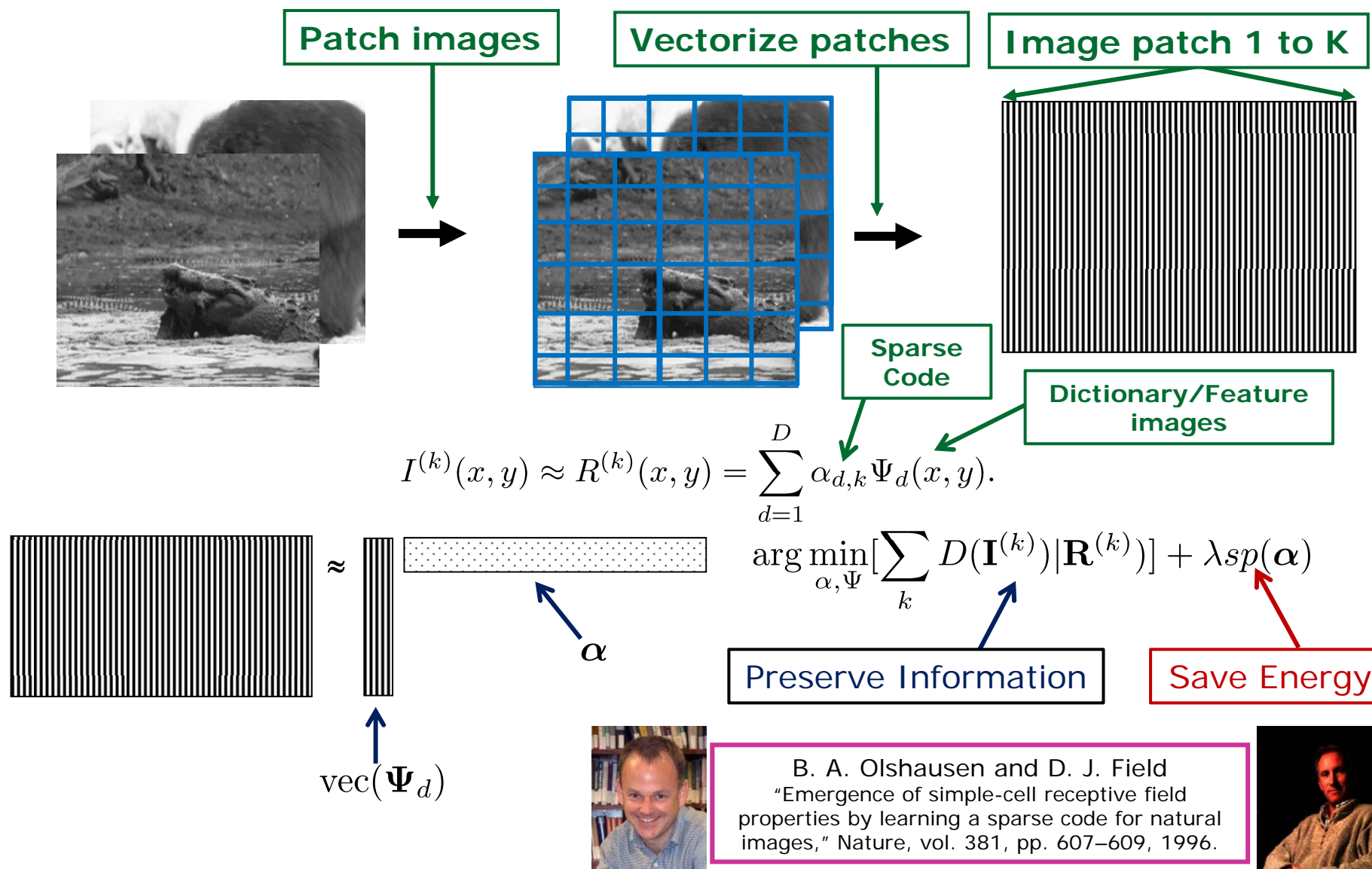
Some kind of *regularity* or *self-similarity*. If there is no regularity, or no repetition caused by self-similarity, then surely there is no pattern. But if there is such regularity or repetition, then *this is a form of redundancy, and offers the opportunity for recoding to reduce it*. Of course the pattern element can be completely arbitrary, a sequence of randomly selected digits for example, but if repeated this element will make a pattern. Thus it seems to me that *the importance of redundancy is almost a tautology and follows simply from the nature of pattern*.



