HMAX-S: DEEP SCALE REPRESENTATION FOR BIOLOGICALLY INSPIRED IMAGE CATEGORIZATION

Christian Theriault , Nicolas Thome , Matthieu Cord

Université Pierre et Marie Curie, UPMC-Sorbonne Universities, LIP6, 4 place Jussieu, 75005, Paris, France

ABSTRACT

This paper presents an improvement on a biologically inspired network for image classification. Previous models have used a multiscale and multi-orientation architecture to gain robustness to transformations and to extract complex visual features. Our contribution to this type of architecture resides in the building of complex visual features which are better tuned to images structures. We allow the network to build complex features with richer information in terms of the local scales of image structures. Our classification results show significant improvements over previous architectures using the same framework.

Index Terms— Image classification, scale, biological network visual cortex

1. INTRODUCTION

The task of visual classification is a cornerstone of computer vision. A good classification system shall respond invariably to objects within the same class and differently between classes. The current State-of-the-art in this field revolves around *Bag-of-features* methods which rely on building vector signatures from a dictionary of features [1]. The vector signatures in these models are essentially obtained in one step using a flat architecture.

These flat architectures are in sharp contrast with biological visual systems which are characterized by deep multilevel processing. Research on biological visual systems has been and important field of work since the famous work of T.Hubel & D.Wiesel [2]. One biologically inspired network which has received a great deal of attention comes from the HMAX model of Riesenhuber&al [3]. This networks alternates layers of features extraction with layers of maximum pooling. The feature extraction layers gives the network its discriminative power while the pooling layers allow for partial invariance to scale and to the exact configurations of features. Serre &al [4] extended the original HMAX model to add multi-scale representations as well as more complex visual features. Shuangping&Lianwen [5] combined these biologically inspired features with other descriptors to gain more discrimination.

Mutch&al [6] extended the network of [4] by tuning the complex visual features to the dominant local orientations. This improvement yielded visual prototypes which are better tuned to the image structures and thus less sensitive to noise and clutters. However, the visual prototypes used in [6] are shallow in scale and therefore not optimally tuned to local image structure.

In this article, we propose a new architecture (HMAX-S) by allowing the network to represent visual prototypes with multiple local scales. By using a range of available local scales at each position, our model tunes itself to the local scale of images thereby increasing its discriminative power when compared to the shallow prototypes used in [6]. Integrating local discriminative features into global features helps recognition of object in complex situations [7]. Our network is not only able to represent more precisely the local structures of images, it also provides a mean to integrate the relative scales of local visual features into to the global scale of objects. We report extensive experimentation on the Caltech image data set in order to validate our model.

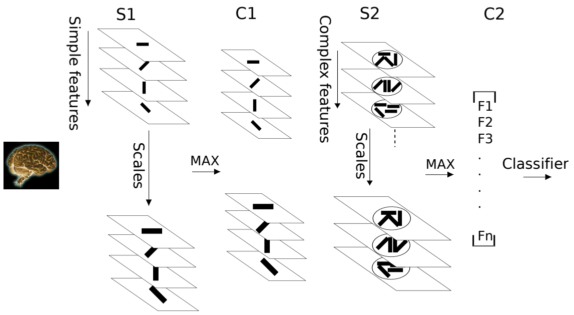

Fig. 1. Global architecture of the network used in [6].

2. GENERAL FRAMEWORK

Our image classification model follows the general 4 layer architecture used in [4, 6] and illustrated in figure 1.

S1 layer. The S1 layer extracts simple features by taking the normalized dot product of Gabor filters G of multiple orientations θ and scales s with local patches X at each position in the image plane such that

$$R_{\theta,\sigma}(G,X) = \frac{\langle G|X\rangle}{\|X\|} \tag{1}$$

with

$$G(x,y) = \exp(\frac{x_o^2 + \gamma y_o^2}{2s^2}) \cdot \cos(\frac{2\pi}{\lambda}x_o) \tag{2}$$

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. All parameters have been setup to provide biologically plausible features, as in [4]. Each filter is normalized to zero mean and unit length.

C1 layer. The C1 layer selects the local maximum value of each S1 orientation over two adjacent scales. As in [4], the spatial area of maximum selection is proportional to the scale of the corresponding Gabor filter. The C1 layer operation corresponds to a spatial down sampling where only the maximum values are kept.

