On the Origins of Hierarchy in Visual Processing

Angelo Franciosini, Victor Boutin and Laurent U Perrinet
INT, Aix Marseille Univ, CNRS, Marseille, France

Motivation

It is widely assumed that visual processing follows a forward sequence of processing steps along a hierarchy of laminar sub-populations of the neural system. Taking the example of the early visual system of mammals, most models are consequently organized in layers from the retina to visual cortical areas, until a decision is taken using the representation that is formed in the highest layer. Typically, features of higher complexity (position, orientation, size, curvature, ...) are successively extracted in distinct layers [2, 3]. This is prevalent in most deep learning algorithms and stems from a long history of feed-forward architectures [2]. One of the most successful paradigm to achieve such a representation relies on algorithms performing alternately sparse coding and dictionary learning. As shown in previous studies [4, 6], such an algorithm converges to a set of kernels that has strong analogies with the receptive fields of simple cells located in the Primary Visual Cortex of mammals (V1). Using the Multilayer Convolutional Sparse Coding (ML-CSC) from [8] we unsupervisedly trained a simple two-layer convolutional neural network on a set of natural images with a growing number of neurons in the second layer. By doing this, we could quantitatively manipulate the complexity of the representation emerging from such learning and analyze the sub-populations composed by the combination of the simple-cell-like oriented kernels found in the first layer.
Within the ML-CSC, [8] gives theoretical guarantees of stability and recovery for the learning and coding problem. Given an input signal $y_k \in \mathbb{R}^N$, this problem consists in finding a set of sparse maps $\{\gamma_i^k\}_{i=1}^L$ and dictionaries $\{D_i\}_{i=1}^L$ that fit the the Lasso formulation:

$$\begin{cases} \min_{\{\gamma_i^k\}\{D_i\}} \sum_{k=1}^{K} \|y^k - D^{(L)} \circledast \gamma_L^k\|_2^2 + \sum_{i=2}^{L} \zeta_i \|D_i\|_1 + \lambda_L \|\gamma_L^k\|_1 \\ D_i = [d_i^1, d_i^2, ..., d_i^J] \\ s.t. \quad \forall i, j \quad \|d_i^j\|_2 = 1, \end{cases} \tag{1}$$

where d_i^j is the j^{th} atom of the i^{th} dictionary.

ML-CSC algorithm

```
Input: training set $\{y_k\}_{k=1}^K$, initial dictionaries $\{D_i\}_{i=1}^L$
for $k = 1$ to $K$ do
    $D^{(L)} = D_1 \circledast D_2 \circledast ... \circledast D_{L-1}$
    $\hat{D}^{(L)} \leftarrow D^{(L)}/Norm(D^{(L)})$
    $\gamma_L = SparseCoding(y_k, \hat{D}^{(L)}, \lambda_L)$
    $\gamma_L \leftarrow \gamma_L/Norm(D^{(L)})$
    for $i = L$ to $2$ do
        $D_i \leftarrow \mathcal{S}_{\zeta_i}(D_i - \eta\nabla f(D_i))$
        $D_i \leftarrow D_i/Norm(D_i)$
    end
    $D_1 \leftarrow D_1 - \eta\nabla f(D_1)$
    $D_1 \leftarrow D_1/Norm(D_1)$
end
Output: $\gamma_L, \{D_i\}_{i=1}^L$
```