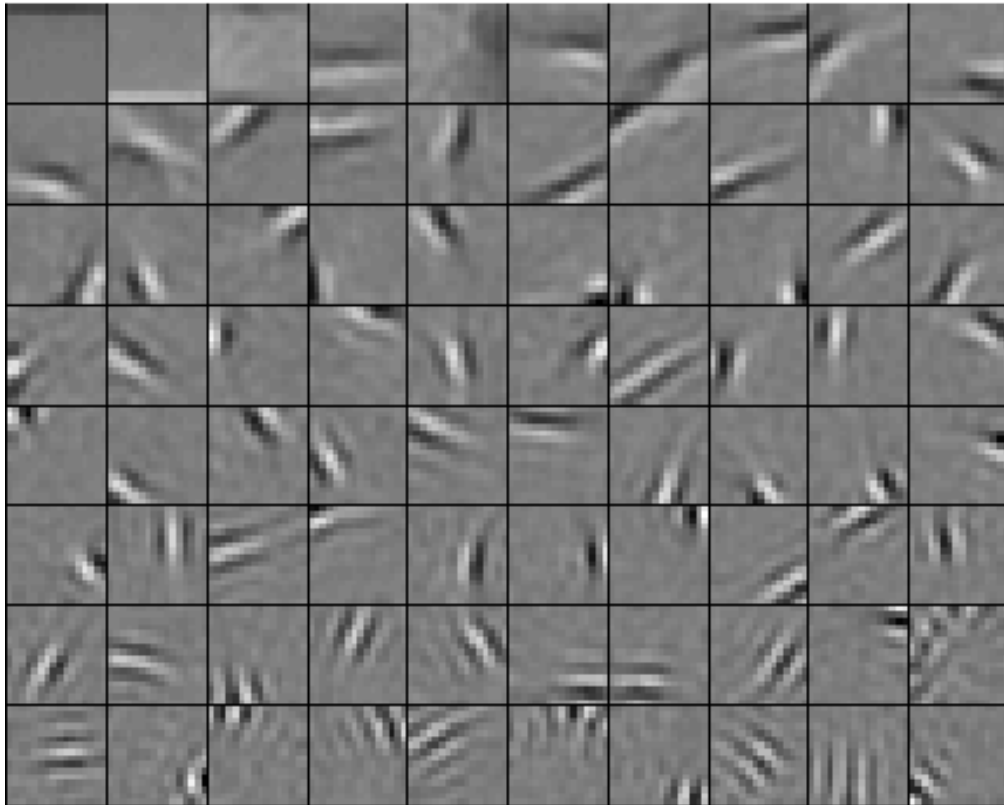
洵洋溢于德比氏之心兮  
 至顯哉照臨下土  
 發乎帝之心兮  
 愛生如衆寡  
 洵充沛于德比氏之心兮  
 願基督重來大地  
 簡在帝心之則兮  
 實德比氏之所循  
 匪全知之主是依兮  
 民將多艱而遠巡  
 嗟摩王以萬類兮  
 因罔鈞之所秉  
 將愛與先之所範兮  
 以墨魯慈之門  
 曰惟先與愛與力兮  
 其使大法重張弘乎寰宇



# Sparse Coding (Dictionary Learning)



# Gabor like features $\Psi$ resembling simple cell behavior of V1 in the brain extracted! – Nature 1996



the brain might employ sparse coding since:

- *it allows for increased storage capacity in associative memories*
- *it makes the structure in natural signals explicit*
- *it represents complex data in a way that is easier to read out at subsequent level of processing*
- *it is energy efficient.*



B. A. Olshausen and D. J. Field  
"Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

B. A. Olshausen and D. J. Field  
"Sparse coding of sensory inputs," *Current Opinion in Neurobiology*, vol. 14, pp. 481–487, 2004.



**These features are however very redundant particular in terms of shift, rotation and scaling**

# Sparse coding has therefore been extended to general transformation invariance of the feature images.



$$I(x, y) \approx \sum_{d,m} \alpha_{d,m} (T_m \Psi_d)(x, y)$$

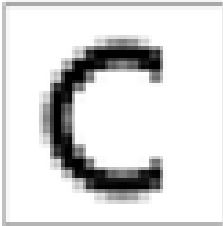
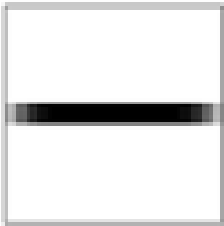
H. Wersing, J. Eggert, and E. Körner,  
"Sparse coding with invariance constraints," Proc. Int.  
Conf. Artificial Neural Networks ICANN, pp. 385–392,  
2003.

