Extended coding and pooling in the HMAX model

Christian Thériault, Nicolas Thome, *Member, IEEE,* and Matthieu Cord, *Member, IEEE*

Université Pierre et Marie Curie, UPMC-Sorbonne Universities, LIP6, 4 place Jussieu, 75005, Paris, France
firstname.lastname@lip6.fr

***Abstract*—This paper presents an extension of the HMAX model: a neural network model for image classification. The HMAX model can be described as a four-level architecture with a first level consisting of multi-scale and multi-orientation local filters. We introduce two main contributions to this model. First, we improve the way the local filters at the first level are integrated into more complex filters at the last level, providing a flexible description of object regions, combining local information of multiple scales and orientations. These new filters are discriminative and yet invariant, two key aspects of visual classification. We evaluate their discriminative power and their level of invariance to geometrical transformations on a synthetic image set. Second, we introduce a multi-resolution spatial pooling. This pooling encodes both local and global spatial information to produce discriminative image signatures. Classification results are reported on three image data sets, Caltech101, Caltech256 and Fifteen Scenes. We show significant improvements over previous architectures using a similar framework.**

I. Introduction

The task of visual classification is a cornerstone of image processing and computer vision. This remains one of the most challenging problems of the field since it implies identifying complex categories inside images, such as scenes or objects. A good classification system should respond invariably to objects within the same class and differently between classes. One key aspect of such a system is the ability to define and learn representations with a proper balance between discriminability and invariance.

In the field of computer vision, some developments have pointed towards this goal. One is the design of discriminative low-level local features, such as SIFT [1] and HoG [2]. These local descriptors provide a discriminative signature of image patches, and are invariant to various image degradations such as geometric and photometric transformations. Another development in computer vision is the emergence of mid-level representations based on the Bag of Words (BoW) model [3]. The BoW model, inspired from the text retrieval community, leads to state-of-the art performances in most standard databases. To achieve human performance level, the ultimate solution to image classification remains unclear and alternative avenues, such as biological vision, can be explored in order to define image representations.

When considering models of visual recognition it is difficult to ignore the level of performance achieved by biological vision. The mammalian visual system displays recognition abilities that cannot be matched by any artificial system and it seems wise to consider insights about the functioning of the visual cortex.

Models of the mammalian visual system mainly originate with the Nobel prize work of Hubel & Wiesel [4]. A key point of their discovery is that neurons in the visual cortex describe a manifold of localized filters organized into columns of spatial frequencies and orientations. Their work gave biological support to early psychophysical theories stating that the visual system analyzes patterns into multiple and independent frequency channels [5], [6]. More recent studies [7], [8] provide mathematical models to the work of [4]. In [8] it is shown that the point spread function of neurons in the mammalian visual cortex can be modeled by Gaussian derivative filters (i.e band-pass filter) of multiple orientations and scales. In [7] it is shown how such filters emerge by learning statistics of natural images. All these considerations regarding the local receptive fields of visual neurons are also given a strong theoretical setting in the scale-space theory of vision [9], [10], [11], [12], [13]. The scale-space theory describes the visual front end of the cortex with a family of local and scaled gaussian operators. This formulation of the visual cortex can be implemented in multi-layer neural networks composed of simple units with local receptive field profiles [14].

According to the biological model, the low level operations of these multi-layer networks are defined by local oriented filters at multiple scales (i.e. Gaussian derivatives or Gabor filters). These networks combine the low level representations into object level representations suitable for recognition tasks [15], [16], [14], [17], [18], [19]. Different types of feature combinations in the hierarchy can be considered and produce different performances [20], [21]. Physiological studies suggest that feedforward activation, with little or no feedback, produces the early recognition response while sustained feedback mechanisms generate a more attentive response [22], [14]. When modeling a purely feedforward activation, the challenge is to produce high level representations which are both discriminative and invariant. Indeed, by building complex and global representations from simple and localized features these networks face the problem of finding a balance between object specific representations and invariant representations to ensure differentiation between classes of object and invariance inside each class.

In this paper, we introduce an architecture for image classification which extends on previous work based on basic operations of the visual cortex [19], [1]. In particular, our network pools over oriented and scaled filters at the lowest level, which correspond to early operations of *simple cells* and *complex cells* in the V1 cortical area [14]. The work of [23] also presents an extension of [19] by integrating sparsity, a refined pooling strategy and a feedback mechanism to

select relevant representations. We keep this basic framework of *simple cells* and *complex cells* operations, and improve previous networks [19], [1], [24] by refining the filters on the last level of the network which integrate simple local filters into more complex filters covering larger and more complex image regions.

Specifically, we introduce two main contributions that improve the classification capacity of previous similar networks.

First, the coefficients of each filter on the last layer are trained to better discriminate the image content. Importantly, this gain in terms of discriminability is also coupled with an increase in terms of invariance. This joint discriminability and invariance improvement is achieved due to the ability to extract relevant image structures while being tolerant to various degradations such as geometric transformations or occlusions.

Second, we present a flexible multi-resolution radial approach to pool the outputs of filters across the image. Neurons in the inferior temporal visual cortex (IT) are known to have limited receptive fields of various sizes [14]. These can be interpreted as pooling over local regions of various sizes on the visual field, which results in partial invariance to spatial position. In this spirit, our multi-resolution pooling corresponds to matching a given filter inside spatial neighborhoods with different pooling radii, yielding different levels of invariance. The optimal level of invariance, for a classification task, can then be learned by a classifier in a supervised manner at the highest level of the network.

The remainder of this paper is organized as follows. Section II presents state-of-the-art methods that are the most connected to ours. The general HMAX network architecture is depicted in section III. Section IV gives the details of our contributions, while sections V and VI give supporting experimental results. Finally, section VII concludes the paper and gives directions for future works.

II. Related Work

In this section, we review the approaches which are the most relevant to our approach.

A. Bag of Words (BoW) Methods

BoW models have extensively been studied in the last decade due to their good performances for classifications in many object or scene databases. In the BoW model [3], a set of local and accurate descriptors (*e.g.* SIFT) is first computed, forming the so-called "Bag of Features" (BoF) for a given image. The BoF is then transformed to a constant-size image representation to generate the Bag of Words (BoW). The BoW can be interpreted as an occurrence histogram of visual words, where the visual codebook (dictionary) has been trained from a set of local descriptors. The mapping of visual codebooks against image descriptors can be decomposed into a coding phase followed by a pooling step, as formalized in [25]. In the original BoW model [3], a simple vector quantization stage is applied for coding, and the codes are aggregated with an average pooling strategy. Several improvement have been proposed to improve coding and pooling steps. To reduce the quantization errors induced by vector quantization, one may rely on soft assignment [26], or sparse coding techniques [27], [25], that explicitly minimize reconstruction error. The Restricted Boltzman Machine (RBM) model has been used to produce fast sparse coding inferences [28]. Regarding pooling, max pooling has recently been studied and proved to be a good alternative to sum pooling, especially when linear classifiers are used. An extension of the BoW formalism uses a pooling which encodes the distance-to-codeword distribution [29]. Another extension of BoW models using Fisher kernels which benefits from both generative and discriminative approaches has also shown good classification results [30]. Finally, since the BoW model ignore spatial information, most of the approaches integrate the Spatial Pyramid Scheme (SPM) [31], also extended in the context of photographic style image classification [32]. Other approaches, based on the BoW model, also encode the relative spatial distributions between visual words [33], [34]. Learning algorithms have also been used to learn efficient feature combinations [35].

B. Deep & biologically inspired architectures

Multi-layer networks or the convolutional networks introduced by LeCun *et al.* [17], [36] are certainly amongst the pioneer works of this type of architecture. The main idea is to learn each layer representations from data. In the original convolutional networks, parameters of the whole network are trained in a supervised manner using the error backpropagation algorithm. Ranzato *et al.* [17] focus on unsupervised learning of features at every layer of a standard convolutional neural network, while Lee *et al.* [37] propose to use a Convolutional Restricted Boltzman Machine (CRBM) for image categorization, and report promising performances. A key aspects of these type of models consists in learning a hierarchical composition of filters [38], [17]. The depth of these models, although appealing, implies a very large number of coefficients to be learned and often require to solve complex and highly non convex optimization problems.

Other biologically inspired models focus on building networks of simple and complex features based on physiological data about the mammalian visual pathways [4], [15], [39], [14], [40], [16]. Beginning with low level filters, matching the physiological recordings in the early steps of the visual pathway, the challenge is to organize such low level representations into a coherent robust object level representation. The low level representations are often fixed but supported by physiological studies, while higher level representations can be learned and driven by a specific task. The pioneer work of [15], [41], [36] has shown how a deep architecture can be trained to merge simple visual features into a more complex whole while retaining some degree of invariance to basic visual transforms. In [40] effort is put in statistical learning of the relative position of simple features and carrying this information into a global discriminative representation. The work in [16] uses temporal correlations, between views of a transforming object, to learn a multi-layer architecture with invariant properties to various visual transformations. In [39] the emphases is put on the unsupervised learning of relevant object level features using the temporal aspect of neural encoding.