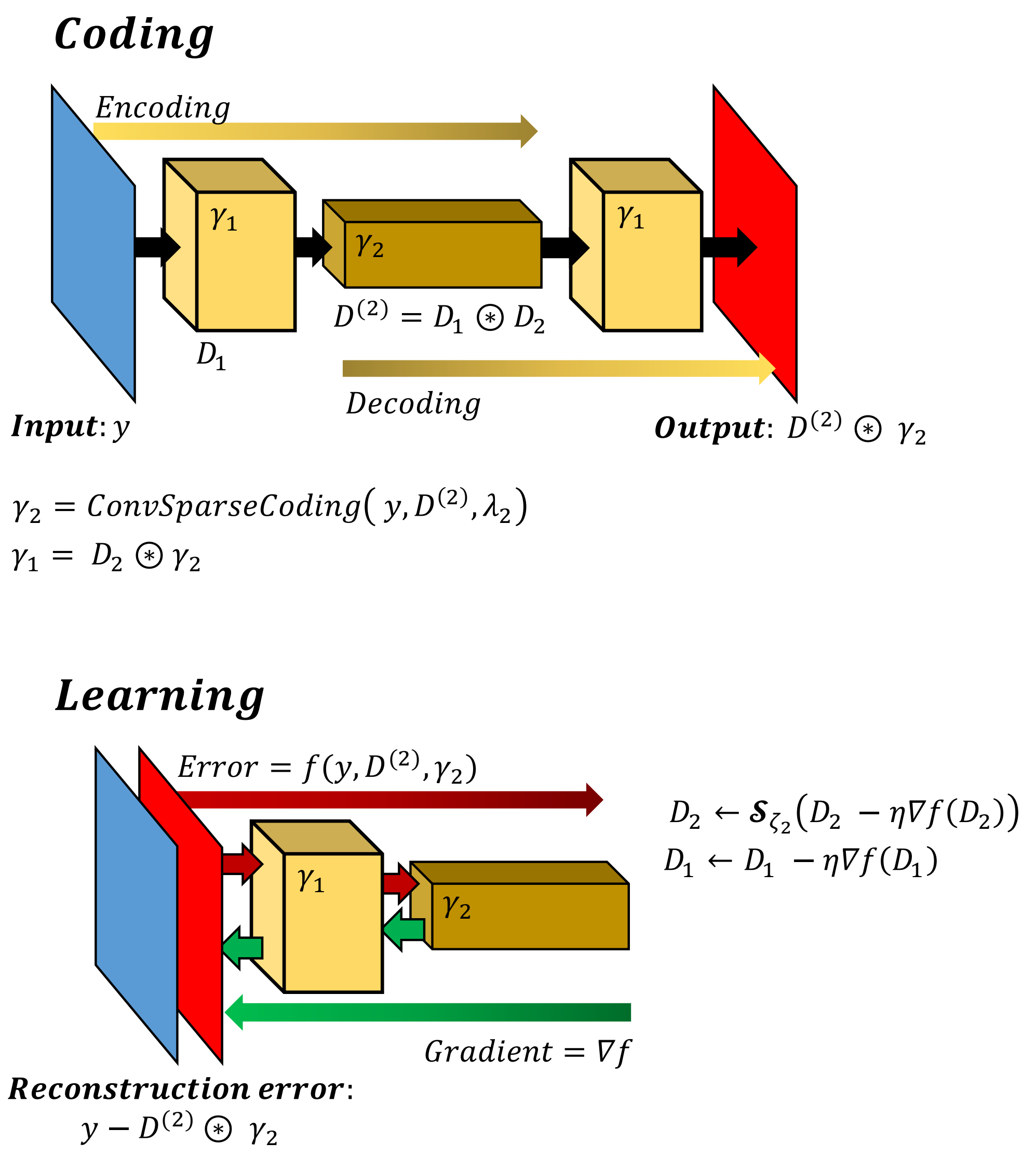
Coding
Encoding
γ_1
γ_2
γ_1
D_1
$D^{(2)} = D_1 \circledast D_2$
Decoding
Input: y
Output: $D^{(2)} \circledast \gamma_2$
$\gamma_2 = ConvSparseCoding(y, D^{(2)}, \lambda_2)$
$\gamma_1 = D_2 \circledast \gamma_2$
Learning
$Error = f(y, D^{(2)}, \gamma_2)$
$D_2 \leftarrow \mathcal{S}_{\zeta_2}(D_2 - \eta\nabla f(D_2))$
$D_1 \leftarrow D_1 - \eta\nabla f(D_1)$
γ_1
γ_2
$Gradient = \nabla f$
Reconstruction error:
$y - D^{(2)} \circledast \gamma_2$

Fig. 1:
Scheme of ML-CSC: Schematic view of the Sparse Dictionary Learning in the ML-CSC framework for $L = 2$. First the inner representation has been inferred via Convolutional Sparse Coding, in this case a modified ISTA [1]. Then the reconstruction error is propagated trough the network forward and backward in order to calculate the gradient of the optimization function. The first layer dictionary D_2 is then updated trough a gradient descent algorithm, while a proximal operator, in this case a soft thresholding is used to induce sparseness in the second layer dictionary D_2.