J. Eggert, H. Wersing, and E. Körner,  
"Transformation invariant representation and nmf," in  
Neural Networks, 2004,  
vol. 4, pp. 2535– 2539.

- These operators,  $T_m$ , account for any desired transformation within each patch, such as shift, rotation and scaling.

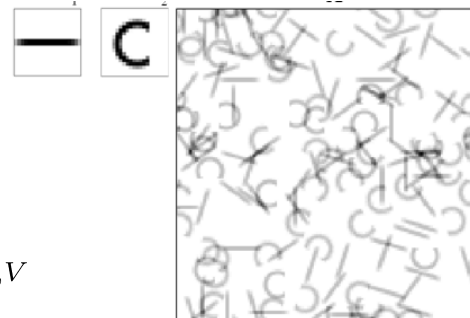
**However, subdividing the images into image patches creates loss of information between the patches at boundaries...**

# Let's consider a simple example of redundancy

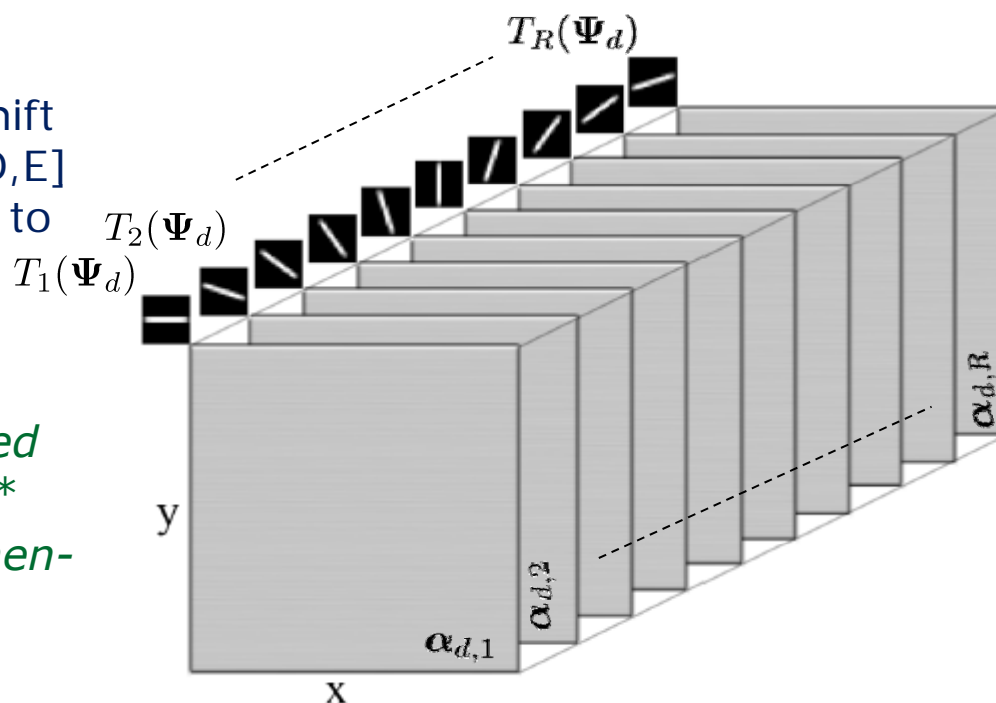


# Our proposed model

$$\mathbf{I} \approx \mathbf{R} = \sum_{d=1}^D \sum_{r=1}^R \alpha_{d,r} * T_r(\Psi_d). \quad \alpha_{d,r} \in \mathbb{R}^{X+U, Y+V}, \Psi_d \in \mathbb{R}^{U, V}$$



- The above model is related to shift invariant sparse coding [A,B,C,D,E] with the extension of invariance to general transformations.
- *The model does not depend on subdividing images into patches*
- *Shifts can be efficiently calculated by 2D convolutions denoted by \**
- *General transformations implemented through the operator  $T_r$*



- [A] M. N. Schmidt, M. Mørup "Non-negative Matrix Factor 2D Deconvolution for Blind Single Channel Source Separation", ICA2006
- [B] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," In Proceedings of the Neural Information Processing Systems (NIPS), vol. 19, 2007.
- [C] T. Blumensath and M. Davies, "On shift-invariant sparse coding," International Conference on Independent Component Analysis and Blind Source Separation, vol. 26, pp. 1205–1212, 2004.
- [D] M. Mørup, M. N. Schmidt, and L. K. Hansen, "Shift invariant sparse coding of image and music data," 2008.
- [E] P. Smaragdis, B. Raj, and M. Sashanka, "Sparse and shiftinvariant feature extraction from non-negative data," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2069–2072, 2008.