S2 layer. The S2 layer combines the C1 features into more complex features using a Radial Basis Filter (RBF) defined by

$$Y_i = \exp(-\beta\|X - P_i\|^2) \tag{3}$$

Contact: theriaultc@poleia.lip6.fr

where β is a tuning parameter and $i = 1...N$

These N filters compute a Gaussian distance between C1 patches X and visual prototypes P_i. In [6] a large number of prototypes was used (i.e $N = 4000$). The prototypes P_i are defined during a training phase. This training phase is critical since it impacts the type of complex features represented by the network.

C2 layer. The C2 layer selects the maximum output of each S2 filter. In [4], the maximum is selected on the whole image. In order to incorporate spatial information in our model, we choose to select the maximum value of each S2 filter in a neighborhood surrounding the position where the prototype P_i was learned during training [6].

Classifier. The C2 output vectors are used to train one-against-one linear classifiers (SVM) [8].

3. OUR MODEL: HMAX-S

Our HMAX-S model is based on the previous (S1-C1-S2-C2) architecture with two major improvements. As opposed to the RBF (equation 3) used in [4, 6] we instead redefined the S2 filters with a normalized dot product such that

$$Y_i = \frac{\langle P_i | X \rangle}{\|P_i\| \|X\|} \tag{4}$$

Our S2 filters compute the cosine similarity between C1 patches X and visual prototypes P_i. We found that a normalized dot product, as opposed to RBF, not only speeds up computations but also improves classification scores. A normalized dot product is invariant to light intensity changes, whereas the RBF used in [6] is not. Indeed, even if the S1 dot product is normalized, the S2 filters still cover larger regions than the S1 filters. Using a normalized dot product at the S2 level maintains the S1 normalization which is otherwise lost when using RBF.

Our next improvement is in the important training phase of the prototypes P_i. Prototypes P_i are patches of $n \times n$ units selected on the C1 layer of training images. According to equations 1 the prototypes selected during training define the visual features extracted by the S2 filters. In [6], each position (p, q) in the $n \times n$ patch defining a prototype was set to the strongest C1 unit across all orientations. Figure 2a) shows the visualization of a prototype as defined in [6]. The yellow square indicates the spatial area covered by the prototype. The red ellipses indicate the corresponding S1 filters with respect to scale and orientation.

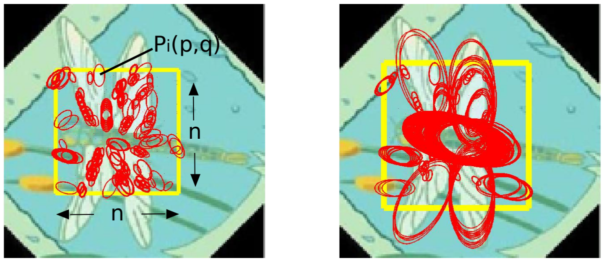

Fig. 2. a) Prototype as in [6]. b) Our multi-scale prototype.

In HMAX-S the prototypes P_i are also $n \times n$ patches selected from the C1 layers of training images. However, as shown by figure 2b), they are not limited to one scale. This means that a prototype represents features over multiple orientations and multiple scales. This is in contrast with [6] where the prototypes have no inner scale variations. As a result, our S2 filters are more adapted to the local structures of training images and thus less sensitive to noise and clutter.

3.1. HMAX-S S2 prototype definition

A prototype P_i is defined in a local C1 coordinate frame centered on a central scale s_c. At every position (p, q), each value $P_i(p, q)$ in a prototype is set to the value of the strongest C1 unit across all orientations $\theta \in \Theta$ and across a local range of scales $s \in S = [s_c - \Delta s, s_c + \Delta s]$ such that:

$$P_i(p, q) = \text{C1}(p, q, s^*, \theta^*) \tag{5}$$

where

$$(\theta^*, s^*) = \underset{\theta \in \Theta, s \in S}{\text{argmax}} \quad \text{C1}(p, q, s, \theta) \quad \forall (p, q) \tag{6}$$

This is illustrated in figure 3, with $\Delta s = 1$, $s_c = 3$, and $\Theta = \{0\ , \frac{\pi}{3}\ , \frac{2\pi}{3}\ , \pi\ \}$. The S2 prototypes $P_i(p, q)$ at position (p, q) defined by equation 5, and corresponding to the optimal scale and orientation (θ^*, s^*) defined by equation 6, are represented in red. These local connexions are established during training. Once training is done, when computing the S2 layer for new images (eq 4), the S2 filters P_i are blind to other orientations and scales.