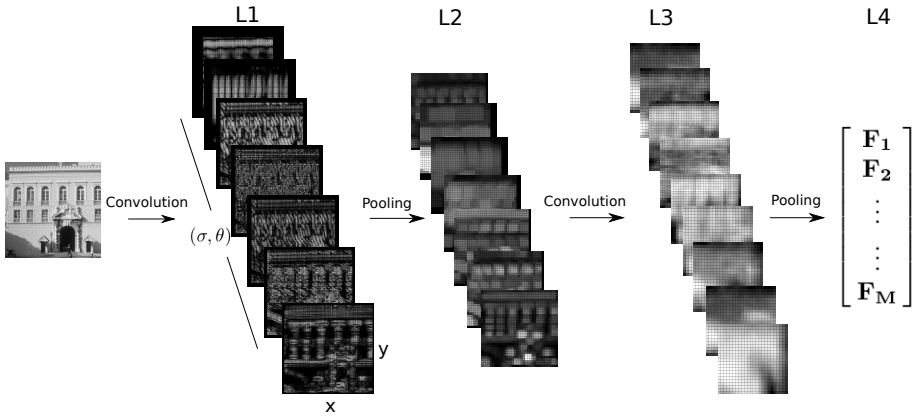

Fig. 1. General convolution network: the network alternates layers of feature mapping (convolution) and layers of feature pooling. The convolution layers generate specific feature information whereas the pooling layers generate invariance by relaxing the configuration of these features.

Another important contribution to biologically inspired models comes from the HMAX model [18], which focuses less on learning and more on designing simple operations inspired by the visual cortex. This networks alternates layers of features extraction with layers of maximum pooling, mimicking known data about the ventral pathway of visual cortex. Serre *et al.* [19] extend the original HMAX model to add multi-scale representations as well as more complex visual features. Huang *et al.* [23] also improved the HMAX model with sparsity constraints, a different pooling strategy and a feedback mechanism to improve feature learning. The model proposed by Mutch *et al.* [1] is the most closely related to ours: they improve the network of [19] by tuning the filters to the dominant local orientations. In our previous work [24], we further improved this idea to local scales. In [42] it is also shown how HMAX filters can outperform state-of-the-art filters such as SIFT under various controled invariance tasks on synthetic images.

We extend on the properties of this particular family of models. We use two levels of filters in which the second level filters are trained to optimally fit the dominant local geometry of images. Our contributions can be summarized as follows:

- Training of filters which generate representations that are more discriminative and more invariant
- Design of a novel multi-resolution pooling that codes the spatial distribution of each category.
- Experimental validations of the discriminative power and invariance level of the network with respect to previous models

Additionally, we highlight the complementarity between our representations and the local descriptors used in BoW models. We combine both strategies to reach or outperform state-of-the art results. The local descriptors used in the BoW models are usually fine-grained description of an image, which correspond to small image areas, *e.g.* 16×16 pixel patches. The descriptors presented in this paper operate on a different scale and correspond to features with larger spatial extent, and constitute therefore complementary representations from those extracted with BoW models.

III. General HMAX model

The general HMAX model follows a basic alternating convolution/pooling scheme as in [19], [18] and illustrated in figure 1. Each convolution step yields a set of feature maps and each pooling step provides robustness to variations in these feature maps. Below we describe the operations of each layer as done in [19].

Layer 1. Each feature map $\mathbf{L1}_{\sigma,\theta}$ can be obtained by convolution of the input image with a set of Gabor filters $g_{\sigma,\theta}$ with orientations θ and scales σ. These filters are used to model *simple cell* activation in the V1 area of the visual cortex [14]

$$g_{\sigma,\theta}(x,y) = \exp\left(\frac{x_o^2 + \gamma y_o^2}{2\sigma^2}\right) \cdot \cos\left(\frac{2\pi}{\lambda} x_o\right), \tag{1}$$

where $x_o = x\cos\theta + y\sin\theta$ and $y_o = -x\sin\theta + y\cos\theta$. The parameter γ indicates the aspect ratio of the filter and λ its wavelength.

Given an image I, Layer 1 at orientation θ and scale σ is given by the absolute value of the convolution product

$$\mathbf{L1}_{\sigma,\theta} = |g_{\sigma,\theta} * I|. \tag{2}$$

Layer 2. Each feature map $\mathbf{L2}_{\sigma,\theta}$ is a dimension reduction of $\mathbf{L1}_{\sigma,\theta}$ obtained by selecting maxima on local neighborhoods. A well known effect of maximum pooling over local neighborhood is the invariance to local translations and thereby to global deformations [18], [15].

Specifically, the second layer partitions each $\mathbf{L1}_{\sigma,\theta}$ map into small neighborhoods $\mathbf{u}_{i,j}$ and selects the maximum value inside each $\mathbf{u}_{i,j}$ such that

$$\mathbf{L2}_{\sigma,\theta}(i,j) = \max_{\mathbf{u}_{i,j} \in \mathbf{L1}_{\sigma,\theta}} \mathbf{u}_{i,j}. \tag{3}$$

Some degree of scale invariance is also achieved by keeping only the maximum output over two adjacent scales at each position (i,j).

Layer 3 Layer $\mathbf{L3}$ at scale σ is obtained by convolving filters $\boldsymbol{\alpha}^m$, which we call HL filters, against layer $\mathbf{L2}_\sigma$

$$\mathbf{L3}_\sigma^m = \boldsymbol{\alpha}^m * \mathbf{L2}_\sigma. \tag{4}$$

HL filters represent visual descriptors of "mid-level" areas in the image which combine "low-level" Gabor filters of multiple orientations at a given scale. To compute equation 4, HL filters must first be trained as described below.

Training In the basic HMAX framework [19], as shown in figure 2, the HL filters $\boldsymbol{\alpha}^m$ are the result of a sampling process over the layer $\mathbf{L2}$ of training images. This sampling process has three parameters: scale, spatial position, and spatial size. Specifically, HL filters are generated by randomly sampling *prototype blocks* of $\mathbf{L2}$ coefficients of spatial size $n \times n$ at position (x, y) and scale σ, covering all orientations θ. In [19], $M \sim 1000$ *prototypes blocks* are sampled over the training set to create M HL filters. For illustration, in figure 2, the shaded blue represents the entire $\mathbf{L2}$ layer with all scales and orientations concatenated together along the z axis. The shaded red, $\mathbf{L2}_\sigma$, illustrates one slice of the layer at scale σ and containing all the orientations.

Layer activation As shown by the right part of figure 2, each sampled block defines one HL filter which can be later matched against the $\mathbf{L2}$ layer of new images. In [19] each HL filter is matched against layer $\mathbf{L2}$ at all spatial positions and all scales. Specifically, as given by equation 4, each HL filter is convolved over each scale map $\mathbf{L2}_\sigma$ to produce the feature maps $\mathbf{L3}_\sigma^m$.

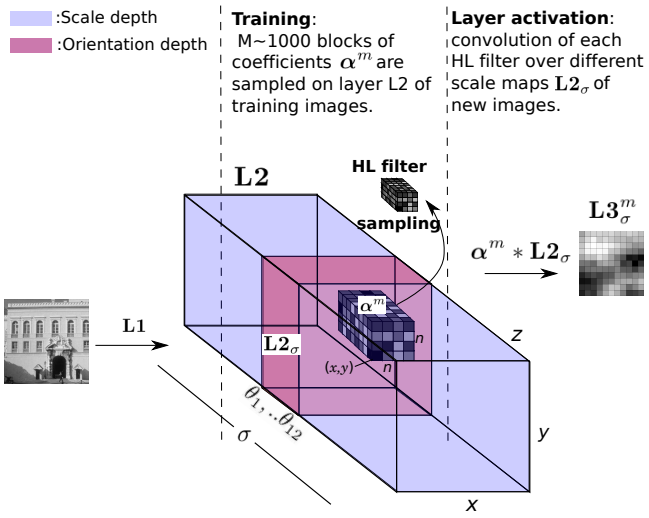

Fig. 2. HMAX: level-3 operations. Training: $M \sim 1000$ HL filters are defined by sampling *prototype blocks* of $\mathbf{L2}$ coefficients from training images. Layer activation: Given a new image, each HL filters is convolved over all positions of each scale map $\mathbf{L2}_\sigma$.

Layer 4 To gain global invariance, the final signature is computed by selecting the maximum output of $\mathbf{L3}_\sigma^m$ across all position and scales. The final layer is thus a vector of dimension $M \sim 1000$ where each coefficient gives the maximum output of each HL filter across scales σ and positions (x, y).

$$\mathbf{L4} = \begin{bmatrix} \max\limits_{(x,y),\sigma} \mathbf{L3}_\sigma^1(x,y) \\ \vdots \\ \max\limits_{(x,y),\sigma} \mathbf{L3}_\sigma^M(x,y) \end{bmatrix}. \quad (5)$$

IV. Advanced coding and spatial pooling strategy

Here we describe our contribution to the HMAX model and give parameter details for each layer.