Why sparse coding?

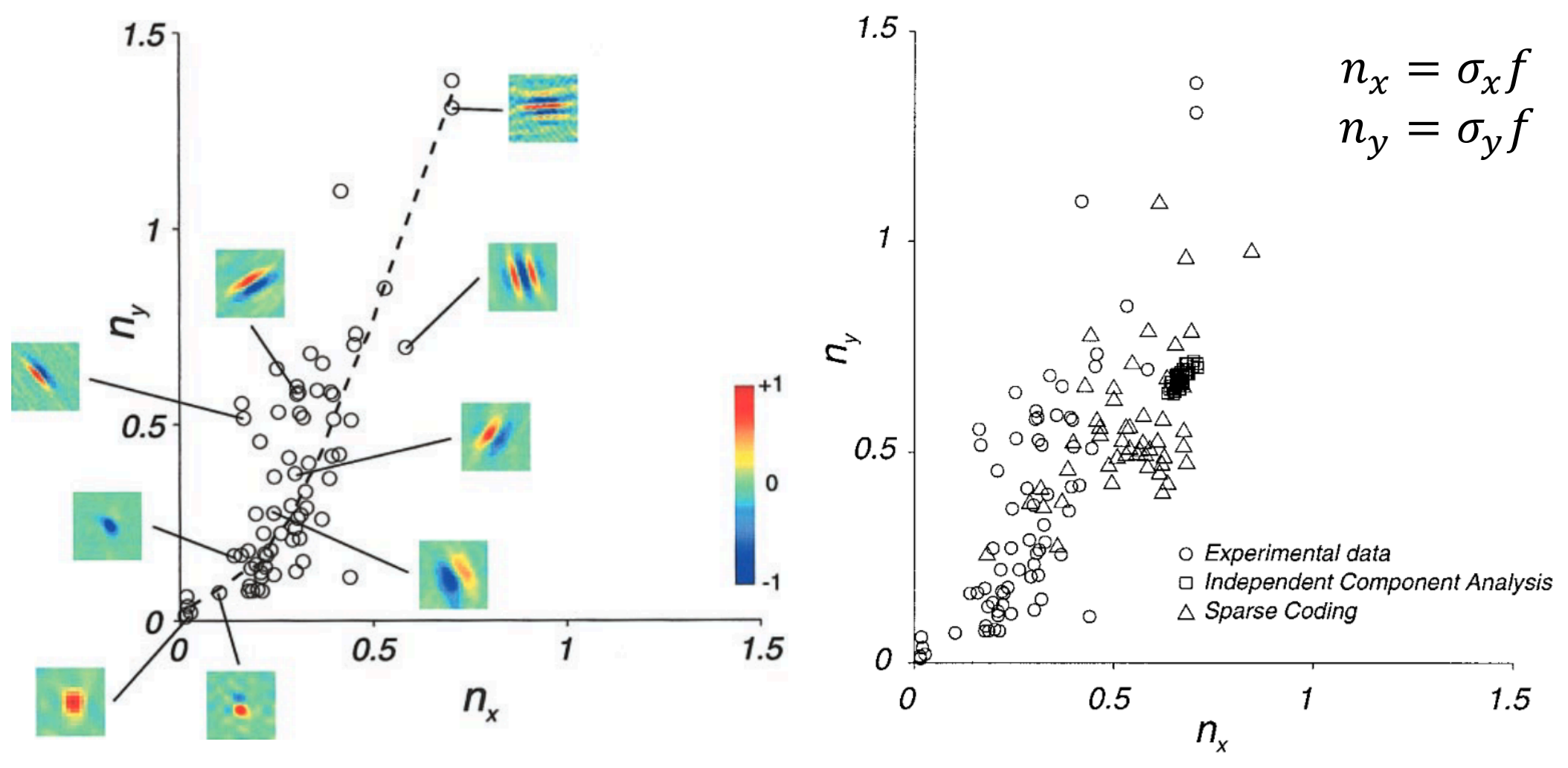

Fig. 2:
Resemblance between RCs found in V1 and those predicted by sparse coding: Example of the ability of Sparse Dictionary Learning to predict bandpass, localized, oriented filters similar to the receptive fields (RCs) of simple cells in the primary visual