# Parameter estimation for analysis of N-images

$$\mathbf{I}^{(n)} \approx \mathbf{R}^{(n)} = \sum_{d=1}^D \sum_{r=1}^R \alpha_{d,r}^{(n)} * T_r(\Psi_d).$$

$$\arg \min_{\alpha, \Psi} \left[ \sum_n D(\mathbf{I}^{(n)} | \mathbf{R}^{(n)}) \right] + \lambda \sum_{d,r} sp(\alpha_{d,r})$$

Our specific choice of  $D(\cdot | \cdot)$  and  $sp(\cdot)$ :

$$\arg \min_{\alpha, \Psi} \left[ \sum_n \|\mathbf{I}^{(n)} - \mathbf{R}^{(n)}\|_F^2 \right] + \lambda \sum_{d,r} |\alpha_{d,r}|_1$$

Each alternating update of  $\alpha$  and  $\Psi$  form two convex subproblems, however, the joint optimization as for regular sparse coding is non-convex.

Hessian of the problem very memory intensive, thus, we use first order methods to solve for  $\alpha$  and  $\Psi$ .

# Update of $\Psi$

We constrain  $\|\Psi_d\|_F = 1$  in order to avoid trivial regularized solutions where  $\alpha \rightarrow \mathbf{0}$  and  $\Psi \rightarrow \infty$  based on the normalization invariant gradient descent procedure proposed in

J. Eggert and E. Körner, "Sparse coding and nmf," in Neural Networks, 2004, vol. 4, pp. 2529–2533

$$\Psi_d = \frac{\tilde{\Psi}_d}{\|\tilde{\Psi}_d\|_F}$$

$$\tilde{\Psi}_d \leftarrow \Psi_d - \mu(\nabla_{\Psi_d} - \langle \nabla_{\Psi_d}, \Psi_d \rangle \Psi_d)$$

Where gradient for our model is given by:

$$\nabla_{\Psi_d} = \sum_{n=1}^N \sum_{r=1}^R T_r^{-1}(\mathbf{I}^{(n)} - \mathbf{R}^{(n)}) * T_r^{-1}(T_\pi(\alpha_{d,r}^{(n)}))$$

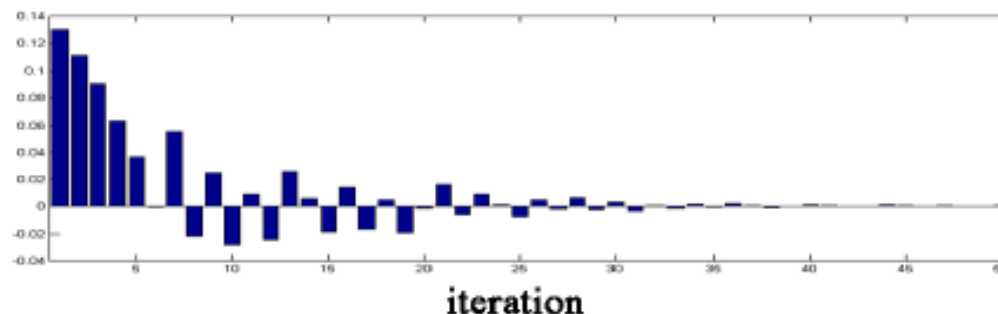
$T_\pi$  denotes rotation by 180 degree

## Update for $\alpha$

- Solving for  $\alpha$  can be rewritten as a standard sparse regression (LASSO) problem, i.e. 
$$\arg \min_{\mathbf{s}} L(\mathbf{s}) = D(\mathbf{x}|\mathbf{A}\mathbf{s}) + \lambda sp(\mathbf{s}),$$

- A naive gradient descent procedure will however due to the sparsity term oscillate around zero, resulting in very slow convergence!

$$\mathbf{s} \leftarrow \mathbf{s} - \mu \nabla_{\mathbf{s}} = \mathbf{s} - \mu(\nabla_{\mathbf{s}} D(\mathbf{x}|\mathbf{A}\mathbf{s}) + \lambda \nabla_{\mathbf{s}} sp(\mathbf{s}))$$



**To avoid this property we propose the following simple gradient based sparse coding (GB-SC) procedure:**

- Take a gradient step according to  $\nabla_{\mathbf{s}} D(\mathbf{x}|\mathbf{A}\mathbf{s})$ :

$$\mathbf{s}^* \leftarrow \mathbf{s} - \mu(\nabla_{\mathbf{s}} D(\mathbf{x}|\mathbf{A}\mathbf{s}))$$