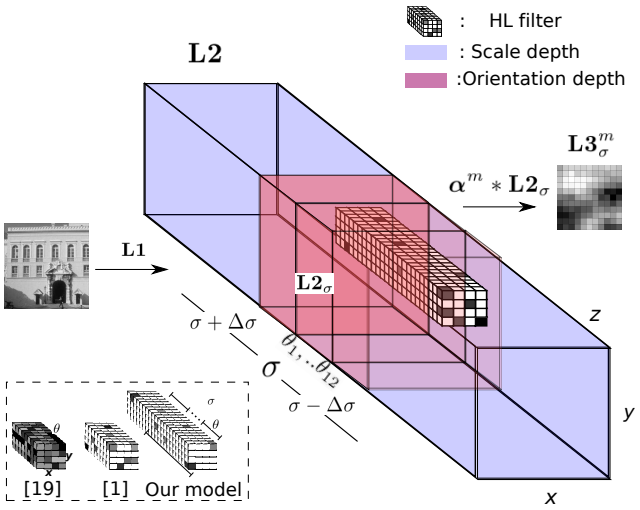

Fig. 3. Our model: level-3 operations. In our model each HL filter spans over multiple scales simultaneously. In this example, the filter is convolved simultaneously over multiple scales centered at scale σ At training the coefficients corresponding to weak scales and orientations are set to zero making our filter more discriminative, ignoring weaker scale and orientations during testing.

Layer 1. As done in [19], we use equation 2 to define each $\mathbf{L1}_{\sigma,\theta}$ map. The scale range of σ varies with grid size according to table I. We use a range of 12 orientations $\Theta = \{\frac{k\pi}{12}\}, k \in \{0 \dots 11\}$ and 8 scales $\mathbb{S} = [\sigma_1, .., \sigma_8]$. The aspect ratio was set to $\gamma = 0.3$ to match the settings in [19]. To ensure scale invariance, each filter is normalized to zero mean and unit length. To obtain invariance to light intensity, each pixel patch during the convolution product in equation 2 is normalized to unit length before being multiplied by the filter.

Scale	Filter size	σ	λ
1	7×7	2.8	3.5
2	11×11	4.5	5.6
3	15×15	6.7	7.9
4	19×19	8.2	10.3
5	23×23	10.2	12.7
6	27×27	12.3	15.5
7	31×31	14.6	18.2
8	35×35	17.0	21.2

TABLE I
LAYER 1 FILTERS PARAMETERS. AS DONE IN [19] THE SCALE AND WAVELENGTH OF EACH FILTER IS CHOSEN TO MATCH PHYSIOLOGICAL RECORDINGS.

Layer 2. Each feature map $\mathbf{L2}_{\sigma,\theta}$ is obtained as described for the HMAX model in section III. Similarly to [1], we applied a competition to both the orientation and scale coefficients setting to zero the weaker coefficients at each position (i, j). We used pooling neighborhoods $\mathbf{u}_{i,j}$ of sizes proportional to the scale of processing as in [19] and given by table II.

Layer 3 Our first principal contribution is in way the HL filters are trained and used to generate layer $\mathbf{L3}$.

Scale	Neighborhood $\mathbf{u}$ sizes
1	8×8
2	10×10
3	12×12
4	14×14
5	16×16
6	18×18
7	22×22
8	24×24

TABLE II
LAYER 2 MAXIMUM POOLING NEIGHBORHOOD SIZE. THE POOLING SIZE IS PROPORTIONAL TO THE FILTERING SCALE OF THE PRECEDING LAYER.

Training Our methods contrasts with earlier models [18], [19], [1] since our HL filters are not limited to a single scale. This difference with the original HMAX model is illustrated when comparing figure 2 with figure 3. There are two main differences to be noted between the two figures:

1) Modeling : our HL filters cover a range $\sigma \pm \Delta\sigma$ of scales simultaneously. This gives more representation power to each HL filter. By increasing its scale range, each HL filter can represent "mid-level" structures containing multiple scales inside the same spatial neighborhood. This is an improvement over the representation in [19], [1], where the HL filters span a single scale, limiting the possibility of each filter to match the local scale of image structures.

2) Robustness : our HL filters are trained to optimally match the dominant local scales and orientations, discarding weaker training scales and orientations. By setting its coefficients on Gabor filters which produce strong training outputs, each HL filter gains robustness to interfering orientations and scales (i.e noise and clutter) when presented with a new image. This principle has been introduced in [1] for the case of orientations. As shown at the bottom left of figure 3, the HL filters in [1] are a refined version of [19], where the HL filter coefficients corresponding to weak orientations are set to zero (white cubes set to zero). This increased discriminative power reduces interference caused by weak orientations during testing. We extend the principle of [1] by also specializing each filter on the dominant local scale, setting to zero coefficients corresponding to weaker scales. This makes our HL filters even more discriminative by ignoring weaker scales and orientation on test images.

Algorithm 1 summarizes the steps for training the HL filters. As for the HMAX our HL filters are trained by sampling *prototype* blocks from the $\mathbf{L2}$ layer of training images. Specifically, we sample *prototype* blocks $B^m \in \mathbb{R}^{n\times n\times|S|\times|\Theta|}$ from layer $\mathbf{L2}$. The dimension $n \in \{4, 8, 12, 16\}$ define the spatial size of the HL filter, $|S| \in \{1, 3, 5, 7\}$ its scale range, and $|\Theta| = 12$ its orientation range. Using the sampled *prototype* block B^m, the coefficients $\alpha^m_{i,j,\sigma,\theta}$ are set such that

$$\alpha^m_{i,j,\sigma^*,\theta^*} = \begin{cases} B^m_{\sigma^*,\theta^*}(i,j) & \text{if } (\sigma^*,\theta^*) = \arg\max B^m_{\sigma,\theta}(i,j) \\ 0 & otherwise \end{cases}. \quad (6)$$

Equation 6 can be thought of as a learning rule which sets to zero the connections on weak orientations and scales (white cubes in figure 3) after being presented with a single *prototype example* B^m. For each HL filter, the coordinate $s_m = (x_m, y_m, \sigma^m_s)$ at which *prototype* B^m is sampled is memorized. This memorized coordinate is used at layer $\mathbf{L4}$ to encode spatial information about each HL filter when presented with new images.

Algorithm 1 HL filter training

Require: M: number of filters to train
$|\Theta| = 12$: number of orientations
$|S| \in 1, 3, 5, 7$: number of scales
$n \in 4, 8, 12, 16$: spatial size
for $m = 1$ to M **do**
  Select one training image
  Activate up to layer L2
  Select a random coordinate $s_m = (x_m, y_m, \sigma^m_s)$ on L2
  Extract a random sample $B^m \in \mathbb{R}^{n\times n\times|S|\times|\Theta|}$ at position s_m
  Apply equation 6 to get α^m
end for
return s_m , α^m

Layer activation Each feature map $\mathbf{L3}^m_\sigma$ is a convolution product of the $\mathbf{L2}$ layer with filter $\boldsymbol{\alpha}^m$ centered at scale σ. Specifically, the output $\mathbf{L3}^m_\sigma(x, y)$ is given by the dot product of $\boldsymbol{\alpha}^m$ with the block $\mathbf{L2}_\sigma(x, y)$ at spatial position (x, y), centered on scale σ.

$$\mathbf{L3}^m_\sigma(x, y) = \langle \alpha^m, \mathbf{L2}_\sigma(x, y) \rangle. \quad (7)$$

For our most basic network we use $M = 4080$ filters to generate M maps $\mathbf{L3}^m_\sigma$. We normalize to unit length each components of equation 7 so that it gives the cosine between both components. In [19], [1] a radial basis function (RBF) is used for layer $\mathbf{L3}$. After experimental verification showing better performances we chose a normalized dot product as opposed to a RBF. We observe near 3% improvement in classification score using the normalized dot product defined in equation 7 when compared to the RBF function used in [1], [19]. One possible advantage of using normalization here is to ensure that geometrical similarities of features are kept invariably with respect to light intensity variations.

The toy example (synthetic image) in figure 4 shows how our HL filters adapts differently to local image structures when compared with [1], [19]. In the figure, the red ellipses correspond to the local scales and orientations of Gabor filters selected by the HL filter $\boldsymbol{\alpha}^m$. In [1] the randomly chosen scale of the filter is misadapted to the local image scale. This results in sub-optimal Gabor filters selected along the edges and not corresponding to the optimal local scale of the training image. In [19] the randomly chosen scale is again misadapted to the local image scale and all orientations are trained. As shown, our HL filter adapts to the optimal local scale and orientation of the training image.

Layer 4. Our second main contributions is in the way the outputs of HL filters $\boldsymbol{\alpha}^m$ are spatially pooled together to create

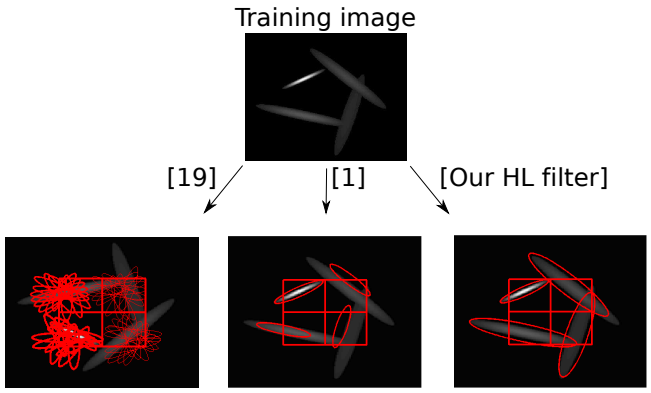

Fig. 4. Toy examples. The red ellipses indicates the local scale and orientation of Gabor filters selected by the HL filter. On the left, the HL filter is trained at a randomly chosen scale on all orientations as done in [19]. In the middle the HL filter is trained on a randomly chosen scale and learns the optimal orientation as done is [1]. On the left, our HL filter adapts to the optimal scale and orientation.

a full image signature at layer **L4**. As done in [19] (figure 5) one can store the maximum global output of each HL filter into one vector signature. In [1], spatial information is represented by memorizing the training position of each HL filter and then taking the maximum output for each test image in the neighborhood of the training position. In [31] a pyramidal pooling approach (SPM) is used to code spatial information.