cortex. Other Linear Methods like ICA fail in such task. Adapted from [6]

Learning on natural images

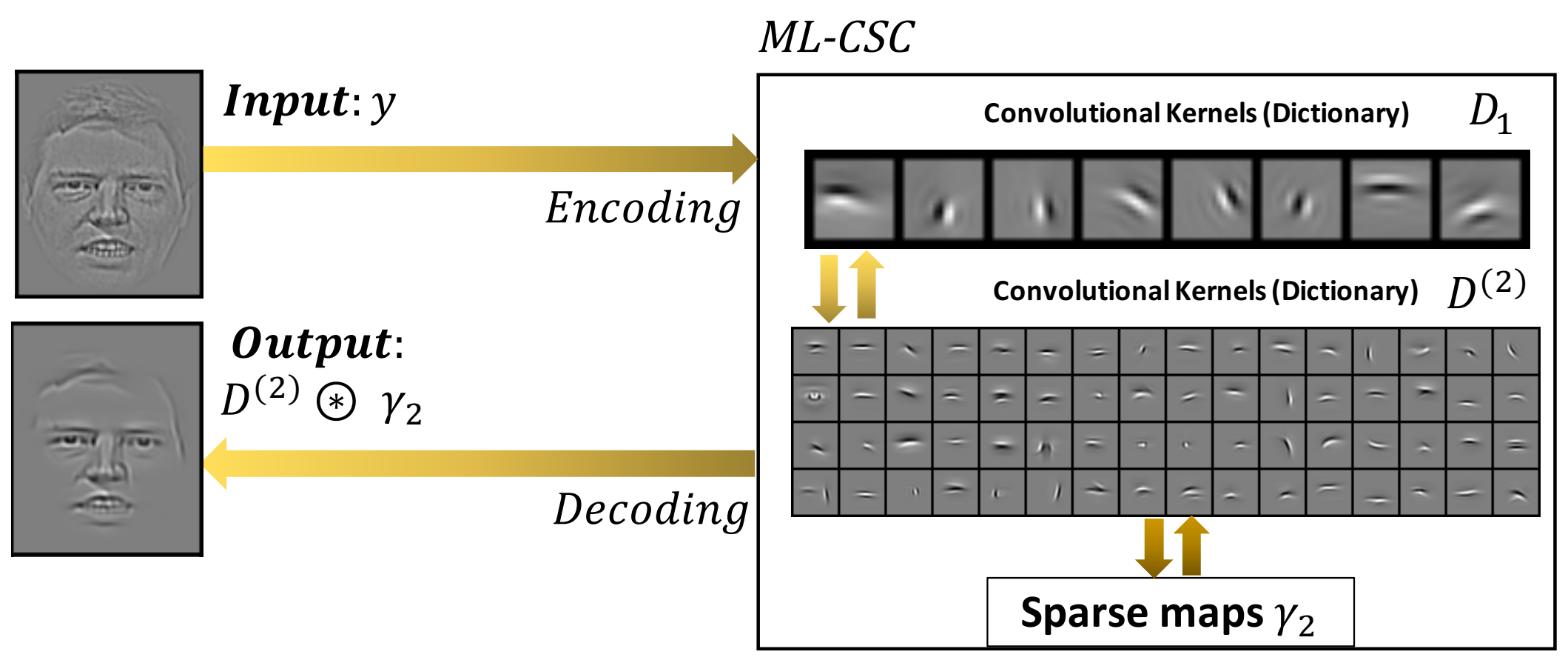
ML-CSC
Input: y
Encoding
Convolutional Kernels (Dictionary) D_1
Convolutional Kernels (Dictionary) $D^{(2)}$
Output:
$D^{(2)} \circledast \gamma_2$
Decoding
Sparse maps γ_2