- Take a gradient step according to  $\lambda \cdot \nabla_{\mathbf{s}} sp(\mathbf{s})$  and truncate to zero if crossing zero:

$$\mathbf{s} \leftarrow \mathbf{s}^* - \mu(\lambda \cdot \nabla_{\mathbf{s}^*} sp(\mathbf{s}^*)), \quad \text{if } \text{sgn}(s_t) \neq \text{sgn}(s_t^*) \text{ then } s_t = 0.$$

# For the choice of the Frobenius norm and $L_1$ -regularization we find:

$$\begin{aligned} L(\mathbf{s}) &= D(\mathbf{x}|\mathbf{A}\mathbf{s}) + \lambda sp(\mathbf{s}), \\ &= \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_F^2 + \lambda|\mathbf{s}|_1 \end{aligned}$$

We now have for the gradients:

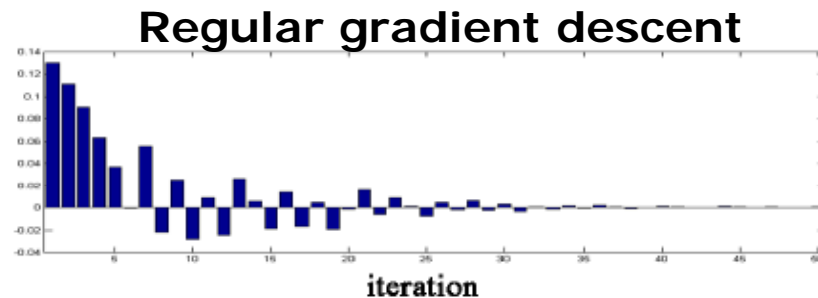
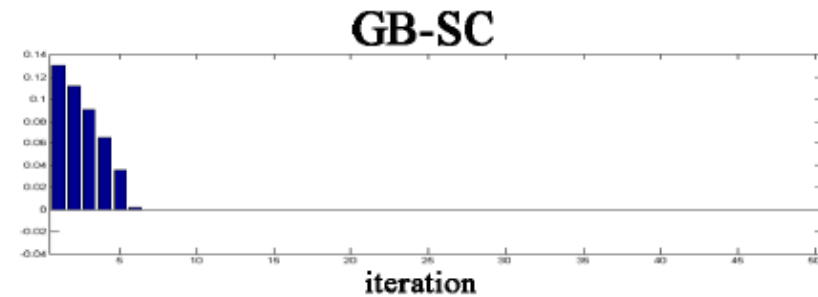
$$\begin{aligned} \nabla D(\mathbf{x}|\mathbf{A}\mathbf{s}) &= \mathbf{A}^\top (\mathbf{A}\mathbf{s} - \mathbf{x}) \\ \nabla sp(\mathbf{x}) &= \text{sgn}(\mathbf{s}) \end{aligned}$$

Which equivalently for our original problem corresponds to:

$$\begin{aligned} \nabla_{\alpha_{d,r}^{(n)}} D(\mathbf{I}^{(n)}|\mathbf{R}^{(n)}) &= (\mathbf{I}^{(n)} - \mathbf{R}^{(n)}) * (T_\pi(T_r(\Psi_d))) \\ \nabla_{\alpha_{d,r}^{(n)}} sp(\alpha_{d,r}^{(n)}) &= \text{sgn}(\alpha_{d,r}^{(n)}) \end{aligned}$$

Plugging this into the GB-SC we have:

$$\begin{aligned} \alpha_{d,r}^{(n)*} &\leftarrow \alpha_{d,r}^{(n)} - \mu(\nabla_{\alpha_{d,r}^{(n)}} D(\mathbf{I}^{(n)}|\mathbf{R}^{(n)})) \\ \alpha_{d,r}^{(n)} &\leftarrow \alpha_{d,r}^{(n)*} - \mu(\lambda \cdot \nabla_{\alpha_{d,r}^{(n)*}} sp(\alpha_{d,r}^{(n)*})), \quad \text{if } \text{sgn}(\alpha_{d,r}^{(n)*}) \neq \text{sgn}(\alpha_{d,r}^{(n)}) \text{ then } \alpha_{d,r}^{(n)} = 0. \end{aligned}$$