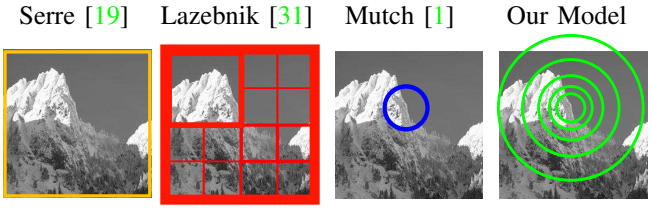

Fig. 5. Image partitions for pooling. In [19] the entire image is used to pool each HL filter. In [31] a pyramidal (SPM) partition of the image is used to code spatial information. In [1] a localized pooling regions is defined for each HL filter. Our pooling is localized with multiple spatial resolutions.

Here, we introduce a spatial pooling which merges aspects of the pyramidal pooling in [31] and the localized pooling in [1]. Using these principles, our HL filters perform a maximum pooling over image regions of various sizes. Specifically, for each HL filter $\boldsymbol{\alpha}^m$ a set of concentric search regions $\mathbf{S}_i$ is established around the coordinate $s_m = (x_m, y_m, \sigma_s^m)$ which was memorized at training. To retain some scale information, the search region is also established at ± 1 scale around σ_s^m.

Search Radius	% of image size
R_1	5
R_2	10
R_3	30
R_4	50
R_5	70
R_6	100

TABLE III
LAYER 4 POOLING RADII. THE POOLING RADII ARE EMPIRICALLY CHOSEN TO COVER THE WHOLE IMAGE WITH A SUFFICIENTLY FINE SPATIAL RESOLUTION.

Each search region $\mathbf{S}_i$ is defined by a radius R_i as shown in table III and centered on the memorized coordinate s_m . We chose 6 levels of spatial pooling resolution. The lowest level ($R_1 = 5\%$) corresponds to the level of resolution used in [1]. The highest level ($R_6 = 100\%$) ensures that the entire image is covered, not discarding any feature. The remaining 4 resolutions are chosen to ensures a sufficiently fine resolution to encode variations in spatial positions.

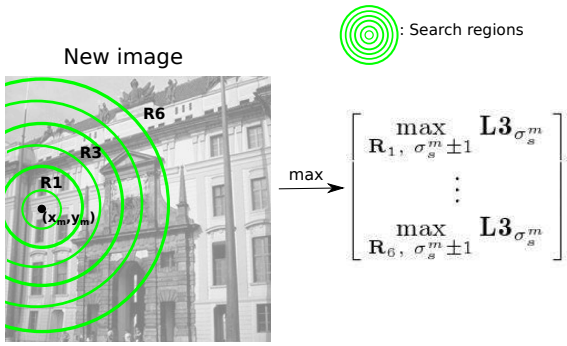

Fig. 6. Multi-resolution pooling. The training coordinate $s^m = (x_m, y_m, \sigma_s^m)$ of each HL filter is memorized. For each new image, a concentric series of 6 search spatial regions are centered on this coordinate and spanning $\sigma_s^m \pm 1$ scales . The maximum value is pooled from each search region. This generates both rich and localized spatial information.

By defining a local pooling region centered on the training position of each HL filter we take advantage of spatial regularities inside a given image class as done in [1]. By also varying the pooling radius (figure 6) we allow a more precise encoding of spatial relations. But unlike [31], our pooling is centered on each HL filter and is therefore feature specific.

For example as shown in figure 7, two categories might share a similar feature (i.e peak) but consistently positioned around the same position inside each category and at different positions between categories. It is therefore essential for classification to encode the spatial position of these features while capturing some variations, which is accomplished by using multiple radii off pooling.

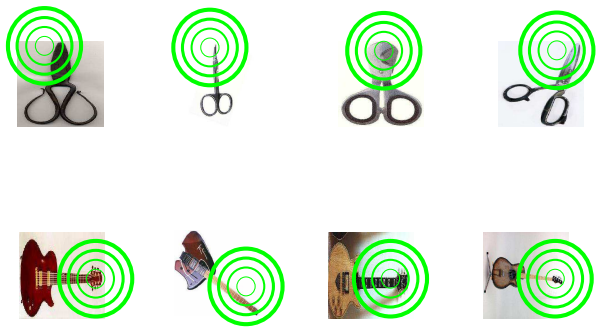

Fig. 7. Multi-resolution pooling. Both categories contain a similar feature (i.e peak). The feature is located in a limited but different region in both categories, with some variations inside each category. Both the category specific position and the variations within it are coded by our multi-resolution pooling.

This maximum pooling procedure is applied for each HL filter and the results are concatenated into a final **L4** vector signature (equation 8). Each element of the **L4** vector repre-

sents the maximum activation level of each HL filter inside each search region.

$$\mathbf{L4} = \begin{bmatrix} \max\limits_{\mathbf{R}_1,\ \sigma_s^1 \pm 1} \mathbf{L3}^1_{\sigma_s^1} \\ \vdots \\ \max\limits_{\mathbf{R}_1,\ \sigma_s^M \pm 1} \mathbf{L3}^M_{\sigma_s^M} \\ \vdots \\ \max\limits_{\mathbf{R}_6,\ \sigma_s^1 \pm 1} \mathbf{L3}^1_{\sigma_s^1} \\ \vdots \\ \max\limits_{\mathbf{R}_6,\ \sigma_s^M \pm 1} \mathbf{L3}^M_{\sigma_s^M} \end{bmatrix}. \quad (8)$$

Classifier. The layer 4 signature vector of each image are used to train one-against-all classifier, using a gaussian kernel with L^2 norm [43].

V. Classification Experiments

We give classification results for three images sets and we breakdown the improvements according to our contributions.

A. Data sets

To evaluate our network on classification tasks, we use three natural image data sets (figure 9). The first two, Caltech101 and Caltech256, are composed of various objects classes whereas the second one, Fifteen Scenes, corresponds to indoor/outdoor scenes.

1) Caltech101 and Caltech256: The Caltech101 [44] image set is composed of 102 categories for a total of 9144 natural images. The Caltech256 [45] image set is composed of 257 categories or a total of 30607 natural images. For both, each category represents a particular object against either a plain background or a natural scene.

2) Fifteen scenes: The Fifteen Scenes data set [44] is composed of 15 categories of urban and rural scenes for a total of 4885 images.

B. Classification results

1) Caltech101: Our basic architecture trains 40 HL filters per category for a total of $M = 4080$ HL filters. We used the standard classification procedure with 15 and 30 training examples per class, as done in all models presented in table IV. A *mean comparison Student t-test*, with a risk $\alpha = 0.05$, shows that our score of 69.52% is significantly above all biologically inspired architectures reported in [19], [1], [23], [24], [17], [37], [46], [38], [40] and compares with the highest scores in [47]. All these architectures use a similar generic framework of alternating convolution/pooling. Our highest score of 76.32% reaches state-of-the-art level for 30 training examples when compared to benchmark models using BoW methods with mono feature descriptors. Our architecture generates a total increase of 15% over the results in [1], the model most closely related to ours.

We reimplemented the two models presented in [19], [1], which are the most closely related to ours, and were able to reproduce or go above their published scores of 42% and 54% respectively. We used these reimplementations as our first baselines. Note that our basic score of 59% is already above the score reported in [1]. This is explained by the fact that we use a L^2 norm gaussian kernel instead of a linear classifier and also by our choice of implementing a multi-resolution of Gabor filters as opposed to an image pyramid with a fixed filter size. Beginning with our most basic setup, table IV shows the increase in classification scores observed when adding step by step the various aspects of our contributions.

First, when adding a normalized dot product on layer **L3** (equation 7) instead of the RBF function used in [1] we observe an increase near 3% in classification scores.