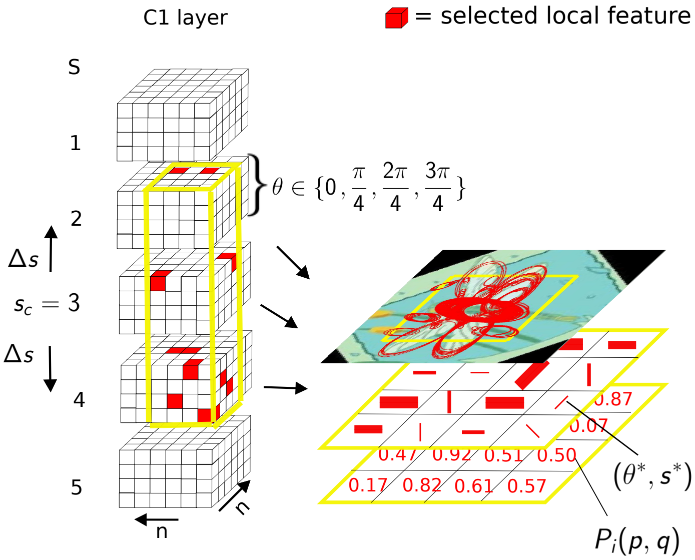

Fig. 3. HMAX-S multi-scale S2 prototype.

Algorithms 1 summarizes the steps for training one S2 prototype.

3.2. Multi-scale image signature

When an image is presented to the network, each S2 prototypes P_i is searching for a match in a local neighborhood U_i. As shown by figure 4, the search region U_i of prototype P_i is a limited region over scale and space. The search region U_i is centered on the central scale s_c and on the spatial position P_i^c at which P_i was selected during training. More specifically the search region U_i of each prototype P_i is defined by shifting the prototype's local frame (yellow box) over a limited spatial region and over the 2 adjacent scales, as illustrated by the blue box in figure 4. The i^{th} element of the C2 signature, that we denote C2$_i$, is obtained by taking the best match in U_i of prototype P_i, determined by the maximum Y_i value (eq 4).