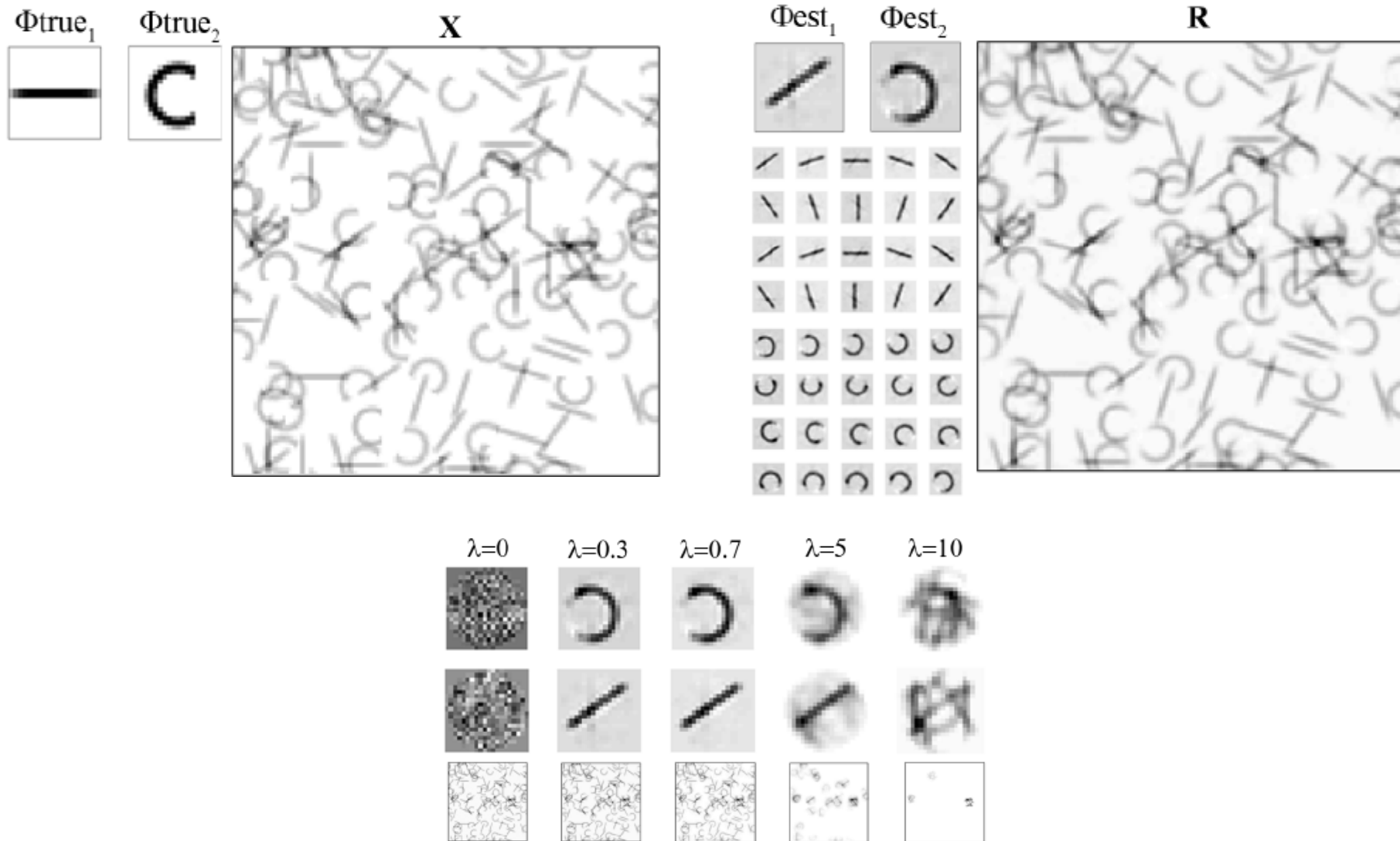
## Evaluation of our proposed GB-SC to second order methods for $L_1$ - sparse regression (LASSO)

$$L(\mathbf{s}) = \|\mathbf{x} - \mathbf{A}\mathbf{s}\|_F^2 + \lambda|\mathbf{s}|_1$$