Second, we observe a jump of near 4% when using HL filters with multiple scales $|S| \in \{1, 3, 5, 7\}$. There is indeed a trade-off between the precision at which the HL filter fits the training data (discriminative power) and the level of scale invariance it can achieve. To account for both discriminative

	15 images	30 images
Our model		
$\|S\| = 1$	53	59
+ normalized dot product	**56.17 ± 0.48**	**63.00 ± 0.9**
$\|S\| = 1 \cup 7$	**59.21 ± 0.18**	**66.84 ± 1.05**
+ multi-resolution pooling	**60.1 ± 0.5**	**69.52 ± 0.39**
+ pixel level gradient	**68.49 ± 0.75**	**76.32 ± 0.97**
Deep biologically inspired architectures		
Serre *et al.* [19]	35	42
Mutch&Lowe [1]	48	54
Huang *et al.* [23]	49.8 ±1.25	
Theriault *et al.* [24]	54 ±0.5	61 ±0.5
Lecun *et al.* [17]	-	54±1.0
Lee *et al.* [37]	57.7±1.5	65.4±0.5
Jarret *et al.* [46]	-	65.6±1.0
Zeiler *et al.* [38]	58.6±0.7	66.9±1.1
Fidler *et al.* [40]	60.5	66.5
Zeiler *et al.* [47]		71.0 ±1.0
BoW architectures		
Lazebnik *et al.* [31]	56.4	64.6±0.7
Zhang *et al.* [48]	59.1±0.6	62.2±0.5
Wang *et al.* [27]	64.43	73.44
Yang *et al.* [49]	67.0±0.5	73.2±0.5
Boureau *et al.* [25]	-	75.7±1.1
Sohn *et al.* [50]	-	77.8

TABLE IV
Classification results in Average Precision on Caltech101

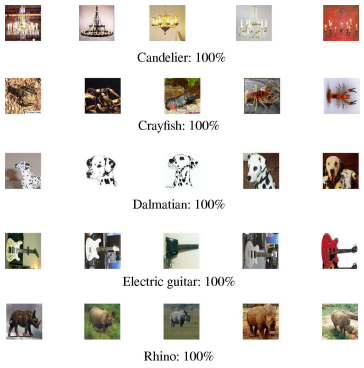

Fig. 8. Our best five classification accuracies on Caltech101

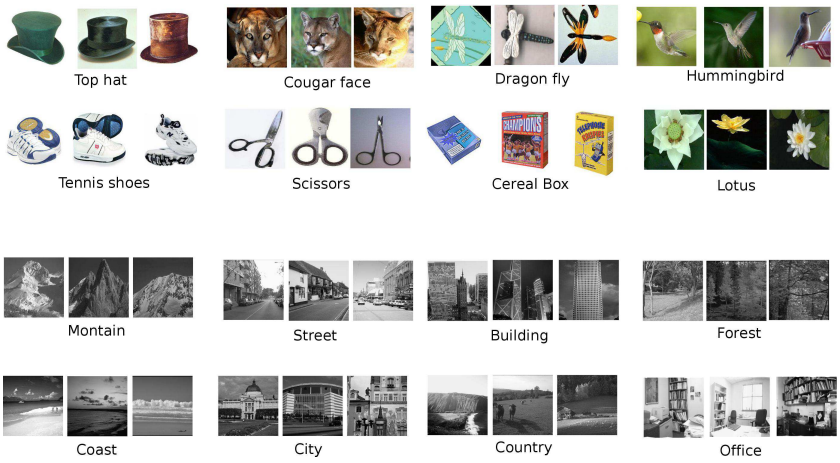

Fig. 9. Top: samples from the Caltech101 and Caltech256 image sets. Bottom: samples from the Fifteen Scenes image set

power and invariance we can train HL filters on all values of $|S|$.

Third, when adding our multi-resolution pooling at the final level of the network, an additional 2% increase is observed. For certain objects the spatial position relative to other features in the image can be very informative. To encode spatial training positions as well as spatial relations between features our final $\mathbf{L4}$ image signature concatenates all pooling radii $\{\mathbf{R}_1, ..., \mathbf{R}_6\}$ defined in table III.

The independent effects of our two main contributions, the multi-resolution pooling and the deeper HL filters are shown in table V. When combining HL filters signatures using all 7 scales with signatures using a single scale, a 4% increase is observed. A *paired sample Student t-test* on 10 independent splits shows this increase to be significant with a risk $\alpha = 0.05$. When adding the multi-resolution pooling alone a near 2% increase is observed, also significant with a risk $\alpha = 0.05$.

Our model	
$\|S\| = 7$	**62.82** $\pm$ **0.90**
$\|S\| = 1 \cup 7$	**66.84** $\pm$ **1.05**
$\|S\| = 7$ + multi-resolution pooling	**64.58** $\pm$ **1.05**

TABLE V
IMPROVEMENT OBTAINED INDEPENDENTLY FOR OUR TWO MAIN CONTRIBUTIONS TESTED ON CALTECH101 FOR 30 TRAINING EXAMPLES. DIFFERENCES ARE SIGNIFICANT ON A *paired sample Student t-test* WITH A RISK $\alpha = 0.05$.

By being composed of large arrays of Gabor filters, our HL filters give spatial information spanning large image areas (object level) as shown in figure 10. Although Gabor filters are defined as local frequency operators they respond similarly to an oriented second order derivative and they share relations with the family of gaussian derivative operators [10], [11], [12]. Therefore it is fair to say that our HL filters fall into the general category of descriptors composed of local band-pass filters, measuring or approximating various orders of spatial derivatives.

One such descriptor used for image classification is the SIFT descriptor [25], [49] which relies on local gradients. Contrarily to a SIFT descriptor, one single HL filter pools multiple spatial scales inside the same patch simultaneously (red ellipses in figure 10). For this reason one single HL filter can describe an object using multiple scales across one large image region (i.e 90x90 pixels). Different SIFT descriptors can operate at different scales but the scale of each individual descriptor is the same across the patch which usually covers smaller image regions (16x16 up to 48x48 pixels). These are two different levels of representation. Combining these descriptors allows for fine gradient descriptions as well as more macro, object level, descriptions. When combining the pixel level encoding of SIFT descriptors as done in [49] with the larger spatial span of HL filters our model generates 76.32% in classification score reaching state-of-the-art level on models using mono features descriptors (i.e. derivative filters, band-pass filters). A *mean comparison Student t-test* shows the increase over [49] to be significant with a risk $\alpha = 0.05$.

2) Caltech256: Table VI shows classification results on the Caltech256 data set. As for Caltech101, the basic architecture trains 40 HL filters per category for a total of $M = 10280$ HL filters. Again our model reaches state-of-the-art scores on similar architectures using a 4 layer architecture with convolution and maximum pooling as in [47]. When combining with the pixel level gradient of [49] our score reaches near state-of-the-art and improves the scores in [49] by close to 7%, clearly above statistical significance at a risk $\alpha = 0.05$.

Our model	
$\|S\| = 1 \cup 7$	
+ multi-resolution pooling	**31.23** $\pm$0.38
+ pixel level gradient	**40.56** $\pm$0.28

Deep biologically inspired architectures	
Zeiler *et al.* [47]	33.2$\pm$0.8
Bow architectures	
Yang *et al.* [49]	34.02$\pm$0.35
Wang *et al.* [27]	41.19
Boureau *et al.* [51]	41.7 $\pm$0.8

TABLE VI
CLASSIFICATION RESULTS IN AVERAGE PRECISION CALTECH256 FOR 30 TRAINING EXAMPLES.

3) Fifteen Scenes: For the Fifteen Scenes data set, our basic architecture trains 300 HL filters per category for a total of $M = 4500$ HL filters. Table VII shows the confusion matrix of our network applied to the Fifteen Scenes set. As for the Caltech101 image set we combine the complementary mid-level descriptions of our HL filters with the pixel level descriptions of SIFT as in [49]. Our global average classification score of 82.94% is above or close to benchmark results obtained in [31], [25]. A *mean comparison Student t-test* shows the increase over [49] to be significant with a risk $\alpha = 0.05$. More importantly, as shown in table VIII our standard model improves our reimplementation of [1],[19] by over 10% and 20% respectively.

	bedroom	calsuburb	industrial	kitchen	livingroom	coast	forest	highway	insidecity	mountain	country	street	building	paroffice	store
bedroom	0.89			0.01			0.03					0.03			
calsub..		0.71	0.05			0.02				0.13	0.04			0.01	
indust..		0.03	0.76			0.01			0.01	0.06	0.02			0.05	
kitchen				0.93					0.02	0.01				0.01	
living..	0.01				0.81		0.14					0.01			0.01
coast	0.02	0.01		0.06		0.6	0.01		0.04	0.01			0.01	0.18	0.01
forest	0.01				0.1		0.79	0.1				0.03			0.01
highway				0.01				0.91	0.03						0.02
insid..			0.01	0.04		0.02		0.03	0.84						
mount..	0.02	0.12	0.10					0.01	0.01	0.64	0.03			0.06	
country			0.03							0.03	0.91				
street	0.02											0.95			
build..										0.01			0.98		
paroff..	0.01	0.01	0.01			0.01		0.02	0.07	0.05		0.01		0.76	
store					0.02		0.01		0.01						0.91

TABLE VII
CONFUSION MATRIX FOR *15 scenes* IMAGE SET

Our model	
$\|S\| = 1 \cup 7$	
+ multi-resolution pooling	**74.35** ±0.83
+ pixel level gradient	**82.94** ±0.57

Deep biologically inspired architectures	
Mutch&Lowe [1]	63.5
Serre *et al.* [19]	53.0
Bow architectures	
Lazebnik *et al.* [31]	81.4 ±0.45
Yang *et al.* [49]	80.4 ±0.45
Boureau *et al.* [25]	84.3±0.45

TABLE VIII
CLASSIFICATION RESULTS IN AVERAGE PRECISION ON FIFTEEN SCENES FOR 100 TRAINING EXAMPLES. SCORES FOR [1],[19] ARE OBTAINED BY OUR OWN REIMPLEMENTATION.