Fig. 3:
Example of ML-CSC applied to the AT&T face dataset [7]

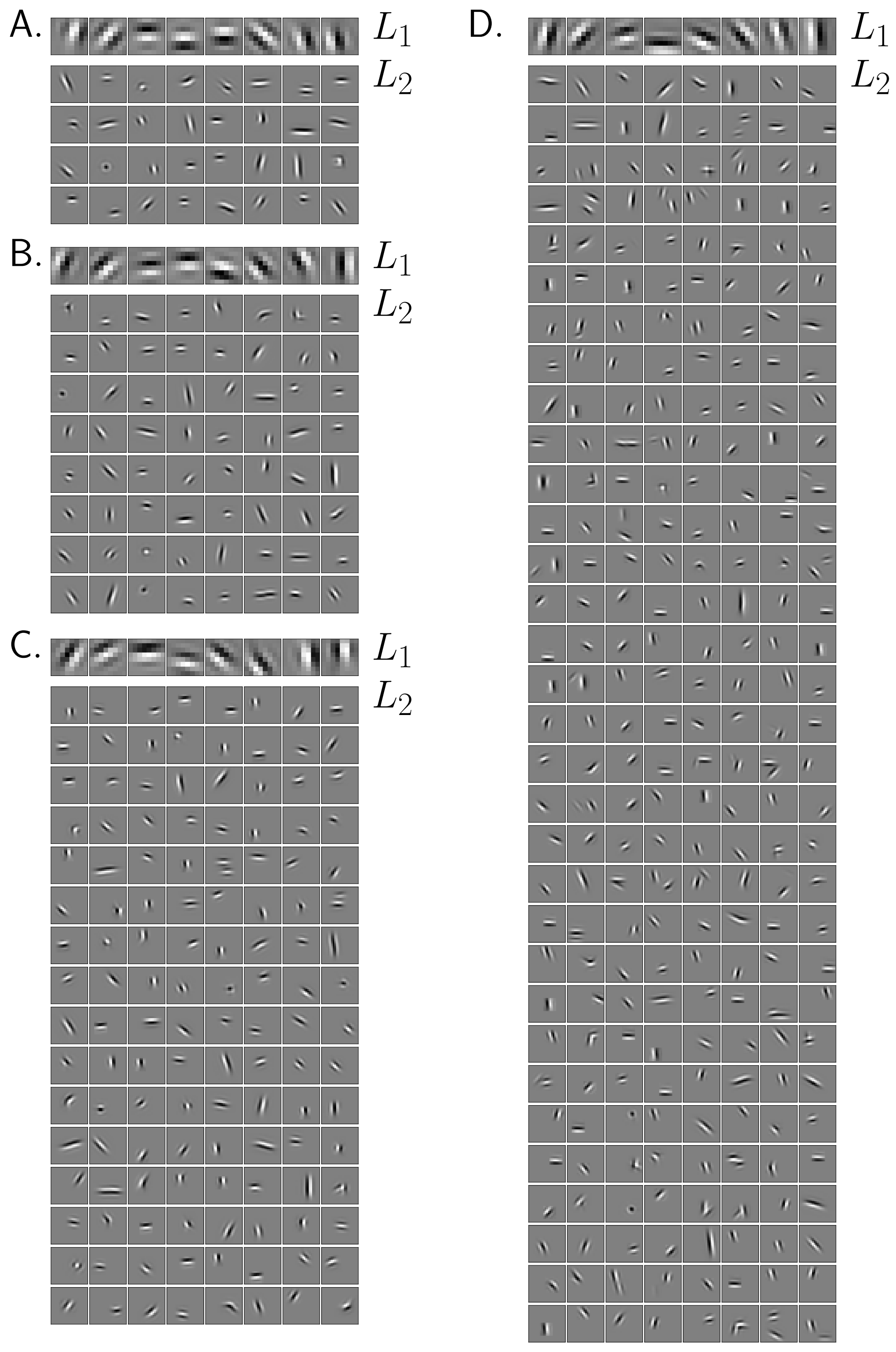

Fig. 4:
Results for different architectures with increasing complexity: We show the result of the optimization function (1) applied to a dataset of patches (64×64 pixels) extracted from a dataset of natural with a level of sparsity $\lambda_L = 10$. All the networks were trained with a first layer dictionary, D_1, composed of 8 convolutional kernels (atoms d_1^j) of 8×8 pixels (L_1). We trained 4 different networks with an increasing number of atoms d_2^j in the second layer dictionary: (A), (B), (C) and (D), corresponding to second layer dictionaries composed of respectively 32, 64, 128 and 256 atoms d_2^j. A single atom d_2^j was always composed of 9×9 pixels and 8 channels, with an effective dimension of 16×16 pixels (as in Fig. 1, L_2).