	$256 \times 100$	$256 \times 256$	$256 \times 1000$	$256 \times 2500$
BD-SC	$0.3641 \pm 0.3044$	$11.6250 \pm 4.4922$	—	—
SignSearch	$0.0750 \pm 0.0359$	$0.1984 \pm 0.1342$	<b><math>0.3734 \pm 0.1759</math></b>	<b><math>1.6969 \pm 0.6441</math></b>
Conjugate gradient	$0.4172 \pm 0.0651$	$1.1219 \pm 0.2560$	$9.0297 \pm 1.8055$	$45.6297 \pm 12.0142$
LARS	$0.0453 \pm 0.0226$	<b><math>0.1313 \pm 0.0787</math></b>	$0.4313 \pm 0.1477$	$1.9813 \pm 0.6342$
BPD	$0.5703 \pm 0.0696$	$0.9313 \pm 0.0748$	$2.8719 \pm 0.1389$	$15.5047 \pm 0.7882$
GB-SC	<b><math>0.0125 \pm 0.0066</math></b>	$0.3172 \pm 0.2121$	$2.0688 \pm 1.0760$	$22.8828 \pm 12.2846$

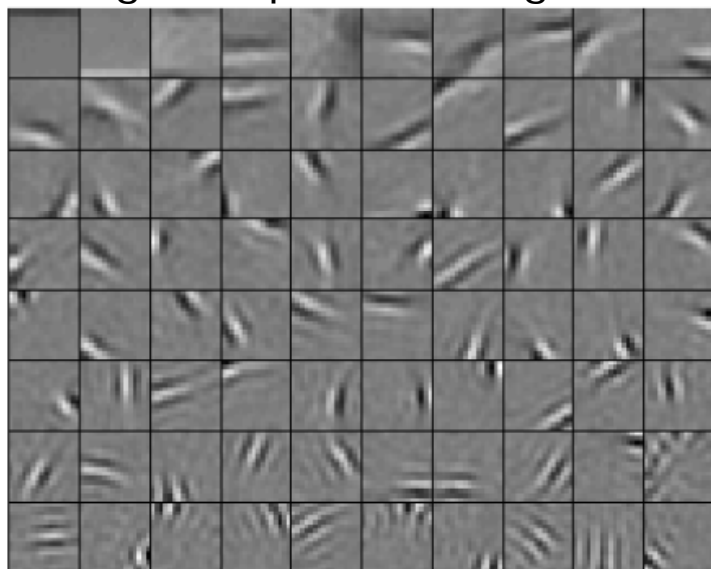
Table 1. Comparison of the CPU time for various sparse coding algorithms on different problem sizes.

# Synthetic Data Example



# Sparse coding of natural images

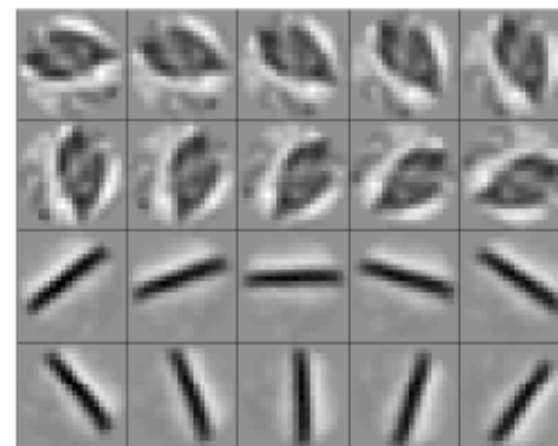
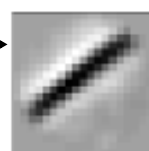
Regular Sparse Coding