The high discriminative power of our deepest HL filters is made obvious for certain categories in the Fifteen Scenes image set. For instance, our model consistently gives near perfect classification score for the Building category. This category showcases our model's ability to fit arrays of simple, scaled and oriented local structures. Categories such as Building are mostly composed of well defined local structures of multiple scales and orientations. Our most discriminative HL filters ($|S| = 7$) are well suited for this type of image stimuli. Indeed, our HL filters are expected to find close to optimal fit on organized patterns of clear-cut structures. This translates quite well into the high classification score obtained for the Building category, also illustrated in figure 11

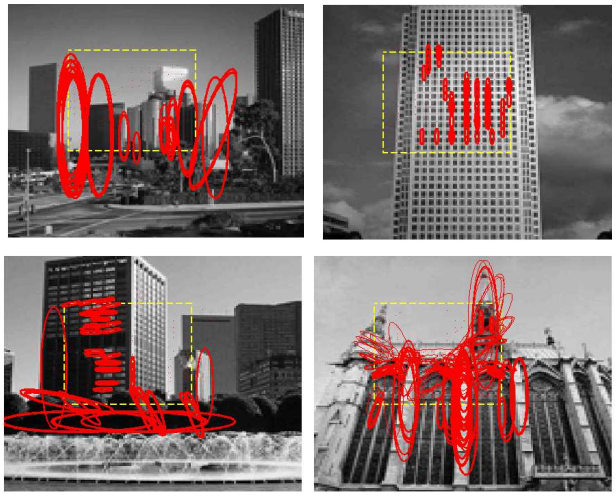

Fig. 11. The building categories showcases our model's ability to consistently fit arrays of simple, scaled and oriented local structures.

C. Computation time

We used a 8 core PC at 3.47MHz, using a simple and non-optimized Matlab code. Generating image signatures is faster then a regular sparse code BoW model thanks to our feedforward architecture. Activating layers $\mathbf{L1} - \mathbf{L2}$ is fast (∼1 sec) using simple convolution functions. Activating layer $\mathbf{L3}$ is more computationally demanding since it requires convolution of, for example, $M \sim 4080$ HL filters which are coded as high dimension matrices. The total feedforward activation of one image through the network takes roughly 4 seconds. Training the HL filters, is fast since it only requires activation of layers $\mathbf{L1} - \mathbf{L2}$. For example, on Caltech101, training $M \sim 4080$ HL filters takes close to 1 hour. Training and testing the classifier can be costly since we used a classifier with a L^2 norm gaussian kernel. Depending on the size of the data set, computing the kernel can be time demanding: close to 1 hour for Caltech101. Timing is proportional for the other data sets.

VI. FURTHER ANALYSIS

Here we give quantitative and qualitative explanations with respect to the improvements gained from our HL filters with multiple scales

Recall that the two key properties of our HL filters are their discriminative power and their invariance level. A good balance between discriminative power and invariance is necessary for classification tasks. As illustrated in figure 10 different objects can be matched by HL filters with different scale depths. While some objects are matched using a single scale HL filter (more freedom for scale invariance), others are matched using HL filters with multiple scales (more discriminative).

To evaluate more precisely the invariant and discriminative properties of our HL filters, we generated 1000 synthetic

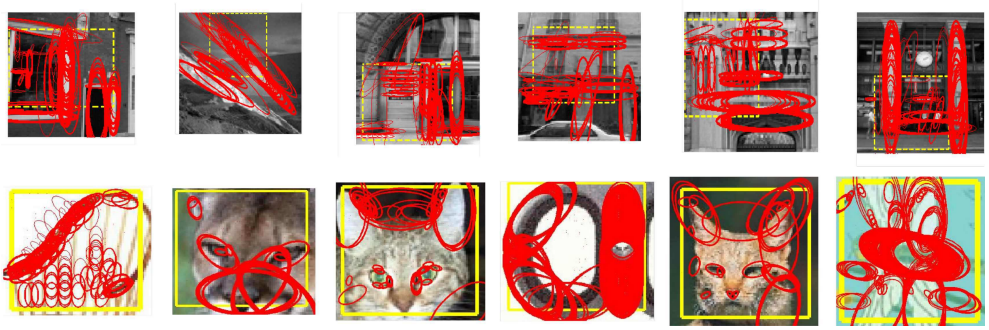

Fig. 10. HL filters visualization. Trained HL filters on some of the images in figure 9. The yellow bounding box defines the spatial range of the HL filter. The red ellipses indicate the local scales and orientations of the Gabor filters selected by the HL filter.

toy images. These synthetic images are generated from the superposition of thresholded Gabor filters with varying scales, orientations and positions as shown in figure 12. More precisely, the images consists of keeping the central part of Gabor filters (equation 1) corresponding to a full wavelength λ. This is a simple way of generating images where the local structures correspond to ideal inputs. Although the Caltech101,Caltech256 and the Fifteen Scenes sets are widely used for image classification, such image set make it difficult to evaluate the invariant properties of filters under various geometrical transformations [42], [52]. By using synthetic images we are able to control parameters such as translation, rotation and scaling and evaluate the invariance level of our filters under such transformations.

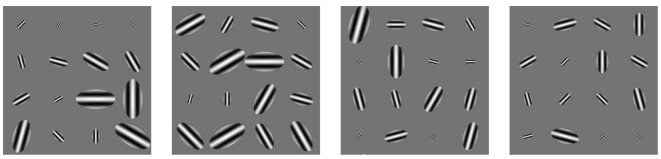

Fig. 12. Synthetic images: composed of non rigid arrays of Gabor filters. Each local feature is thresholded to cover one wavelength as defined in equation 1

A. Discriminative power of HL filters: fitness error

This section evaluates the discriminative power of our HL filters. To do so we measure the fitness error of our HL filters on training images and compare the results to our reimplementation of previous models [19], [1].

Since $\alpha^m \in \mathbb{R}^{n\times n\times|S|\times|\Theta|}$ and because $0 \leq \alpha^m_{i,j,\sigma,\theta} \leq 1$, it follows from equation 6 that $\sum_{i,j,\sigma,\theta} \alpha^m_{i,j,\sigma,\theta} \leq n^2$. We can then define the fitness error of each HL filter by

$$e = n^2 - \sum_{i,j,\sigma,\theta} \alpha^m_{i,j,\sigma,\theta}. \quad (9)$$

Figure 13 shows the average fitness error with respect to the scale range $|S|$ of HL filters as defined in section IV. The graph shows that HL filters with deeper scales are on average better tuned to the training images structures and thus more discriminative.

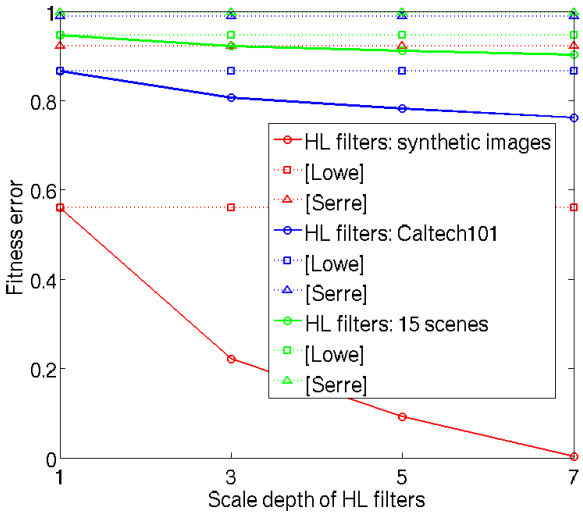

Fig. 13. Training image fit. Fitness error is lower with increasing scale range $S \subseteq \mathbb{S}$ of HL filters. All of our HL filters give better fit to local image structures than HL filters used in [19], [1]. As expected the more scales are available for training the HL filters the more they fit local image structures.

Figure 14 illustrates 2 different synthetic image patches composed of 4×4 Gabor features. For clearer visualization, the Gabor features are spread at a constant distance from each other. The red ellipses indicates the learned scales and orientations after training a 4×4 HL filter over each patch. The left image shows the fitting obtained for a HL filter using all available scales ($|S| = 7$). Clearly the local scale and orientation learned by the HL filters fit the local image structure. The right image shows the training misfit obtained when using a single-scale HL filter as it is the case for our reimplementation of [1].

B. Invariance level of HL filters

As mentioned above a good balance between discriminative power and invariance is necessary for classification tasks. This section evaluates the invariance level generated by our HL filters under various local transformations. For this purpose we evaluate the invariance of the HL filters outputs with respect to local geometrical transformations such as translation,

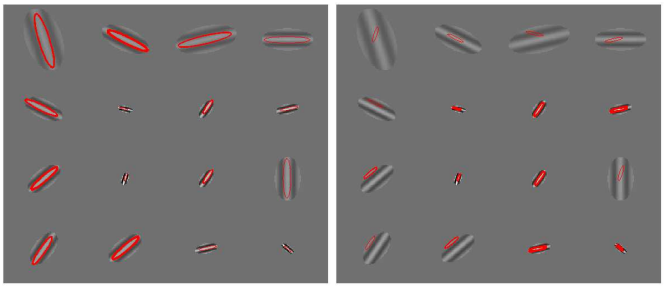

Fig. 14. Toy example: trained HL filter one synthetic image patch. On the left image a filter with full scale depth ($|S| = 7$) is trained. The red ellipses indicate the local scales and orientations learned by the HL filter. The filter clearly matches the local image structures. The right image shows the local misalignment of the filter when limited to one randomly chosen scale as in [1].

rotation and scaling. All these local transformations lead to global deformations at object level. Robustness to these local transformations is therefore a key aspect to good classification.