Quantitative analysis

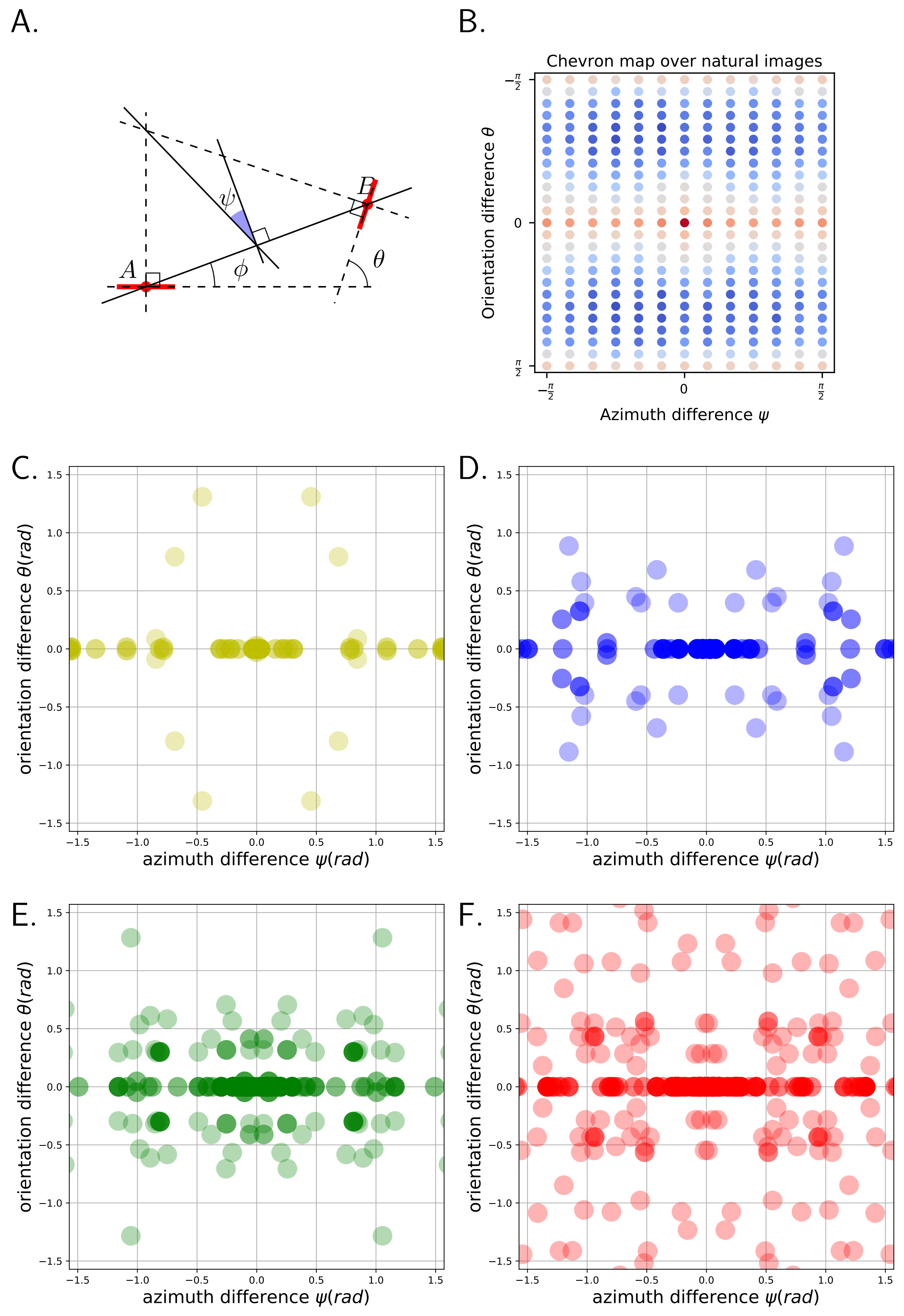

Fig. 5:
Scatter plot of orientation co-occurrences: The analysis of co-occurrences was performed using two measures of angular distance between each pair of edges, (the localized oriented filters in D_1), in the effective representation of the second layer dictionary $D^{(2)}$ [5]. The two measures are the orientation difference θ and the azimuth difference ψ (A), in particular, the axis $\psi = 0$ corresponds to co-circular edges configuration, while $\theta = 0$ to co-linear configurations, an example of this statistics extracted by a set of natural images is given in (B). We show how co-linear features are dominant in all the 4 tested architectures: (C), (D), (E) and (F) corresponding to second layer dictionaries composed of respectively 32, 64, 128 and 256 atoms d_2^j. The first architecture (C) shows almost uniquely co-linear configurations, right angles and parallel configurations start emerging for higher second later dimensions (D), (E). The greatest variety is present in the last architecture (F) where, though co-linearity is still dominant, a greater number of complex combination can be observed: co-circular, right angles and ”Y” shaped configurations.

References