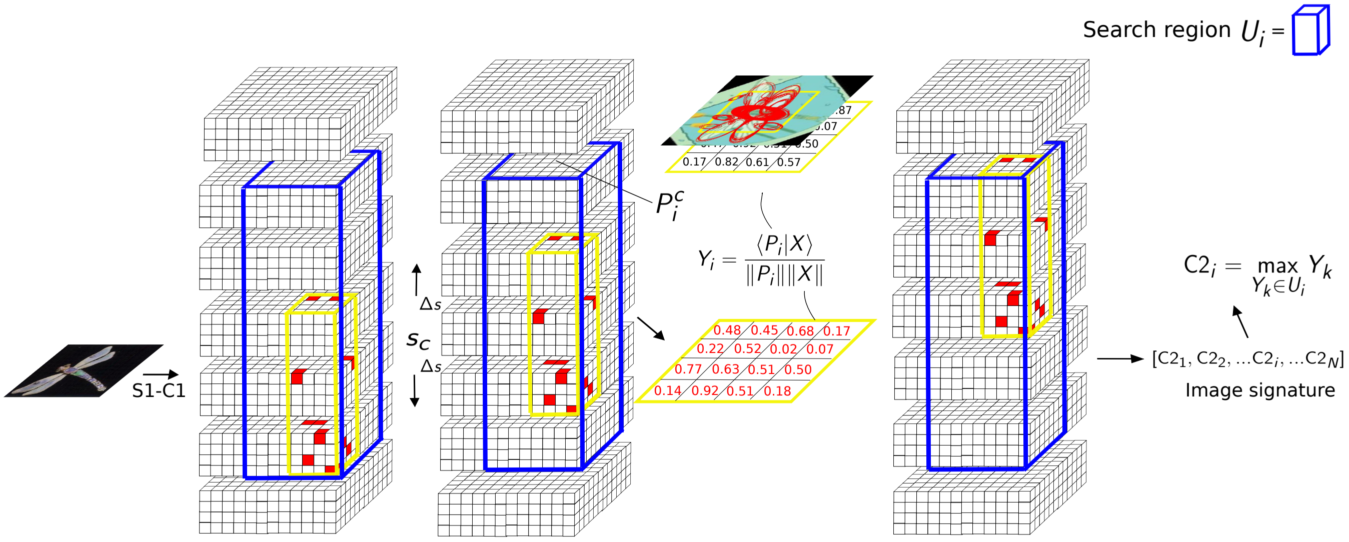

Fig. 4. HMAX-S multi-scale signature.

```
Algorithm 1: Prototype selection
//Input C1: $\Theta \times S' \times a \times b$  // (a,b):size $\Theta$:orientations $S'$:scales
//Random initialization:
$n \leftarrow rand(4, 8, 12, 16)$, $P^c \leftarrow rand(a, b)$, $s_c \leftarrow rand(S')$,
  $\Delta s \leftarrow rand(1, 2, 3)$, $S \leftarrow [s_c - \Delta s, s_c + \Delta s]$
//Computing the prototype
for $p = P_x^c$ to $P_x^c + n - 1$
    for $q = P_y^c$ to $P_y^c + n - 1$
        $(\theta^*, s^*) = \underset{\theta \in \Theta, s \in S}{\text{argmax}}\ \text{C1}(p, q, s, \theta)$
        $P_i(p, q) = \text{C1}(p, q, s^*, \theta^*)$
    end for
end for
Return $P_i$, $(\theta^*, s^*)$
```

Searching for each prototype in a limited region over scale and space, provides partial invariance to scale and position while retaining some spatial and scale information [1, 6].

4. VISUALIZATION AND CLASSIFICATION EXPERIMENTS

To test the improvements gained by HMAX-S we used the Caltech 101 images data set. The data set consists of 9,197 images split into 102 categories each containing from 45 to 800 images. We used standard classification protocol [6] with 15 and 30 training examples. We used 4 different spatial sizes of prototypes with $n \in \{4, 8, 12, 16\}$ and a total of 8 scales and 12 orientations. We trained 4080 S2 prototypes.

4.1. Classification results with multi-scale visual prototypes

First we give quantitative results showing the improvement gained from our two main contributions. We tested the improvement of HMAX-S in terms of the number of scales used by the prototypes P_i. As shown by table 1 up to 4% increase in classification scores is obtained when using prototypes with deeper scale range. This increase in classification scores over the the results obtained in [6] comes from our S2 prototypes which are are better adapted to the structures of training image which increases the discriminative power of the network.

Model	$\|S\|$	15 images	30 images
Mutch&Lowe [6]	1	48	54
HMAX-S (eq.3)	3	50	58
	5	51	58
HMAX-S (eq.4)	3	53	58
	5	53	59

Table 1. Classification results in Average Precision, done over three runs with $std \simeq 0.5$ for both methods.

Figure 5 shows the average image reconstruction error with respect to the scale depth of S2 prototypes. The reconstruction error for each prototypes P_i measures how well it matches the local region of the training image. Not surprisingly the graph gives a quantitative validation of our initial intuition stating that S2 prototypes with deeper scales are on average better tuned to the training images structures and thus more discriminant.

Our second main contribution was to replace equation 3 with equation 4. An improvement of 2% was obtained when using normalized dot product as opposed to RBF. As explained in section 3 a normalized dot product computes the similarity between prototypes and image patches robustly with respect to light intensity whereas a RBF is sensitive to lighting conditions. Using equation 4 with multi-scale prototypes yields a gain in classification of 5%.