1) Local translation: One property gained by training HL filters with a lower fitness error is to increase the network invariance level to basic local geometrical transformations. For example, global deformations observed at object level can be decomposed into local translations [15] which can be minimized by our HL filters. To evaluate the effect of local translations on the output of our network we first generate 1000 synthetic image patches and train one HL filter on each patch according to equation 6. This generates 1000 HL filters optimally fitting the corresponding image patch as shown on the left in figure 14. After training we apply local translations (see figure 15) of increasing amplitude on each training image and measure the output of each corresponding HL filter on the transformed image.

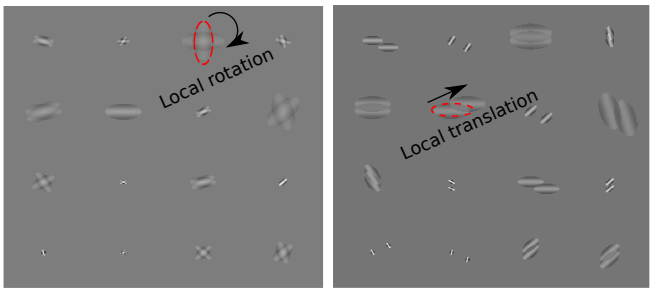

Fig. 15. Local transformations. Rotation and translation are applied individually to each local feature in the image patch to create a globally distorted image. The distorted images are then used to measure the robustness of HL filters.

As shown in figure 16, the output of HL filters decreases with increasing local translations. The decrease is less pronounced in the case of HL filters with a deep scale range, showing more local translation invariance (i.e. global deformation). Indeed by using the optimal local scale, the deeper HL filters pool from a corresponding local neighborhood as defined in equation 3 and table II in such a way that larger structures are pooled invariably from larger spatial regions at the **L2** level.

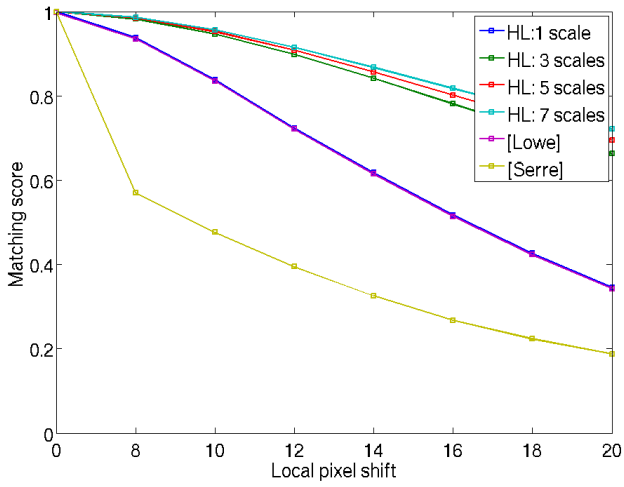

Fig. 16. Effect on local translations with respect to scale depth of HL filters. When compared to HL filters has the ones used in [19], [1], our HL filter with deeper scale range are less sensitive to image deformations created by local translations

2) Local Rotation: Yet another advantage of training HL filters with a lower fitness error is that they are locally aligned with the axis of relevant image structures and are thus less sensitive to local perturbations around this axis. To evaluate the level of invariance of HL filters with respect to local rotation we use the same procedure as for local translation but instead apply local rotation of increasing amplitude on the training images. Figure 17 shows a lesser decrease in matching score for deeper our HL filters compared with the ones used in [1], [19]. Deeper HL filters are less sensitive to image deformations created by local rotations. This is explained, as shown in figure 14, by the fact that deep HL filter are better aligned, or centered, with the local structures of images. Consequently a local rotation will not affect the HL filter response as much as it is the case for misaligned HL filters as obtained for our reimplementation of [1], [19].

3) Local Scaling: Local scaling in our network underlines particularly well the necessary balance between discriminative power and invariance for classification. When local scaling is applied in the same way as for translation and rotation a different pattern is observed. HL filters with maximum scale depth ($|S| = 7$) are more discriminative as shown in section VI-A but are less robust to local scaling than other HL filters ($|S| = 3, 5$). The deepest HL filter uses all available network scales to fit the image patches. It optimally matches the image local scales and as result has more discriminative power. However it leaves no freedom to find a perfect match at other scales. The high level of discriminative power gained by using all scales makes the HL filter with $|S| = 7$ less invariant to scaling in comparison to HL filters with less depth. On the other hand too much invariance to local transformations compromises the discriminative power of the network. As shown in section V-B a balance between discriminative and invariant properties of HL filters generates better classification performances.

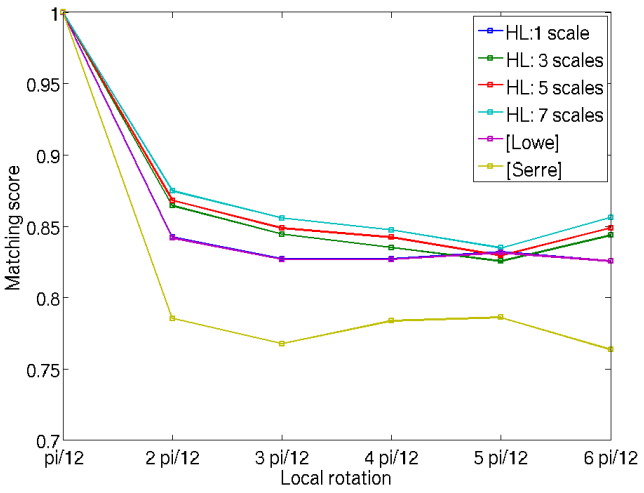

Fig. 17. Effect on local rotation with respect to scale depth of HL filters. Our HL filter with deeper scale range are more robust to image distortion created by local rotation.

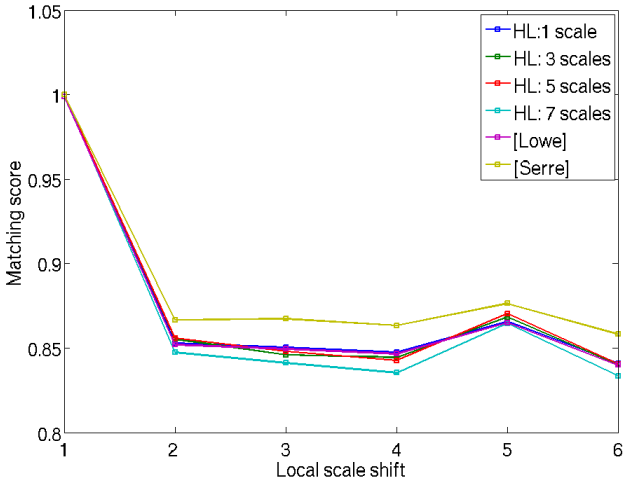

Fig. 18. Effect on local scaling. HL filters with more scale depth are less robust to local scaling. Too much invariance as in [19] leeds to less discrimination, reducing classification power of the network. Optimally, our network combines the invariant and discriminative aspects of all HL filters to increase classification results in section V-B.

C. Discussion regarding potential relationships with biological systems

Neurophysiological studies [53] suggest that the spatial receptive field profiles of neurons observed in cortical area V4 is constructed by pooling from specific orientation and spatial spatial frequency channels from more peripheral stages of visual processing. In particular, other studies [54], [55] also suggest that many neurons in area V4 are sensitive to boundary information (i.e. orientation,scales) at a specific position relative to the object center. These findings share some principles with our HL filters which are sensitive not only to multiple orientations at different positions from the center of their receptive fields but also to multiple scales. Also, recordings of neurons in the inferior temporal visual cortex (IT) show that these neurons have limited receptive fields of various sizes [14]. In particular neurophysiological [56], [57] studies suggest that information about the relative spatial positions of objects at different eccentricities from the fixation point, is coded by a population of IT neurons with various receptive field sizes. Moreover, the sizes of these receptive fields vary in the presence of other objects and background. These neurons can be seen as pooling over local regions of different sizes on the visual field, which results in partial invariance to spatial position. These findings also share some principles with our model where pooling HL filters over multiple radii at layer **L4** encodes the relative spatial position of features.

VII. Conclusion

The architecture presented in this paper allows for the manipulation of two crucial variables for image classification: discriminability and invariance. Our filters are modeled and trained to optimally fit local image structures and as a results, generate a good balance between discriminative representations and invariance. In particular, our results on three natural image sets and one synthetic image set highlight the increase in discriminative power of our network as well as its robustness to local geometrical transformations. Moreover, spatial organization of local features into a global representation is a key aspect to image recognition. In this regard, the multi-resolution pooling introduced in this paper provides rich spatial information, resulting in improved classification scores.