[1] Amir Beck and Marc Teboulle. “A Fast Iterative Shrinkage-Thresholding Algorithm”. In: *Society for Industrial and Applied Mathematics Journal on Imaging Sciences* 2.1 (2009), pages 183–202. ISSN: 1936-4954. DOI: 10.1137/080716542.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), page 436.

[3] DA Mely and Thomas Serre. “Towards a system-level theory of computation in the visual cortex”. In: *Computational and Cognitive Neuroscience of Vision* (2016), pages 1–28. DOI: 10.1007/978-981-10-0213-7.

[4] Bruno A. Olshausen and David J. Field. “Sparse coding with an overcomplete basis set: A strategy employed by V1?” In: *Vision Research* 37.23 (1997), pages 3311–3325. ISSN: 00426989. DOI: 10.1016/S0042-6989(97)00169-7.

[5] Laurent U. Perrinet and James A. Bednar. “Edge co-occurrences can account for rapid categorization of natural versus animal images”. en. In: *Scientific Reports* 5 (2015), page 11400. ISSN: 2045-2322. DOI: 10.1038/srep11400.

[6] Dario L Ringach et al. “Spatial Structure and Symmetry of Simple-Cell Receptive Fields in Macaque Primary Visual Cortex Spatial Structure and Symmetry of Simple-Cell Receptive Fields in Macaque Primary Visual Cortex”. In: (2014), pages 455–463.

[7] Ferdinando S Samaria and Andy C Harter. “Parameterisation of a stochastic model for human face identification”. In: *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on.* IEEE. 1994, pages 138–142.

[8] Jeremias Sulam et al. “Multi-Layer Convolutional Sparse Modeling: Pursuit and Dictionary Learning”. In: (2017), pages 1–29.