On center  
off surround



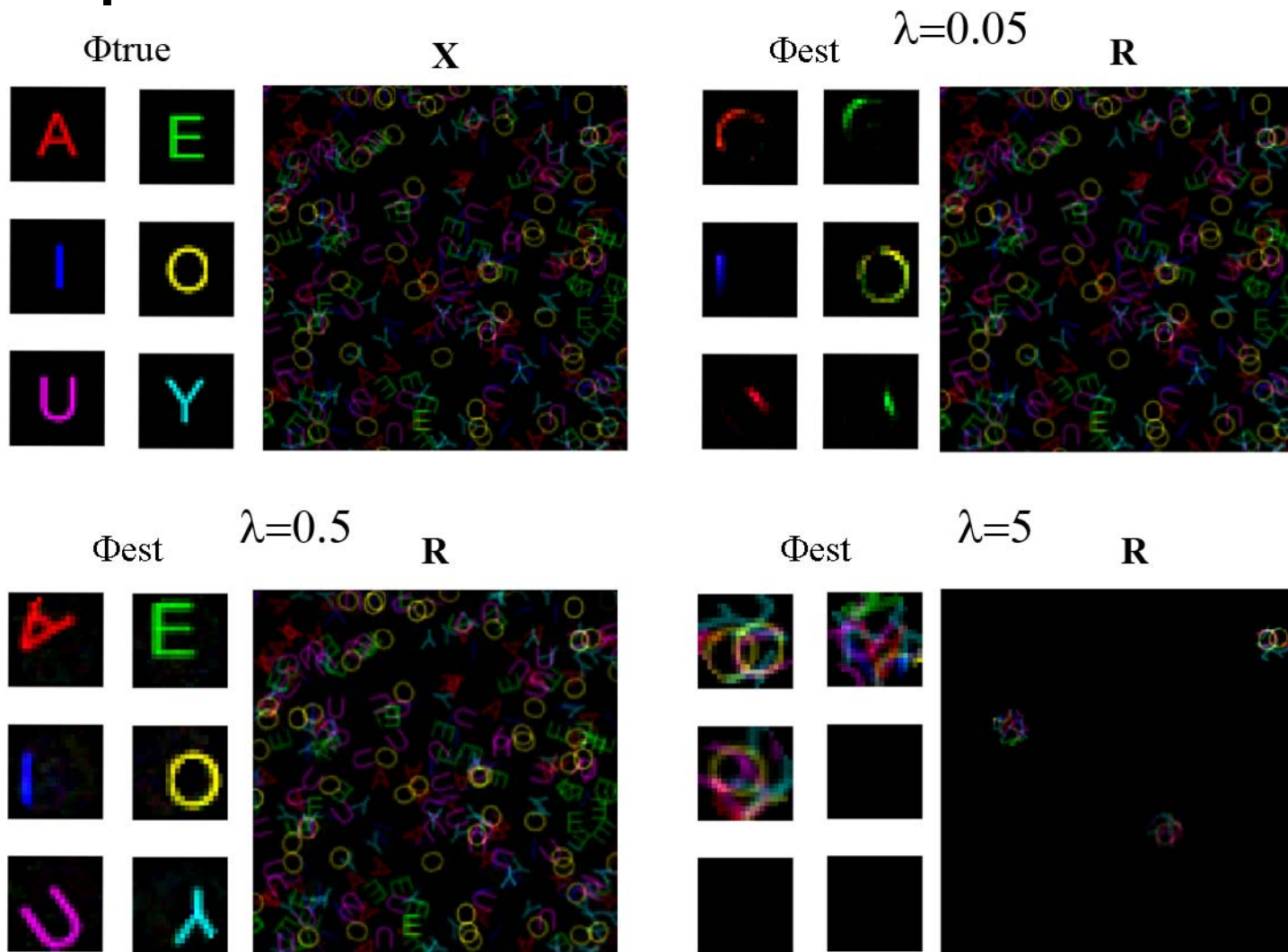
Edge detector



Shift Invariant Sparse Coding



# Rotation and shift invariant analysis of synthetic example where features also code for color



# Shift invariant analysis of a house



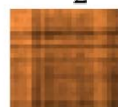
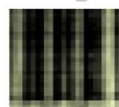
Real image



Reconstructed image

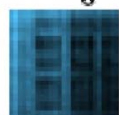
$\Psi_1$

$\Psi_2$

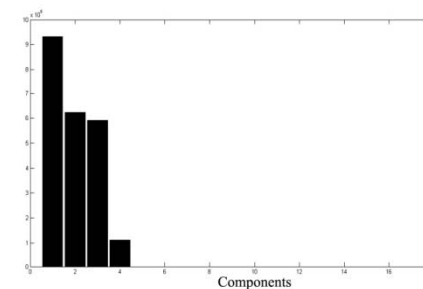


$\Psi_3$

$\Psi_4$



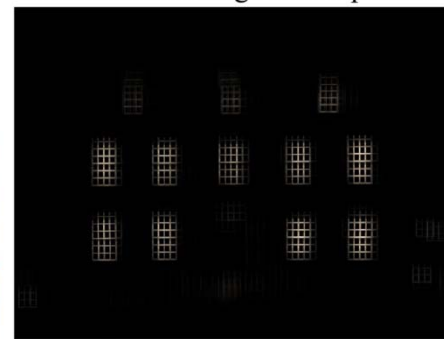
Feature images



Reconstructed Image of component 2



Reconstructed image of component 4



## Outlook and open problems

- Presently we considered shift and rotation invariance, but other types of transformations such as **scale invariance** can be efficiently implemented using resampling of the feature images  $\Psi_d$  which can be efficiently implemented by **zero padding in the fourier domain**.
- An open problem is the **estimation of  $\lambda$**  as well as the **number of components,  $D$** . We envision that automatic relevance determination (ARD) within Bayesian inference can be employed by **imposing a component wise hyperparameter  $\lambda_d$  that is also optimized**.

**Thank you!**