References

[1] Mutch.J and Lowe.D.G, "Object class recognition and localization using sparse features with limited receptive fields," *Int. J. Comput. Vision*, vol. 80, pp. 45–57, October 2008.
[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
[3] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, 2003, vol. 2, pp. 1470–1477.
[4] Hubel.D and Wiesel.T, "Receptive fields of single neurones in the cat's striate cortex," *Journal of physiology*, pp. 574–591, 1959.
[5] Campbell F. W. and Robson J. G., "Application of fourier analysis to the visibility of gratings," *Journal of Physiology*, vol. 197, pp. 551–566, 1968.
[6] D. E. Broadbent G. B. Henning, B. G. Hertz, "Some experiments bearing on the hypothesis that the visual system analyses spatial patterns in independent bands of spatial frequency," *Vision Research*, vol. 15, pp. 887–897, 1975.
[7] D. J. Field B. A. Olshausen, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
[8] R. A. Young, "The gaussian derivative model for spatial vision: I. retinal mechanisms," *Journal of Physiology*, vol. 2, pp. 273–293, 1987.
[9] A.P. Witkin, "Scale-space filtering," 1983, vol. 8 of *Proc. 8th Int. Joint Conf. Art. Intell*, pp. 1019–1022.
[10] Luc M J Florack, Bat M Ter Haar Romeny, Jan J Koenderink, and Max A Viergever, "Scale and the differential structure of images," *Image and Vision Computing*, vol. 10, pp. 376–388, 1992.
[11] J.J. Koenderink, "Operational significance of receptive field assemblies," *Biological Cybernetic*, vol. 58, pp. 163–171, 1988.
[12] J.J. Koenderink and A.J. van Doorn, "Representation of local geometry in the visual system," *Biological Cybernetic*, vol. 55, pp. 367–375, 1987.
[13] T.Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, pp. 77–116, 1998.
[14] E.T Rolls and G Deco, *Computaitonal neuroscience of vision*, Press:Oxford, 1 edition, 2006.

[15] Kunihiko Fukushima and Sei Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern Recognition*, vol. 15, no. 6, pp. 455–469, 1982.

[16] Edmund T. Rolls and T. T. Milward, "A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures," *Neural Comput.*, vol. 12, pp. 2547–2572, November 2000.

[17] Marc'Aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 1–8, 2007.

[18] Riesenhuber.M and Poggio.T, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 1999.

[19] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 411–426, 2007.

[20] Lior Wolf, Stanley M. Bileschi, and Ethan Meyers, "Perception strategies in hierarchical vision systems," in *CVPR (2)*, 2006, pp. 2153–2160.

[21] Stanley Bileschi, "Image representations beyond histograms of gradients: The role of gestalt descriptors," in *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[22] Saul Hochstein and Merav Ahissar, "View from the top: Hierarchies and reverse hierarchies in the visual system," *Neuron*, vol. 36, pp. 791–804, 2002.

[23] Yongzhen Huang, Kaiqi Huang, Dacheng Tao, Tieniu Tan, and Xuelong Li, "Enhanced biologically inspired model for object recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 41, no. 6, pp. 1668–1680, 2011.

[24] M. Cord C. Theriault, N. Thome, "Hmax-s: Deep scale representation for biologically inspired image categorization," in *IEEE International Conference on Image Processing*, 2011.

[25] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce, "Learning mid-level features for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2559–2566.

[26] Jan C. van Gemert, Cor J. Veenman, Arnold W.M. Smeulders, and Jan-Mark Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1271–1283, 2010.

[27] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 3360–3367, 2010.

[28] Hanlin Goh, Nicolas Thome, Matthieu Cord, and Joo-Hwee. Lim, "Unsupervised and Supervised Visual Codes with Restricted Boltzmann Machines," in *European Conference on Computer Vision (ECCV 2012)*, 2012.

[29] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de Albuquerque Araújo, "BOSSA: extended BoW formalism for image classification," in *18th IEEE International Conference on Image Processing (ICIP 2011)*, Sept. 2011, pp. 2966–2969.

[30] Florent Perronnin, Jorge Snchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *IN: ECCV*, 2010.

[31] Ponce.J Lazebnik.S, Schmid.C, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," 2006, vol. 2 of *CVPR*, pp. 2169–2178.

[32] Jan C. van Gemert, "Exploiting photographic style for category-level image classification by generalizing the spatial pyramid," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, New York, NY, USA, 2011, ICMR '11, pp. 14:1–14:8, ACM.

[33] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, Washington, DC, USA, 2006, CVPR '06, pp. 2033–2040, IEEE Computer Society.

[34] H.Ling and S. Soatto, "Proximity distribution kernels for geometric context in category recognition," in *Proceedings of the Tenth IEEE International Conference on Computuer Vision (ICCV)*, 2007.

[35] D. Picard, N. Thome, and M. Cord, "An efficient system for combining complementary kernels in complex visual categorization tasks," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 3877–3880.

[36] Yann Lecun, Lon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[37] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, pp. 609–616, ACM.

[38] M Zeiler, Dilip Krishnan, G Taylor, and Rob Fergus, "Deconvolutional networks for feature learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[39] Timothe Masquelier and Simon J. Thorpe, "Unsupervised learning of visual features through spike timing dependent plasticity," .

[40] S. Fidler, B. Boben, and A. Leonardis, "Similarity-based cross-layered hierarchical representation for object categorization," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Alaska, USA, June 2008.

[41] Kunihiko Fukushima, "Neocognitron for handwritten digit recognition," *Neurocomputing*, vol. 51, pp. 161–180, 2003.

[42] David D. Cox James J. DiCarlo Nicolas Pinto, Youssef Barhomi, "Comparing state-of-the-art visual features on invariant object recognition tasks," in *IN: IEEE Workshop on Applications of Computer Vision*, 2011.

[43] Cord.M and Cunningham.P, *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval (Cognitive Technologies)*, Springer-Verlag TELOS, 1 edition, 2008.

[44] R. Fergus L. Fei-Fei and P. Perona, "Learning generativevisual models from few training examples: an incremental bayesian approach tested on 101 object categories.," in *In CVPR Workshop on Generative-Model Based Vision*, 2004.

[45] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep. 7694, California Institute of Technology, 2007.

[46] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun, "What is the best multi-stage architecture for object recognition?," in *Proc. International Conference on Computer Vision (ICCV'09)*. 2009, IEEE.

[47] Matthew D Zeiler, Graham W Taylor, and Rob Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *International Conference on Computer Vision*, 2011, pp. 2018–2025.

[48] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 2126–2136, 2006.

[49] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *in IEEE Conference on Computer Vision and Pattern Recognition(CVPR*, 2009.

[50] Kihyuk Sohn, Dae Yon Jung, Honglak Lee, and Alfred Hero III, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition," in *Proceedings of 13th International Conference on Computer Vision*, 2011.

[51] Y-Lan Boureau, Nicolas Le Roux, Francis Bach, Jean Ponce, and Yann LeCun, "Ask the locals: Multi-way local pooling for image recognition.," in *ICCV*, Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc J. Van Gool, Eds. 2011, pp. 2651–2658, IEEE.

[52] Nicolas Pinto, David D Cox, and James J DiCarlo, "Why is real-world visual object recognition hard?," *PLoS Comput Biol*, vol. 4, no. 1, pp. e27, 01 2008.

[53] Jack L. Gallant Stephen V. David, Benjamin Y. Hayden, "Spectral receptive field properties explain shape selectivity in area v4.," *Journal of neurophysiology*, vol. 96, no. 6, pp. 3492–3505, 12 2006.

[54] Anitha Pasupathy Charles E. Connor Maximilian Riesenhuber Charles Cadieu, Minjoon Kouh and Tomaso Poggio, "A model of v4 shape selectivity and invariance," *Journal of neurophysiology*, vol. 98, pp. 1733–1750, 2007.

[55] Anitha Pasupathy and Charles Connor, "Shape representation in area V4: Position-specific tuning for boundary conformation," *Journal of Neurophysiology*, vol. 86 (5), pp. 2505–2519, 2001.

[56] Nikolaos C. Aggelopoulos and Edmund T. Rolls, "Scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene.," *European Journal of Neuroscience*, vol. 22, pp. 2903–2916, 2005.

[57] Edmund T. Rolls, Nicholas C. Aggelopoulos, and Fashan Zheng, "The receptive fields of inferior temporal cortex neurons in natural scenes," *Journal of Neuroscience*, vol. 23, pp. 339–348, 2003.

Christian Thériault received a B.Sc from McGill university in Montréal where he studied physiology and psychology. He also received a B.Sc in mathematics (2008) and a Ph.D in psychology (2006) from Université du Québec à Montréal as well as a Master's degree in applied mathematics from université Paris Descartes in 2010.

Nicolas Thome received the diplôme d'Ingénieur from the École Nationale Supérieure de Physique de Strasbourg, France, the DEA (MSc) degree from the University of Grenoble, France, in 2004 and, in 2007, the PhD degree in computer science from the University of Lyon, France. In 2008, he was a postdoctoral associate at INRETS in Villeneuve d'Ascq, France. Since 2008 is an assistant professor at Université Pierre et Marie Curie (UPMC) and Laboratoire d'Informatique de Paris 6 (LIP6). His research interests are in the area of Computer Vision and Machine Learning, particularly in the design and learning of complex image representations and similarities, with applications to image and video understanding.

Matthieu Cord received the Ph.D. degree in Computer Science in 1998 from the University UCP, France, before working in the ESAT lab. at KUL University, Belgium, and in the ETIS lab, France, as Assistant Professor from 2000 to 2006.

He is currently a full Professor of Computer Science at UPMC Paris 6 Sorbonne Universities. His research interests include Computer Vision, Image Processing, and Pattern Recognition. He developed several systems for content-based image and video retrieval, focusing on interactive learning-based approaches. He is also interested in Machine Learning for Multimedia processing, Digital preservation, and Web archiving. Prof. Cord has published a hundred scientific publications and participated in several international projects (European FP6 and FP7, Singapore, Brazil) on these topics. He is a member of the IEEE and was nominated in 2009 at the IUF (French Research Institute) for a 5 years delegation position.