4.2. Optimal parametrization

The lower reconstruction error obtained when using more scales translates into more discriminative power which results in better representation of local image structures. As shown by table 2 a large improvement in performance was obtained for certain categories when

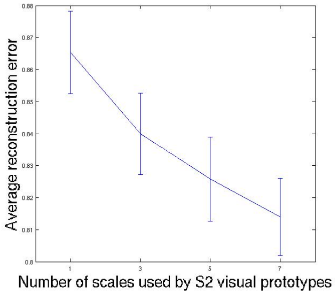

Fig. 5. Reconstruction error is lower with respect to scale depth of S2 prototypes.

using prototypes with 7 scales.

	Wild cat	Dalmatian	Llama	Wrench	Scissors
APD	+25	+13	+15	-11	-11

Table 2. Average Precision difference (APD): 5 scales prototypes vs 1 scale prototypes.

Certain categories are better classified when using single scale prototypes. This can be explained by the fact that when using one scale to define the prototypes there are more possibilities of finding a match when shifting the U_i search region which results in more scale invariance. Conversely, when using multiple scales to define a prototype there are fewer scales available to shift the U_i search region at other scales which results in less scale invariance.

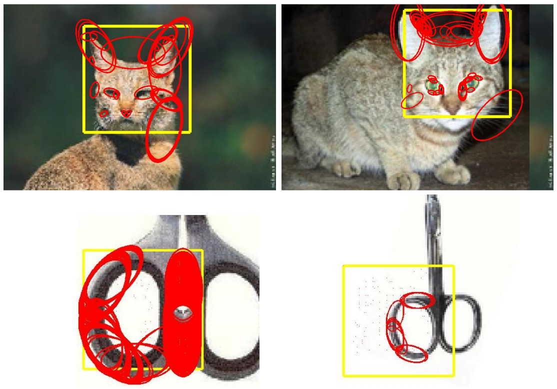

Fig. 6. The Wild Cat category is well represented using multi-scale prototypes while the Scissors category is well represented with single scale prototypes.

It thus seems that for good classification, a compromise must be made between discrimination and invariance. –Too much discrimination results in less invariance whereas to much invariance results in less discrimination between categories–. To this regard, different categories seem to require a different balance between invariance and discrimination.

One idea to exploit the optimal balance in discrimination and invariance for each category was to combine all prototype depths $|S| \in \{1, 3, 5, 7\}$ into the model to get C2 signatures containing features of all scale depth. A linear SVM will naturally perform selection on scale by allocating more weight to the discriminant features. After training the SVM, the discriminative scales for each category should be allocated more weights thereby giving more discriminative power to the network while retaining sufficient invariance for categorization. The score of 61% clearly indicates the advantage of using a more refined local representation, not only in terms of orientations as in [6] but also in terms of scale. 7% improvement in classification scores is obtained when using S2 filter prototypes of all scale depths. This allows the formation of C2 signatures which represent the image structure more faithfully, giving more discriminative power to the network and thus alleviating the work of the classifier.

Model	$\|S\|$	15 images	30 images
Mutch&Lowe [6]	1	48	54
HMAX-S (eq.4)	1,3,5,7	54	61

Table 3. Classification results in Average Precision, done over three runs with $std \simeq 0.5$.

5. CONCLUSIONS

Deep architectures inspired by the visual cortex certainly offer a great deal of promises when considering the level of performance achieved by the human visual system. A balance between complex and invariant visual information is essential for image classification. The HMAX-S architecture presented in this paper allows for the manipulation of these two crucial variables. Our contribution was focused on increasing the complexity variable of the network by building S2 filters with richer information. Our future work shall focus on the learning of visual prototypes in order to optimize the balance between invariance and complexity of visual representations.

6. REFERENCES

[1] 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.

[2] Hubel.D and Wiesel.T, “Receptive fields of single neurones in the cat’s striate cortex,” *Journal of physiology*, pp. 574–591, 1959.

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

[4] Bileschi.S Riesenhuber.M Poggio.T Serre.T, Wolf.L, “Robust object recognition with cortex-like mechanisms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 411–426, 2007.

[5] Shuangping.H and Lianwen.J, “Enhanced visual categorization performances by incorporation of simple features into bim features,” in *ICIP*, 2010, pp. 3865–3868.

[6] 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.

[7] Miguet.S Thome.N, Merad.D, “Learning articulated appearance models for tracking humans: a spectral graph matching approach,” *Signal Processing: Image Communication*, vol. 23, pp. 769–786, December 2008.

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